ISSI
2019

SAPIENZA
UNIVERSITÀ DI ROMA

17th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2019

with a Special STI Indicators Conference Track

2-5 September 2019
Sapienza University of Rome, Italy

PROCEEDINGS

VOLUME II

EdizioniEfesto

PROCEEDINGS OF THE 17TH CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS

© Edizioni Efesto - ISBN 978-88-3381-118-5 - August 2019
Printed in Italy

Editors: *Giuseppe Catalano, Cinzia Daraio, Martina Gregori, Henk F. Moed and Giancarlo Ruocco*

Graphic cover design: *Francesco Manzo* | *graframan.com*
Cover photo: ©*Fayee - stock.adobe.com*

INDEX OF PAPERS (FULL PAPERS AND RESEARCH IN PROGRESS)

INDEX OF POSTERS

Robustness of journal classifications in SSH: an empirical analysis from Italy

Tindaro Cicero[1] and Marco Malgarini[2]

[1] *tindaro.cicero@anvur.it*
(ANVUR) National Agency for the Evaluations of Universities and Research Institutes, 00153 Rome (Italy)

[2] *marco.malgarini@anvur.it*
(ANVUR) National Agency for the Evaluations of Universities and Research Institutes, 00153 Rome (Italy)

Abstract

In Humanities and Social Sciences, a two-tier journal classification implemented by selected scientific experts has been recently used in Italy in the context of the National Habilitation programme; for the same period, peer review is available for a large number of journal articles, developed in the framework of two national evaluation exercises (VQR 2004-2010 and VQR 2011-14). We take advantage of this rich dataset in order to check if journals classified as top class by scientific experts show higher impact and if articles published in those journals receive higher marks in peer-reviewed evaluation exercises. Results are rather mixed: the impact indicator does not seem to be significantly higher for top-class journals with respect to those classified in the second tier; on the other hand, median marks received by individual articles in the two evaluation exercises are significantly higher for top journals. We can draw two main conclusions from the analysis: the first is that journal classification performed by experts does not necessarily rely on impact measures, depending more instead on different criteria for journal selection. However, journal classification based on experts' judgement seem to proxy quite well on average the quality of individual articles as measured by peer review.

Introduction

Peer review is generally considered as the best method for evaluating research (Moed, 2008); however, it can be costly and time consuming, and may be affected by various kind of biases (Lee et al, 2013). For these reasons, some countries use ratings or rankings of journals as a way to inform peer evaluation (see for instance Walker et al, 2018). Ratings may be either based on indicators or on experts' opinions: typically, indicators are widely used in Science, Technology, Engineering and Mathematical (STEM) areas, whilst experts judgements are mostly used in Humanities and Social Sciences (SSH[1]). In Italy, journal indicators have been used in STEM areas in the country' two main evaluation exercises, the VQR 2004-2010 and VQR 2011-14 (see Anfossi et al, 2016); in SSH, instead, peer review was not supported by external information concerning journal classification, even if reviewers had the knowledge of the venue in which the article under evaluation was published[2]. However, in those areas journal ratings have been used in the context of the procedures concerning the National Habilitation of candidates for professorships. The co-existence in the same county and for the same period of time of two distinct evaluation procedures, one based on pure peer review of journal articles and the other on journal classification, makes it possible to test the robustness of experts' journal classification: this may be done contrasting journal ratings with the results obtained on the basis of the peer review of individual articles (Bonaccorsi et al., 2018). Moreover, robustness of classification may also be tested against the information concerning impact indicators of journals extracted from the Scopus database. In the following, we firstly describe with some detail the journals database used by ANVUR in the National Habilitation procedures. We hence present the results of the comparison among peer review at the article

[1] For a more complete discussion on the various methods of journal rating and ranking, see Ferrara and Bonaccorsi (2016).
[2] For a description of the Italian evaluation exercise, see Ancaiani et al (2015).

and journal level and among journal peer review and the indicators extracted from Scopus. Some consideration on the results obtained concludes the paper.

The classification of journals for the National Habilitation Programme

According to Ministerial Decree 76/2012, updated with the Ministerial Decree 120/2016, in SSH candidates for professorship are selected on the basis of various indicators of research activity, including the number of articles in scientific journals and in A-Class journals. According to the decrees, ANVUR is in charge of producing the two-tier classification of scientific and A-Class journals, with the possible assistance of external scientific experts. ANVUR has issued a regulation establishing the formal requirements for a journal to be classified as "scientific" or "A-Class" (see ANVUR, 2017a). More precisely, journals should possess specific process and product requirements. As for process requirements, a scientific journal should target an academic audience, possess a ISSN code and employ at least a single-blind method of peer review. On top of that, A-Class journals should employ a double-blind method of peer review; moreover, for A-Class journals, articles that have been submitted for evaluation in the VQR should have received on average a better evaluation than that received by scientific journals in the same scientific sector. As product requirements, journals should comply with pre-defined rules in terms of regularity of publication, Board composition, diffusion in the scientific community, rules of access, scientific content and international scope. According to ANVUR rules, journals in Economics and Statistics should also have to comply with pre-determined requirements in terms of impact, as measured by journal indicators extracted from international databases. Currently, a group of 42 experts specifically appointed by ANVUR on the basis of a public call is in charge of verifying the aforementioned requirements (see http://www.anvur.it/gruppo-di-lavoro-ric/gruppo-di-lavoro-riviste-e-libri-scientifici-2017/ for the list of experts). Overall, a total of over 20,000 and 5,500 journals are classified as scientific and A-Class, respectively, with the sectoral distribution shown in table 1.

Table 1. Rated journals by ANVUR (number of Scientific and Class A journals)

Areas	*N. of Scientific*	*N. of A-Class*	*% of A-Class*
Architecture	2,115	305	12.6%
Antiquities, philology, literary studies, art history	6,488	1,992	23.5%
History, philosophy and pedagogy	6,873	1,420	17.1%
Law	2,330	422	15.3%
Economics and Statistics	7,150	1,031	12.6%
Political and social science	3,988	1,059	21.0%
Total	**20,016**	**5,476**	**21.5%**

Source: Authors elaboration on ANVUR data

Figure 1 presents the world distribution of scientific journals classified by ANVUR (information on publisher's country is available for 70% of journals): most of the journals are concentrated in Europe and the US, with a good representation also for journal edited in Australia, Brazil, Canada and India; despite the strong growth in scientific publishing in the last decade, China is still under represented in the Italian classification, together with the rest of Asia, the Latin America and Africa.

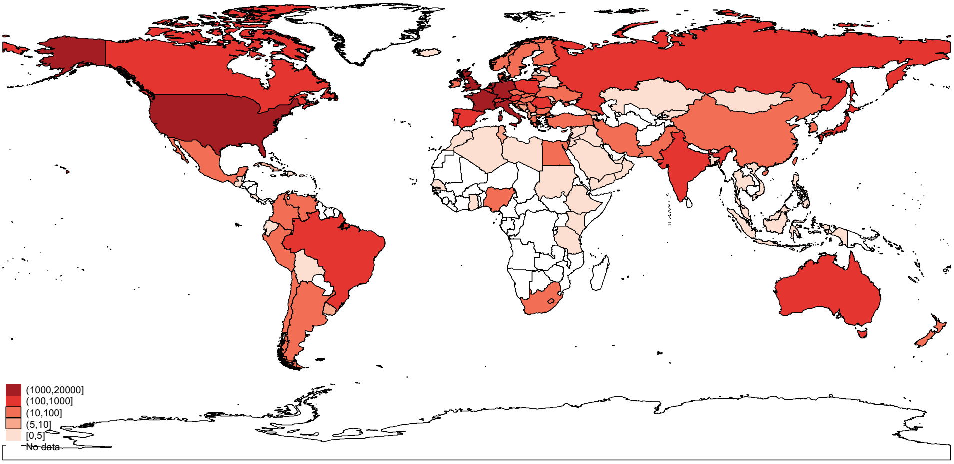


Figure 1. World distribution of journals rated by ANVUR

Source: Authors elaboration on ANVUR data

As for the language of publication, almost 50% of the classified journals are published in English and 14% in Italian (see table 2). There is also a remarkable share of journals that are published in more than one language, while a small but not negligible share of publications is issued in various other languages.

Table 2. Language of classified journals

Language	*Frequency*	*Percentage*	*Cumulate percentage*
English	9,816	49,04	49,04
Multilanguage	3,384	16,91	65,95
Italian	**2,847**	**14,22**	**80,17**
French	1,011	5,05	85,22
Spanish	774	3,87	89,09
German	461	2,30	91,39
Portuguese	321	1,60	93,00
Russian	84	0,42	93,42
Romanian	74	0,37	93,78
Flemish	63	0,31	94,10
Chinese	44	0,22	94,32
Polish	41	0,20	94,52
Croatian	39	0,19	94,72
Others	334	1,67	96,39
not available	723	3,61	100,00

Source: Authors elaboration on ANVUR data

Journal classification and Scopus indexation

ANVUR classification, with the only aforementioned exception of the classification in Economics and Statistics, is not influenced by standard bibliometric indicators such as the impact factor. However, it is possible to evaluate ex post whether classified journals are indexed in the major international journal databases, and if the frequency of classification is higher for A-Class with respect to scientific journals: indeed, if this is the case, the evidence may be considered as a very preliminary confirmation of the robustness of the ANVUR classification, also helping shedding some light on the influence of impact indicators on journal classification. In this respect, table 3 reports the share of classified journals that are indexed in the Scopus[3] database, distinguished by scientific area; on average, 43% of scientific journals and over 63% of top rated journals are also indexed in Scopus. However, percentages differ remarkably among areas: in Economics, over 61% of scientific journals and almost 96% of A-Class journals are included in the Scopus database, while on the other hand in Law only 20.3% of scientific journals and 27% of A-Class journals are indexed. Similar results are obtained by Siversten (2014) on Norway's higher education institutions.

Table 3. Classified journals indexed in Scopus

Areas	*N. of Scientific*	*N. of Scopus*	*% Scopus*	*N. A-Class*	*N. Scopus A*	*% Scopus A*
8 - Architecture	2,115	742	35.1%	305	188	61.6%
10 - Antiquities, philology, literary studies, art history	6,488	1,945	30.0%	1,992	925	46.4%
11 - History, philosophy and pedagogy	6,873	2,912	42.4%	1,420	1012	71.3%
12 - Law	2,330	473	20.3%	422	114	27.0%
13 - Economics and Statistics	7,150	4,369	61.1%	1,031	989	95.9%
14 - Political and social science	3,988	1,892	47.4%	1,059	869	82.1%
Total	**20,016**	**8,642**	**43.2%**	**5,476**	**3475**	**63.5%**

Source: Authors elaboration on ANVUR and Scopus data

Journals classified by ANVUR pertain to a large number of Scopus ASJCs': in this respect, table 4 shows the list of the Scopus subject categories including at least 100 journals classified by ANVUR. ASJC are ordered with respect to the number of journals included in the ANVUR classification; the table also reports the number of journals for each ASJC and for each ANVUR scientific Area (see also table 3 for the area definition). Overall, most common ASJCs' are those that are more specific to Humanities and Social Sciences: for instance, 780 journals classified by ANVUR are indexed in the ASJC "Sociology and Political Science" and 735 in "History". However, a non-negligible 426 and 258 journals respectively are indexed in "Medicine (All)" and "Applied Mathematics", respectively, an indication that Italian SSH scholars tend also to publish outside of their more obvious fields of interest.

Table 4. Mapping of asjc covered by ANVUR classification, by Area (only asjc with over 100 journals are reported)

Asjc	*Denominaton asjc*	*8*	*10*	*11*	*12*	*13*	*14*	*Total*
3312	Sociology and Political Science	22	127	159	52	169	251	780
1202	History	59	341	254	14	21	46	735
2002	Economics and Econometrics	17	9	52	23	461	10	572

[3] Scopus has a larger coverage for journals classified by ANVUR with respect to possible alternatives; this result is in line with the literature (see for instance Siversten, 2014) which shows that the coverage of humanities and social sciences is larger in Scopus than in other databases.

1208	Literature and Literary Theory	10	516	18	2	3	4	553
3310	Linguistics and Language	8	508	20	1	7	4	548
1203	Language and Linguistics	8	471	18	1	7	6	511
3316	Cultural Studies	23	319	101	5	22	46	516
3304	Education	15	67	278	5	56	26	447
3305	Geography, Planning and Development	124	37	154	7	71	45	438
1211	Philosophy	9	113	240	14	8	19	403
2700	Medicine(all)	18	35	131	3	232	7	426
3320	Political Science and International Relations	5	33	65	45	74	119	341
3308	Law	3	19	34	191	34	45	326
1408	Strategy and Management	22	4	15	6	225	3	275
2604	Applied Mathematics	12	5	42	0	199	0	258
1403	Business and International Management	11	5	12	7	209	3	247
1213	Visual Arts and Performing Arts	67	137	13	1	7	1	226
1706	Computer Science Applications	33	26	44	1	119	5	228
1201	Arts and Humanities (miscellaneous)	13	59	74	3	37	22	208
2003	Finance	3	0	8	9	187	1	208
2613	Statistics and Probability	4	0	14	0	175	0	193
1212	Religious studies	5	97	52	12	1	15	182
3314	Anthropology	6	76	66	2	11	25	186
3315	Communication	2	73	29	0	25	41	170
3301	Social Sciences (miscellaneous)	3	11	41	5	60	45	165
2308	Management, Monitoring, Policy and Law	53	0	25	12	56	12	158
2739	Public Health, Environmental and Occupational Health	21	6	41	1	79	19	167
1204	Archaeology	28	114	5	0	0	3	150
3303	Development	19	10	37	1	62	19	148
1405	Management of Technology and Innovation	13	0	8	2	121	1	145
3300	Social Sciences(all)	14	24	41	8	30	25	142
2600	Mathematics(all)	6	1	44	1	89	1	142
3302	Archaeology	27	108	7	1	1	3	147
2000	Economics, Econometrics and Finance(all)	4	2	15	7	101	3	132
1712	Software	14	16	23	0	79	0	132
1400	Business, Management and Accounting(all)	6	4	9	5	104	0	128
3204	Developmental and Educational Psychology	1	19	69	0	22	13	124
1407	Organizational Behavior and Human Resource Management	1	3	11	3	96	4	118
1804	Statistics, Probability and Uncertainty	3	0	5	0	109	0	117
1406	Marketing	6	2	2	3	102	1	116
1402	Accounting	2	1	4	6	109	0	122
2719	Health Policy	3	1	35	3	58	12	112
1710	Information Systems	4	11	19	1	68	7	110
1207	History and Philosophy of Science	5	25	63	2	7	6	108
1803	Management Science and Operations Research	6	1	2	0	98	0	107
1105	Ecology, Evolution, Behavior and Systematics	13	17	43	0	31	2	106
3207	Social Psychology	3	14	35	1	28	30	111
3306	Health (social science)	6	3	41	2	22	35	109
1702	Artificial Intelligence	12	21	27	1	42	0	103
2611	Modelling and Simulation	14	4	12	0	72	0	102
1205	Classics	2	94	3	1	0	0	100

Source: Authors elaboration on ANVUR and Scopus data.

Journal classification and the Italian research evaluation exercise

As a next step in our analysis, we check for the use of classified journals in the two main Italian evaluation exercises: table 5 shows the number of scientific journals classified by

ANVUR for which at least 1 article was submitted for evaluation in the VQR 2004-2010 and VQR 2011-14. As it may be seen, the share of journals used in the two evaluation exercises with respect to the total of Scientific journals was on average equal to 23.4% in the first VQR, dropping to 21.1% in the second exercise. Indeed, the fall of the share of classified journals used in the second VQR is a direct result of the fact that the second exercise covered a shorter time span (4 years instead of 7); as consequence, in the second exercise each Italian researcher was asked to submit only two publications instead of three. The share of journals with at least one article evaluated in the two VQRs' reaches 38.5 and 34.9% respectively in the two evaluation exercises in Law, while it is equal to only 27.4% and 24.9% respectively in History, philosophy and pedagogy. Journal articles are a more common media to disseminate knowledge in Law than in History and Philosophy: in the second VQR exercise, the share of journal articles on the total number of publications submitted for evaluation was in fact equal to 39.6% in Law and to 34.9% in History and Philosophy. In this respect, however, Economics and statistics was the area with the highest share of articles with respect to total submissions (72.8%), followed by Political and Social Sciences (40.9%); in Antiquities and literary studies the share was equal to 32.3%, in Architecture to 26.4%[4].

Table 5. Journals classified by ANVUR whose scientific articles was submitted to VQR 2004-2010 and VQR 2011-2014

Areas	*N. of Scientific journals*	*VQR 2004-2010*	*% VQR 2004-2010*	*VQR 2011-2014*	*% VQR 2011-2014*
Architecture	2,115	679	32.1%	579	27.4%
Antiquities, philology, literary studies, art history	6,488	1,898	29.3%	1,542	23.8%
History, philosophy and pedagogy	6,873	1,886	27.4%	1,714	24.9%
Law	2,330	896	38.5%	813	34.9%
Economics and Statistics	7,150	2,077	29.0%	1,991	27.8%
Political and social sciences	3,988	1,376	34.5%	1,284	32.2%
Total	**20,016**	**4,675**	**23.4%**[5]	**4,217**	**21.1%**

Source: Authors elaboration on ANVUR data

A test for robustness of Journal classification

In the previous sessions, we have provided first evidence that the share of indexed Journals is larger for A-Class than for merely scientific journals. This result may be considered as a very preliminary evidence that the ANVUR two-tier classification is robust with respect to the probability for a journal of being indexed in Scopus, which in turn can be considered as a proxy for journal quality. However, it may also be the case that the inclusion in Scopus can have a positive influence on journal classification: in fact, according to the ANVUR requirements, indexation is considered as an explicit criteria of classification in economics and statistics, but we cannot exclude that it may have been used as a reference also in other disciplines. In order to dig deeper into the relationship among ANVUR classification and journal indexation, for every academic discipline (in Italy called "Settore Concorsuale") in which candidates for professorship are selected[6] we compute the mean of the SJR indicator[7] for the two subsets of scientific journals that are classified as No-A Class[8] and A-Class.

[4] See also ANVUR (2017b), table 2.6

[5] The total percentage of scientific journal submitted in the two VQR is lower than the percentages of six areas, because a large number of journal was classified in more than one area.

[6] The list of academic discipline is available at http://attiministeriali.miur.it/media/265754/allegato_a.pdf

Figure 2 shows the box plot distributions of the SJR indicator in the 78 academic disciplines of interest grouped into six scientific areas. The wideness of the box represents the dispersion among academic disciplines in terms of SJR. In area 8 and 12, experts approved a mutual recognition of A-class evaluations among different academic fields; for this reason, boxplot shows no variance among them.

In general, the median value of SJR is clearly higher for A-Class than for merely scientific journals in Economics and Statistics and, even if only marginally, in Architecture; it should however be remembered that, according to ANVUR rules, in Economics and Statistics impact was one of the criteria to be used for journal classification: hence, the fact that A-Class journals show a higher impact (as measured by the SJR) does not come as a surprise. In other areas, the median value of SJR is instead quite similar for the two categories of journals, being indeed marginally higher for scientific rather than for A-Class journals in Antiquities, philology, literary studies and art history. Overall, there is no clear evidence that A-Class journals are characterised by a higher impact with respect to Scientific journals, at least if impact is measured in terms of the value of the SJR indicator.

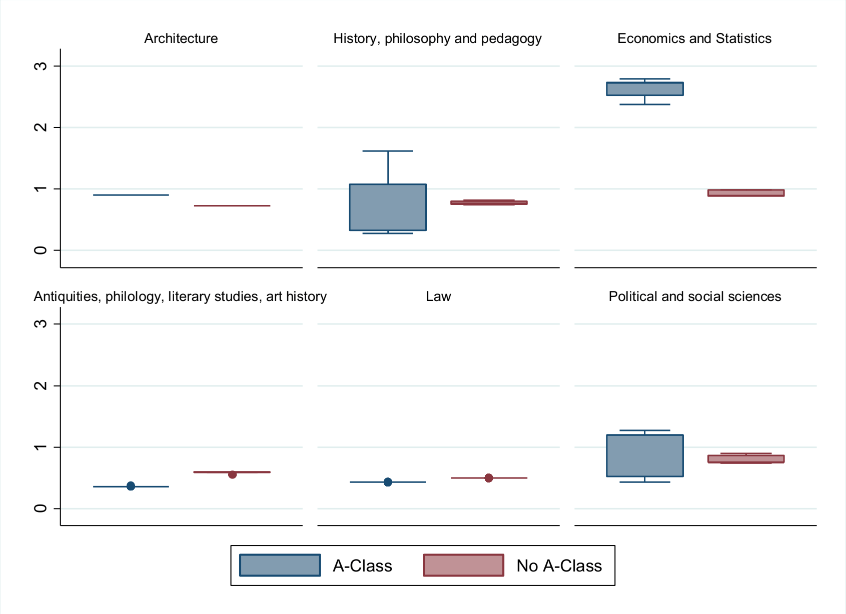


Figure 2. Box plot of SJR (2017) average value of ANVUR journals covered in Scopus in the 78 academic disciplines of interest grouped into six scientific areas

Source: Authors elaboration on ANVUR data

In the two Italian VQRs', articles in SSH were evaluated with peer review. Peer evaluation was coordinated by the 170 members of the disciplinary Groups of Evaluation Experts (GEV

[7] We prefer to use the SJR indicator with respect to possible alternatives like the impact factor or the h index because the SJR allows to take into account the importance or prestige of the journals where citations come from.

[8] The list of scientific journals also includes A-Class journals; in the remaining of the paper, we compare A-Class journals only with the scientific journals that are not also comprised in the A-Class list.

in the Italian acronym) in SSH, assisted by a very large group of national and international reviewers (considering only SSH, 7.263 and 8.343 reviewers were respectively used in the two evaluation exercises). Reviewers were asked to assess the merit of each article against the criteria of originality, methodological rigour and impact: in this sense, VQR evaluation is based on criteria that are quite different from those adopted in the evaluation of journals, as reported in the first section of the paper. The only exception is the area of economics and statistics, were VQR evaluation was explicitly based on journal impact, a parameter that in this area is also used in journal evaluation. The independence of the VQR evaluation with respect to the journal evaluation is also confirmed by the fact that the group of experts in charge of journal classification is only a small subset of the much larger group of VQR GEV and reviewers. Moreover, of the 42 experts forming the group in charge of journal classification, six out of them were also part of the GEVs. However, it should be considered that experts and reviewers had access to the venue of the article they valued, hence we cannot completely rule out that it may have influenced their opinion somehow. Figures 3 and 4 show the box plot distribution of the marks received in the two VQR exercises by the articles published in A-Class and Non-A Class journals. In total, we consider 34.929 articles (19.323 in the VQR 2004-10 and 15606 in the VQR 2011-14). In this case, there is quite a clear evidence, both in the VQR 2004-2010 and in the VQR 2011-14, that the papers published in A-Class journals received an higher mark than those published in no-A Class journals. This appear to be particularly true in Economics and Statistics, but a remarkable difference does emerge also for Architecture, History and Philosophy and Law. Differences tend to be quite similar in the two VQR exercises.

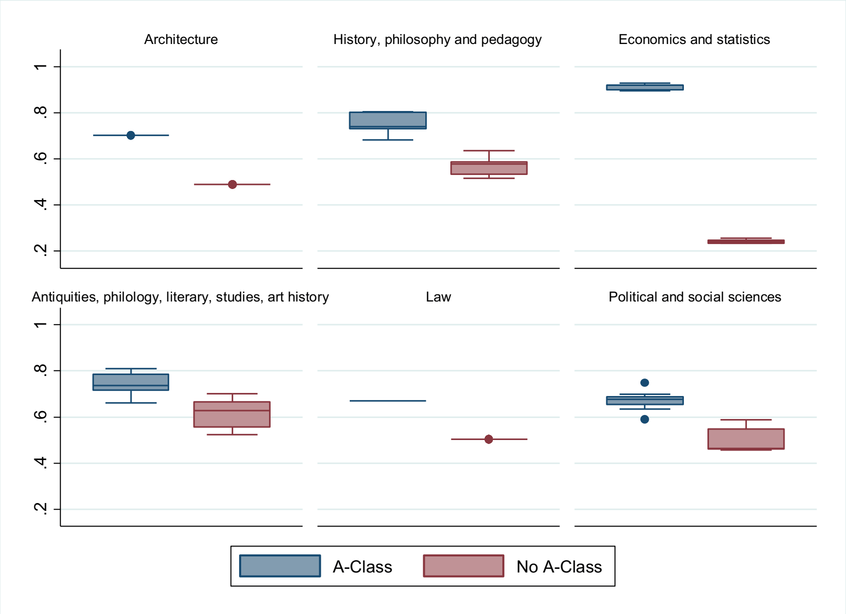


Figure 3. Box plot of VQR average value of ANVUR journals submitted to VQR 2004-2010 in the 78 academic disciplines of interest grouped into six scientific areas

Source: Authors elaboration on ANVUR data

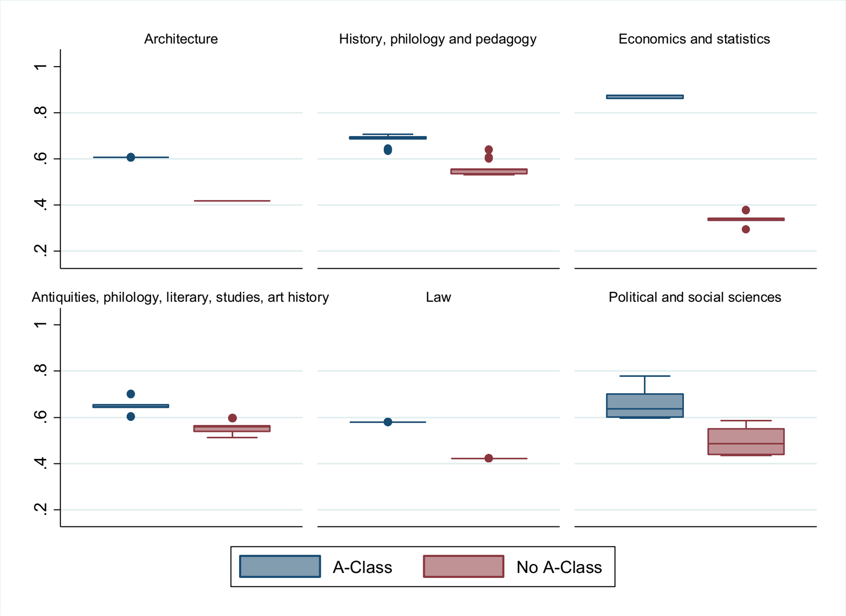


Figure 4. Box plot of VQR average value of ANVUR journals submitted to VQR 2011-2014 in the 78 academic disciplines of interest grouped into six scientific areas

Source: Authors elaboration on ANVUR data

As a final step, we propose a simple test for evaluating the statistical significance of the differences among A-Class and no-A Class journals in terms of the value of the SJR indicator and of the marks received in the two VQR exercises. The null hypothesis is that the score (SJR or VQR) of A-Journals is equal to that of No Class A journals, and this is tested against the one-side alternative that the score is higher than that of No Class A journals:

$$H_1 = A\ rated\ journals\ average\ score > No\ A\ rated\ journals\ average\ score$$

The test is calculated as follows:

$$t_{calculated} = \frac{A\ rated\ journals\ average\ score - No\ A\ rated\ journals\ average\ score}{s_{pooled}\sqrt{\frac{1}{n.of\ A\ rated\ journals} + \frac{1}{n.of\ NO\ A\ rated\ journals}}}$$

Table 6 shows the results: the difference among VQR results for Class A and non Class A journals is almost always statistically significant, for all sectors and Area and for both VQR exercises; the only exceptions are a sector in History and Philosophy and two in Political and Social Sciences. On the other hand, the difference among Class A and non Class A journals in terms of SJR is never statistically significant in Law and in Antiquities, philology, literary studies art and history; mixed results emerge in other areas.

Table 6. Statistical test (test t) between average scores of A rated journals and not A rated journals in the 78 academic disciplines of interest grouped into six scientific areas

Areas		IrVQR1	IrVQR2	SJR2017
Architecture	n. of significative test	5 out of 5	5 out of 5	5 out of 5
	(min) t test	5.383	6.280	1.772
	(max) t test	5.383	6.280	1.772
	(min) p-value	**0.000**	**0.000**	**0.038**
	(max) p-value	**0.000**	**0.000**	**0.038**
Antiquities, philology, literary studies, art history	n. of significative test	20 out of 20	20 out of 20	0 out of 20
	(min) t test	2.354	2.335	2.936
	(max) t test	5.286	3.529	4.107
	(min) p-value	**0.000**	**0.000**	0.998
	(max) p-value	**0.010**	**0.011**	1.000
History, philosophy and pedagogy	n. of significative test	12 out of 13	12 out of 13	8 of 13
	(min) t test	0.850	0.053	0.088
	(max) t test	6.775	5.633	4.992
	(min) p-value	**0.000**	**0.000**	**0.000**
	(max) p-value	0.199	0.521	1.000
Law	n. of significative test	17 out of 17	17 out of 17	0 out of 17
	(min) t test	6.722	5.984	0.526
	(max) t test	6.722	8.170	0.754
	(min) p-value	**0.000**	**0.000**	0.700
	(max) p-value	**0.000**	**0.000**	0.775
Economics and Statistics	n. of significative test	15 out of 15	15 out of 15	15 out of 15
	(min) t test	9.525	7.966	17.835
	(max) t test	27.443	26.376	22.674
	(min) p-value	**0.000**	**0.000**	**0.000**
	(max) p-value	**0.000**	**0.000**	**0.000**
Political and social science	n. of significative test	6 out of 8	7 out of 8	5 out of 8
	(min) t test	0.048	1.562	0.887
	(max) t test	6.318	6.361	5.194
	(min) p-value	**0.000**	**0.000**	**0.000**
	(max) p-value	0.481	0.061	0.999

Source: Authors elaboration on ANVUR data

Concluding remarks

Using a large dataset of over 20 thousands journals and 35 thousands articles, we have checked for the robustness of the journal classification performed by ANVUR in the framework of National Habilitation programme with respect to both journal impact (as measured by SJR) and the results of peer evaluation of individual articles, performed by ANVUR in the context of two consecutive national evaluation exercises. First of all, we find that there is no direct relationship among journal classification and journal impact: in other words, experts in charge of journal evaluation (with the only exception of those in the area of Economic and Statistics) does not seem to rely on external information on journals impact to perform their task, rather preferring to base their ratings on other process and product requirements (including, among other things, regularity of publication, Board composition, diffusion in the scientific community, rules of access, scientific content and international scope). The resulting classification, however, seems to offer on average a reliable proxy for the quality of individual articles: indeed, the marks obtained in two independent evaluation exercises by articles published in top tier journals are significantly higher – regardless of the scientific field – with respect to those obtained by articles published in non-top journals. Given the fact that peer evaluation is rather costly and time consuming, the latter result seems to encourage the possibility of using some measure of journal classification in order to support peer review in the context of general systemic research evaluation exercises, even in those disciplines (like Humanities and Social Sciences) where "pure" peer evaluation is usually considered as the only method of choice.

References

Ancaiani, A. et al. (2015) 'Evaluating Scientific Research in Italy: The 2004–10 Research Evaluation Exercise', Research Evaluation, 24/3: 242–55.

Anfossi A., Ciolfi A., Costa F., Parisi G., Benedetto S. (2016), Large-scale assessment of research outputs through a weighted combination of bibliometric indicators, Scientometrics, vol. 107, issue 2, pp. 671-683

ANVUR (2017a), *Regolamento per la Classificazione delle Riviste nelle aree non bibliometriche*, http://www.anvur.it/wp-content/uploads/2017/10/RegolamClassificazRiviste~.pdf

ANVUR (2017b), *Evaluation of Research Quality 2011-14, ANVUR Final Report*, http://www.anvur.it/wp-content/uploads/2017/06/VQR2011-2014_Final%20Report.pdf

Bonaccorsi A., Ferrara A., Malgarini M. (2018), "Journal Ratings as Predictors of Article Quality in Arts, Humanities and Social Sciences: an Analysis Based on the Italian Research Evaluation Exercise", in Bonaccorsi A. (Ed.), *The Evaluation of Research in Social Sciences and Humanities*, Springer

Ferrara A., Bonaccorsi A. (2016), *How Robust is Journal Ratings in Humanities and Social Sciences? Evidence from a Large-scale, Multi-method Exercise*, *Research Evaluation*, Volume 25, Issue 3, 1 July 2016, Pages 279–291,

Lee C.J., Sugimoto C.R., Zhang G., Cronin B., *Bias in Peer Review*, Advances in Information Science, Vol. 64, Issue 1, pp. 2-17

Moed, H. F. (2008) 'Research Assessment in Social Sciences and Humanities - Evaluation in the Human Sciences', Paper presented to the Bologna Conference, December 12–13.

Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science. In *Proceedings of the science and technology indicators conference* (pp. 598-604).

Walker, J., Fenton, E., Salandra, R., & Salter, A. (2018). What influences business academics' use of the Association of Business Schools' (ABS) list? Evidence from a survey of UK academics. *British Journal of Management*. https://doi.org/10.1111/1467-8551.12294

Should I move to diversify my scientific network?
A panel analysis of chemists' careers

Marine Bernard[1], Bastien Bernela[1], Marie Ferru[1] and Béatrice Milard[2]

[1] *University of Poitiers, CRIEF EA2249, Poitiers, France*
marine.bernard@univ-poitiers.fr - bastien.bernela@univ-poitiers.fr – marie.ferru@univ-poitiers.fr

[2] *University of Toulouse Jean Jaurès, LISST UMR5193, Toulouse, France*
milard@univ-tlse2.fr

Abtract

This study assesses the impact of geographical mobility of academic researchers on the diversification of their co-authorship network, though a bibliometric analysis, at the micro-scale. We selected 80 chemists (CNRS Research Directors and Full Professors) from two French labs: we collected their 9310 publications, in which they collaborated 41628 times with 14783 co-authors. After manually searching for the location of chemists and their co-authors for each publication, we create a database of the 80 personal co-authorship networks, in which each line represents one collaboration between the researcher and one co-author. To analyze the formation and dynamics of these networks over time and the geographic mobility of researchers, we adopted a panel approach to take into account generations and career stages. Results suggest that the social and geographical patterns of scientific network is not so dependent on the mobility of researchers. We point the stability of the volume of new co-authors and of the researcher's lab colleagues per article (whatever generations and career stages). Mobility has only an instant effect on network openness and restricted to the new city attended; it also weakens the interpersonal relations from previous attended cities.

Introduction

A literature interested in the geography of science (Livingstone, 1995; Eckert and Baron, 2013) has been developed in recent years to better understand the spatial distribution of scientific activities. A crucial issue deals with the question of scientists and inventors' mobility, that is considered as essential for knowledge diffusion and widely encouraged by public authorities, especially in the European Union, through various mobility programs (Morano-Foadi, 2005).

Behind the mobility-knowledge diffusion relationship, the underlying hypothesis is that knowledge is attached to individuals (Cañibano, 2006) and it therefore travels along with people who master it. In this context, it is logically assumed that by being mobile, researchers will meet and collaborate with new researchers with whom they will exchange knowledge. This scientific network diversification driven by mobility, would be therefore the indispensable ingredient for production of new knowledge and its circulation across space. In this article, we propose to question this thesis and to answer the following question: should researchers move to diversify their network? Network diversification could be understood in terms of persons/skills (new partners) and in terms of geography (new cities).

Moving in this direction suggests studying the formation and dynamics of personal scientific networks of researchers and identifying the contribution of researchers' geographical mobility to this dynamic. Whereas scientific network dynamics is largely studied in the framework of geography of science into a macro level perspective (most often at the level of cities and countries), it is rarely considered at the micro level of individual careers. Authors have mainly focused on the problem of "brain drain", from Bhagwati (1976) to Breschi *et al*. (2017): the question of mobility has been addressed through the analysis of migratory flows between territories, and nodes in networks are no longer individuals but organizations, cities or countries. In parallel, some research has developed with a greater focus on individuals, testing mainly the impact of researchers' mobility on their performance (Hoisl, 2007; Lawson and Shibayama, 2015). More marginally, some authors have sought to account for the link between mobility and scientific network. Fontes (2012), based on the study of doctoral students' trajectories, shows that international mobility during the PhD thesis enriches the researcher knowledge network. Woolley *et al.* (2009) verifies this result for the post-doctoral mobility of Asian researchers, while Melin (2004, 2005) nuances it ("the dark side of mobility") through the study of Swedish researchers' mobility. Our research is in line with these latter questions. We seek to test the role of mobility in building and developing the scientific network by focusing at a micro level; however, we consider that mobility cannot be studied in isolation. It seems crucial to reintegrate mobility' impact in the context of the ongoing career of the researchers. To explain the scientific network diversification, we therefore question the impact of career dynamics, including mobility.

Empirically, we focus on scientific publications of researchers since it constitutes a relevant indicator of successful collaborations even if they do not capture all researchers' collaborations. Such bibliometric data generally gives reliable information on institutional (laboratory) scale, however those related to the individual scale of researchers remain problematical. Aware of these limits, we use precise data at the individual level from a cohort of eighty chemists located in two French labs. To refine the bibliometric data, we notably have collected the affiliation and the location of each co-author to understand the geography of co-authorship. Thus, we can analyze the formation and dynamics of co-authorship personal network over time. Under which conditions do they diversify their co-authorship, or do they tend to renew collaborations? To what extent is the geography of co-authorship sensitive to the spatial trajectories of chemists? Are scientific collaborations maintained (or not) over time despite geographical distance?

We focus on chemistry, since the co-authorship in this field is particularly dense due to the way research is practiced (experimental dimension, need for technical skills, importance of teams, etc.). Even so, it is not a scientific field in which the average number of co-authors (between 4 and 5) is tremendous. The eighty studied chemists have published around 10000 articles with 15000 co-authors since the beginning of their career. The paper is structured as follows. The next section describes the data, gives descriptive statistics, and present the methodology used to assess the mobility impact on the co-authorship dynamics. The empirical results are commented on in the fourth section. The final section summarizes the main findings and draws some conclusions.

Data and Method

In order to resolve the literature issues identified in the previous part, we choose to grasp the researcher career of 80 chemists and their scientific collaborations *via* co-authorship (Katz, 1994) thanks to a bibliometric analysis at the micro-scale. We therefore analyze the social and geographic diversification of the co-authorship network of researchers on a fine scale along the career.

The case study of 80 chemists' careers

We selected 80 researchers from two French academic laboratories in chemistry, who were CNRS[1] Research Director or Full Professor in spring 2018 (Table 1). These positions will be referred in the paper as "senior researchers" to simplify the vocabulary. This population was chosen for its comparability. It guarantees a certain work experience and a minimum career longevity: all studied researchers are at an advanced stage of their careers and well-established in their scientific field[2].

[1] French National Scientific Research Center (Centre National de la Recherche Scientifique)

[2] The French system can be characterized as a centralized and regulated labor market controlled by disciplines, close to the Italian and Spanish ones and very different from the US (Pezzoni *et al.*, 2012). Academic careers in France can be divided into three main phases: i) PhD thesis and postdoctoral positions (fixed-term positions) ; ii) "Maître de Conférences" and "Chargé de Recherche CNRS" which corresponds to "assistant professor"; access

In France in 2012, the average age of recruitment for assistant professor in chemistry was around 31 years old and those of full professor was around 42 years old[3].

The two studied labs are IC2MP (Institut de Chimie des Milieux et Matériaux) located in Poitiers, and LCC (Laboratoire de Chimie de Coordination) located in Toulouse (two cities in Western France). These structures share some similarities. Firstly, they benefit from an old history and have acquired an international reputation in their respective scientific specialties. IC2MP is the result of the merger in 2012 of five co-located laboratories, that were active for decades. LCC was born in the 1974 and has grown over decades without major changes in terms of organizational and scientific patterns. Secondly, the two labs are about the same size: each of them gathers around 250 members, including around 100 permanent ones. Although the total number of senior researchers is almost the same for the two labs (38 for IC2MP and 42 for LCC), their internal structure is quite different: IC2MP has few but big teams (5 research teams, with an average of 7.6 senior researchers) whereas the LCC has many but small teams (16 research teams, with an average of 2.6 senior researchers). In addition, IC2MP is a joint research center (CNRS-University) whereas LCC is a unit specific to the CNRS: this justifies the overrepresentation of full professors in Poitiers (29 out of 38 senior researchers) and of CNRS research directors in Toulouse (26 out of 42 senior researchers). In terms of gender, 72,5% are male, which is consistent with national distribution. 62% of the French academics in chemistry were male in 2018[4]: this share moves from 56% for assistant professors to 75% for full professors, highlighting the glass ceiling in academic careers (Sabatier, 2010).

Table 1. Descriptive about the population

	CNRS Research Director	*Full professors*	*Total*
IC2MP - Poitiers	Men: 6	Men: 26	Men: 32
	Women: 3	Women: 3	Women: 6
	Total: 9	*Total: 29*	*Total: 38*
LCC – Toulouse	Men: 14	Men: 12	Men: 26
	Women: 12	Women: 4	Women: 16
	Total: 26	*Total: 16*	*Total: 42*
Total	Men: 20	Men: 38	Men: 58
	Women: 15	Women: 7	Women: 22
	Total: 35	*Total: 45*	*Total: 80*

We collected all the publications (articles, reviews, letters, notes and editorial materials[5]) of the 80 senior researchers from the Web of Science Core Collection. Thomson Reuters' Web of Science is a well-known and reliable source in science studies to collect publications, frequently used to collect co-publication data. An initial cleaning was performed to avoid namesake bias: we kept 9310 publications, including 9206 co-publications. Most of variables in the publication database have been kept on its own (for instance the title, the year, the scientific journal, the list of co-authors), and some other variables needed an in-depth cleaning, as geography (see later). From the publication list of a scientist, we split co-authored articles into co-authorship pairs (Melin and Person, 1996). As we stand on a personal network perspective, we only interest in *ego-alter* relationships and do not consider relationships between *alter*: *ego* is one of the 80 studied

requires the person to hold a PhD; iii) "Professeur des Universités" and "Directeur de Recherche CNRS" which corresponds to "full professor", access requires the person to hold a "Habilitation à Diriger des Recherches" (HDR).

[3] According to the French Ministry of Higher Edcation and Research : https://cache.media.enseignementsup-recherche.gouv.fr/file/statistiques/05/3/orig2012_302053.pdf

[4] According to the French Ministry of Higher Edcation and Research : http://www.enseignementsup-recherche.gouv.fr/cid129560/fiches-demographiques-des-sections-de-sciences-2017.html

[5] Articles represent 99,4% of the total number of publications.

researchers and *alter* are all the co-authors. Concerning scales of analysis, it is important to distinguish between "cosignature" and co-author: in cases where *ego* co-signs ten times with an *alter*, this relation weights one co-author and ten cosignatures.

Finally, as we focus on the publication and collaboration dynamics, we use the year of the first publication as the beginning of work career of each researcher. Although the 80 senior researchers all occupy the same position in spring 2018, they have unequal career length and started publishing at different decades: the longest career is 51 years and the shortest one is 12. We generated cohorts, gathering together researchers in periods of 10 years, according to their first year of publication. This allows integrating time in two complementary dimensions: the generation and the career stage. Indeed, senior researchers that occupy a senior position today started their careers in different historical contexts: the way science is conducted, and the place of publication activity have hugely changed since 1960's. Also, thanks to a panel approach, the career of each researcher is sequenced in years (the first year of publication is considered as t_0, the second year t_1, and so on) since the propensity of a researcher to develop networks is not the same according to its career stage.

Table 2. Publication and collaboration patterns according to gender, position, generation

			Gender		**Position**		**Generation**			
		All researchers	*Male*	*Female*	*Full professors*	*Research directors*	*1967-1976*	*1977-1986*	*1987-1996*	*1997-2006*
	Number of researchers	**80**	58	22	45	35	13	18	28	21
	Average length of career (years)	**31**	31	29	30	31	46	37	28	19
Publication patterns	*Number of publications*	**9310**	7316	1994	4208	5102	2865	2406	2634	1405
	Number of co-publications	**9206**	7219	1987	4137	5069	2832	2342	2634	1398
	Average number of publications per researcher	**116**	126	91	94	146	220	133	94	66
	Average number of publications per year and researcher	**3,78**	4,01	3,13	3,09	4,64	4,78	3,61	3,33	3,48
Cosignature patterns	*Number of cosignature*	**41628**	31 869	9 759	18 181	23 447	11 157	9 796	13 265	7 410
	Average number of cosignature per researcher	**520**	549	444	404	670	858	544	473	352
	Average number of cosignature per co-publication	**4,52**	4,41	4,89	4,39	4,63	3,93	4,18	5,04	5,3
	Number of co-authors	**14 783**	11 132	3 651	6 908	7 875	3 727	3 317	4 880	2 859
	Average number of co-authors per researcher	**185**	192	166	153	225	286	184	174	136
	Average number of new co-authors per co-publication	**1,61**	1,54	1,84	1,67	1,55	1,32	1,42	1,85	2,05
	'One-off' co-authors	**8574 (58%)**	6361 (57%)	2213 (61%)	4033 (58%)	4541 (58%)	2147 (58%)	1899 (57%)	2822 (58%)	1706 (60%)

Table 2 details descriptive statistics regarding their publication activities and co-authorship: the 80 senior researchers co-published 9206 articles in which they cosigned 41 628 times with 14 783 different co-authors. At the individual scale, each senior researcher published in average 116 publications in their career and 3,78 publications per year. For each publication, they collaborate in average with 4,52 co-authors, including 1,61 new co-authors. These descriptive statistics demonstrate a relative high rhythm of publication activities, and a very collaborative research. Many co-signatures consist of bringing together different scientific expertise not all held by one single person or team. In terms of volume, the majority of co-authors (58%) are 'one-off' partners, *i.e.* *alter* with whom *ego* collaborated only once. Bernela and Milard (2016) showed these 'one-off' co-authors often correspond to "second-rank" co-authors, that are often non-permanent positions and unknown by *ego*: the dominance of team level implies that some co-authors from partner teams are involved in the publication without interaction with *ego*.

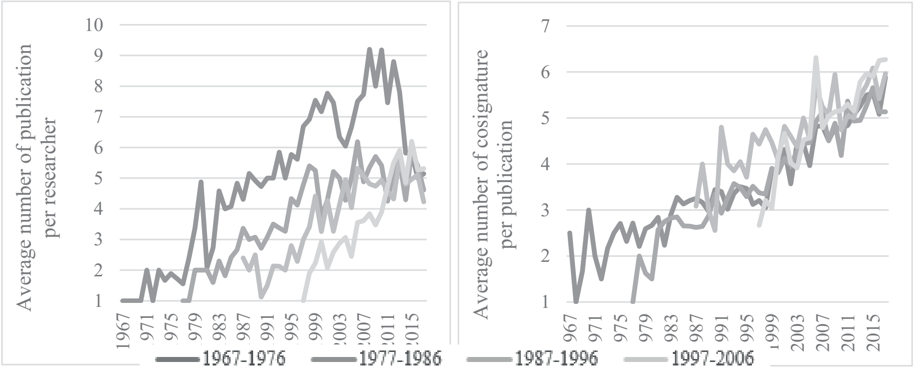


Figure 1: Dynamics of publication and cosignature patterns

Then, Figure 1 illustrates the dynamics of publication and cosignature patterns by generation. Although senior researchers did not start their career in the same historical context, all generations follow the same rhythm of annual publication all along their career. The annual rhythm of publication varies according to the career stage of the senior researcher and not to the generation. However, the collaborative dimension of publications depends on both the career stage and the generation of the senior researcher. Indeed, the average number of cosignature per publication increases throughout the career of each senior researcher. At the same time, the recent historical context encourages senior researchers to publish with more co-authors. For instance, in the 2000's, although the fourth generation just started to publish and the first generation started it more than thirty years ago, their publications had the same collaborative dimension.

From spatial information in publication data to mobility patterns and the geography of co-authorship

To reach our article objective - capturing the effects of mobility on the diversification of co-authorship - we need to define i) mobility patterns of the 80 senior researchers and ii) the geographical distribution of their scientific network (through the location of their *alter*) from the spatial information available in the WoS.

Before 2008, basic spatial information contained in WoS refers to the affiliations of co-authors in the form of a list of addresses, and there is no possibility to identify the address to which each author belongs. Moreover, the number of addresses does not necessarily correspond to the number of authors (several co-authors belonging to the same team, multi-affiliations of an author, etc.). Methodologically speaking, this absence of author-location matching implies that most studies focus on the address level only - how many cities collaborate in a publication? – to provide an understanding of the geography of science (Grossetti *et al.*, 2014; Maisonobe, 2015). It becomes a hindrance when studying geographical patterns of scientific trajectories. In our case, we used web search engines to collect all the locations of co-authors listed on a publication, at the scale of laboratory, city and country. This represents a time-consuming task of disambiguation, standardization and coding of database, necessary to obtain reliable geographical information. To our knowledge, it is the only way to question the geography of collaborations at the micro-level of researchers and has not be done yet, excepted Bernela and Milard (2016).

Among the corpus of 9206 co-publications, we resorted to web searches to allocate the location of *ego* and *alter* for had to manually search on the web for the 5 301 co-publications done before 2008. This job is already done for the 3905 publications after 2008. For each address, we kept three scales of geographical information: the laboratory, the city and the country. Concerning the laboratory, there was an important work of disambiguation: for instance, in the case of the Poitiers lab, “IC2MP” can also be registered as “UMR 7285” or the various names of old laboratories before merger. Finally, we kept multi-affiliations cases, i.e. when *ego* and/or *alter* is located in different laboratories in the same publication. In other words, a same couple *ego-alter* can be repeated several times for a publication but with a different geography, ensuring that the sum of co-signatures weights is equal to 1. Once locations of *ego* and their co-authors were cleaned for each publication, we can code the geographical distance between each pair of *ego-alter*. We break down this measure of distance into four increasing levels: lab, city, national, international. According to affiliations mentioned on publication, is *ego* located in the same laboratory, the same city, or the same country of *alter*? It allows to capture to what extent *ego* and *alter* share a common geographical environment. This micro-scale collection of data, consisting in coding the geographic distance between a researcher and its co-authors for each dyadic relationship makes our database an original tool for the geography of science.

Then, mobility is simply considered as a change in *ego* affiliation from a year to another, at the city level[6]. We are able to distinguish between infra-national and international moves. We identify the number of mobility along researchers’ careers (Figure 2): 19 out of 80 can be considered as sedentary researchers, i.e. without any change in location all over their career. We observe a concentration of mobility at the beginning of the career, highlighting the central role of postdoctoral positions in chemistry. In volume, there is as much international mobility as national ones. Some has never experienced mobility whereas others have moved a lot over their career, but it would not be satisfying to build binary categories of mobile/non-mobile scientists, as done in many articles (Aksnes *et al.*, 2013; Franzoni *et al.*, 2014; Chinchilla-Rodriguez *et al.*, 2018). Even mobile people were sedentary before their first mobility justifying the necessity to contextualize these events all along researchers’ careers to evaluate properly the impact of mobility on the diversification of co-authorship network. Moreover, we increment information about all the cities attended by *ego* in the past and are thus able to contextualize each publication (and co-authorship pair) in the geographic trajectory of the researcher. In a dynamic point of view, we can assess the continuance of relationships with *alter* from cities where *ego* was located before.

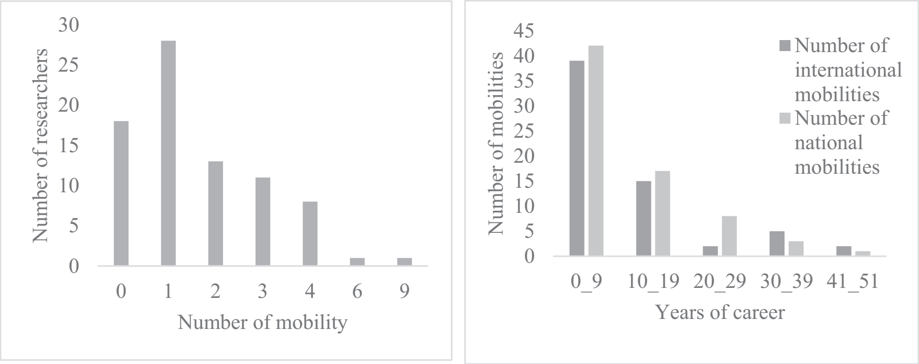


Figure 2: Number of mobility per researcher and during the career

[6] We do not observe institutional mobility within a same city, excepted changes of the lab name in the case of the merger in Poitiers.

Social and geographical dynamics of co-authorship over time

Before testing the impact of mobility on network diversification, it is important to understand the social and geographical patterns of co-authorship over the career (Figure 3): Figure 3a refers to the balance between new *alter* and former ones (*i.e.* with whom *ego* renews co-signatures) while Figure 3b refers to the spatial distribution of *ego-alter* relationships. On both figures, the upper curve refers to the growing trend of the annual mean of co-authors per publication over the career. Whatever generations, as researchers advance their career, the number of co-authors per publication rises.

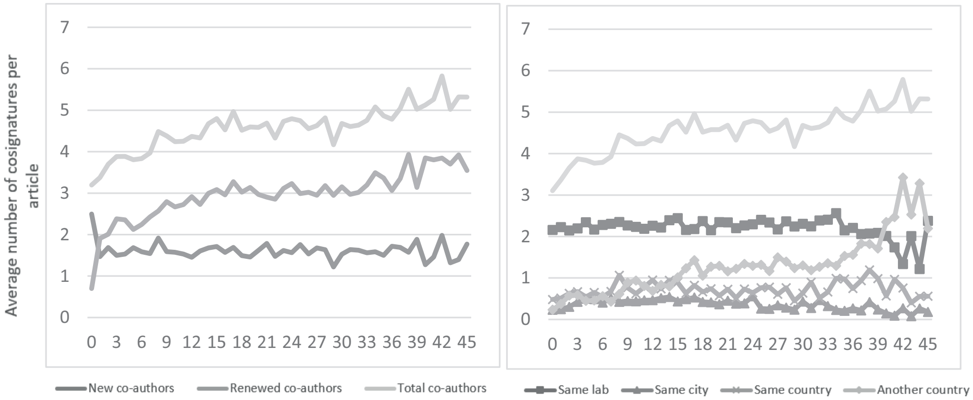


Figure 3: Renewal of co-authorship throughout careers / Geography of ego-alter relationships over time

The rise of co-authors per article throughout career stages is due to the continuance of relationships with some former co-authors, while the average number of new co-authors per article is very stable (around 1.5) over the career. Therefore, when they pursue their career, researchers do not increasingly collaborate with new co-authors but mostly with former ones as already observed by Cabanac *et al.* (2015). Concerning the spatial distribution of co-authors over the career and the distance separating *ego* and *alter* (Figure 3b) we note the important and stable weight of the *ego*'s lab colleagues (local co-authors) in the publication activity all along the career. Although non-local *alter* are a minority, their weight raises throughout the career: geographical openness of co-authorship occurs progressively with a frank internationalization for all generations. The densification of co-authorship is characterized by both a reinforcement of lab colleagues' core and an internationalization of co-authorship.

Table 3: Geography of *ego-alter* relationships

		Geography of *ego-alter* (%)						**Location of *alter* compared to *ego* trajectory (%)**				
		Same lab	Same city	Same country	Other country	Not found	Total	Current city	Attended city	Non-attended city	Not found	Total
New co-authors	n=14783	32,6	9,9	22,8	33,9	0,8	100,0	42,5	3,1	53,5	0,9	100,0
Total co-signatures	n=41628	49,8	8,7	16,1	25,1	0,3	100,0	58,5	3,3	37,9	0,3	100,0

Table 3 proposes a cross-analysis of the geography of co-authorship and the geographical trajectory of *ego*. If we look at the geographical context of the first cosignature between *ego* and *alter*, 42,5% of them happened with *alter* who was in the current city of *ego*. Only 3,1% are *alter* located in a former city of *ego*: after a mobility, *ego* does not seem maintain relationships with cities attended in the past to integrate new *alter* in its network. The main configuration (53,5%) is *alter* from a city where *ego* was never affiliated before, suggesting that mobility is far from being a prerequisite for the opening of the scientific network. Secondly, we observe the high “contribution” of local relationships: *alter* located in the city of *ego* at the moment of the publication represent 42,5% of co-authors and 58,5% of cosignatures: *ego* maintains relationships longer with local co-authors, with whom he cosigns on average more articles than non-local ones.

Econometric estimation and results

Our objective is to capture the impact of mobility on the diversification of network. The panel approach seems to be the most appropriate way to study individual evolutions based on our longitudinal data. We introduce an individual dimension, that represents each *ego* ($i \in N\,[0;80]$) and a time dimension ($t \in N\,[0;51]$), that represents each year of career of ego (the longest career is 51 years). After carrying out a Hausman test we chose to adopt a fixed-individual effect model. Indeed, the “within model” seems appropriate to focus on intra-individual evolutions: the focus on dependant and explicative variables at the individual scale permits to capture significant variations of the 80 trajectories whatever heterogeneity of publication and co-signatures patterns of researchers.

In order to capture the diversification of the co-authorship network, we measure annual flows of entrant and outgoing co-authors in the network. These two complementary dimensions allow us to evaluate the potential gains and losses of mobility experience on the network of senior researchers. Also, we chose to measure these flows as regards the whole network of *ego* (all *alter*) and more specifically *alter* coming from non-attended cities by *ego*. Then, we are able to evaluate the impact of mobility on *alter* who were integrated through the geographic trajectory of *ego* or not. So, our models estimate these following variables: $Y^{1}_{i,t}$ (average number of <u>new *alter*</u> per publication of the *ego i* in the year *t*), $Y^{2}_{i,t}$ (average number of <u>outgoing *alter*</u> per publication of the *ego i* in the year *t*), $Y^{1'}_{i,t}$ (average number of <u>new *alter*</u> coming from a non-attended city by *ego* per publication of the *ego i* in the year *t*) and $Y^{2'}_{i,t}$ (average number of <u>outgoing *alter*</u> coming from a non-attended city by *ego* per publication of the *ego i* in the year *t*). Table 4 details descriptive statistics of these variables.

Table 4: Descriptive statistics of dependent variables

	Direction of the flows	Variables	N	Mean	Std Error	Min	Max
All alter	Entry	$Y^{1}_{i,t}$	2099[7]	1,749	1,391	0	12
	Exit	$Y^{2}_{i,t}$	2099	1,728	1,654	0	15
Alter coming from a non-attend city	Entry	$Y^{1'}_{i,t}$	2099	0,857	1,101	0	10
	Exit	$Y^{2'}_{i,t}$	2099	0,822	1,187	0	10,5

Thanks to two temporal nominal variables, we characterize the mobility of the senior researcher in each year along the career. The first variable describes the dimension of the flow of mobility and the second variable reminds the year of the last event of mobility. These variables are complementary because they evaluate two effects of the mobility: immediate and delayed effects. In the results, we name models using the variable $X^{1}_{i,t}$ as model (a) and models using the variable $X^{2}_{i,t}$ as model (b). Table 5 describes the different modalities of these variables. We control our panel

[7] This value corresponds to the sum of the length of the 80 careers.

models with temporal variables dealing with patterns of publication and co-signatures of *ego i* : $Cosign_{i,t}$ (moving average of the number of co-signature per article of ego *i* between *(t-1), t* and *(t+1)), $Publis_{i,t}$* (moving average of the articles per year of ego *i* between *(t-1), t* and *(t+1)), $Geo\_net_{i,t}$* (number of cities of *alter* with which ego *i* has cosigned until (*t-1*)), $Social\_net_{i,t}$ (number of *alter* with which ego *i* has cosigned until (*t-1*)) and $Citat_{i,t}$[8] (number citations *ego i* received in (*t-1*)).

Table 5: Mobility variables

Variables	Modalities	Description of modalities
$X^1_{i,t}$ What kind of mobility ego *i* did in the year *t* ?	"not_mob"	Ego *i* kept the same city affiliation between *(t-1)* and *t*
	"National"	Ego *i* changed his city affiliation in *t* related to *(t-1)*
	"Internat"	Ego *i* changed his country affiliation in *t* related to *(t-1)*
$X^2_{i,t}$ When did happen the last event of mobility before the year *t* in the career of ego *i* ?	"sedentary"	Ego *i* never has been mobile until *t*
	"this_year"	Ego *i* has been mobile in *t*
	"1_to_3"	Ego *i* has been mobile between (*t-1)* and *(t-3)*
	"4_to_6"	Ego *i* has been mobile between (*t-4)* and *(t-6)*
	"7_to_9"	Ego *i* has been mobile between (*t-7)* and *(t-9)*
	"10+"	Ego *i* has been mobile before *(t-10)*

Table 6: Descriptive statistics of explanatory and control variables

			N	Mean	Min	Max
Mobility trajectory: explanatory variables	$X^1_{i,t}$	not_mob	1965	0,94	0	1
		National	71	0,03	0	1
		Internat	61	0,03	0	1
	$X^2_{i,t}$	sedentary	801	0,38	0	1
		this_year	130	0,06	0	1
		1_to_3	271	0,13	0	1
		4_to_6	213	0,10	0	1
		7_to_9	173	0,08	0	1
		10+	511	0,24	0	1
Publication and cosignature patterns: control variable	$Cosign_{i,t}$		2099	4,31	1,00	17,00
	$Publis_{i,t}$		2099	4,27	0,33	24,67
	$Geo\_net_{i,t}$		2099	17,26	0,00	116,00
	$Social\_net_{i,t}$		2099	81,19	0,00	673,00
	$Citat_{i,t}$		2099	7038,19	0,00	267733,00

Results

Firstly, we observe that the mobility event (national or international one) stimulates the integration of new *alter* during the year of mobility. Indeed, in the model (1.a), the mobility gives a higher probability to have co-signatures with new *alter* (in comparison to years of non-mobility). However, this positive effect of mobility in the network diversification has a very limited duration. In the model (1.b), the mobility has a positive and significant impact only the year the mobility event occurs. The first econometric result is that integration of a new city has only an instant impact on the integration of new *alter*.

The model (2.b) shows that the mobility event has also a positive and significant impact on *alter* exits. Mobility has a more lasting effect on exit of co-authors than on entry, as the positive significant effect is observable for a mobility that occurred until three years before (model (2.b)). Mobility is both synonym of integration and destruction of relationships.

[8] This variable is based on the assumption of a linear distribution of annual citations, from the stock of citations of each publication received in 2018.

After testing our panel models on all *alter*, we restricted the analysis on *alter* located in cities non-attended by *ego*. In the last four columns of Table 7, we observe that mobility has no longer significant impact on the entry and exit of co-authors. So, the creation and rupture of relationship following a mobility only concern *alter* located in cities attended by *ego*. Although mobility allows diversification of the network, its geographical impact is limited to the cities where ego has been located all over its career: mobility does not provide extra bonus of diversification of scientific nerwork.

Discussion and conclusion

The objective of this article is to evaluate the impact of the mobility on the diversification of the co-authorship network. We used 9206 publications of 80 chemists to analyze their 41 628 co-signatures with 14 783 co-authors, at micro-scale. The goal is to evaluate the impact of changing city affiliation (mobility) on the entry and exit of co-authors in the network of researchers over their career.

Descriptive statistics suggest that the social and geographical patterns of co-authorship in chemistry are relatively stable whatever individual trajectories. Although the number of co-authors by publication grows over the career, the volume of new co-authors (social opening) is stable and the volume of former co-authors (continuance of relationships) raises. Moreover, the weight of the *ego*'s lab colleagues (local co-authors) in the publication activity is central and structuring all along. The densification of co-authorship goes with an internationalization process with career progress. However, the mobility does not seem having a premium effect on the diversification of the network. The share of *alter* located in cities attended by *ego* in the past is very low (interpersonal relations are hard to maintain in an experimental practice of science) and mobility does not seem to be a prerequisite for internationalization of co-signatures. Researchers tend to maintain relationship with local co-authors (in terms of co-signatures).

In the panel models, we evaluate the impact of mobility at the intra-individual scale. We see the experience of mobility increases the integration of new co-authors per article. However, this effect is only instantaneous: we do not observe a *premium* effect of mobility over time (Jonkers and Tijssen, 2008). The effect of integration is also restricted to the cities attended by *ego*: there is no "bonus" openness linked to the arrival in a new city. Chinchilla-Rodriguez *et al.* (2017) showed that the international trajectory of researchers is not responsible to his international co-signatures: the international network of mobility is three times smaller that the international co-authorship network. The second effect is that mobility weakens previous relationships and the "exit" effect lasts longer than the "entry" one.

These effects have to be taken into account when thinking about research policy and mobility incentives. Encouraging permanent/job-to-job mobility to expand the network of co-authorship seems counterproductive for two reasons. Firstly, mobility has a significant effect on the network, both on entry and exit: it destroys as many relationships as it creates. Secondly, these effects on the network only concerns co-authors integrated in the geographic trajectory of the researcher. Integrating a new city, the researcher creates only relationships with new lab colleagues and replace former ones. Ackers (2008) shows that in skilled scientific countries, mobility is considered as a sign of excellence only if the researcher had access to local skilled co-authors and had cosigned with them. However, the ability to collaborate with various cities does not depend on the mobility of the researcher. Therefore, short-term mobility should be encouraged because it seems so much powerful to create linkage between scientists.

Table 7: Econometric estimations

	Entry of *alter*		Exit of *alter*		Entry of *alter* from non-attended cities		Exit of *alter* from non-attended cities	
	Model (1.a)	Model (1.b)	Model (2.a)	Model (2.b)	Model (1'.a)	Model (1'.b)	Model (2'.a)	Model (2'.b)
(Intercept)	0,172 * (0,09)	0,150 . (0,09)	-0,167 . (0,10)	-0,237 * (0,10)	-0,442 *** (0,08)	-0,444 *** (0,08)	-0,491 *** (0,08)	-0,497 *** (0,08)
$X^1_{i,t}$ **(ref : not_mob)**								
National	0,627 *** (0,15)		0,206 (0,16)		0,004 (0,12)		-0,036 (0,12)	
Internat	0,563 *** (0,15)		0,286 . (0,17)		0,053 (0,13)		-0,047 (0,13)	
$X^2_{i,t}$ **(ref : sedentary)**								
this_year		0,616 *** (0,12)		0,415 ** (0,13)		0,022 (0,10)		- 0,006 (0,10)
1_to_3		0,108 (0,09)		0,408 *** (0,10)		0,054 (0,08)		0,100 (0,08)
4_to_6		0,144 (0,10)		0,135 (0,11)		-0,026 (0,08)		- 0,033 (0,09)
7_to_9		0,077 (0,11)		0,147 (0,12)		-0,079 (0,09)		- 0,063 (0,09)
10+		- 0,034 (0,09)		0,189 . (0,10)		-0,010 (0,08)		0,087 (0,08)
$Cosign_{i,t}$	0,438 *** (0,02)	0,436 *** (0,02)	0,460 *** (0,02)	0,452 *** (0,02)	0,300 *** (0,02)	0,300 *** (0,02)	0,276 *** (0,00)	0,274 *** (0,02)
$Publis_{i,t}$	-0,059 *** (0,01)	- 0,062 *** (0,01)	-0,153 *** (0,01)	-0,159 *** (0,01)	-0,010 (0,01)	-0,010 (0,01)	-0,049 *** (0,01)	- 0,050 ***
$Geo\_net_{i,t}$	0,004 (0,00)	0,005 (0,00)	0,015 ** (0,00)	0,014 ** (0,00)	0,012 ** (0,00)	0,012 ** (0,00)	0,026 *** (0,00)	0,025 *** (0,00)
$Social\_net_{it}$	-0,002 * (0,00)	- 0,002 . (0,00)	0,004 *** (0,00)	0,004 *** (0,00)	-0,002 . (0,00)	-0,002 (0,00)	-0,001 (0,00)	- 0,001 (0,00)
$Citat_{i,t}$	0,000 (0,00)	0,000 (0,00)	- 0,000 * (0,00)	- 0,000 * (0,00)	-0,000 (0,00)	-0,000 (0,00)	-0,000 (0,00)	- 0,000 (0,00)

References

Ackers, L. (2004, August). Managing relationships in peripatetic careers: Scientific mobility in the European Union. In *Women's Studies International Forum* (Vol. 27, No. 3, pp. 189-201). Pergamon.

Ackers, L. (2008). Internationalisation, mobility and metrics: A new form of indirect discrimination? *Minerva*, 46(4), 411-435.

Aksnes, D. W., Rørstad, K., Piro, F. N., & Sivertsen, G. (2013). Are mobile researchers more productive and cited than non-mobile researchers? A large-scale study of Norwegian scientists. *Research Evaluation*, 22(4), 215-223.

Bernela, B., & Milard, B. (2016). Co-authorship Network Dynamics and Geographical Trajectories-What Part Does Mobility Play? *Bulletin of Sociological Methodology*, *131*(1), 5-24.

Bhagwati, J. N. (1976). Taxing the brain drain. *Challenge*, *19*(3), 34-38.

Breschi, S., Lissoni, F., & Miguelez, E. (2017). Foreign-origin inventors in the USA: testing for diaspora and brain gain effects. *Journal of Economic Geography*, *17*(5), 1009-1038.

Cabanac, G., Hubert, G., & Milard, B. (2015). Academic careers in Computer Science: Continuance and transience of lifetime co-authorships. Scientometrics, 102(1), 135-150.

Cañibano, C. (2006). La gestion de la mobilité professionnelle des chercheurs: un défi pour les politiques de recherche et d'innovation. *La Revue pour l'Histoire du CNRS*, (14).

Chinchilla-Rodríguez, Z., Miao, L., Murray, D., Robinson-García, N., Costas, R., & Sugimoto, C. R. (2017). Networks of international collaboration and mobility: a comparative study. In 16th Intl conf on scientometrics & informetrics.

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R., & Sugimoto, C. R. (2018). Travel bans and scientific mobility: utility of asymmetry and affinity indexes to inform science policy. *Scientometrics*, 116(1), 569-590.

Eckert, D., & Baron, M. (2013). Construire une géographie de la science. *M@ ppemonde*, *110*(2).

Franzoni, C., Scellato, G., & Stephan, P. (2014). The mover's advantage: The superior performance of migrant scientists. *Economics Letters*, 122(1), 89-93.

Gauffriau, M., Larsen, P., Maye, I., Roulin-Perriard, A., & von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, *77*(1), 147-176.

Grossetti, M., Eckert, D., Gingras, Y., Jégou, L., Larivière, V., & Milard, B. (2014). Cities and the geographical deconcentration of scientific activity: A multilevel analysis of publications (1987–2007). *Urban Studies*, *51*(10), 2219-2234.

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31-43.

Lawson, C., & Shibayama, S. (2015). International research visits and careers: An analysis of bioscience academics in Japan. *Science and Public Policy*, *42*(5), 690-710.

Livingstone, D. N. (1995). The spaces of knowledge: contributions towards a historical geography of science. *Environment and planning D: society and space*, *13*(1), 5-34.

Maisonobe, M. (2015). *Étudier la géographie des activités et des collectifs scientifiques dans le monde: de la croissance du système de production contemporain aux dynamiques d'une spécialité, la réparation de l'ADN* (Doctoral dissertation, Université Toulouse le Mirail-Toulouse II).

Melin, G. (2004). Postdoc abroad: inherited scientific contacts or establishment of new networks? *Research Evaluation*, *13*(2), 95-102.

Melin, G. (2005). The dark side of mobility: negative experiences of doing a postdoc period abroad. *Research Evaluation*, *14*(3), 229-237.Melin and Persson, 1996

Morano-Foadi, S. (2005). Scientific mobility, career progression, and excellence in the european research area1. *International migration*, *43*(5), 133-162.Sabatier, 2010

Shauman, K. A., & Xie, Y. (1996). Geographic mobility of scientists: Sex differences and family constraints. *Demography*, *33*(4), 455-468.

Woolley, R., Turpin, T., Marceau, J., & Hill, S. (2008). Mobility matters: Research training and network building in science. *Comparative Technology Transfer and Society*, *6*(3), 159-184.

Indicators of Open Access for universities

Nicolas Robinson-Garcia[1], Rodrigo Costas[2] and Thed N. van Leeuwen[3]

[1] *elrobinster@gmail.com*
Delft Institute of Applied Mathematics (DIAM), TU Delft, Netherlands

[2] *rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Netherlands
DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, South Africa

[3] *leeuwen@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Netherlands

Abstract

This paper presents a first attempt to analyse Open Access integration at the institutional level. For this, we combine information from Unpaywall and the Leiden Ranking to offer basic OA indicators for universities. We calculate the overall number of Open Access publications for 930 universities worldwide. OA indicators are also disaggregated by green, gold and hybrid Open Access. We then explore differences between and within countries and offer a general ranking of universities based on the proportion of their output which is openly accessible.

Introduction

The recent announcement by Science Europe of Plan S, an initiative aiming at providing open access to all publications funded by a group of funding agencies (Else, 2018a, 2018b), has refuelled interest on Open Access at all levels. While Open Access (OA) has been on the agenda of the European Commission for quite some time now (Moedas, 2015), their favourable position towards implementing Plan S (Rabesandratana, 2019) invites to believe that it will soon be also mandatory for all EU funded research. The strictness of Plan S requirements, raises doubts on its viability (Frantsvåg & Strømme, 2019). But still, it evidences the need to monitor OA compliance at all levels, including institutional level.

Universities have been supporting OA for many years now. The most common has been by building and maintaining institutional repositories, and introducing mandates that oblige their researchers to deposit their publications (Harnad, 2007; Harnad et al., 2008). Another more recent way by which institutions are promoting OA, is by sponsoring costs derived from the article processing charges (APC) of open journals (Gorraiz & Wieland, 2009; Gorraiz, Wieland, & Gumpenberger, 2012). In any case, institutions are still faced with the challenge of determining how much of the research they produce is actually openly accessible. Initiatives such as the ranking of OA repositories (Aguillo, Ortega, Fernández, & Utrilla, 2010) offer a partial information which, although valuable, is still insufficient. One of the main limitations is that researchers may combine green and gold OA, and even when self-archiving their publications, they may deposit them in different repositories, impeding institutions to track efficiently their output.

Until recently, there were no more than estimates as to the amount of publications which were available in open access. But the development of platforms like CrossRef, DOAJ or even Google Scholar, along with computational advancements on web scraping, have led to a plethora of large-scale analyses to empirically identify OA literature (Archambault et al., 2014; van Leeuwen, Tatum, & Wouters, 2018; Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018; Piwowar et al., 2018). Overall, these studies report that around half of the scientific literature is freely available, but point towards the increasing availability of publications which do not adhere strictly to what is considered OA.

Here we highlight Unpaywall (Piwowar et al., 2018), which has had a great impact after being implemented by most of the main bibliometric data providers (Else, 2018c). Furthermore, the fact that the Unpaywall API can be freely queried allows others to assess on its performance but also to build on it. In this paper, we present a first attempt at analyzing Open Access at the institutional level, not only in general, but also focusing on the two main routes of OA: the green and the gold route; plus hybrid OA. The purpose for doing so is not only to inform university managers and funding agencies on the level of OA implementation of universities, but also to be able to understand and analyse national trends, and institutional strategies to implementing OA. Although we identify bronze OA, we exclude from our analyses due to the issues related with the sustainability of this type of OA, raising doubts as to its viability from a policy perspective (van Leeuwen, Meijer, Yegros-Yegros, & Costas, 2017)

Data and methods

In this paper we use different sets of sources and combine different methods to determine Open Access. The set of universities analysed and the identification of their publications is retrieved from CWTS in-house version of the Web of Science, based on the institutional name disambiguation developed to produce the Leiden Ranking (Waltman et al., 2012). For each publication, we identify if they are openly accessible and the type of Open Access by querying the Unpaywall information. The Unpaywall API does not labels types of OA but describes what information was derived from each record. More information on the information provided is available at their website the User Guide offered for researchers (http://unpaywall.org/data-format).

The labelling of OA types is described in Figure 1 and already highlights some of the difficulties and controversies raised when trying to define what is actually Open Access (Torres-Salinas, Robinson-Garcia, & Moed, 2019).

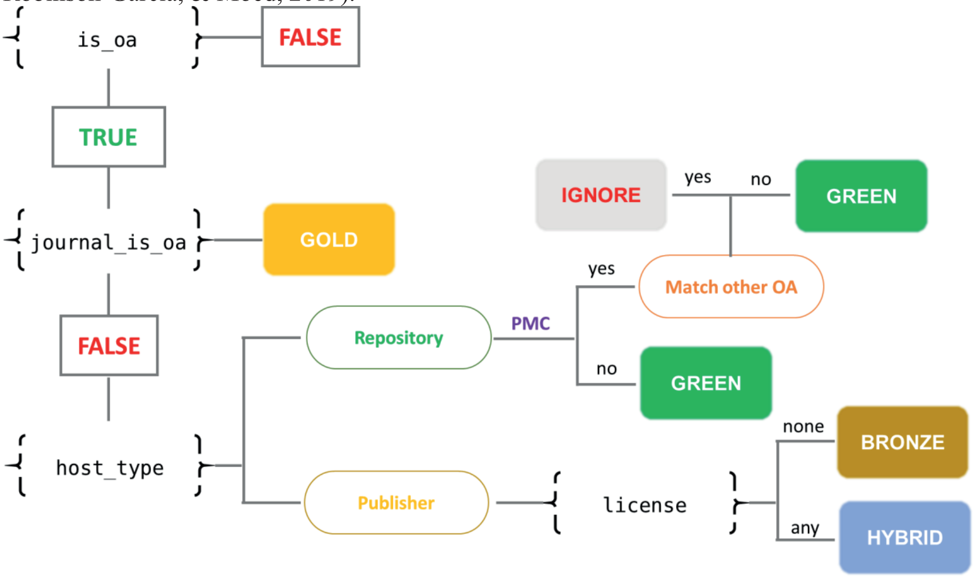
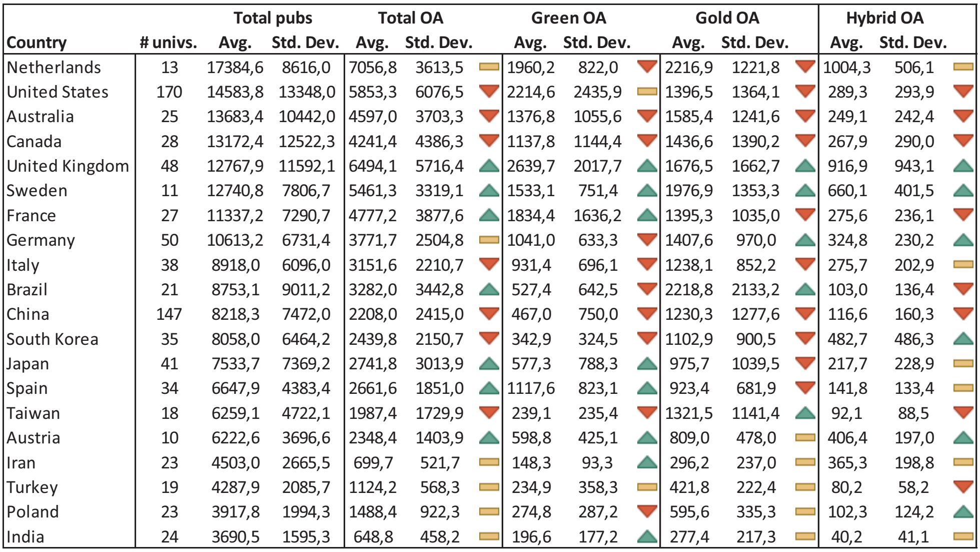


Table 1. Conditions included to determine type of Open Access and total publications retrieved for the 2014-2017 period for our set of 930 universities

	journal_is_oa (Unpaywall)	*evidence* (Unpaywall)	*Gold_OA* (Leeuwen et al. 2017)	Type of OA	Total pubs
∪	TRUE	--	TRUE	Gold	645,547
∩	FALSE	open (via free pdf)	--	Bronze	449,929
∩	FALSE	open (via crossref license, author manuscript)	--	Undetermined	49,697
∩	FALSE	open (via page says license)	FALSE	Hybrid	121,637
		oa repository (…)		Green	562,747

As table 1 shows, four types of OA were considered plus an *undetermined* category which included a segment of publications which were hybrid, bronze and sometimes even gold OA which were impossible to discriminate from by using the fields offered by Unpaywall. These four types of OA are defined as follows:



		Total pubs		Total OA			Green OA			Gold OA			Hybrid OA		
Country	# univs.	Avg.	Std. Dev.	Avg.	Std. Dev.		Avg.	Std. Dev.		Avg.	Std. Dev.		Avg.	Std. Dev.	
Netherlands	13	17384,6	8616,0	7056,8	3613,5	▬	1960,2	822,0	▼	2216,9	1221,8	▼	1004,3	506,1	▬
United States	170	14583,8	13348,0	5853,3	6076,5	▼	2214,6	2435,9	▬	1396,5	1364,1	▼	289,3	293,9	▼
Australia	25	13683,4	10442,0	4597,0	3703,3	▼	1376,8	1055,6	▼	1585,4	1241,6	▼	249,1	242,4	▼
Canada	28	13172,4	12522,3	4241,4	4386,3	▼	1137,8	1144,4	▼	1436,6	1390,2	▼	267,9	290,0	▼
United Kingdom	48	12767,9	11592,1	6494,1	5716,4	▲	2639,7	2017,7	▲	1676,5	1662,7	▲	916,9	943,1	▲
Sweden	11	12740,8	7806,7	5461,3	3319,1	▲	1533,1	751,4	▲	1976,9	1353,3	▲	660,1	401,5	▲
France	27	11337,2	7290,7	4777,2	3877,6	▲	1834,4	1636,2	▲	1395,3	1035,0	▼	275,6	236,1	▼
Germany	50	10613,2	6731,4	3771,7	2504,8	▬	1041,0	633,3	▼	1407,6	970,0	▲	324,8	230,2	▲
Italy	38	8918,0	6096,0	3151,6	2210,7	▼	931,4	696,1	▼	1238,1	852,2	▼	275,7	202,9	▬
Brazil	21	8753,1	9011,2	3282,0	3442,8	▲	527,4	642,5	▼	2218,8	2133,2	▲	103,0	136,4	▼
China	147	8218,3	7472,0	2208,0	2415,0	▼	467,0	750,0	▼	1230,3	1277,6	▼	116,6	160,3	▼
South Korea	35	8058,0	6464,2	2439,8	2150,7	▼	342,9	324,5	▼	1102,9	900,5	▼	482,7	486,3	▲
Japan	41	7533,7	7369,2	2741,8	3013,9	▲	577,3	788,3	▲	975,7	1039,5	▼	217,7	228,9	▬
Spain	34	6647,9	4383,4	2661,6	1851,0	▲	1117,6	823,1	▲	923,4	681,9	▼	141,8	133,4	▬
Taiwan	18	6259,1	4722,1	1987,4	1729,9	▼	239,1	235,4	▼	1321,5	1141,4	▲	92,1	88,5	▼
Austria	10	6222,6	3696,6	2348,4	1403,9	▲	598,8	425,1	▲	809,0	478,0	▬	406,4	197,0	▲
Iran	23	4503,0	2665,5	699,7	521,7	▬	148,3	93,3	▲	296,2	237,0	▬	365,3	198,8	▬
Turkey	19	4287,9	2085,7	1124,2	568,3	▬	234,9	358,3	▬	421,8	222,4	▬	80,2	58,2	▼
Poland	23	3917,8	1994,3	1488,4	922,3	▬	274,8	287,2	▼	595,6	335,3	▬	102,3	124,2	▲
India	24	3690,5	1595,3	648,8	458,2	▬	196,6	177,2	▲	277,4	217,3	▬	40,2	41,1	▬

Figure 1. Institutional output at the country level for those countries with at least 10 universities in the dataset for the 2014-2017 period. Countries are shown based on the number of universities included in the ranking. Arrows show changes in ranking based on average of total number of publications

- **Green Open Access**. Self-archived versions of a manuscript. Here the responsibility lies on the author who is in charge of depositing the document in a repository. This version of the document may not correspond with the final version of the publisher.
- **Gold Open Access**. This refers to journals which publish all of their manuscripts in Open Access regardless of the business model they follow (e.g., publicly sponsored, author pays).

- **Bronze Open Access**. While again journals are the ones offering the publication freely available, this is not subjected to copyright conditions set to be defined as Open Access (i.e., they do not ensure perpetual free access).
- **Hybrid Open Access**. Non-OA journals make specific publications openly accessible usually after the author pays a fee to account for potential losses derived from subscription fees.

In all, a total of 930 universities were analysed for the 2014-2017 period. While we identified some overlap between green and gold OA, the other three categories are exclusive from each other. Finally, in this paper we will consider as OA whichever document which adheres to any of these five types, however we will offer a deeper analysis to those following the green and gold routes. The rationale for this is that these two routes are, in principle, the ones that more closely align with Plan S and with the development of a sustainable movement towards full Open Access.

Results and discussion

The two countries contributing most on the number of universities analysed are United States and China, more tripling the third and fourth countries (Figure 1). The Netherlands have on average the most productive universities followed by United States, Australia and Canada. While this trend is followed on the average number of OA publications, there are important changes on the average output of OA publications they should produce considering their total output. For instance, British universities occupy higher positions when referring to total, green, gold or hybrid OA output on average. However, there seems to be large disparities within the country. On the other side we find countries such as Australia, Canada and China, which are in a lower position based on their average number of OA publications than what they should occupy, according to their overall average number of publications. We find differences on the size of the output of institutions by country. While this is to be expected, we do find more or less disparity when focusing only on OA publications. For instance, there is a greater deviation from the average for Turkish universities when considering only green OA publications, while the deviation is consistently higher for Chinese universities when focusing on any type of OA than when considering all publications. An interesting case is the one observed in The Netherlands, which the country where its universities produce on average the highest number of OA publications, despite falling back on both, green and gold OA. Still, it maintains its first place with regard to the average number hybrid OA publications.

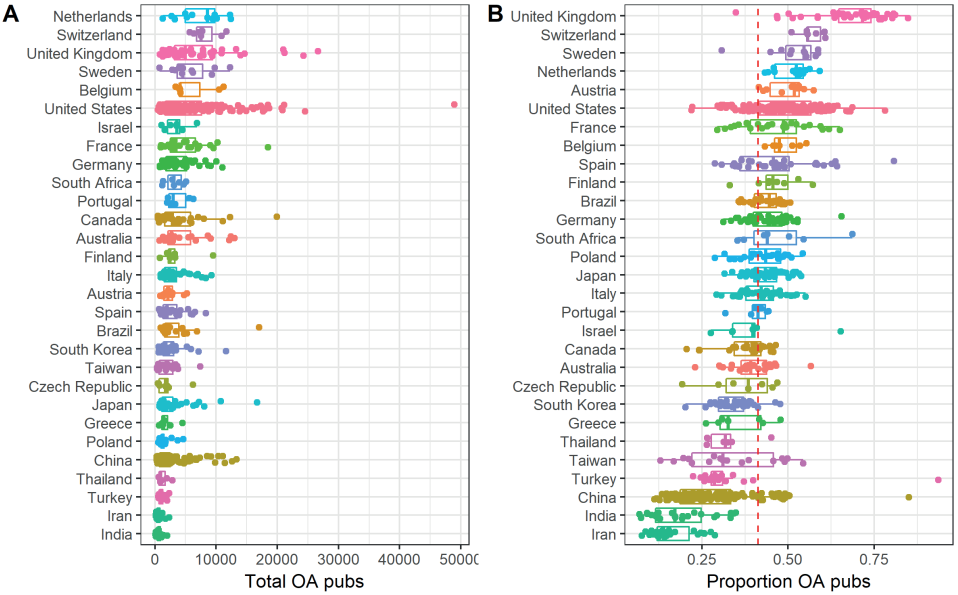


Figure 2. Distribution of universities based on their A) total number and B) proportion of OA publications for countries with > 5 universities included. Countries ordered by median proportion of OA publications. Red dashed line shows world median.

Disparities within countries are further analysed on figures 2 and 3. Here we plot for the distribution of universities by country for their total and proportion of OA publications and by green and gold OA. Overall, we observe that the United Kingdom is the country in which their universities are making a higher proportion of their publications openly available, followed by Switzerland and Sweden (Figure 2B). Furthermore, we find extreme cases in Turkey and China with two universities having virtually all of their output OA. In these two cases, most of the universities show shares of OA lower than the world median. It is also worth noting that most of the countries with higher proportions of OA at the institutional level are European with three notable exceptions. These are Brazil, South Africa and United States. In the case of the two former, we find a very different pattern from the other two BRICS countries shown (China and India), which have a proportion of OA publications below world median. In the latter case, it does not occupy a leading position as it is usually the case with the United States. Russia is not present in these figures as only two universities are included in our dataset.

The ways in which universities are making their publications openly accessible varies greatly when distinguishing between gold and green OA. United Kingdom and Switzerland are again the ones with a higher median on the proportion of green OA their publications have (Figure 3B). Belgium and especially Spain, outstand in third and fourth place with respect to their overall proportion of OA. It is also worth noting the great dispersion on proportion of green OA not only between countries but also within countries. While the world median proportion of green OA is 8.7%. It raises up to 22.1% for the United Kingdom and it is 3.1% for Iran.

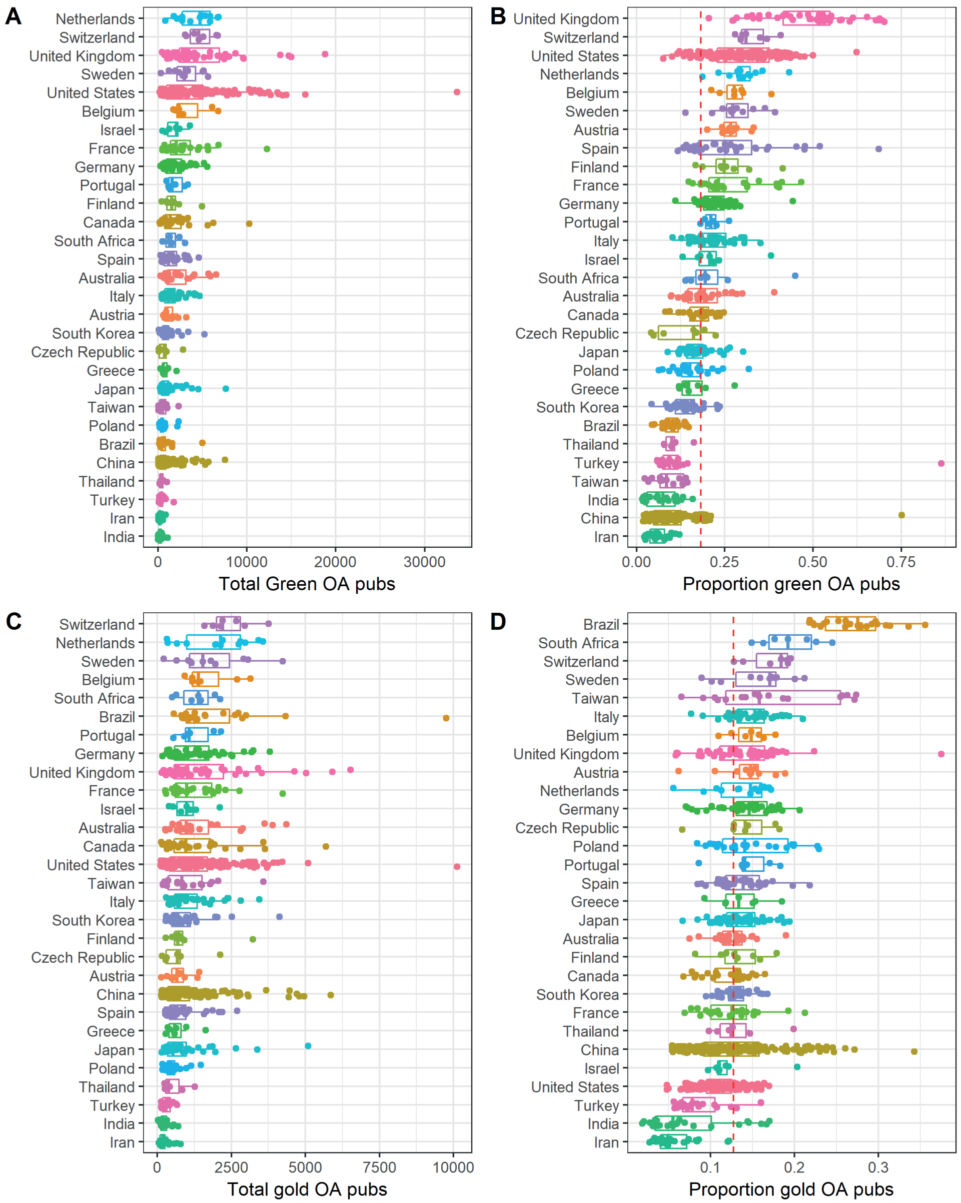


Figure 3. Boxplots by country for countries > 5 universities included based on their A) total number and B) proportion of green OA pubs and C) total number and D) proportion of gold OA pubs. Countries ordered by median proportion of OA publications. Red dashed line shows world median.

In the case of gold OA, the world median institutional share is 11.8%. As observed, there are less disparities within countries than in the case of green OA (figure 3D). In this case, Brazil (26%), South Africa (19%) and Taiwan (17%) are the countries with the largest proportion of their output in gold OA (median values). China is the country with greater disparities between its universities. To better interpret the patterns of these countries we look into the OA journal profile of these four countries, following the three models of gold OA proposed by Torres-Salinas, Robinson-Garcia, & Moed (2019). The first one refers to countries which publish in OA journals owned by publishing firms, preferably mega-journals and with a high Journal Impact Factor. The second model is that of countries which publish in OA journals edited in their own country, preferably in their native language and publicly funded. The third model is a mixed one where gold OA publications are channelled through both OA mega-journals and nationally-oriented OA journals. In all cases but China (where it is second behind *Scientific Reports*), *PLoS One* is the journal with the largest number of publications. In the case of Brazil, the rest of OA journals in the list are in a vast majority Brazilian journals listed in SCielo. In South Africa we observe a combination of regional journals and big OA publishers. Finally, China and Taiwan exhibit a greater reliance on journals from big OA publishers such as Nature Springer, PloS, MDPI or Hindawi.

Finally, we conclude by showcasing in figure 4 the top 50 universities with the largest proportion of OA worldwide. These 50 universities come from 12 different countries. More than half of them come from United Kingdom (26), including major universities such as London City University. Next, Spain positions 6 universities, followed by France and the United States (5 each). While the remaining countries have only one university included in this top 50. Here we note that the two outliers aforementioned from Turkey and China (figure 2B), are actually the top 2 universities on openly accessible literature, mostly relying on green OA (see figure 3B).

Concluding remarks

This paper presents a first attempt at measuring OA uptake by universities worldwide. Europe is hardening its policies towards full OA, and initiatives such as Plan S are being supported by important international funding bodies (e.g., Wellcome Trust, Bill & Melinda Gates Foundation). The introduction of such policies may affect differently across Europe and such policies may expand to other countries. The inclusion of indicators on OA to the Leiden Ranking adds another dimension, which is less traditional, and focused on changes in scholarly communication practices. This can better inform how OA is being implemented and which routes are having a greater implementation. Finally, this contribution allows to study the distribution across the globe of OA uptake, to what extent the initial goals of the OA movement to distribute more equally over the globe reached, especially when looking at OA uptake in e.g., the Global South, effects of regulations (e.g., inclusion of hybrid OA by Plan S), etc.

Here we present a first attempt at developing OA indicators at the institutional level globally. However, there are many issues that still need to be dealt with. For instance, the consideration of Unpaywall as the most important means by which OA is captured, although welcome and remarkable, needs to be better assessed and understood (double occurrences, *undetermined* category, etc.). Also, it is important to understand better and make more distinct in OA analyses gold OA models and specifically publicly-funded gold OA (i.e., SCielo) versus APC models and private publishing firms.

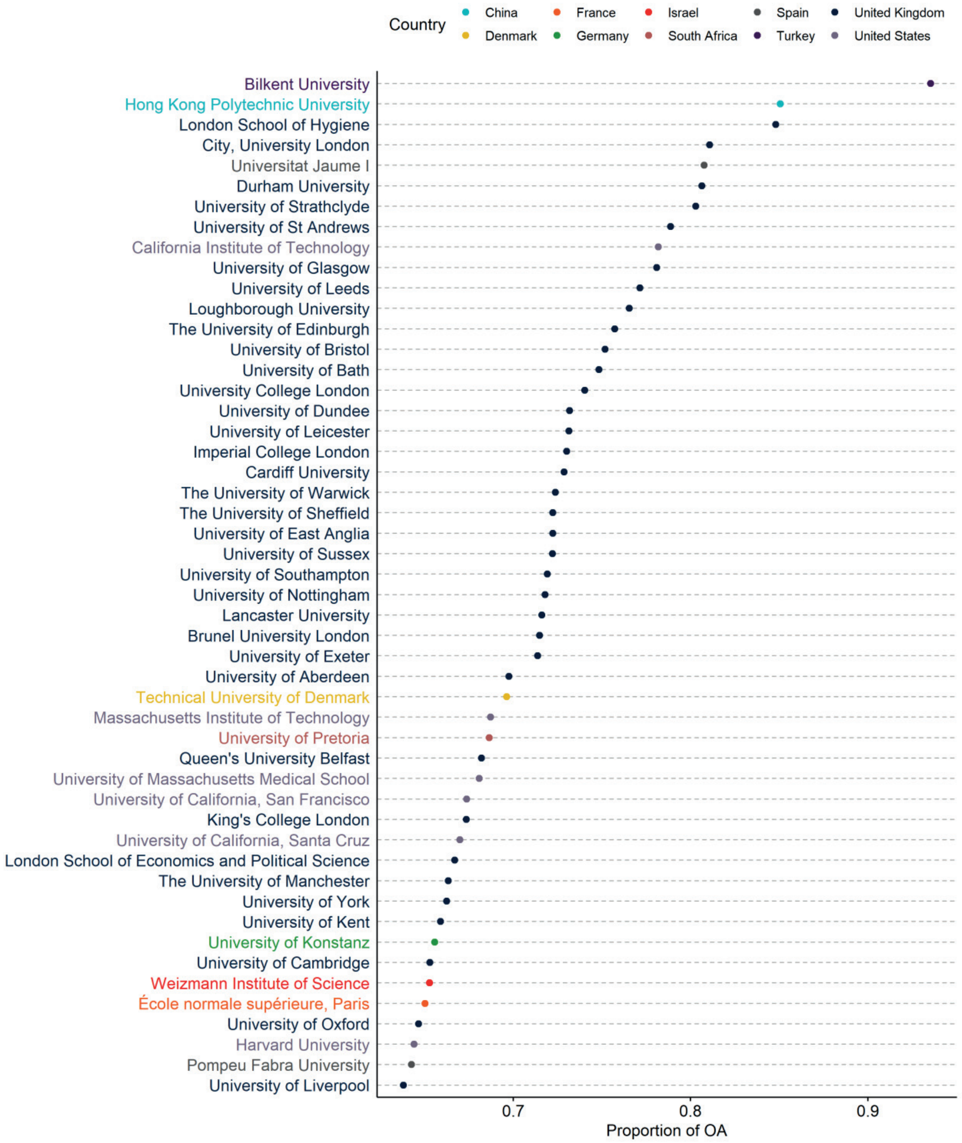


Figure 4. Ranking of top 50 universities based on the proportion of Open Access publications in the 2014-2017 period

Acknowledgments

Nicolas Robinson-Garcia is a Marie Sklodowska-Curie Experienced Researcher in the LEaDing Fellows COFUND programme from the European Commission. This research is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP).

References

Aguillo, I. F., Ortega, J. L., Fernández, M., & Utrilla, A. M. (2010). Indicators for a webometric ranking of open access repositories. *Scientometrics*, *82*(3), 477-486.

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., & Roberge, G. (2014). Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013. Recuperado mayo 10, 2017, a partir de http://digitalcommons.unl.edu/scholcom/8/

Else, H. (2018a). Radical open-access plan could spell end to journal subscriptions. *Nature*, *561*, 17.

Else, H. (2018b). Funders flesh out details of Europe's bold open-access plan. *Nature*. Recuperado enero 16, 2019, a partir de http://www.nature.com/articles/d41586-018-07557-w

Else, H. (2018c). How Unpaywall is transforming open science. *Nature*, *560*, 290.

Frantsvåg, J. E., & Strømme, T. E. (2019). Few Open Access Journals Are Compliant with Plan S. *Publications*, *7*(2), 26.

Gorraiz, J., & Wieland, M. (2009). Multi-authored publications: their influence in the distribution of the financing costs in world licenses. *Research Evaluation*, *18*(3), 215-220.

Gorraiz, J., Wieland, M., & Gumpenberger, C. (2012). Bibliometric practices and activities at the University of Vienna. *Library Management*, *33*(3), 174-183.

Harnad, S. (2007). *Mandates and metrics: How open repositories enable universities to manage, measure and maximise their research assets.* Recuperado a partir de https://eprints.soton.ac.uk/264990/1/openaccess.pdf

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., et al. (2008). The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update. *Serials Review*, *34*(1), 36-40.

Leeuwen, T. N. van, Tatum, C., & Wouters, P. F. (2018). Exploring possibilities to use bibliometric data to monitor gold open access publishing at the national level. *Journal of the Association for Information Science and Technology*, *69*(9), 1161-1173.

van Leeuwen, T., Meijer, I., Yegros-Yegros, A., & Costas, R. (2017). Developing indicators on Open Access by combining evidence from diverse data sources. *arXiv:1802.02827 [cs]*. Presentado en 2017 STI Conference, Paris: Proceedings of the 2017 STI Conference. Recuperado enero 10, 2019, a partir de http://arxiv.org/abs/1802.02827

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819-841.

Moedas, C. (2015). *Open Innovation, Open Science, Open to the World*. European Commission's Directorate-General for Research & Innovation (RTD). Recuperado enero 26, 2017, a partir de http://europa.eu/rapid/press-release_SPEECH-15-5243_en.htm

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., et al. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375.

Rabesandratana, T. (2019). Will the world embrace Plan S, the radical proposal to mandate open access to science papers? *Science | AAAS*. Recuperado enero 16, 2019, a partir de https://www.sciencemag.org/news/2019/01/will-world-embrace-plan-s-radical-proposal-mandate-open-access-science-papers

Torres-Salinas, D., Robinson-Garcia, N., & Moed, H. F. (2019). Disentangling Gold Open Access. En W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Handbook of Science and Technology Indicators*. Cham: Springer. Recuperado a partir de https://arxiv.org/abs/1807.04535

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., et al. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432.

Research on the relationship between citation and altmetrics of Open Access Papers from different geographical regions

Jingda Ding and Jie Guo and Xinjing Liu and Xiangjie Ma and Chao Liu

djdhyn@126.com

Department of Library, Information and Archives, Shanghai University, Shanghai 200444, China

Abstract

With the development of open access, more and more scientific papers are published in journals as *PLoS* and are freely available by online. Analyzing the relationship between citation and altmetrics of open access papers published on *PLoS* in recent 10 years by the 6 countries which are selected in terms of regional distribution, scientific level and native language etc., we induced the following conclusions: First, the different geographical regions have effect on the citation and the altmetrics of 'view' and 'save' (excepting 'share'), because the means of them for developing countries are less than that of developed countries in last 10 years and the curve peaks of these 3 indicators of different countries occur in different years. Second, the citation and the altmetrics of 'view' and 'save' are significantly correlated with each other and have the similar variation patterns and accumulate with year, while the altmetrics of 'share' is just opposite. Therefore, to some extent the altmetrics of 'view' and 'save' can evaluate scientific influence as a complement of citation, and the 'share' of *PLoS* papers seemingly just reflects the attention/mention of user rather than the academic influence because it is transient and uncorrelated with the citation.

Introduction

Traditional bibliometric analysis and peer review have formed the standard methods to assess the 'scientific status of disciplines'(Patel & Chavda, 2016). The number of publications, citation frequency and peer review are traditional science measurement indicators. Since Garfield came up with citation analysis, it has received a lot of popularity. The frequency of citation partly inflects the academic value and influence of research papers through the reference relationship between them. Thus, citations qualify the quantity of publications and this makes it a good indicator to rank publications' value(Lehmann, Jackson, & Lautrup, 2006). As is known, citation analysis also has some deficiencies, but it is still a good method to evaluate the research impact, and some citation indicators such as h-index(Hirsch, 2010) , journal impact factor (Garfield, 1972) have been used for research evaluation though controversial(Garfield, 1999; Leeuwen, 2005). Now, some scholars have suggested that citation is not able to reflect the broader impact of research (Holmberg, Didegah, & Bowman, 2015), can't meet and satisfy the development requirements of the era of web 2.0+.

In the era of Web 2.0, more and more scholars tend to share experiences, express ideas, or disseminate research findings at different social media, such as *Blogs*, *Twitter*, *Facebook* and so on. With the use of these new tools in scientific communication, traditional biblometric analysis and peer review are not enough to evaluate the impact of scientific publications in social media. In this context, alternative metrics, also called "altmetrics", was proposed by Priem and his collaborators at the first time in 2010, and they noted that altmetrics are the creation and research of a new kind of metrology, based on the analysis and dissemination of research production in social network(Priem et al, 2010). Altmetrics are the new metric to evaluate the influence of scientific publications, which discussed, shared, posted, tagged, mentioned or tweeted on social media platforms. According to Work et al.(2015), altmetrics are usually based on activity on online communication platforms, relating to scholars or scholarly content. Typical examples of altmetrics include tweets, mentions in *blog* posts, readership counts on *Mendeley*, posts, likes and shares on social media such as *Facebook* and *Google+*, and recommendations and ratings on *F1000*. Altmetrics are usually considered as the subset of Scientometrics and Webometrics, and they are used to carry out Article-Level Metrics research (Stransky, 2016). According to the collected data of social media platform, altmetrics can evaluate the popularity or social influence of publications(Chavda & Patel, 2016). As Altmetrics, Inllumetrics, Entitymetrics, Usage metrics, Article-level metrics and other terms having been proposed(Glanzel & Gorraiz, 2015), the evaluation of research paper can not only rely on the amount and its citation count(Moed & Halevi, 2015). Multi-source and multidimensional of the measurement in research is the trend.

Based on those, we propose the following two questions and try to solve:

1. What is the relationship between citation and altmetrics of open access papers?

2. Do different geographic regions have an impact on citation and altmetrics of open access papers? If so, what are the key influencing factors?

Literature Review

The researches on relationship between citation and altmetrics for traditional journals are more than that based on the new environment of Open Access (OA), so it is worth spending time to do further study that based on the OA papers from different cultural background. In the context of OA movement, free online availability of scientific literature offers substantial benefits to science and society(Lawrence, 2001), and researchers can join many different websites to publicize their research productions(Thelwall & Kousha, 2017). In the early stage, the research on OA thesis mainly focused on the impact of OA papers. Through the comparison of open access and non-open access(non-OA)(Davis, Lewenstein, Simon, Booth, & Connolly, 2008; Moed, 2007; Norris, Oppenheim, & Rowland, 2008), the value of open access can be understood(Joint, 2009). OA papers have reached or even more than non-OA papers of quality and influence(Hu, 2008). This is a huge advantage in sense of scientific dissemination: each article may receive as wide readership as it deserves, and does not matter whether the journal impact factor is high or low. Yassine et al.(2010) analyzed of 27,197 articles published in internationally peer-reviewed journals, and found out that the open access had significant effect on citations in different scientific disciplines. OA publications obtain more citations compared with those not openly accessible. From the investigation of 15 countries(Shu, 2017), there are 14 countries in which twitter papers cited quantities excess 30% than non-twitter papers. Hence, the influence evaluation of OA papers is more closely related to the new measurement index by comparing with traditional paper. Researches on altmetrics and their possible implication in calculating the influence of publication are becoming widely (Priem et al, 2010; Chavda & Patel, 2016). And scholars are attaching more and more importance to the correlation between citation and altmetrics (Dhiman, 2015; Peters et al.,2016;Waltman & Costas, 2014; Thelwall et al., 2013). Some scholars investigated the correlation from different disciplines like social science, medical science etc., while others analyzed the correlation between different kinds of altmetrics indicators and citation. Costas, Zahedi &Wouters(2015) found that the mentions in blogs and news outlets had a relatively stronger correlation with citation than other altmetrics indicators. Syamili et al.(2017) explored the correlation of citation and altmetrics based on a specific topic about "Ebola", and pointed out that the Twitter count had no correlation with citation frequency while other altmetrics values had the good correlation with it. It seems that altmetrics like tweets, mentions and readership counts (on *Mendeley*) might reflect the impact of publications in society or the public attention, while their connection with the quality of research(often be evaluated by citation frequency) is loosely even none(Bornmann, 2015). Compared to citation, which exists a long time delay, *Twitter* begins to tweet papers in several of days even hours after published, which shows that altmetrics coming from *Twitter* etc. seemingly tends to reflects the attention/mention rather than the academic influence in a short time. And there are more and more scholars to study the impact factors of the correlation between citation and altmetrics (Dhiman, 2015). Peters et al.(2016) surprisingly found that altmetrics had no correlation with citation, which may not correspond to the positive but relatively moderate correlation results in the study of other scholars(Waltman & Costas, 2014; Thelwall et al., 2013). The reasons for different analysis results may be various factors that influence the correlation of citation and altmetrics, such as the year of publication, discipline, user types and habits, different social media platforms and so on. At present, even though the use of altmetrics in measuring the scientific research still in controversial, some areas (Chisolm, 2017; Wang, Alotaibi, Ibrahim, Kulkarni, & Lozano, 2017) have started to bring altmetrics in their own research field to evaluate journals and get the popular magazines of their own field.

In a word, the citation has a positive but is relatively weak correlation with altmetrics, and different kinds of altmetrics indicators have different degree of correlation with citation. However, there is little research that consider whether different cultures or regions have effect on the correlation between them. Researches on regional distribution difference often focus on geographical collaboration at department, institution and national level and the distribution of authors(Abbasi & Jaafari, 2013; Bartneck and Hu, 2010; Gorraiz, Reimann, & Gumpenberger, 2012). There are few of scholars in-depth discussing the differences of the correlation between citation and altmetrics among countries. Based on that, this work analyze the relationship between citation and altmetrics of OA papers from the perspective of different countries.

Research Methodology

PLoS is an open-access journal platform based on peer review, which is about biology, medicine and some diseases. There are 'citation', 'view', 'save' and 'share' in *PLoS ALM* dataset. The 'citation' is the sum of citation count of *Scopus* and *Crossref*, the 'save' is the number of *Mendeley* bookmarks, the 'view' is the total number of page views and downloads of *PLoS* and *PubMed Central*, and the 'share' is the discussion counts by *Twitter* and *Facebook*, which are records of different aspects of the *PLoS* papers. In the article, we selected the papers published in 6 *PLoS* journals by 6 representative countries between 2009 and 2018 as the research sample. These 6 journals are *PLoS Biology*, *PLoS Computation Biology*, *PLoS Genetics*, *PLoS Medicine*, *PLoS Neglected Tropical Diseases and PLoS Pathogens*, which belong to the field of biomedical technology. These 6 countries—Brazil, China, Germany, Japan, Russia and USA are selected, because China and Brazil are developing countries while Germany, Japan, Russia and USA are developed countries. In addition, Germany is located in Europe, Russia spans the Eurasian continent, USA and Brazil pertain to America, and China and Japan belong to Asia, and USA is the English-speaking country while other countries are not. Given that the value of ‘view’, ‘save’, ‘share’ and ‘citation’ would be changed with time, the deadline we collect data is April 19, 2019. Table 1 presents the quantity of papers published on *PLoS* journals by the 6 countries in every year.

Table 1. The number of papers published by the 6 countries in each year

Countrie	*200*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*SUM*
China	55	74	106	126	179	206	207	213	223	259	1648
German	174	211	212	277	334	311	349	338	195	221	2622
Japan	69	77	93	107	103	102	121	130	107	111	1020
Russia	9	9	14	13	22	13	16	14	15	19	144
USA	874	1038	1120	1282	1513	1497	1587	1450	1627	1713	13701
SUM	121	1460	1603	1894	2244	2266	2405	2285	2294	2442	20103

Results

Descriptive Statistics

From Table 1, we can see that the quantity of the United States is much higher than the other countries, and the number of papers published by each country keeps growing though there is a slight fluctuation in recent 5 years. Figure 1 displays the percentage of paper published by each country in every year. Although the number of papers in each country is growing, the proportion of papers published annually (based on the total number of 6 countries in this paper) varies. As can be seen from Figure 1, China and Russia are on the rise, Brazil and Japan remain stable. For the United States and Germany, they show the different change, the former was falling until to 2016, the latter was increasing in stability before 2016 and decreased in latest 2 years. Generally, as a scientifically developed country, the amount of papers published by the United States is far ahead. Germany is the second though still far behind the United States, China is slightly more than half of Germany’s, Japan is close to Brazil with the approximately annual output of 1000 or so, which are fourth and fifth respectively, Russia is the least.

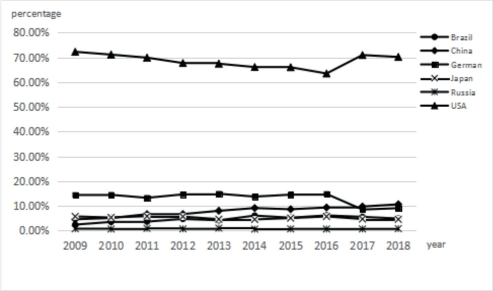


Figure 1. Percentage of papers in each year

Figure 2 is the ratio of the United States and other five countries in the 6 *PLoS* journals, which shows the advantages and disadvantages of each country in biomedical domain. In the 6 countries, Brazil's development is extremely unbalanced, whose number of papers published in *PLoS Neglected Tropical Diseases Journal* is more than the other 5 journals. China is similar to Brazil, whose number of papers published in *PLoS Genetics Journal* is more than the other 5 journals. Germany's paper number in *PLoS Medicine Journal* is less than the other 5 journals, as do China, Japan and Russia. The United States publishes a relatively balanced number of papers in these six journals and has obvious advantage.

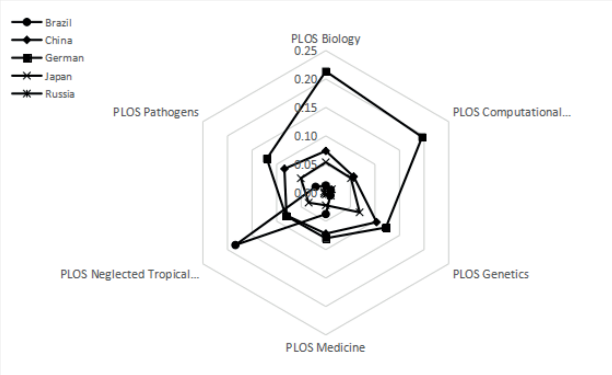


Figure 2. Distribution of publications in the 6 journals of each country

Mean and Coverage Degree

Different kinds of altmetrics indicators of *PLoS* papers had different coverage, which refers to the ratio of a non-zero count of some indicator of papers to the total number of the sample papers, measuring the distribution breadth of the indicator on sample papers. Table 2 shows the mean value and coverage of the 6 countries from 2014 to 2018 and from 2009 to 2018. The mean and coverage of 'citation', 'view' and 'save' in recent 10 years are higher than or equal to that in latest 5 years, while 'share' is just the opposite, which indicates that the life cycle of 'share' is shorter than that of the others. The coverage of 'view' in each country are 100%, which testifies that each paper published by the 6 countries has been viewed and downloaded in *PLoS* and *PubMed Central* at least once. And the coverage of 'citation' is more than 91% in different countries, which presents that open access papers are easily accessible and thus cited. Only the coverage degree of 'share' is relatively lower, which is no more than 85% and there are significant differences between the latest 5 years and the recent 10 years.

Table 2. Mean and coverage of the 'view', 'save', 'share' and 'citation' of each country

Index	*Item*	*Brazil*		*China*		*Germany*		*Japan*		*Russia*		*USA*	
		5 Y	*10Y*	*5 Y*	*10Y*	*5 Y*	*10Y*	*5 Y*	*10Y*	*5 Y*	*10Y*	*5 Y*	*10Y*

View	Mean	5416.51	6281.54	6346.63	7147.29	6499.70	8337.70	6109.38	7424.84	8107.38	10313.30	6797.14	8342.36
	coverage	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Save	Mean	38.68	45.83	32.03	40.95	44.03	62.96	37.39	51.86	53.00	70.42	41.26	59.02
	coverage	0.93	0.95	0.92	0.95	0.95	0.97	0.95	0.97	0.91	0.95	0.93	0.96
Share	Mean	28.62	23.32	11.04	8.62	19.49	13.69	15.70	10.00	29.25	18.68	27.42	19.03
	coverage	0.75	0.68	0.69	0.58	0.78	0.62	0.75	0.55	0.84	0.66	0.80	0.63
Citation	Mean	12.33	19.14	16.30	25.85	15.49	30.78	14.40	27.98	21.40	36.56	14.05	28.27
	coverage	0.92	0.94	0.93	0.95	0.93	0.95	0.93	0.96	0.92	0.96	0.91	0.94

Correlation analysis of Citation and Altmetrics

Table 3 lists the correlation coefficients of citation with ‘view’, ‘save’ and ‘share’ in each country. We can see that the citation has strong correlation with ‘save’ or ‘view’, while its relevancy with ‘share’ is very weak even none.

Table 3. Correlation coefficients between different indexes of the 6 countries

Cor.	*View*	*Save*	*Share*
Citation (Brazil)	0.714556995**	0.756788094**	0.076196384*
Citation (China)	0.755038197**	0.808202254**	0.002956998**
Citation (Germany)	0.735796999**	0.737906617**	-0.192310765**
Citation (Japan)	0.762207350**	0.753003825**	-0.179578338**
Citation (Russia)	0.805850672**	0.799028859**	0.040922930
Citation (USA)	0.688062988**	0.767937599**	-0.145536290**

(Note: ** represent significance at 0.01 level; *represent significance at 0.05 level)

Compare analysis between Citation and Altmetrics for the 6 Countries

Figure 3 depicts the changes of ‘citation’, ‘view’, ‘save’ and ‘share’ with time in each country. The lateral axis expresses the year, the left vertical axis represents the number of ‘view’ and the right vertical axis represents the value of ‘citation’, ‘save’ and ‘share’. The units of vertical axes vary from different countries in order to make the overall change of index clear in each country.

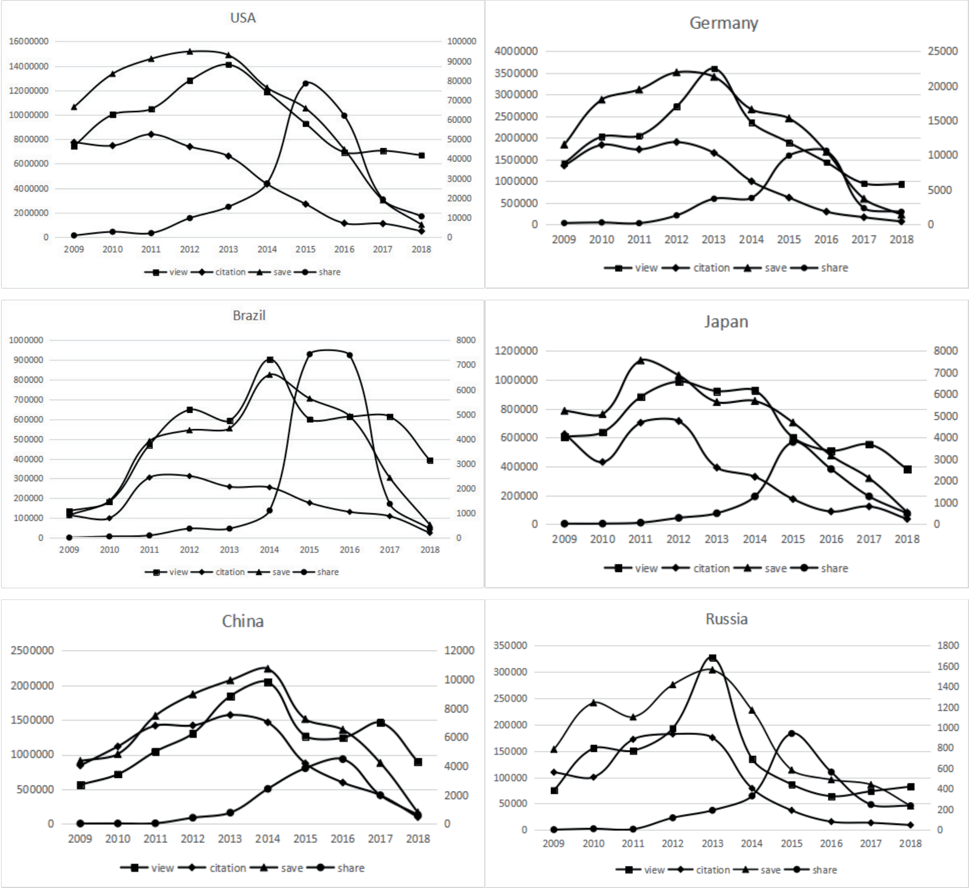


Figure 3. Changes of citation and altmetrics with year for the 6 countries

As can be seen from Figure 3, the curve peaks of 'citation', 'view', 'save' and 'share' of each country occurred in different years. Among them, the 'citation' peak of the United States occurred in 2011 and the maximum citation age of *PLoS* papers published by the United States is 8 years because 2019 is the observational year, i.e. these papers are active and vigorous in the latest 8 years after publication. Similarly, the 'citation' peaks of Germany, Brazil, Japan, China and Russia respectively occurred in 2012, 2012, 2012, 2013, 2012, accordingly whose maximum citation age separately is 7,7,7,6,7 years. Then the 'view' peaks of the United States, Germany, Brazil, Japan, China and Russia respectively occurred in 2013, 2013, 2014, 2012, 2014, 2013, and the 'save' peaks of them respectively occurred in 2012, 2012, 2014, 2011, 2014, 2013, and the curves of 'share' in 6 countries reached their peaks in 2015 or 2016. We can draw Table 4 from Figure 3. As a whole, the *PLoS* papers for each country firstly reach the 'share' peak after publication for 2 years or 3, then reach the 'save' or 'view' peak after 5 or more years, finally reach the 'citation' peak.

Table 4. The year in which the index peak appeared of the 6 countries

Peak	*USA*	*Germany*	*Brazil*	*Japan*	*China*	*Russia*
Citation	2011	2012	2012	2012	2013	2012
View	2013	2013	2014	2012	2014	2013
Save	2012	2012	2014	2011	2014	2013
Share	2015	2016	2015	2015	2016	2015

(Note: the date of collect data is April 19, 2019)

On the other hand, the curves of 'citation', 'view' and 'save' nearly keep the similar fluctuation trend, and exist the significant correlation between 'citation' and 'view' or 'save', which demonstrates that there are greater similarity of change among them. Differently, the curves of 'share' in 6 countries start to grow obviously from 2011 and reach their peaks in 2015 or 2016(until to the date analyzed), and there is no significant correlation between 'citation' and 'share'.

Discussion

This article is mainly to analyze the relationship between citation and altmetrics, and explore whether the different geographical regions have effect on the 'citation', 'view', 'save' and 'share', so as to improve the quality of research evaluation. Hence, we discuss the following 3 points.

The influence of different geographical regions on the 'citation', 'view', 'save' and 'share'

As shown in table 2, the average values of 'citation', 'view' and 'save' of China and Brazil are lower than that of the United States, Germany, Japan and Russia from 2009 to 2018, the former are developing countries and the latter are developed countries. Therefore, the geographical regions have effect on these three indicators because different countries have different cultural backgrounds, information behavior habits and even the quality of papers. But the 'share' is an exception, the 'share' of Brazil are the biggest among the 6 countries in 10 years period, which may due to that Brazil published more papers in *PLoS Neglected Tropical Diseases Journal* than the other 5 journals.

Figure 3 and Table 4 display that the peaks of 'citation', 'view', 'save' and 'share' are different among the 6 countries. For the 'citation', the peak of the United States occurred in 2011 and the maximum citation age of the papers published by the United States is 8 years, i.e. these papers are active and vigorous in recent 8 years after published. The peak of China occurred earliest among the 6 countries, whose maximum citation age is 6 years. These manifest that the different geographical regions also influence the time of active and vigorous of papers.

The relationship among the 'citation', 'view', 'save' and 'share'

The means of the 'citation', 'view' and 'save' of *PLoS* papers in recent 10 years are bigger than that in latest 5 years among the 6 countries in Table 2, which indicates that these three indicators have the common characteristic as their values all increase with the extend of publication time. However, the change of 'share' mean is on the contrary. Although articles will be shared quickly in 1 year or 2 after published, they will not be mentioned anymore as time goes on, i.e., the ‘share’ is transient and timely, whose life cycle is shorter than that of the other 3 indicators.

Table 3 and Figure 3 also show that ‘citation’, ‘view’ and ‘save’ have significant correlations and similar variation patterns and accumulate with time for the 6 countries, i.e. the newly published papers need a maturity period to get more ‘citation’ (Syamili and Rekha, 2017), ‘view’ and ‘save’. Hence, we could induce that the ‘view’ and ‘save’ --- two kinds of altmetrics index, can evaluate the scientific influence as a complement measurement of traditional citation analysis, though they do not reflect the same kind of impact as citation(Costas, Zahedi, & Wouters, 2015). Due to the convenience of open access platforms like *PLoS* and the wide use of online reference managers such as *Mendeley*, *Endnote* and *CiteULike* etc., more and more readers can easily access the scientific papers, and freely view, download or save in *Mendeley* etc. Though the role of open access and social media in promoting dissemination of scientific papers couldn’t be ignored, we should be careful not to overstate the value of altmetrics. Cameron(2015) said that scientific assessor should recognize the limitation of altmetrics when using it to evaluate the influence of country, institution and individual. As analyzed in this paper, the ‘share’ has no significant correlation with citation in this paper.

The analysis of the ‘share’

The ‘share’ of *PLoS* platform means that the total number of a paper discussed or mentioned by *Twitter*, *Facebook* etc. It can be seen from Table 2 that the mean of ‘share’ of papers published in latest 5 years is more than that in recent 10 years, the curves of ‘share’ for the 6 countries start to grow from 2011 and quickly reach their peaks in 2015 or 2016 (the date of collect data is April 19, 2019) in Figure 3. So the articles published in latest 5 years, especially in latest 3 years would be discussed easily by the users of *Twitter*, *Facebook* etc., i.e., the life cycle of ‘share’ of *PLoS* papers is short, which can be instantly tweeted after publication, while their citation need a the long time to accumulate. Moreover, considering the weak correlation between ‘share’ and ‘citation’ according to Table 3, we believe that the ‘share’ of *PLoS* papers seemingly just reflects the attention/mention of user rather than the academic influence.

Conclusion

Through analysis of the Article-Level Metrics data of papers published in *PLoS* series journals by different countries, the following conclusions are shown: Firstly, the geographical region has effect on the citation and the altmetrics of ‘view’ and ‘save’ (excepting ‘share’), because the mean of each of them in China and Brazil is less than that in USA, Germany, Japan and Russia from 2009 to 2018, and the curve peaks occurred in different years among the 6 countries. As is known, the former two are developing countries while the latter four are developed countries, and the United States is the only one whose native language is English of the 6 countries. Hence, we induced that the level of scientific development in different countries has more effect to these 3 indicators than the native language. For example, with the rapid development of science and technology in China, whose mean of citation has been slightly higher than the developed countries in latest 5 years. Secondly, ‘citation’ and ‘view’, ‘save’ are significantly correlated with each other and have the similar variation patterns and accumulate with year, while the ‘share’ shows the difference. Therefore, to some extent the ‘view’ and ‘save’ can evaluate scientific influence of papers as a complement measurement of traditional citation analysis. For example, extending the author-based influence measurement to the reader’s range. And the ‘share’ of *PLoS* papers seemingly just reflects the attention/mention of user rather than the academic influence because it is transient and uncorrelated with the citation.

In addition, it’s necessary to discuss the influence factors of relationship between citation and altmetrics of open access papers in detail, such as time, user, discipline or subject, social media etc., and the quantity of the sample in this paper also limit the universality of conclusions. In future, we will make a further research for the evaluation of open access papers from different platforms, different disciplines and different periods etc.

Funding

The work is supported by Chinese National Social Science Foundation, which is “Research on the Change Law and Model Reconstruction of Academic Information Exchange System in Digital Scientific Research Environment (Code:17BTQ071)”.

References

Abbasi, A. & Jaafari, A. (2013). Research impact and scholars’ geographical diversity. *Journal of Informetrics*, 7(3), 683-692.

Bartneck, C. & Hu, J. (2010). The fruits of collaboration in a multidisciplinary field. *Scientometrics*, 85(1), 41-52.

Borchardt, R. & Roemer, R. C. (2013). Institutional altmetrics and academic libraries. *Information Standards Quarterly*, 25(2), 14-19.

Bornmann, L. (2015). Alternative metrics in scientometrics: a meta-analysis of research into three altmetrics. *Scientometrics*, 103(3), 1123-1144.

Bornmann, L. & Haunschild, R. (2017). Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? *Scientometrics*, 110(2),937-943.

Cameron B. (2015). The use of altmetrics as a tool for measuring research impact. *Australian Academic & Research Libraries*, 46(2), 121-134.

Chavda, J. & Patel, A. (2016). Measuring research impact: bibliometrics, social media, altmetrics, and the BJGP. *British Journal of General Practice*, 66(642), e59-61.

Chisolm, M. S. (2017). Altmetrics for Medical Educators. *Academic Psychiatry*, 41(4), 460-466.

Costas, R., Zahedi, Z. & Wouters, P. (2015). Do "Altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10): 2003-2019

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G. & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: randomised controlled trial. *British Medical Journal*, 337:a568.

Dhiman, A. K. (2015). Bibliometrics to Altmetrics: Changing Trends in Assessing Research Impact. *Desidoc Journal of Library & Information Technology*, 35(4), 310-315.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471-479.

Garfield, E. (1999). Journal impact factor: A brief review. *Canadian Medical Association Journal*, 161(8), 979-980.

Glanzel, W. & Gorraiz, J. (2015). Usage metrics versus altmetrics: confusing terminology? *Scientometrics*, 102(3), 2161-2164.

Gorraiz, J., Reimann, R. & Gumpenberger, C. (2012). Key factors and considerations in the assessment of international collaboration: a case study for Austria and six countries. *Scientometrics*, 91(2), 417-433.

Hirsch, J. E. (2010). An index to quantify an individual's scientific reseach output. *Scientometrics*, 102(3), 16569-16572.

Holmberg, K., Didegah, F. & Bowman, T. D. (2015). The different meanings and levels of impact of altmetrics. Accepted for oral presentation at the 11th International Conference on Webometrics, Informetrics and Scientometrics & 16th COLLNET Meeting, 26-28 November, New Delhi, India.

Hu, D. H. & Chang, X. W. (2008). An Evaluation of the Quality and Impact of the Open Access Journals' Articals. *Library and Information Service*, 52(2), 61-64.(in chinese)

Joint, N. (2009). The Antaeus column: does the ‘open access’ advantage exist? A librarian’s perspective. *Library Review*, 58(7), 476-481.

Lawrence , S. (2001). Free online availability substantially increases a paper's impact. *Nature*, 411(6837), 521-521.

Leeuwen, T. N. V. & Moed, H. F. (2005). Characteristics of journal impact factors: The effects of uncitedness and citation distribution on the understanding of journal impact factors. *Scientometrics*, 63(2), 357-371.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2006). Measures for measures. *Nature*, 444(7122), 1003-1004.

Moed, H. F. (2007). The effect of "Open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13), 2047-2054.

Moed, H. F. & Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, 66(10), 1988-2002.

Norris, M., Oppenheim, C. & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, 59(12), 1963-1972.

Patel, A. & Chavda, J. (2016). Measuring research impact: bibliometrics, social media, altmetrics, and the BJGP. *British Journal of General Practice*, 66(642), 59-61.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C. & Gorraiz, J. (2016). Research data explored: an extended analysis of citations and altmetrics. *Scientometrics*, 107, 723-744.

Priem, J., Taraborelli, D.,Groth, P. et al.(2010). Altmetrics:a manifesto[EB/OL].[2017-11-08].http://altmetrics.org/manifesto.

Shu, F. H., S. (2017). On the citationadvantage of tweeted papers atthe journal level. *Proceedings of the Association for Information Science and Technology*, 54(1), 366-342.

Stransky, S. G. (2016). The fourth amendment and bulk telephone meta data an overview of recent case law. *Saint Louis University School of Law*, 35(1), 3.

Syamili, C. & Rekha, R. V. (2017). Do altmetric correlate with citation? : A study based on PLoS ONE journal. *COLLNET Journal of Scientometrics and Information Management,* 11(1), 103-117.

Thelwall, M., Haustein, S., Larivière, V. & Sugimoto, C.R. (2013). Do altmetrics work? Twitter and ten other social web services. *PloS One*, 8(5), e64841.

Thelwall, M. & Kousha, K. (2017). ResearchGate articles: Age, discipline, audience size, and impact. *Journal of the Association for Information Science and Technology*, 68(2), 468-479.

Waltman, L. & Costas, R. (2014). F1000 Recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, 65(3), 433-445.

Wang, J., Alotaibi, N. M., Ibrahim, G. M., Kulkarni, A. V. & Lozano, A. M. (2017). The Spectrum of Altmetrics in Neurosurgery: The Top 100 "Trending" Articles in Neurosurgical Journals. *World Neurosurg*, 103, 883-895.

Yassine, G., Chawki, H., Larivière Vincent, Yves, G., Les, C. & Tim, B., et al. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE*, 5(10), e13636.

Characterizing the Potential of Being Emerging Generic Technologies: A Bi-Layer Network Analytics-based Prediction Method

Yi Zhang[1], Yihe Zhu[2], Lu Huang[2], Guangquan Zhang[1] and Jie Lu[1]

[1] yi.zhang@uts.edu.au; guangquan.zhang@uts.edu.au; jie.lu@uts.edu.au
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

[2] zhuyihe12@foxmail.com; huanglu628@163.com
School of Management and Economics, Beijing Institute of Technology (China)

Abstract

Despite tremendous involvement of bibliometrics in profiling technological landscapes and identifying emerging topics, how to predict potential technological change is still unclear. This paper proposes a bi-layer network analytics-based prediction method to characterize the potential of being emerging generic technologies. Initially, based on the innovation literature, three technological characteristics are defined, and quantified by topological indicators in network analytics; a link prediction approach is applied for reconstructing the network with weighted missing links, and such reconstruction will also result in the change of related technological characteristics; the comparison between the two ranking lists of terms can help identify potential emerging generic technologies. A case study on predicting emerging generic technologies in information science demonstrates the feasibility and reliability of the proposed method.

Introduction

An early definition of emerging generic technologies can be traced back to the early 1990s, highlighting technologies that enable revolutionary impacts on the economy and society (Martin, 1995), and Maine and Garnsey (2006) moved down on the line and specified the 'generic' nature as benefits on a wide range of sectors and the 'emerging' nature as the potential for innovation. Even though emerging generic technologies conceptually contain overlaps with emerging technologies (Rotolo et al., 2015), it is clear that, compared to technologies that hold disruptive powers to a given technological area – e.g., dye sensitized solar cells (Zhang, Zhou, et al., 2014), the exploitations and applications of emerging generic technologies would create values for fostering innovation in broad disciplines (Coccia, 2017) – e.g., nanotechnology (Maine & Thomas, 2017). During the past decades, investigations on how to measure the impacts of emerging generic technologies on accelerating the economic growth (Bresnahan & Trajtenberg, 1995; Crafts, 2004; Qiu & Cantwell, 2018) and how to transfer technological breakthroughs into impactful innovations (Appio et al., 2017; Sinfield & Solis, 2016) have been conducted in the innovation literature.

The engagement of bibliometrics on assisting in the management of technology has been well observed, e.g., profiling a given technological area (Chakraborty et al., 2015; Guo et al., 2010), identifying emerging topics in science and technology (Glänzel & Thijs, 2012; Small et al., 2014), and tracking the pathways of technological change (Hou et al., 2018; Zhang et al., 2016; Zhou et al., 2014). The use of advanced information technologies, e.g., topic models, streaming data analytics, and machine learning techniques, greatly strengthens the capability of traditional bibliometrics in handling large-scale data analytics (Ding & Chen, 2014; Klavans & Boyack, 2017), discovering hidden relationships (Zhang et al., 2018; Zhang et al., 2017), and visualizing complicated landscapes and structures (Börner et al., 2012; Suominen & Toivanen, 2016).

Even though it has been a long time since the use of network analytics in social science (Borgatti et al., 2009), network analytics was introduced to bibliometric studies in the late 2000s, which were initially used to investigate research collaborations and disciplinary interactions through analyzing bibliographic couplings (Yan et al., 2009; Yang et al., 2010). Its effective combination with citation networks has attracted great attentions on identifying emerging topics and evaluating research impacts (Takeda & Kajikawa, 2009; Yan, 2015). Such advantages have been applied for predicting emerging technologies (Érdi et al., 2013) and discovering technological opportunities (Park & Yoon, 2018). However, despite recognitions from the both communities, concerns are still raised, e.g., bibliometrics is insufficient on 'characterizing the potential of what is detected to be emerging' (Rotolo et al., 2015). Additionally, with the rapid development of natural language processing (NLP) techniques, co-word statistics provide a new angle for bibliometrics, but how to explore insights based on semantics retrieved from co-word-based networks is still elusive. Apparently, such insights are complementary with citation networks.

Aiming to address these concerns, this paper is to propose a bi-layer network analytics-based prediction method for characterizing the potential of being emerging generic technologies. Initially, we refer studies conducted by Maine and Garnsey (2006) and Rotolo et al. (2015), and consider emerging generic technologies as novel and fast-growing technologies with prominent impacts on a relative broad range of disciplines. A co-authorship network and a co-term network are constructed and integrated as a bi-layer network to represent the content of involved disciplines/technologies, and indicators for profiling the topological structure of networks are introduced to identify technological characteristics from three aspects – i.e., fundamentality, connectivity, and externality. Further, a link prediction approach is incorporated to calculate weights of all links (including missing links) in the both networks, and such reconstruction would be the key to capture a potential characteristic change of involved technologies. Thus, investigating such change could be the way of characterizing the potential of being emerging generic technologies. We then demonstrate the feasibility and reliability of the proposed method through a case study, which predicts emerging generic technologies in information science disciplines by analyzing 17,882 articles published in 15 selected journals and conference proceedings in the field between Jan 1, 2000 and Dec 31, 2016.

The rest of this paper is organized as follows: The Methodology section describes the details of the proposed bi-layer network analytics-based prediction method, and the Case Study section follows, presenting the data, results, and empirical insights derived from the case. We then conclude our study and outline potential future directions.

Methodology

The research framework of the proposed bi-layer network analytics-based prediction method for characterizing the potential of being emerging generic technologies is given in Fig. 1.

Technological characteristics

Following the studies conducted by Maine and Garnsey (2006) and Rotolo et al. (2015), we identify new technologies that can be fundamentally applied to a broad range disciplines, with capabilities of connecting diverse technological areas and adaptively transferring among enterprises, as emerging generic technologies, and the characteristics of emerging generic technologies are specifically defined from the following three perspectives:

- Fundamentality is to measure *whether this technology can be applied to a broad range of sectors, disciplines, or research areas*.

- Connectivity is to measure *whether this technology is sharing close relationships with other technologies in the same or different technological areas*.
- Externality is to measure *whether this technology is involved and can be transferred among diverse enterprises and research groups*.

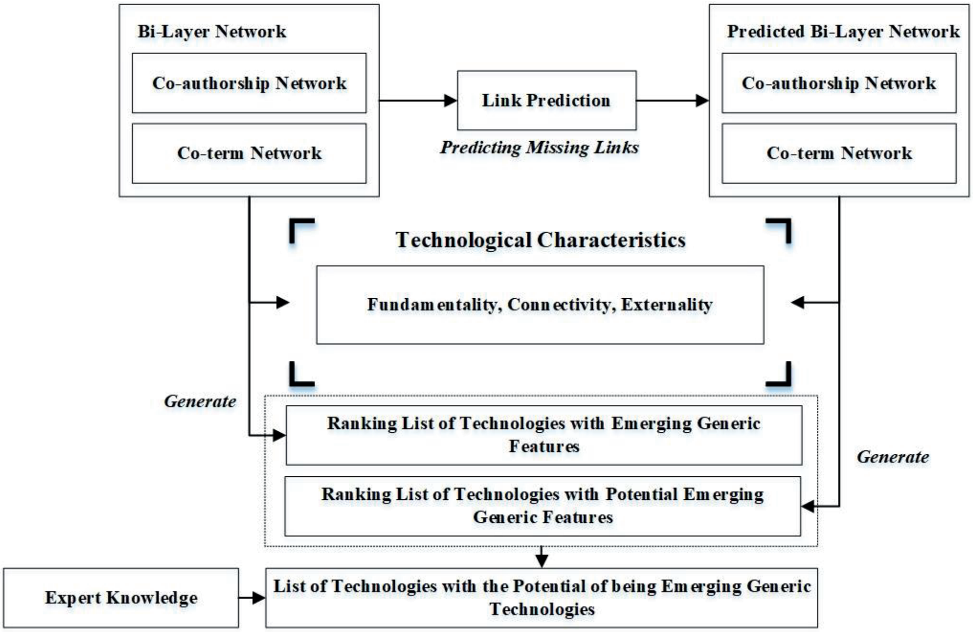

Bi-Layer Network
Co-authorship Network
Co-term Network
Link Prediction
Predicting Missing Links
Predicted Bi-Layer Network
Co-authorship Network
Co-term Network
Technological Characteristics
Fundamentality, Connectivity, Externality
Generate
Generate
Ranking List of Technologies with Emerging Generic Features
Ranking List of Technologies with Potential Emerging Generic Features
Expert Knowledge
List of Technologies with the Potential of being Emerging Generic Technologies

Figure 1. Research framework of the bi-layer network analytics-based prediction method.

Bi-layer network analytics

A bi-layer network includes a co-term network and a co-authorship network. We denote $N = \{(V^t, E^t), (V^a, E^a), E^{at}\}$ as a bi-layer network, in which (V^t, E^t) and (V^a, E^a) are the sets of nodes and links in the co-term network and the co-authorship network respectively and E^{at} is the set of links between the two networks. A sample of a bi-layer network is given in Fig. 2.

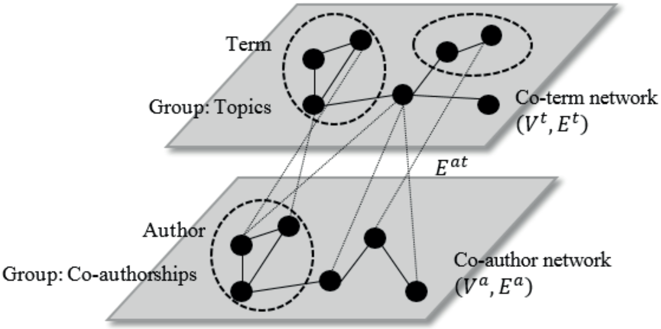


Figure 2. Sample of a bi-layer network.

Specifically, the co-term network is generated based on the co-occurrence statistics of terms derived from the title and abstract fields of collected records, and the co-authorship network is based on the statistics of co-authorship behaviors. The both networks are non-direct graphs, in which 1) each node represents either a term or an author and 2) each link represents the co-occurrence/co-authorship relationships between connected nodes and is weighted by the frequency of such co-occurrence/co-authorship. Significantly, the authorships of terms are used to be the links between the two layers – i.e., the co-term network and the co-authorship network.

When considering each term represents a technological component (e.g., materials, functions, manufacturing processes, and applications), we apply network analytics for investigating the topological structures of the bi-layer network to quantify the three characteristics of emerging generic technologies.

- Fundamentality

The fundamentality of a technology is to measure the breadth and depth of a technology's influence in given technological areas. In the co-term network, centrality, a traditional indicator of measuring network topological structures (Freeman, 1977, 1978), is exploited to quantify the value of the fundamentality. Since a number of centrality-based indicators have been developed for different emphases, three forms of centrality are involved in this study:

1) Degree Centrality – the degree of a node, reflecting the breadth of its potential influence. The degree centrality of node v_i^t in the co-term network can be calculated as follows:
$$DC(v_i^t) = \frac{\sum_{j=1}^{|V^t|} w_{v_i^t,v_j^t}}{|V^t| - 1}$$
where $|V^t|$ is the number of nodes in the co-term network and $w_{v_i^t,v_j^t}$ is the weight of the link between node v_i^t and node v_j^t.
2) Closeness Centrality – the closeness between a node and other nodes in the same network, reflecting its professionalism in a given area, that is, the depth of its potential influence. The closeness centrality of node v_i^t can be calculated as follows:
$$CC(v_i^t) = \frac{|V^t| - 1}{\sum_{j=1}^{|V^t|} d_{v_i^t,v_j^t}}$$
where $d_{v_i^t,v_j^t}$ is the shortest distance between node v_i^t and node v_j^t.
3) Between Centrality – the number of the shortest paths crossing a node, reflecting its role in a cross area, that is, its potential influence in a cross-disciplinary direction. The between centrality of node v_i^t can be calculated as follows:
$$BC(v_i^t) = \frac{2\sum \frac{\sigma(v_i^t)_{v_s^t,v_p^t}}{\sigma_{v_s^t,v_p^t}}}{(|V^t| - 1)(|V^t| - 2)}, v_i^t \neq v_s^t \neq v_p^t$$
where v_s^t and v_p^t are two different nodes in the network, $\sigma_{v_s^t,v_p^t}$ represents the number of the shortest paths between nodes v_s^t and v_p^t, and $\sigma(v_i^t)_{v_s^t,v_p^t}$ is the number of the shortest paths between nodes v_s^t and v_p^t, crossing node v_i^t.

The three forms of centrality exploit different topological structures – e.g., degree centrality concentrates on the number of neighbor nodes, closeness centrality highlights the capability of connecting other nodes, and between centrality emphasizes the importance of a node in the communication of a network. In other words, the three forms of centrality could cover the breadth and depth of a technology's influence in a given technological area, as well as its potential influence in a cross-disciplinary direction. Given the circumstances, we exploit the three for calculating the fundamentality of a technology, and we calculate the fundamentality of a node $F(v_i^t)$ as the average value of the three indicators.

- Connectivity

The connectivity of a technology is considered as its relationships with other technologies and technological groups, indicating its capability of involving diverse sectors, disciplines, and technological areas. In the co-term network, 1) initially, a smart local moving algorithm (Waltman & Van Eck, 2013) is applied for community detection – i.e., identifying technological groups G^t; and then, 2) we calculate the connectivity $C(v_i^t)$ between node v_i^t and its community as follows:

$$C(v_i^t) = \frac{\sum_{j=1}^{|G^t(v_i^t)|} w_{v_i^t,v_j^t}}{|G^t(v_i^t)|}$$

where $|G^t(v_i^t)|$ represents the number of nodes in the community to which node v_i^t belong.

- Externality

The externality of a technology takes both technologies and their owners into considerations – i.e., if a technology is owned by more than one owner (e.g., enterprises and research institutions) and can be easily transferred between those owners, or even between different sectors, we consider this technology is generic in the related fields. Thus, both the co-term network and the co-authorship network in a bi-layer network will be exploited for measuring the externality of a node $E(v_i^t)$ as follows:

$$L(v_m^a) = \sum_{n=1}^{|V^a|} w_{v_m^a,v_n^a}$$

$$E(v_i^t) = \sum_{n=1}^{|V^a|} w_{v_i^t,v_n^a} \times L(v_n^a)$$

where v_m^a is a node in the co-authorship network and $|V^a|$ is the number of nodes in the co-authorship network.

Despite some weighting approaches, e.g., entropy-based and standard deviation-based weights, we decide to use a 3D map to visualize values of the three technological characteristics, highlighting distinctive values based on diverse requirements and preferences.

Link prediction

A common neighbors (CN)-based link prediction approach (Newman, 2001) is exploited to weight all links (including missing links) in the bi-layer network, and such reconstruction of the network would represent possible connections between terms and potential collaborations between authors in future. The basic assumption of the CN-based approach is that if two unlinked nodes have many common neighbors, it is highly possible that a link will appear between the two nodes. Thus, the CN value of each link can be calculated as follows:

$$CN(v_x, v_y) = \sum_{z=1}^{|V(v_x,v_y)|} \left(w_{v_x,v_z} + w_{v_y,v_z}\right)$$

where v_x and v_y are two different and unlinked nodes in a bi-layer network (either the co-term network or the co-authorship network) and $|V(v_x, v_y)|$ is the set of nodes in the bi-layer network, which connect v_x and v_y.

The output of this link prediction approach is a ranking list of all links in the bi-layer network, including missing links in the current network. Thus, a predicted bi-layer network will be generated, reflecting potential technological change in the near future.

Identification of emerging generic technologies

According to the technological characteristics, a ranking list (List A) of technologies with emerging generic features will be generated based on a bi-layer network. With the exploitation of link prediction approaches, missing links in the bi-layer network will be created and existing links will be re-weighted, i.e., a predicted bi-layer network is constructed. Apparently, the change of the topological structure of the existing network will result in the change of the technological characteristics of related technologies, and thus, a new ranking list (List B) will be generated. Therefore, comparing the two lists respectively generated by the two bi-layer networks will help characterize the potential of being emerging generic technologies. Several selection criteria will be highlighted, including:

- A technology only appears in List B and with a high rank;
- Compared to List A, the rank of a technology in List B dramatically increase;
- A technology appears in the top rank of the both lists;

Case Study: What are emerging generic technologies in information science?

It would likely be arguable that information science can only represent an individual discipline and it is critical to identify emerging generic technologies from such one discipline rather than a broad range of disciplines. Our consideration here is that information science has been spearheading a cross-disciplinary direction that bridges fundamental studies (e.g., mathematics, physics, and computer science) with real-world needs raised in disciplines of social science. Therefore, it would be interesting to identify emerging generic technologies from such a cross-disciplinary area, which would originate from other disciplines but build up the foundations of information science and create extensive impacts on and out of the discipline. We followed the search strategy proposed by Hou et al. (2018) and selected 15 journals and conference proceedings, covering 17,445 records between January 1, 1996 and December 31, 2016.

Table 1. List of selected journals and conference proceedings

Journal Name	*Journal Name*
Annual Review of Information Science and Technology	Library Resources & Technical Services
Information Processing & Management	Program: Automated Library and Information Systems
Journal of the Association for Information Science and Technology	Information Research
Journal of Documentation	Journal of Informetrics
Journal of Information Science	Research Evaluation
Library & Information Science Research	The Electronic Library
ASIS&T Annual Meeting Proceedings	Information Technology and Libraries
Scientometrics	

Note that the table only lists the current names of selected journals and conference proceedings, but we fully considered their previous names when collecting data.

We combined the title and abstract fields of the 17,445 records and retrieved 213,031 terms by a natural language processing (NLP) function integrated in the VantagePoint[1]. A term clumping process (Zhang, Porter, et al., 2014) was applied for data cleaning by removing noise and consolidating synonyms, and the stepwise results are given in Table 2. The 25,359 terms were used for constructing the co-term network.

[1] VantagePoint is a software platform for bibliometrics-based text analytics and knowledge management, owned by Search Technology Inc. More details can be found at the website: www.vantagepoint.com.

Table 2. Stepwise results of term clumping

Step	*Description*	*#Terms*
0	Raw terms retrieved by the NLP technique;	213,031
1	Remove single-word terms, e.g., “information”;	189,111
2	Remove terms starting/ending with non-alphabetic characters, e.g., “step 1” and “1.5 m/s”;	180,209
3	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions;	175,488
4	Remove common terms in scientific articles, e.g., “research framework”;	157,041
5	Consolidate synonyms based on expert knowledge, e.g., “co-word analysis” and “word co-occurrence analysis”;	135,967
6	Consolidate terms with the same stem, e.g., “information system” and “information systems”	109,115
7	Remove terms appearing less than 3 times;	25,359

Note that: 1) Expert knowledge in Step 5 were mostly based on previous experiments and experiences; and 2) we usually remove terms appearing once in the dataset, but we decided to increase the threshold to keep the scale of terms at a relatively small level in Step 7.

Regarding to author names, we collected 5349 distinctive authors from a raw list of 18,882 authors, and the cleaning process includes: 1) a light author name disambiguation function integrated in the VantagePoint was applied to consolidate potential variations – e.g., “Eugene Garfield”, “Garfield, Eugene”, and “E Garfield”; and 2) authors who only published one paper in our dataset were removed. The co-authorship network was then constructed.

Thus, a bi-layer network was built up by connecting the co-term network and the co-authorship network with links, representing the authorships of terms and weighted by the frequency. a demonstration of the bi-layer network in VOSViewer (Waltman et al., 2010) is given in Fig. 3. Note that links between the co-term network and the co-authorship network are not given.

Network analytics were applied to quantify the three technological characteristics of the 25,359 terms, and the descriptive statistics of the results are given in Table 3. Based on the mean of the three characteristics, we selected 1000 terms and generated one 3D map in Fig. 4 (Left), visualizing and locating distinctive terms in a 3D solution.

Table 3. Descriptive statistics for technological characteristics.

No.	*Characteristics*	*Sub-characteristics*	*Max*	*Min*	*Mean*	*S.D.*
1	Fundamentality	Degree Centrality	1	0	0.014	0.023
		Closeness Centrality	1	0	0.690	0.068
		Between Centrality	1	0	0.001	0.008
		Average	1	0	0.235	0.028
2	Connectivity	N/A	7.47	0	0.050	0.220
3	Externality	N/A	20853	0	154.3	437.5

Note that regarding to fundamentality, we used the average of the three sub-characteristics as the value of fundamentality in further analytics.

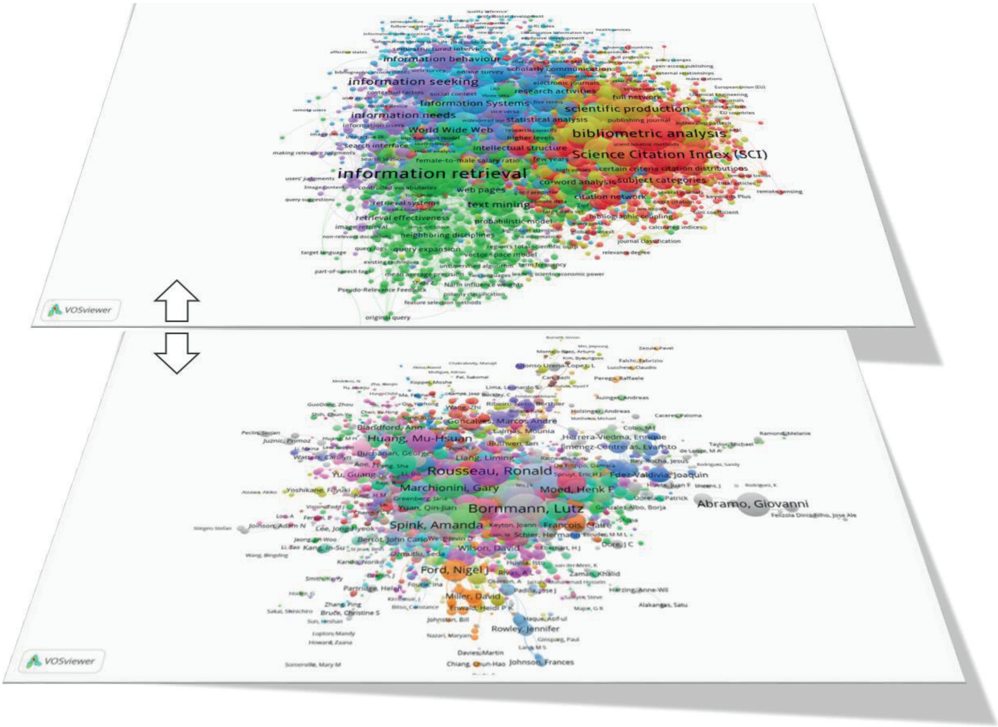


Figure 3. A bi-layer network for information science.

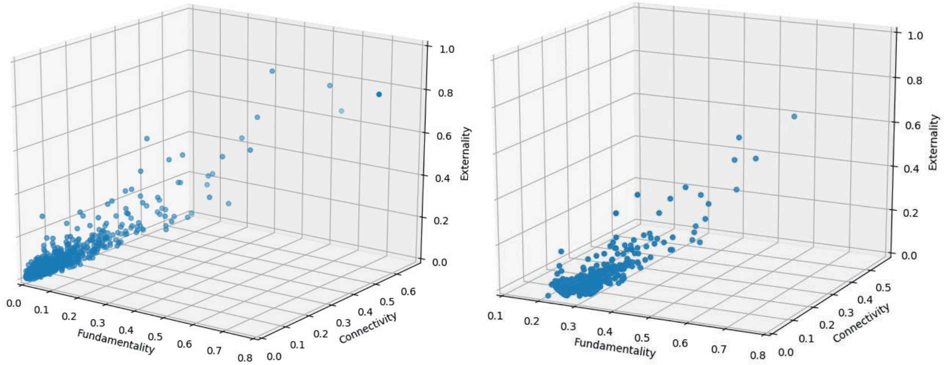


Figure 4. 3D map for 1000 terms with technological characteristics – Left for the current bi-layer network and Right for the predicted bi-layer network.

The common neighbor-based link prediction approach was then applied to calculate the CN values of all links, including missing links. With such values, the structure of the bi-layer network was changed and the technological characteristics of all nodes could be re-calculated. The descriptive statistics for technological characteristics in the predicted bi-layer network are given in Table 4, and a 3D map for visualizing selected 1000 terms with technological characteristics is given in Fig. 4 (Right).

Table 4. Descriptive statistics for technological characteristics in the predicted bi-layer network.

No.	*Characteristics*	*Sub-characteristics*	*Max*	*Min*	*Mean*	*S.D.*
1	Fundamentality	Degree Centrality	1	0	0.010	0.036
		Closeness Centrality	1	0	0.001	0.014
		Between Centrality	1	0	0.693	0.063
		Average	1	0	0.235	0.031
2	Connectivity	N/A	1	0	0.001	0.015
3	Externality	N/A	1	0	0.027	0.023

We exploited the receiver operating characteristic (ROC) analysis and the value of area under the curve (AUC) to validate the performance of the link prediction approach (Fawcett, 2006). Briefly, in the ROC analysis the applied dataset was randomly divided into a training set and a test set, then, the ranking list generated by the link prediction approach in the training set would be compared with the true ranking list in the test set, and an AUC value can be calculated. The AUC values for links in the co-term network, in the co-authorship network, and between the two networks are given in Fig. 5.

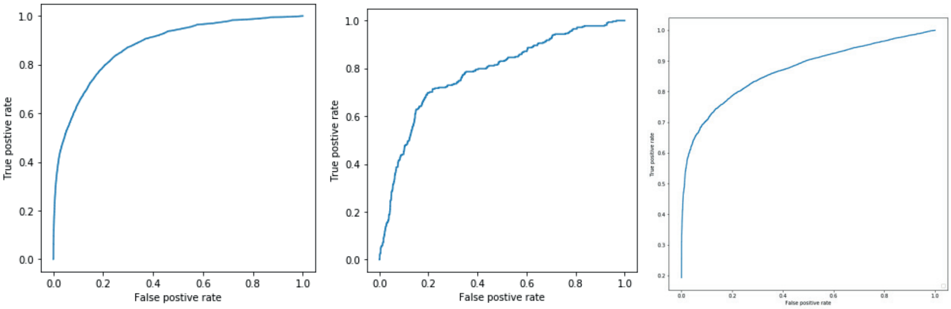


Figure 5. AUC values for validating the link prediction approach – Left for links in the co-term network, Middle for links in the co-authorship network, and Right for links between the both.

As shown in Fig. 5, AUC values for links in the co-term network, in the co-authorship network, and between the two networks are 0.89, 0.79, and 0.95 respectively, indicating an acceptable result of the link prediction approach.

We then compared the difference between the two ranking lists (i.e., the ones generated by the bi-layer network and the predicted bi-layer network respectively) and picked up a list of terms (given in Table 5) whose ranking is within the Top 50 in the one generated by the predicted bi-layer network but largely different from the previous one, indicating its potential of being emerging generic technologies in information science.

Table 5. Selected terms indicating the potential of being emerging generic technologies

No.	*Terms*	*Rank Change*	*No.*	*Terms*	*Rank Change*
1	Information retrieval	63 - 1	6	Text mining	383 - 24
2	Information seeking	15 - 4	7	Social network analysis	52 - 28
3	Digital libraries	23 - 9	8	Science policy	54 - 36
4	Information systems	45 - 17	9	Co-authorship network	79 - 43
5	H index	169 - 22			

Note that compared to emerging generic "technologies", we would prefer to extend this concept and consider these selected terms as emerging generic "research topics" to highlight their emphasis on scientific research.

Terms appearing in Table 5 are coherent with the study conducted by Zhang et al. (2018), where several key topics in bibliometrics were identified. Several insights are summarized below:

- As a fundamental toolkit, the involvement of *information retrieval* (e.g., *text mining*) and *information systems* techniques has significantly changed the information science discipline, but with the rapid development of information technologies, especially artificial intelligence, the involvement would be further enhanced and become an emergent direction in information science.
- *Information seeking* and *digital libraries* would be considered as two mainstream tasks of information science and library science, and the boom of social media would become a key to dramatically extend its current research areas and generate new topics.
- *Social network analysis* and *co-authorship network* are the applications of complex network analytics for analyzing science maps, which could be a cross-disciplinary direction and have attracted great attention in the past decades.
- *H index* is a traditional indicator for research evaluation in bibliometrics and could be considered as the application of complex network analytics as well. How to modify h index to evaluate researchers and research institutions from comprehensive aspects is still a hot topic in bibliometrics.
- *Science policy* could be a practical area of information science (e.g., bibliometrics). Even though such applications have appeared in the literature for decades, new problems in the area of science, technology, and innovation policy (STIP), and new solutions for existing STIP problems are still challenging researchers in information science.

Conclusions and Future Studies

This paper provides a bi-layer network analytics-based prediction method for characterizing the potential of being emerging generic technologies, in which 1) three technological characteristics are identified and then quantified by topological indicator, and 2) a common neighbor-based link prediction approach is applied for reconstructing networks with weighted missing links. Comparison between the ranking lists of terms indicating the potential of being emerging generic technologies, which are respectively generated by the current and the reconstructed networks, is used to identify potential emerging generic technologies. A case study on predicting emerging generic technologies in information science demonstrates the feasibility and reliability of the proposed method.

Future directions can be conducted to address limitations of this study from the following aspects: 1) it would be crucial to build up the conceptual foundation of this study in our further studies, which might provide solid theoretical support for the proposed method; 2) a modified link prediction approach can be developed to better adapt to a bi-layer network, and comparisons with baselines can be applied as well; 3) it is more convincing to quantitatively or qualitatively validate the results based on different indicators and with diverse practical needs; and 4) examining the proposed method in cases with relatively broad disciplines would further help demonstrate its reliability.

Acknowledgments

This work is supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994.

References

Appio, F. P., Martini, A., & Fantoni, G. (2017). The light and shade of knowledge recombination: Insights from a general-purpose technology. *Technological Forecasting and Social Change, 125*, 154-165.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323*(5916), 892-895.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., . . . Boyack, K. W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS One, 7*(7), e39464.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'? *Journal of econometrics, 65*(1), 83-108.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2015). On the categorization of scientific citation profiles in computer science. *Communications of the ACM, 58*(9), 82-90.

Coccia, M. (2017). The source and nature of general purpose technologies for supporting next K-waves: Global leadership and the case study of the US Navy's Mobile User Objective System. *Technological Forecasting and Social Change, 116*, 331-339.

Crafts, N. (2004). Steam as a general purpose technology: A growth accounting perspective. *The Economic Journal, 114*(495), 338-351.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C‐word, and cocitation methods. *Journal of the Association for Information Science and Technology, 65*(10), 2084-2097.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics, 95*(1), 225-242. doi:10.1007/s11192-012-0796-4

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Glänzel, W., & Thijs, B. (2012). Using "core documents" for detecting and labelling new emerging topics. *Scientometrics, 91*(2), 399-416. doi:10.1007/s11192-011-0591-7

Guo, Y., Huang, L., & Porter, A. L. (2010). The research profiling method applied to nano‐enhanced, thin‐film solar cells. *R&D Management, 40*(2), 195-208.

Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: a document co-citation analysis (2009–2016). *Scientometrics, 115*(2), 869-892.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984-998.

Maine, E., & Garnsey, E. (2006). Commercializing generic technology: The case of advanced materials ventures. *Research Policy, 35*(3), 375-393.

Maine, E., & Thomas, V. (2017). Raising financing through strategic timing. *Nature nanotechnology, 12*(2), 93.

Martin, B. R. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management, 7*(2), 139-168.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025102.

Park, I., & Yoon, B. (2018). Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. *Journal of Informetrics, 12*(4), 1199-1222.

Qiu, R., & Cantwell, J. (2018). General purpose technologies and local knowledge accumulation—A study on MNC subunits and local innovation centers. *International Business Review, 27*(4), 826-837.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827-1843.

Sinfield, J., & Solis, F. (2016). Finding a lower-risk path to high-impact innovations. *MIT Sloan management review, 57*(4), 79.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450-1467.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human‐assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(19), 2464–2476.

Takeda, Y., & Kajikawa, Y. (2009). Optics: A bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics, 78*(3), 543-558.

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B, 86*(11), 471.

Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629-635.

Yan, E. (2015). Research dynamics, impact, and dissemination: A topic‐level analysis. *Journal of the Association for Information Science and Technology, 66*(11), 2357-2372.

Yan, E., Ding, Y., & Zhu, Q. (2009). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics, 83*(1), 115-131.

Yang, C., Park, H., & Heo, J. (2010). A network analysis of interdisciplinary research relationships: The Korean government's R&D grant program. *Scientometrics, 83*(1), 77-92.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099-1117.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179-191.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Science evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925-1939.

Zhang, Y., Zhou, X., Porter, A. L., Gomila, J. M. V., & Yan, A. (2014). Triple Helix innovation in China's dye-sensitized solar cell industry: Hybrid methods with semantic TRIZ and technology roadmapping. *Scientometrics, 99*(1), 55-75.

Zhou, X., Zhang, Y., Porter, A. L., Guo, Y., & Zhu, D. (2014). A patent analysis method to trace technology evolutionary pathways. *Scientometrics, 100*(3), 705-721.

Crowdsourcing open citations with CROCI
An analysis of the current status of open citations, and a proposal

Ivan Heibi[1], Silvio Peroni[1] and David Shotton[2]

[1] *{ivan.heibi2, silvio.peroni}@unibo.it*
Digital Humanities Advanced Research Centre (DHARC), Department of Classical Philology and Italian Studies, University of Bologna, via Zamboni 32, 40126 Bologna (Italy)

[2] *david.shotton @oerc.ox.ac.uk*
Oxford e-Research Centre, University of Oxford, 7 Keble Rd, Oxford OX1 3QG (United Kingdom)

Abstract

In this paper, we analyse the current availability of open citations data in one particular dataset, namely COCI (the OpenCitations Index of Crossref open DOI-to-DOI citations; http://opencitations.net/index/coci) provided by OpenCitations. The results of these analyses show a persistent gap in the coverage of the currently available open citation data. In order to address this specific issue, we propose a strategy whereby the community (e.g. scholars and publishers) can directly involve themselves in crowdsourcing open citations, by uploading their citation data via the OpenCitations infrastructure into our new index, CROCI, the Crowdsourced Open Citations Index.

Introduction

The availability of open scholarly citations – i.e. citation data that are *structured*, *separate*, *open*, *identifiable* and *available* (Peroni and Shotton, 2018) – is a public good, which is of intrinsic value to the academic world as a whole (Shotton, 2013; Peroni et al. 2015; Shotton, 2018), and is particularly crucial for the scientometrics and informetrics community, since it supports reproducibility (Sugimoto et al., 2017) and enables fairness in research by removing such citation data from behind commercial paywalls(Schiermeier, 2017). Despite the positive early outcome of the Initiative for Open Citations (I4OC, https://i4oc.org/), namely that almost all major scholarly publishers now release their publication reference lists, with the result that more than 500 million citations are now open via the Crossref API (https://api.crossref.org), and despite the related ongoing efforts of sister infrastructures and initiatives such as OpenCitations (http://opencitations.net) and WikiCite/Wikidata (https://www.wikidata.org), many scholarly citations are not freely available. While these initiatives have the potential to disrupt the traditional landscape of citation availability, which for the past half-century has been dominated by commercial interests, the present incomplete coverage of open citation data is one of the most significant impediments to open scholarship (van Eck et al., 2018).
In this work, we analyse the current availability of open citations data within one particular dataset, namely COCI (the OpenCitations Index of Crossref open DOI-to-DOI citations; http://opencitations.net/index/coci). This dataset is provided by OpenCitations (Peroni et al., 2015), a scholarly infrastructure organization dedicated to open scholarship and the publication of open bibliographic and citation data by the use of Semantic Web (Linked Data) technologies. Launched in July 2018, COCI is the first of the Indexes proposed by OpenCitations (http://opencitations.net/index) in which citations are exposed as first-class data entities with accompanying properties. It has already seen widespread usage (over nine hundred thousands API calls since launch, with half of these in January 2019), and has been adopted by external services such as VOSviewer (van Eck and Waltman, 2010).
In particular, in this paper we address the following research questions (RQs):

1. What is the ratio between open citations vs. closed citations within each category of scholarly entities included in COCI (i.e. journals, books, proceedings, datasets, and others)?

2. Which are the top twenty publishers in terms of the number of open citations received by their own publications, according to the citation data available in COCI?
3. To what degree are the publishers highlighted in the previous analysis themselves contributing to the open citations movement, according to the data available in Crossref?

The results of these analyses show a persistent gap in the coverage of the currently available open citation data. To address this specific issue, we have developed a novel strategy whereby members of the community of scholars, authors, editors and publishers can directly involve themselves in crowdsourcing open citations, by uploading their citation data via the OpenCitations infrastructure into our new index, **CROCI, the Crowdsourced Open Citations Index**.

Methods and material

To answer the RQs mentioned above, we used open data and technologies coming from various parties. Specifically, the open CC0 citation data we used came from the CSV dump of most recent release of COCI dated 12 November 2018 (OpenCitations, 2018), which contains 449,840,503 DOI-to-DOI citation links between 46,534,705 distinct bibliographic entities. The Crossref dump we used for the production of this most recent version of COCI was dated 3 October 2018, and included all the Crossref citation data available at that time in both the 'open' dataset (accessible by all) and the 'limited' dataset (accessible only to users of the Crossref Cited-by service and to Metadata Plus members of Crossref, of which OpenCitations is one – for details, see https://www.crossref.org/reference-distribution/).

We additionally extracted information about the number of closed citations to each of the 99,444,883 DOI-identified entities available in the October 2018 Crossref dump. This number was calculated by subtracting the number of open citations to each entity available within COCI from the value "is-referenced-by-count" available in the Crossref metadata for that particular cited entity, which reports all the DOI-to-DOI citation links that point to the cited entity from within the whole Crossref database (including those present in the Crossref 'closed' dataset).

Furthermore, we extracted the particular publication type of each entity, so as to identify it either as a journal article, or as a book chapter, etc. We determined these publication types for all the DOI-identified entities available in the Crossref dump we used. We then identified the publisher of each entity, by querying the Crossref API using the entity's DOI prefix. This allowed us to group the number of open citations and closed citations to the articles published by that particular publisher, and to determine the top twenty publishers in terms of the number of open citations that their own publications had received.

Finally, we again queried the Crossref API, this time using the DOI prefixes of the *citing* entities, to check the participation of these top twenty publishers in terms of the number of open citations they were themselves publishing in response to the open citation movement sponsored by I4OC. Details of all these analyses are available online in CC0 (Heibi et al., 2019).

Results

First (RQ1) we determined the numbers of open citations and closed citations received by the entities in the Crossref dump. All the entity types retrieved from Crossref were aligned to one of following five categories: journal, book, proceedings, dataset, other – the mapping between Crossref types and the five types we used in our analysis is illustrated in the description of the table "croci_types.csv" in (Heibi et al., 2019). The outcomes are summarised in Figure 1, where it is evident that the number of open citations available in COCI is always greater than the number of closed citations to these entities within the Crossref database to which COCI does not have access, for each of the publication categories considered, with the categories *proceedings* and *dataset* having the largest ratios.

Analysis of the Crossref data show that there are in total ~4.1 million DOIs that have received no open citations and at least one closed citation. Conversely, there are ~10.7 million DOIs that have received no closed citations and at least one open citation in COCI. Most of the papers in both these categories have received very few citations.

The outcome of the second analysis (RQ2) shows which publishers are receiving the most open citations. To this end, we considered all the open citations recorded in COCI, and compared them with the number of closed citations to these same entities recorded in Crossref. Figure 2 shows the top twenty publishers that received the greatest number of open citations. Elsevier is the first publisher according to this ranking, but it also records the highest number of closed citations received (~97M vs. ~105.5M). The highest ratio in terms of open citations vs. closed citations was recorded by IEEE publications (ratio 6.25 to 1), while the lowest ratio was for the American Chemical Society (ratio 0.73 to 1).

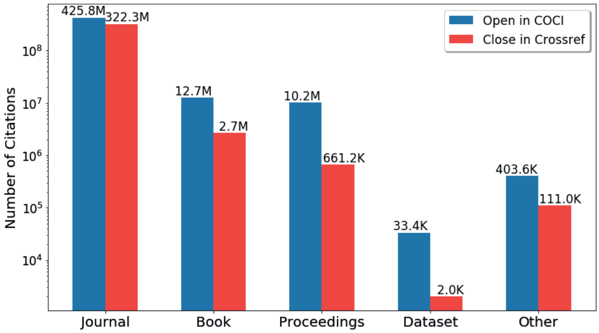


Figure 1. The number of open citations (available in COCI) vs. closed citations (according to Crossref data) received by the cited entities within COCI, analyzed and grouped according to five distinct categories. [Note that the vertical axis has a logarithmic scale].

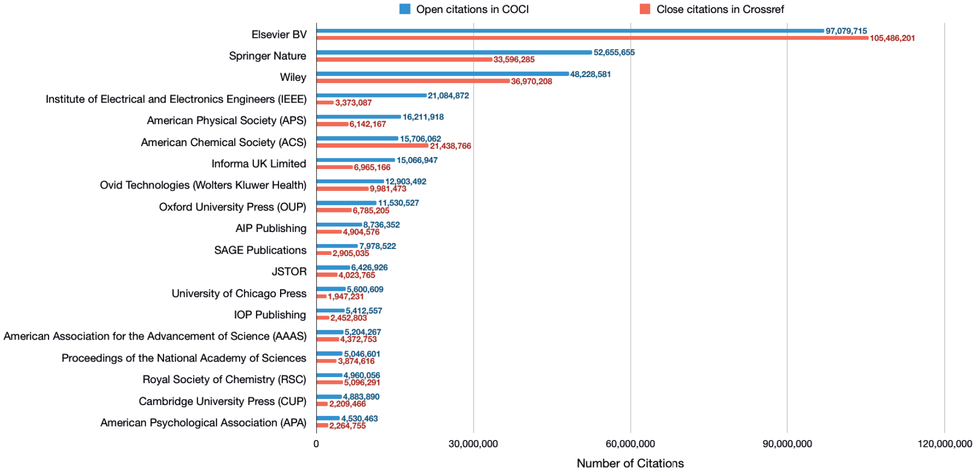
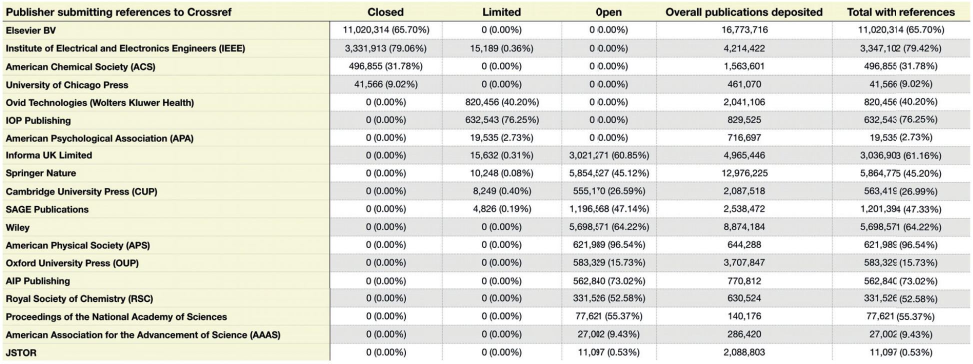


Figure 2. The top twenty publishers sorted in decreasing order according to the number of open citations the entities they published have received, according to the open citation data within COCI. We accompany this count with the number of closed citations to the entities published by each of them according to the values available in Crossref.

Considering the twenty publishers listed in Figure 2, we wanted additionally to know their current support for the open citation movement (RQ3). The results of this analysis (made by querying the Crossref API on 24 January 2019) are shown in Figure 3. Among the top ten publishers shown in Figure 2, i.e. those who themselves received the largest numbers of open citations, only five, namely Springer Nature, Wiley, the American Physical Society, Informa UK Limited, and Oxford University Press, are participating actively in the open publication of their own citations through Crossref.

It is noteworthy that JSTOR contributes very few references to Crossref, while the many citations directed towards its own holdings place JSTOR twelfth in the list of publishers receiving open citations (Figure 2). However, as the last column of Figure 3 shows, *all* the major publishers listed here are failing to submit reference lists to Crossref for a large number of the publications for which they submit metadata, that number being the difference between the value in the last column for that publisher and the combined values in the preceding three columns. JSTOR is the worst in this regard, submitting references with only 0.53% of its deposits to Crossref, while the American Physical Society is the best, submitting references with 96.54% of its publications recorded in Crossref.

Additional information about these analyses, including the code and the data we have used to compute all the figures, is available as a Jupyter notebook at https://github.com/sosgang/pushing-open-citations-issi2019/blob/master/script/croci_nb.ipynb.



Publisher submitting references to Crossref	Closed	Limited	Open	Overall publications deposited	Total with references
Elsevier BV	11,020,314 (65.70%)	0 (0.00%)	0 0.00%)	16,773,716	11,020,314 (65.70%)
Institute of Electrical and Electronics Engineers (IEEE)	3,331,913 (79.06%)	15,189 (0.36%)	0 0.00%)	4,214,422	3,347,102 (79.42%)
American Chemical Society (ACS)	496,855 (31.78%)	0 (0.00%)	0 0.00%)	1,563,601	496,855 (31.78%)
University of Chicago Press	41,566 (9.02%)	0 (0.00%)	0 0.00%)	461,070	41,566 (9.02%)
Ovid Technologies (Wolters Kluwer Health)	0 (0.00%)	820,456 (40.20%)	0 0.00%)	2,041,106	820,456 (40.20%)
IOP Publishing	0 (0.00%)	632,543 (76.25%)	0 0.00%)	829,525	632,543 (76.25%)
American Psychological Association (APA)	0 (0.00%)	19,535 (2.73%)	0 0.00%)	716,697	19,535 (2.73%)
Informa UK Limited	0 (0.00%)	15,632 (0.31%)	3,021,271 (60.85%)	4,965,446	3,036,903 (61.16%)
Springer Nature	0 (0.00%)	10,248 (0.08%)	5,854,527 (45.12%)	12,976,225	5,864,775 (45.20%)
Cambridge University Press (CUP)	0 (0.00%)	8,249 (0.40%)	555,170 (26.59%)	2,087,518	563,419 (26.99%)
SAGE Publications	0 (0.00%)	4,826 (0.19%)	1,196,568 (47.14%)	2,538,472	1,201,394 (47.33%)
Wiley	0 (0.00%)	0 (0.00%)	5,698,571 (64.22%)	8,874,184	5,698,571 (64.22%)
American Physical Society (APS)	0 (0.00%)	0 (0.00%)	621,989 (96.54%)	644,288	621,989 (96.54%)
Oxford University Press (OUP)	0 (0.00%)	0 (0.00%)	583,329 (15.73%)	3,707,847	583,329 (15.73%)
AIP Publishing	0 (0.00%)	0 (0.00%)	562,840 (73.02%)	770,812	562,840 (73.02%)
Royal Society of Chemistry (RSC)	0 (0.00%)	0 (0.00%)	331,526 (52.58%)	630,524	331,526 (52.58%)
Proceedings of the National Academy of Sciences	0 (0.00%)	0 (0.00%)	77,621 (55.37%)	140,176	77,621 (55.37%)
American Association for the Advancement of Science (AAAS)	0 (0.00%)	0 (0.00%)	27,002 (9.43%)	286,420	27,002 (9.43%)
JSTOR	0 (0.00%)	0 (0.00%)	11,097 (0.53%)	2,088,803	11,097 (0.53%)

Figure 3. The contributions to open citations made by the twenty publishers listed in Figure 2, as of 24 January 2018, according to the data available through the Crossref API. The counts listed in the first three results columns of this table refers to the number of publications for which each publisher has submitted metadata to Crossref that include the publication's reference list, the categories *closed*, *limited* and *open* referring to publications for which the reference lists are not visible to anyone outside the Crossref Cited-by membership, are visible only to them and to Crossref Metadata Plus members, or are visible to all, respectively. Additional information on this classification of Crossref reference lists is available at https://www.crossref.org/reference-distribution/. The fourth results column in the table shows the total number of publications for which the publisher has submitted metadata to Crossref, whether or not those metadata include the reference lists of those publications, and the fifth results column shows the total number of publications for which the publisher *has* submitted the reference list with the other metadata. The percentage values given in parentheses show the percentage of publications in each category whose metadata submitted to Crossref includes the reference lists, these percentages being obtained by dividing the values in each column by the total number of publications for which that publisher has submitted metadata to Crossref shown in the fourth results column.

It should be stressed that a very large number of potentially open citations are totally missing from the Crossref database, and consequently from COCI, for the simple reason that many publishers, particularly smaller ones with limited technical and financial resources, but also all the large ones shown in Figure 3 and most of the others, are simply not depositing with Crossref the reference lists for any or all of their publications.

Discussion

According to the data retrieved, the open DOI-to-DOI citations available in COCI exceed the number of closed DOI-to-DOI citations recorded in Crossref for every publication category, as shown in Figure 1. The journal category is the one receiving the most open citations overall, as expected considering the historical and present importance of journals in most areas of the scholarly ecosystem. However, the number of closed citations to journal articles within Crossref is also of great significance, since these 322 million closed citations represent 43% of the total.

It is important to note that about one third of these closed citations to journal articles (according to Figure 2) are references to entities published by Elsevier, and that references from within Elsevier's own publications constitute the largest proportion of these closed citations, since Elsevier is the largest publisher of journal articles. Thus, Elsevier's present refusal to open its article references is contributing significantly to the invisibility of Elsevier's own publications within the corpus of open citation data that is being increasingly used by the scholarly community for discovery, citation network visualization and bibliometric analysis.

It is also worth mentioning the discrepancy between the citations available in COCI, which comes from the data contained in the open and limited Crossref datasets as of 3 October 2018, and those available within those same Crossref datasets as of 24 January 2019. The most significant difference relates to IEEE. While the citations present in COCI include those from IEEE publications to other entities prior to November 2018 (since in October 2018 its article metadata with references were present within the Crossref *limited* dataset), in November 2019 this scholarly society decided to close the main part of its Crossref references, and thus from that moment they became unavailable to Crossref Metadata Plus members such as OpenCitations, as highlighted in Figure 3. Thus, IEEE citations from articles whose metadata was submitted to Crossref after the date of this switch to *closed* can no longer be automatically ingested into COCI.

To date, the majority of the citations present in Crossref that are not available in COCI comes from just three publishers: Elsevier, the American Chemical Society and University of Chicago Press (Figure 3). In fact, considering the average value of 18.6 DOI-to-DOI citation links for each citing entity – calculated by dividing the total number of citations in COCI by the number of citing entities in the same dataset – these three publishers are holding more than 214 million DOI-to-DOI citations that could potentially be opened. (The IEEE citation data which was in the Crossref 'limited' category as of October 2018 are actually included in COCI, although those from that organization's more recent publications will no longer be, as mentioned above).

We think it is deeply regrettable and almost incomprehensible that any professional organization, learned society or university press, whose primary mission is to serve the interests of the practitioners, scholars and readers it represents, should choose *not* open all its publications' reference lists as a public good, whatever secondary added-value services it chooses to build on top of the citations that those reference lists contain.

CROCI, the Crowdsourced Open Citations Index

The results of the Initiative for Open Citations (I4OC) have been remarkable, since its efforts have led to the liberation of millions of citations in a relatively short time. However, many more citations, the lifeblood of the scholarly communication, are still not available to the general public, as mentioned in the previous section. Some researchers and journal editors, in particular, have recently started to interact with publishers that are not participating in I4OC, in attempts to convince them to release their citation data. Remarkable examples of these activities are the petition promoted by Egon Willighagen (https://tinyurl.com/acs-petition) addressed to the American Chemical Society, and the several unsuccessful requests made to Elsevier by the Editorial Board of the Journal of Informetrics, which eventually resulted in the resignation of the entire Editorial Board on 10 January 2019 in response to Elsevier's refusal to address their issues (http://www.issi-society.org/media/1380/resignation_final.pdf).

To provide a pragmatic alternative that would permit the harvesting of currently closed citations, so that they could then be made available to the public, we at OpenCitations have created a new OpenCitations Index: **CROCI, the Crowdsourced Open Citations Index**, into which individuals identified by ORCiD identifiers may deposit citation information that they have a legal right to submit, and within which these submitted citation data will be published under a CC0 public domain waiver to emphasize and ensure their openness for every kind of reuse without limitation. Since citations are statements of fact about relationships between publications (resembling statements of fact about marriages between individual persons), they are not subject to copyright, although their specific textual arrangements within the reference lists of particular publications may be. Thus, the citations from which the reference list of an author's publication has been composed may legally be submitted to CROCI, although the formatted reference list cannot be. Similarly, citations extracted from within an individual's electronic reference management system and presented in the requested format may be legally submitted to CROCI, irrespective of the original sources of these citations.

To populate CROCI, we ask researchers, authors, editors and publishers to provide us with their citation data organised in a simple four-column CSV file ("citing_id", "citing_publication_date", "cited_id", "cited_publication_date"), where each row depicts a citation from the citing entity ("citing_id", giving the DOI of the cited entity) published on a certain date ("citing_publication_date", with the date value expressed in ISO format "yyyy-mm-dd"), to the cited entity ("cited_id", giving the DOI of the cited entity) published on a certain date ("cited_publication_date", again with the date value expressed in ISO format "yyyy-mm-dd"). The submitted dataset may contain an individual citation, groups of citations (for example those derived from the reference lists of one or more publications on a particular topic), or entire citation collections. Should any of the submitted citations be already present within CROCI, **these duplicates will be automatically detected and ignored**.

The date information given for each citation should be as complete as possible, and minimally should be the publication years of the citing and cited entities. However, if such date information is unavailable, we will try to retrieve it automatically using OpenCitations technologies already available. DOIs may be expressed in any of a variety of valid alternative formats, e.g. "https://doi.org/10.1038/502295a", "http://dx.doi.org/10.1038/502295a", "doi: 10.1038/502295a", "doi:10.1038/502295a", or simply "10.1038/502295a".

An example of such a CVS citations file can be found at https://github.com/opencitations/croci/blob/master/example.csv. As an alternative to submissions in CSV format, contributors can submit the same citation data using the Scholix format (Burton et al., 2017) – an example of such format can be found at https://github.com/opencitations/croci/blob/master/example.scholix.

Submission of such a citation dataset in CSV or Scholix format should be made as a file upload either to Figshare (https://figshare.com) or to Zenodo (https://zenodo.org). For provenance purposes, the ORCID personal identifier of the submitter of these citation data should be explicitly provided in the metadata or in the description of the Figshare/Zenodo object. Once such a citation data file upload has been made, the submitter should inform OpenCitations of this fact by adding an new issue to the GitHub issue tracker of the CROCI repository at https://github.com/opencitations/croci/issues.

OpenCitations will then process each submitted citation dataset and ingest the new citation information into CROCI. These CROCI citations will be made available at http://opencitations.net/index/croci using a REST API and a SPARQL endpoint, and will additionally be published periodically as data dumps in Figshare, all releases being under CC0 waivers. We propose in future to enable combined searches over all the OpenCitations indexes, including COCI and CROCI.

We are confident that the community will respond positively to this proposal of a simple method by which the number of open citations available to the academic community can be increased, in particular since the data files to be uploaded have a very simple structure and thus should be easy to prepare. In particular, we hope for submissions of citations from within the reference lists of authors' green OA versions of papers published by Elsevier, IEEE, ACS and UCP, and from publishers not already submitting publication metadata to Crossref, so as to address existing gaps in open citations availability. We look forward to your active engagement in this initiative to further increase the availability of open scholarly citations.

References

Burton, A., Fenner, M., Haak, W. & Manghi, P. (2017). Scholix Metadata Schema for Exchange of Scholarly Communication Links (Version v3). Zenodo. DOI: 10.5281/zenodo.1120265

Heibi, I., Peroni, S. & Shotton, D. (2019). Types, open citations, closed citations, publishers, and participation reports of Crossref entities. Version 1. Zenodo. DOI: 10.5281/zenodo.2558257

OpenCitations (2018). COCI CSV dataset of all the citation data. Version 3. Figshare. DOI: 10.6084/m9.figshare.6741422.v3

Peroni, S., Dutton, A., Gray, T. & Shotton, D. (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation, 71: 253-277. DOI: 10.1108/JD-12-2013-0166

Peroni, S. & Shotton, D. (2018). Open Citation: Definition. Version 1. Figshare. DOI: 10.6084/m9.figshare.6683855

Schiermeier, Q. (2017). Initiative aims to break science's citation paywall. Nature News. DOI: 10.1038/nature.2017.21800

Shotton, D. (2013). Open citations. Nature, 502: 295-297. DOI: 10.1038/502295a

Shotton, D. (2018). Funders should mandate open citations. Nature, 553: 129. DOI: 10.1038/d41586-018-00104-7

Sugimoto, C. R., Waltman, L., Larivière, V., van Eck, N. J., Boyack, K. W., Wouters, P. & de Rijcke, S. (2017). Open citations: A letter from the scientometric community to scholarly publishers. ISSI. http://www.issi-society.org/open-citations-letter/ (last visited 26 January 2018)

van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2): 523-538. DOI: 10.1007/s11192-009-0146-3

van Eck, N.J., Waltman, L., Larivière, V. & Sugimoto, C. R. (2018). Crossref as a new source of citation data: A comparison with Web of Science and Scopus. CWTS Blog. https://www.cwts.nl/blog?article=n-r2s234 (last visited 26 January 2018)

Merits and Limits: Applying open data to monitor open access publications in bibliometric databases

Aliakbar Akbaritabar[1] and Stephan Stahlschmidt[2]
akbaritabar@dzhw.eu
DZHW, German Centre for Higher Education Research and Science Studies, (Germany)
stahlschmidt@dzhw.eu
DZHW, German Centre for Higher Education Research and Science Studies, (Germany)

Abstract

Identifying and monitoring Open Access (OA) publications might seem a trivial task while practical efforts prove otherwise. Contradictory information arise often depending on metadata employed. We strive to assign OA status to publications in Web of Science (WOS) and Scopus while complementing it with different sources of OA information to resolve contradicting cases. We linked publications from WOS and Scopus via DOIs and ISSNs to Unpaywall, Crossref, DOAJ and ROAD. Only about 70% of articles and reviews from WOS and Scopus could be matched via a DOI to Unpaywall. Matching with Crossref brought 53 distinct licences, which define in many cases the legally binding access status of publications. But only 53% of publications hold only a single licence on Crossref, while more than 42% have no licence information submitted to Crossref. Contrasting OA information from Crossref licences with Unpaywall we found contradictory cases overall amounting to more than 13%, which might be partially explained by (ex-)including green OA. A further manual check found about 17% of OA publications that are *not accessible* and about 15% non-OA publications that are *accessible* through publishers' websites. These preliminary results suggest that identification of OA state of publications denotes a difficult and currently unfulfilled task.

Introduction

Open access (henceforth OA) in scholarly communication describes unrestricted access to published peer-reviewed documents written by and addressed to researchers. These documents have traditionally been disseminated via publications in scientific journals, which charge for access to the respective content. Stimulated by a call for greater openness and transparency in general ("open science"), the OA movement has nowadays been accepted as one, though not the only, alternative for the dissemination of scholarly documents. Even publishers seem to embrace this new model as providing a suitable infrastructure while at the same time securing their own economic interests.

This inter-mixture of interests has resulted not only in one, but several forms of OA publications such as Gold, Hidden Gold, Hybrid, Green, Delayed, Bronze and Black which are mainly based on right to access and pay to publish models depending on venues where the OA publication is accessible.

Due to the individual ascription of single publications to one or several of these categories and the decentralized structure of the scientific publishing market with a variety of diverse publishers, the identification of OA is less trivial than it might seem. Even large bibliometric data provider rely on external information to provide information on OA[1] and most large scale undertakings by the scientometric community to obtain reliable information on OA prevalence rely on the use of web crawlers (Archambault et al., 2013; Piwowar et al., 2018)

Inspired by the Hybrid OA Dashboard (Jahn, 2017) we applied licensing information detailing the legally binding access state supplied by publishers to the publisher association Crossref to identify OA publications. We determined the OA status of all publications retrieved from Web of Science (henceforth WOS) and Scopus in-house databases of 2017 by confronting them to two sources of OA information, i.e., Unpaywall and Crossref. In Section 2, we present our data and methods. In Section 3 we present our findings, while we discuss our main results in Section 4.

[1] https://clarivate.com/blog/easing-access-to-open-access-clarivate-analytics-partners-with-impactstory/

Data & Method

We queried all publications from Scopus and WOS in in-house databases of 2017. Data included article's unique ID from database and DOI. We matched those DOIs with Unpaywall database from April 18th 2018 to determine the OA status for each single publication. In parallel, we matched those DOIs with Crossref data (using snapshot of the data from April 2018 based on plus service described in Crossref (2018)) and retrieved the available information on the licences of publications[2].

Additionally, we used the journals' ISSNs provided by Wohlgemuth, Rimmert, & Winterhager (2016) (and the updated version in Rimmert, Bruns, Lenke, & Taubert (2017)) to identify Gold OA publications. They use different known OA indexes (e.g., DOAJ[3] (Directory of Open Access Journals) and ROAD[4] (Directory of Open Access scholarly Resources) and determine if the respective ISSN is listed in those databases. They differentiate between ISSN and ISSNL which is more fine-grained by adding a specific ISSN to some special issues. We tried both ISSN and ISSNL, since the latter had higher matching records, therefore in our analysis presented in the Results section we use the ISSNL.

It is necessary to note that some publications had multiple licence URLs in Crossref database, we followed a procedure with four steps to ensure using only one licence per publication (see Table 2 for the frequencies of these publications):

1. If a publication had only one record in Crossref database, whether it had an *OA*, *non-OA*, *unclear* licence or *no licence information (i.e. NA)*, we used this status and categorized the publication as a unique one.
2. If a publication had multiple *OA* licence URLs, we removed the duplicates and categorized it as *OA*.
3. If a publication had a mixture of *OA* and *non-OA* licence URLs, we removed the duplicates and categorized it as *OA*.
4. If a publication had multiple *non-OA* licence URLs, we removed the duplicates and categorized it as *non-OA*.

A research assistant controlled the unique licences (a total of 56) we extracted from Crossref with available information online to categorize them as *OA* and *non-OA*. We used this categorization in parallel to established OA identification procedures (e.g., searching for journal's ISSN in DOAJ and ROAD in Gold OA identification) to ensure a higher level of robustness in our results.

In OA Identification process and in order to determine if a publication was OA or not, we applied a multi-category view separating Gold, Hidden Gold, Hybrid and Delayed OA, while doing so, we reached a new category of Probable Hybrid OA. Our investigation strategy for each category was as follows:

- **Gold OA**: As described earlier, we used the ISSNs provided by Rimmert et al. (2017) to determine Gold OA. We matched the respective ISSN (from both WOS and Scopus) with DOAJ and ROAD. If the respective ISSN was listed in one of those directories, the publication is categorized as *Gold OA*. We confronted Gold OA from DOAJ and ROAD with our research assistant's categorization of Crossref licences after the manual check of unique licence URLs.
- **Hidden Gold OA**: we used metadata from WOS and Scopus to determine the journal issue and looking at the licences of all publications in a single issue, if all publications had *OA*

[2] It is neccessary to note that our effort to send large number of requests to Crossref API (even while using plus service and through both rcrossref package in R and more fine-grained httr requests directly to Crossref API) faced timeout and response time errors and alternatively we chose to use the in-house snapshot of the Crossref data to circumvent the above error. This meant parsing large corpus of JSON files which can be time consuming depending on the goals of the analysis. Any effort on automating the proposed OA identification procedure needs to overcome the technical issues like this.

[3] https://doaj.org/

[4] https://road.issn.org/

licences, but the ISSN was not indexed in DOAJ or ROAD we categorized it as *Hidden Gold OA*.

- **Hybrid OA**: If an issue had at least one *non-OA* publication while having one or more *OA* publications, we categorized the OA publications as *Hybrid OA*.
- **Probable Hybrid OA**: If an issue did not have a *non-OA* publication while having one or more *OA* publications and some publications in the issue didn't have licence information, we categorized them as *Probable Hybrid OA*.
- **Delayed OA**: In all of the above cases, we looked into delays based on Crossref metadata (a difference in terms of days from day of publication and the date licence was assigned to the publication as described in CrossRef-API (2019), this is the time period known as *embargo time*) to determine if they were *Delayed*, therefore each of the above categories were split to two groups, *delayed* and *not-delayed*. If a publication had multiple licence URLs on Crossref, we controlled their respective delay times, if any of those were *not-delayed* we categorized the publication as such, while if any of the licences were *delayed*, the publication is identified as a *delayed* one.
- **Closed Access**: Strictly speaking, if the number of publications in an issue was equal to the number of *non-OA* publications and the ISSN was not indexed in DOAJ or ROAD, we categorized them as *Closed Access*.
- **NA (Not available)**: A publication that was not fitting in any of the above categories or did not have a licence URL to determine its condition was categorised as *NA*. Number of NAs are higher than *Closed Access* publications, since we aimed to keep the definitions as strict as possible.

Results

We present the results in two main sections, one regarding *Unpaywall* and the other on licences extracted from *Crossref*. We then present the *comparison between Unpaywall and Crossref* and the results of our *manual checks on random samples* for robustness of the results.

Table 1 shows the number of articles and review papers from WOS and Scopus with an equivalent record in Unpaywall database. It presents also the total number of articles and review papers in WOS/Scopus to provide a baseline for comparison. Unpaywall has close to 70% coverage in both cases while coverage of WOS is slightly higher (can be due to different indexing philosophy or DOIs completeness). In the following tables (in Unpaywall results), publications are limited to only articles and review papers published in 2000-2017.

Table 1: All articles and reviews from WOS (2000-2016) and Scopus (2000-2016) that could be matched to Unpaywall database via DOIs

Data Source	Frequency	Percent
WOS (Unpaywall match)	13,875,946	69.75%
WOS (total)	19,894,531	•
Scopus (Unpaywall match)	17,820,375	68.67%
Scopus (total)	25,951,839	•

Figure 1 presents the distribution of journals and publications indexed in WOS (top) and Scopus (bottom) matched with Unpaywall database and crosschecked the ISSNs with DOAJ. *Missing on DOAJ* in these Figures refer to those journals whose ISSN was missing from Rimmert et al. (2017) data, therefore we could not check if the ISSN is listed in DOAJ or not while *Others* means the ISSN was existing in Rimmert et al. (2017) but it was not listed as *OA* in DOAJ. Share of publications which don't have a matching ISSN in DOAJ (meaning they are not Gold OA) and are identified as OA in Unpaywall is interesting on both Figures (designated with “Missing on DOAJ | Unpaywall OA” as label). They could be other OA types (green, hybrid, hidden gold).

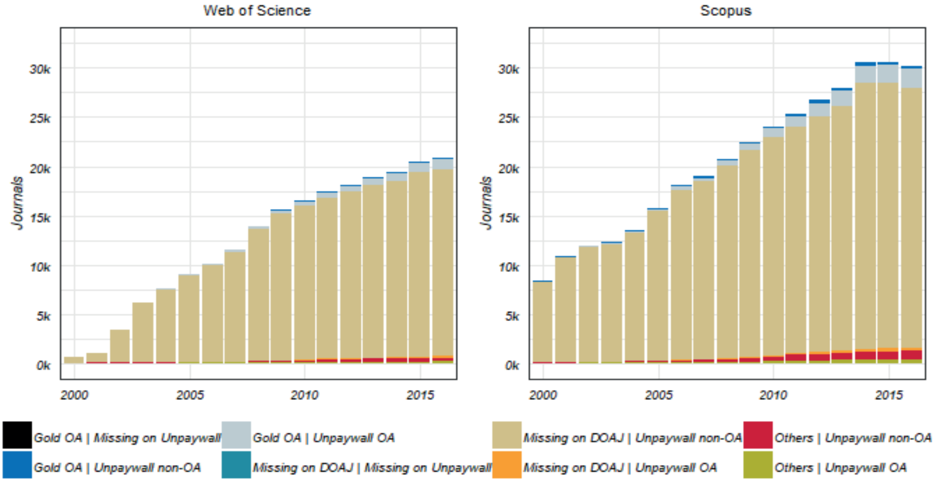


Figure 1: Journals indexed in WOS and Scopus matched with Unpaywall database and crosschecked the ISSNs with DOAJ (Gold OA) between 2000 and 2016

We matched publications to Crossref data from April 2018 and found 56 distinct licence types for all of the publications. Table 2 presents a descriptive view on whether publications have licence information recorded in Crossref. It shows that about 43% of publications from WOS or Scopus with a matching DOI indexed in Crossref do not have a licence URL. Some of the publications had more than one licence information in Crossref (as an example, the number of DOIs that each have 6 licence records on Crossref are 1,152). In case of multiple licences, if a publication had at least one OA licence, we categorized it as OA.

Table 2: Number of licences per DOI found in Crossref for articles and reviews indexed in either WOS, Scopus or both between 2000 and 2016

Number of licences per DOI	Frequency of DOIs	Percent
0	6,571,079	42.74
1	8,143,752	52.97
2	655,729	4.26
3	3,472	0.02
4	17	0.00
6	1,152	0.01

Figure 2 present the Gold, Hidden Gold, Hybrid and Delayed OA status of the publications from WOS (top) and Scopus (bottom), which is presented as trends over the years. We limited the years to 2000-2017 to show the most recent trends. To make these Figures more readable, we removed NA (those without a matching DOI or without a licence information on Crossref).

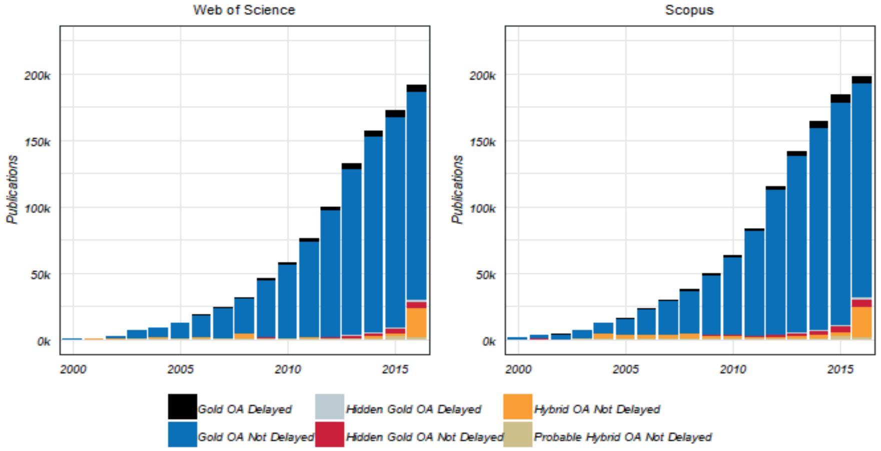


Figure 2: Count of gold and hybrid OA publications between 2000 and 2016 based on Crossref licence information, DOAJ and ROAD)

Tables 3 and 4 present the OA status comparison between Unpaywall and Crossref in WOS and Scopus publications, respectively. Note, Crossref OA status in the Tables is the categorization we developed using respective licence URLs. We double checked the contradictory cases and improved our while-list of OA licences, while some of the contradictions still remain (e.g., Unpaywall declares those publications as OA while they are closed access or vice versa, in case of licences on Crossref that are open access while the publication is declared as non-OA on Unpaywall). Overall contradictory cases amount to 22.98 % in WOS and 22.91% in Scopus which might partly be explained by the wider scope of Unpaywall including also green OA publications that might not be identified via license information only.

Table 3: OA status comparison between Unpaywall and Crossref on WOS publications

Crossref OA Status	Unpaywall OA Status	Frequency	Percent
Closed Access	Closed Access	4,767,019	35.26
NA	Closed Access	4,395,218	32.51
NA	Open Access	2,168,747	16.04
Closed Access	Open Access	1,649,674	12.20
Open Access	Open Access	438,100	3.24
Open Access	Closed Access	99,062	0.73
NA	NA	20	0.00
Closed Access	NA	10	0.00

Table 4: OA status comparison between Unpaywall and Crossref on Scopus publications

Crossref OA Status	Unpaywall OA Status	Frequency	Percent
Closed Access	Closed Access	5,890,312	40.75
NA	Closed Access	4,055,736	28.06
NA	Open Access	1,991,393	13.78
Closed Access	Open Access	1,879,773	13.01
Open Access	Open Access	506,106	3.50
Open Access	Closed Access	130,398	0.90
Open Access	NA	4	0.00
Closed Access	NA	1	0.00

Tables 5 and 6 present the result of our research assistant’s manual check for accessibility to article’s PDF file from publishers websites compared to the respective licence in Crossref and the OA status we manually assigned to those URLs in contrast to OA status from Unpaywall. It is interesting to see there are publications defined as Non-OA while their PDF is accessible from the publisher (14.42% in WOS and 14.54% in Scopus) or vice versa, OA publications (based on either Unpaywall, Crossref or both) that are not accessible online (17.57% in WOS and 16.74% in Scopus). Note also the contradictory cases between Crossref and Unpaywall, where metadata from one shows OA and the other Closed, which requires further probes (22.98% in WOS and 22.91% in Scopus, these percentages are quite close to contradictions observed in the overall sample presented in Tables 3 and 4). Our effort to complement these databases proves that none of them could be used in isolation. We aim to follow-up and use PDF URLs provided by Unpaywall in large scale to control the ratio of publications which can be accessed.

Table 5: Random sample OA status check on publications from WOS

PDF Manually accessible?	Licence status	Pub OA?	Frequency	Percent
PDF Accessible	Open Access	Unpaywall OA	104	46.85
No Access to PDF	Closed Access	Unpaywall non-OA	45	20.27
No Access to PDF	Open Access	Unpaywall non-OA	18	8.11
No Access to PDF	Closed Access	Unpaywall OA	16	7.21
PDF Accessible	Closed Access	Unpaywall OA	16	7.21
PDF Accessible	Closed Access	Unpaywall non-OA	14	6.31
No Access to PDF	Open Access	Unpaywall OA	5	2.25
NA	Closed Access	Unpaywall non-OA	1	0.45
PDF Accessible	NA	Unpaywall non-OA	1	0.45
PDF Accessible	Open Access	Unpaywall non-OA	1	0.45
PDF Accessible	NA	Unpaywall OA	1	0.45

Table 6: Random sample OA status check on publications from Scopus

PDF Manually accessible?	Licence status	Pub OA?	Frequency	Percent
PDF Accessible	Open Access	Unpaywall OA	104	45.81
No Access to PDF	Closed Access	Unpaywall non-OA	48	21.15
PDF Accessible	Closed Access	Unpaywall OA	17	7.49
No Access to PDF	Open Access	Unpaywall non-OA	17	7.49
No Access to PDF	Closed Access	Unpaywall OA	16	7.05
PDF Accessible	Closed Access	Unpaywall non-OA	14	6.17
No Access to PDF	Open Access	Unpaywall OA	4	1.76
PDF Accessible	NA	Unpaywall OA	2	0.88
No Access to PDF	Closed Access	Missing on Unpaywall	1	0.44
PDF Accessible	Open Access	Missing on Unpaywall	1	0.44
NA	Closed Access	Unpaywall non-OA	1	0.44
PDF Accessible	NA	Unpaywall non-OA	1	0.44
PDF Accessible	Open Access	Unpaywall non-OA	1	0.44

Conclusions

It is clear that publishing as OA is on the rise in recent years. This trend is observed similarly in WOS and Scopus (while Scopus has higher raw publication counts but trends are identical) and based on OA identification stemming from both Unpaywall and Crossref. But still the majority of publications are closed access. We observed that despite the high coverage of Unpaywall (close to 70% of *articles* and *reviews* in both WOS and Scopus), it doesn't provide enough metadata (as of April 2018) for OA categorization thus could be limiting for large scale OA monitoring in the leading bibliometric databases. Licence information from Crossref is more detailed and it gives a good possibility to complement Unpaywall metadata. Although we overcame the downsides by complementing these databases, we still found further contradictions between them with manual random checks. Some publications were OA (based on their licences or Unpaywall status) while their PDF files were *not accessible* through publishers' websites. Some publications were closed access, while their PDF files were *accessible*. We found that the issue of multiple records for some publications or multiple licence information is something that needs to be seriously considered in OA monitoring. While we tried to test different scenarios in OA identification, still there are publications that won't fit into any of the scenarios and we had to categorize them as *NA* (since we wanted to keep the *Closed Access* definition as strict as possible), these are the publications that need to be further studied and usually the metadata of the OA databases are lacking for them. We propose OA monitoring activities to try to benefit from our approach in complementing the metadata from OA databases, i.e. Unpaywall and Crossref, while noting that there are contradictions between these sources. Our effort to complement these databases proves that none of them could be used in isolation.

References

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Rebout, L., & Roberge, G. (2013). Peer-reviewed papers at the European and world levels—2004-2011. *Info@ Science*, *1*, 495–6505.

Crossref. (2018, October). Crossref. Retrieved from https://www.crossref.org/

CrossRef-API. (2019, October). CrossRef API. Retrieved from https://github.com/CrossRef/rest-api-doc#filter-names

Jahn. (2017, January). About the hybrid OA dashboard. Retrieved from https://subugoe.github.io/hybrid_oa_dashboard/about.html

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., , Farley, A., West, J. and Haustein, S., (2018). The state of OA: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, *6*, e4375.

Rimmert, C., Bruns, A., Lenke, C., & Taubert, N. C. (2017). ISSN-matching of gold OA journals (ISSN-gold-OA) 2.0.

Wohlgemuth, M., Rimmert, C., & Winterhager, M. (2016). ISSN-matching of gold OA journals (ISSN-gold-OA).

Google Search results as an altmetrics data source?

Kim Holmberg[1] and Timothy D. Bowman[2]

[1] *kijohol@utu.fi*
Research Unit for the Sociology of Education, 20014 University of Turku (Finland)

[2] *timothy.d.bowman@wayne.edu*
School of Information Sciences, Wayne State University (USA)

Abstract

This research-in-progress investigates currently available means to collect data relevant for altmetrics research from search engines. Content analysis is used to study accuracy and relevance of the collected data and to assess the applicability of Google Search results for altmetrics research. The results show that although search engines contain a vast amount of data that could be useful for altmetrics research at a small scale, the data does not come without problems. A high number of false positives in the retrieved data showed that the data cannot be used as such and further research is required into improving the data collection strategy.

Introduction

Altmetrics research has investigated alternative online data sources for traces of scholarly communication and the use of scholarly outputs (e.g., documents, presentations, data and videos), with the intention of studying what these events could indicate about the attention research has received or what kind of impact research has had within, and outside of, these contexts. While altmetrics has become a popular research area within the domain of information science, the idea of altmetrics is not a new one. Webometrics, which preceded altmetrics, primarily used hyperlinks to mine information from the web in order to understand social and structural connections between websites and the organizations and people they represent. The original idea behind webometrics was that hyperlinks could provide valuable information about the connections between webpages or the websites they connect to, in a similar way that citations can provide information about the use of earlier work, which is cited by later work. The success of Google has demonstrated the value of hyperlinks and what they indicate, as counting hyperlinks has been at the core of Google's ranking algorithm (Brin & Page, 1998) from the beginning of the search engine's launch. The operators some search engines used to offer made it possible to, for instance, retrieve all hyperlinks that targeted a specific website or that originated from a specific website, or a combination of the two.

For altmetrics this would mean that it would have been possible to find all the webpages that contained a hyperlink to a specific scientific document or to all documents from a specific journal, thus providing insights into the use and popularity of specific scientific outputs. To the best of the knowledge of the authors, none of the major search engines offer currently any means to track links to a website. There are, however, commercial services that provide link data (e.g. https://moz.com/link-explorer that allows 10 free queries per month), mainly intended for Search Engine Optimization and benchmarking of competitors' web presence.

This research focuses on currently available free options for link data collection. This research will set out to investigate if, and how, hyperlinks relevant for altmetrics research could be identified and collected using the current means available. This research will test various methods to retrieve link data about scholarly outputs using Google to find DOIs (Digital Object Identifiers) assigned to a specific sample of scientific articles. In order to verify the relevance and accuracy of the methods, a content analysis will be conducted to find out what kind of results Google returns and in what context the DOI has been mentioned. The authors will also present the unsuccessful approaches that were tested and conclude with recommendations on how Google Search results could be used an altmetrics data source.

Literature review

Citations acknowledge the use of earlier work and thus they are generally thought to be recognition of valuable work. In a similar way the original idea of Google was that hyperlinks were created to link to some valuable work; whoever creates the hyperlink has judged the target webpage to have some value. The similarities between citations and hyperlinks led researchers to investigate hyperlinks under the term webometrics. It can be said that webometrics research started with the seminal paper on Web Impact Factors (WIF) by Ingwersen (1998). Much like Journal Impact Factors, the original idea with WIFs was that the number of hyperlinks per webpages in a website would give a comparable factor of the quality or impact of the organization behind the website. While different modifications of the original WIF were introduced (e.g. Thelwall, 2001; 2003), impact factors based on hyperlinks were found to be mostly unreliable due to (in contradiction to citations) non-standardized practices of creating links and the multitude of possible motivations for creating them. A great deal of early webometrics research focused on academic linking. This line of investigations showed that linking between universities was strongly connected to productivity (Tang & Thelwall, 2004) and that while higher ranked scholars attracted more links from their peers (Thelwall & Harries, 2003), this too was mainly due to higher productivity rather than quality. Data collection for webometrics research was typically conducted using one of two methods; for small scale research projects a web crawler could be set to crawl all the pages and collect link data, but for larger projects link data retrieved from open search engine APIs were used. Unfortunately, all the major search engines gradually ceased providing the retrieval of link data, which led webometrics researchers to investigate a range of alternative data sources. Vaughan and Shaw (2003) investigated the possibilities to search for title mentions, or so-called web citations, to scientific articles. This approach was further developed by Sud and Thelwall (2014) so that the resulting webpage was queried for links to the mentioned scientific paper, not just the mention of a title. A link to the correct webpage would then be further evidence of a relevant mention or citation. Kousha and Thelwall (2006) introduced URL citations, in which mentions of URLs are searched for in the text of a webpage, which could be present in either text or hyperlink form. The three approaches above, web citations, linked title mentions, and URL citations all overlap, as links can be identified with each approach but are not a requirement. These approaches have disadvantages including that they may generate false matches if the titles do not uniquely identify the article and that the volume of matches may report as low if the webpages don't mention the URL to the article when discussing it. This project will continue this line of investigations and study whether hyperlink data could be retrieved efficiently using the currently available data collection methods and whether the collected link data could be used to enhance altmetrics research.

Data and methods

A thorough studying of the current search operators suggested that Google's *allinanchor:* operator may return relevant results for altmetrics purposes. According to its documentation (goo.gl/MqS1jJ) the operator should return webpages that contain all the queried terms in a backlink or an anchor linking to another webpage. For instance, a query using -- allinanchor:"cats and dogs" -- would return all pages that have all the queried terms in an anchor on the page. Our hypothesis was that this approach could work as a work-around for the *link:* operator, which no longer exists. The authors utilized the Top 10 papers from the 2017 Altmetric.com listing as a starting point to test available methods for data collection.

The *allinanchor:* operator was tested using both titles of scientific articles and their DOIs. When using the operator to search for titles, many of the webpages listed in the search results contained the queried terms or the title phrase in the title of the page or in text form somewhere in the content, instead of finding them in anchor text in the content of the page.

The operator worked even worse when using DOIs as search terms, returning some false results for the tested queries. This may be due to the special characters commonly found in DOIs. These findings suggest that in certain cases the use of the *allinanchor:* command includes results that do not meet the parameters of the command. Based on these initial results, the authors argue that there seems to be minimal added value when using the *allinanchor*: operator over the approach using title mentions (Vaughan & Shaw, 2003).

Next, the authors examined searching for so-called URL citations, or for the URLs in text format (Kousha & Thelwall, 2006). This was found to not be a viable option as the results contained partial matches in many cases; for instance, a search for: **https://doi.org/10.1016/s0140-6736(17)**32252-3 would return webpages containing results such as **https://doi.org/10.1016/S0140-6736(17)**31352-1 and **https://doi.org/10.1016/S0140-6736(17)**33001-5, which matched only the first part of the query (in bold). Leaving out the domain and using only the unique DOIs returned similar results in subsequent tests. A search for: 10.1016/j.respol.2017.02.008 returned both webpages that contained the queried DOI and webpages (or other online content) with only partial matches. This approach does, however, work better for queries that do not include any special characters, such as '(' or ')'. For instance, a query for: 10.1038/nature23305 appears to return pages matching the query.

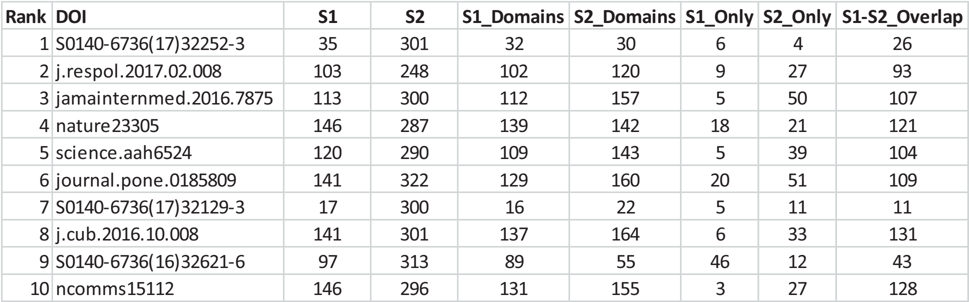
The third method tested was to search for partial DOIs (leaving out the journal identifier component of the DOI). In other words, for the DOI mentioned above the first part would be left out and the search would be done for *nature23305* only. While this strategy seemed to return a higher number of results, the approach was not without issues. First of all, the estimated numbers of hits that Google provides for all the queries made were found to change even between consecutive repeated queries. Google presents an estimate of the number of hits for any given query and it is well known and admitted by Google that this number is not accurate (https://support.google.com/gsa/answer/2672285?hl=en). For instance, Google might state that a particular query has 50,000 results, but as one goes through the results page by page the estimate reported by Google may change. These changing estimates may result from the accuracy of the estimate increasing as Google retrieves more and more results. Furthermore, after navigating through a number of result pages Google states that "in order to show you the most relevant results, we have omitted some entries very similar to the [number] already displayed" and the number of estimated results decreases significantly. Choosing to include the omitted results again changes the number of estimated results and the number of presented results. Moreover, the listing of results stops after navigating through a variable number of pages. Based on these tests, it appears that Google does not provide more than approximately 300 hits, at least for the type of queries tested here. It is possible that the algorithm determines that the relevance of the results is significantly lower after the first 300 results, thus providing more results would not be cost-effective. These variations in result estimates, omitted results, and the ceasing of displaying results after a number of pages have been viewed suggests that Google estimates should not be used as an indicator of any type.

After testing the various data collection strategies, the authors used the partial, article identifying DOIs and performed a Google search, separately collecting the results that both included and excluded the results that Google had first omitted. The two datasets were compared and the first set of results was ultimately used as a data set for content analysis. This decision assumed that most Google users would utilize this set of results, ignoring the results that Google omits from their searches. A content analysis approach was used to determine whether the results contained any false positives, i.e. results that did not contain a reference to the searched DOI. The authors also determined what type of webpage or website the DOI was found within and in what context the DOI was mentioned. The classification scheme developed for this study utilized a grounded theory approach (Glaser and Strauss, 1967) to determine the type of page and context in which the DOI was discovered. The initial classification of the type of page resulted in 119 categories. Categories that were judged to overlap significantly were merged, resulting in a total of 18-page categories. Cohen's κ was run to determine if there was agreement between two authors on the type of pages returned in the search results for 10 percent (n=31) of the coded results. There was good agreement between the two authors' categorizations, κ = .646 (95% CI, .300 to .886), p < .0005. The classification of the context resulted in an initial set of 14 categories; after merging similar and/or overlapping categories the final number of context categories was 4. Cohen's κ was run to determine if there was agreement between two authors on the context of the DOI found in the page for 10 percent (n=31) of the coded results. There was moderate agreement between the two authors' categorizations, κ = .595 (95% CI, .300 to .886), p < .0005.

Results

The results from the data collection identified several issues that require further analysis (Table 1). It would appear that if the search term includes special characters such as parenthesis, the results may be lower. It would also appear that when including the results first omitted by Google the results are constantly capped at approximately 300 results. It is also clear that the results first omitted by Google do not only contain similar pages from the same domain; in all but two of the cases the results including the omitted results included more unique domains.

Table 1. Results from the data collection (S1 = Search results, excluding the results omitted by Google, S2 = Search results, including the results omitted by Google, S1_Domains = Unique domains in S1, S2_Domains = Unique domains in S2, S1_Only = Domains only present in S1, S2_Only = Domains only present in S2, S1-S2_Overlap = Domains found in both S1 and S2)



Rank	DOI	S1	S2	S1_Domains	S2_Domains	S1_Only	S2_Only	S1-S2_Overlap
1	S0140-6736(17)32252-3	35	301	32	30	6	4	26
2	j.respol.2017.02.008	103	248	102	120	9	27	93
3	jamainternmed.2016.7875	113	300	112	157	5	50	107
4	nature23305	146	287	139	142	18	21	121
5	science.aah6524	120	290	109	143	5	39	104
6	journal.pone.0185809	141	322	129	160	20	51	109
7	S0140-6736(17)32129-3	17	300	16	22	5	11	11
8	j.cub.2016.10.008	141	301	137	164	6	33	131
9	S0140-6736(16)32621-6	97	313	89	55	46	12	43
10	ncomms15112	146	296	131	155	3	27	128

A random sample of 300 results from the first set of search results (S1) was taken for content analysis. The content analysis included analysis of 1) the types of pages that contained a mention of the selected scientific papers, and 2) the context in which the paper was mentioned or the intention of mentioning it.

Type of page. As seen in Table 2, most of the resulting pages were scientific articles or pages of scientific journals (16.0%), followed by pages of university repositories for scientific papers or other databases or types of listings of publications (15.0%), and finally mainstream news sites (12.3%). As evidence of the dynamic nature of the web, 8% of the resulting pages could no longer be accessed, as they have most likely been taken down sometime between Google's last crawl of the page and the date when the data was analyzed. Companies and organizations of different kinds, research centers and scientific societies, and blogs constituted for approximately 6% of the mentions each. Websites advocating different causes and social media postings captured by Google constituted approximately 5% of the mentions.

Table 2. Type of page

Type of page	**n**	**%**
Scientific text or journal	48	16.0
Repositories, databases and listings of scientific texts	45	15.0
News site	37	12.3
Page not found or access denied	24	8.0
Companies, hospitals, governmental agencies	19	6.3
Blogs (personal, company, news, course, etc)	18	6.0
Research centres and scientific societies	18	6.0
Advocacy and initiative site	16	5.3
Social media postings	16	5.3
Press releases and newsletters	12	4.0
Online CVs and professional profiles	9	3.0
Powerpoint	9	3.0
Online magazines	7	2.3
File sharing	6	2.0
Google Books	6	2.0
Online video	4	1.3
Course sites	3	1.0
Reports and meeting protocols	3	1.0
Total	**300**	**100.0**

Intention of linking. In the majority of the investigated websites and webpages the intention of mentioning or linking to the scientific articles (Table 3) was to reference them (67.7 %, n = 203). These references were, for instance, in other scientific articles, news stories, blog entries, other social media, press releases, and magazines. In the category labelled Listing, the investigated article was found in various listings of a number of scientific articles (6.7 %, n = 20). In about 8% (n=25) of the cases the DOI was found on a webpage about the investigated article, such as the publisher's page for that specific article. In 17.3% (n=54) of the cases no mention of the investigated article could be found on the page. This category includes the 24 cases where the page was no longer available, but also includes 30 cases where the page was visited and no mention or reference could be found. It is possible that the page had been updated and the reference had been removed, but it may also be possible that the page had been indexed and resulted in a false positive to the set query, as was discovered to be the case in 10 of the results. Two of the DOIs in the sample were found to result in false positives in the search results: j.respol.2017.02.008 and j.cub.2016.10.008. A careful examination of the search results revealed that partial hits of the search terms could be found among the results. In these cases all the parts of the DOI could be found on the page, but not in a single string, for instance, "j.respol" could be found somewhere on the page, while "2017.02.008" could be found in connection with another journal identifier somewhere else on the page.

Table 3. Intention of the mention/link

Row Labels	**n**	**Percent**
Reference	203	67.7
Not found	52	17.3
Investigated paper	25	8.3
Listing	20	6.7
Grand Total	**300**	**100.0**

Discussion

This research set out to investigate Google Search results as a potential data source of altmetrics. The investigation included identifying and testing potential data collection methods and an assessment of the accuracy of the results of these tests. The results indicate that Google Search can be used to discover websites and webpages that mention specific scientific articles, but due to the lack of open APIs the approach can only be recommended for small scale projects where the obtained results can be manually investigated. In addition, the results suggest that collecting altmetrics data using Google Search is not without problems.

Google omits some search results, possibly judging them to contain duplicate content to a page already shown in the results or to be of low quality or irrelevant to the search, resulting in algorithmic penalty. Based on the findings from this study, the omitted results contained domains that were not already among the results listed, thus ruling out the cause being duplicate content in some instances. In addition, for an unknown reason Google appears to cut off the results at approximately 300 listed results. Further research is required to understand the reasons and impact these algorithmic decisions have on the applicability of Google Search for altmetrics. Furthermore, the number of false positives in the data is a concern and some automated approach should be developed to minimize the number of these occurrences.

Future research should take the limitations of this research into account, specifically that this research only analyzed the DOI search results from the top 10 papers from the 2017 Altmetric Top 100 list; different results may be obtained from an expanded data set. However, the high number of false positives for the relatively small dataset used in this research can already be seen as critical evidence against the usefulness of search results for altmetrics. Google was the only search engine used and other search engines may provide different results. While the results of this investigation showed that search data can contain relevant data for altmetrics research, future research should further determine the applicability of search results for altmetrics research by investigating possible automated approaches for large scale data collection, filtering, and analysis, possibly using the data collected by Common Crawl (http://commoncrawl.org/). Future work could also include a content analysis of the omitted search results, which may reveal new information about Google's ranking algorithm. As many users most likely use the results that Google provides, the ranking algorithm has great power in dictating what people find and what is omitted from them. It would be highly important to better understand how Google decides what to, and what not to, display as results.

References

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN systems*, 3(1-7) (http://infolab.stanford.edu/~backrub/google.html).

Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.

Kousha, K., & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. *Scientometrics*, 68(3), 501-517.

Sud, P. & Thelwall, M. (2014). Linked title mentions: A new automated link search candidate. *Scientometrics*, 101(3), 1831-1849.

Tang, R. & Thelwall, M. (2004). Patterns of national and international Web inlinks to US academic departments: An analysis of disciplinary variations. *Scientometrics*, 60(3), 475-485.

Thelwall, M. (2001). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52(13), 1157-1168.

Thelwall, M. (2003). Web use and peer interconnectivity metrics for academic Web sites. *Journal of Information Science*, 29(1), 1-10.

Thelwall, M. & Harries, G. (2003). The connection between the research of a university and counts of links to its Web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science and technology*, 54(7), 594-602.

Vaughan, L. & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? *Journal of the American Society for Information Science and Technology*, 54(14), 1313-1324.

How Public Investment Fuels Innovation: Clues from Government-subsidized Patents 1980-2017

Lin Zhang[1], Yujie Peng[2], Wenjing Zhao[3], Lixin Chen[4] and Ying Huang[5,*]

[1] *zhanglin_1117@126.com*
1 School of Information Management, Wuhan University, Wuhan 430072 (China)
Department of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046 (China)
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, B-3000 Leuven (Belgium)

[2] *pyj8176@126.com*
2 Department of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046 (China)

[3] *cady_zhao0827@163.com*
3 School of Information Management, Wuhan University, Wuhan 430072 (China)

[4] *lynnchenlixin@163.com*
4 School of Information Technology, Shangqiu Normal University, Shangqiu 476000 (China)

[5*] *huangying_work@126.com*
5 School of Information Management, Wuhan University, Wuhan 430072 (China)
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, B-3000 Leuven (Belgium)

Abstract

How does government activity connect to innovation? This question remains unanswered in the literature, leaving a huge gap for policymakers and academics. In the US, a government-interest statement in a patent indicates that the US federal government holds a financial interest or a right over the innovation being patented. This offers a window to examine the possible relationships between federal funding and technological advancement. Hence, this paper addresses the following research questions: 1) How does federal agency support for utility inventions accelerate innovation? 2) Which sectors are most helpful in communicating the relationships between funding and invention? 3) Is there variation in the technological fields federal agencies fund? This analysis of US patents granted between 1980 and 2017 seeks to shed light on the pattern of federal agency support for innovation. The results should be instructive and useful for strategic planning by managers and decision-makers.

Introduction

In an era of economic globalization where knowledge is the main currency, innovation has been an inexhaustible driving force of national sustainable development and progress. It is, therefore, no wonder that policymakers and academics have shown great interest in ways to improve the innovation capacity of a nation. Among all the innovation paradigms, technological innovation has become a leading factor in the competition among nations, regions, and enterprises. Technology is becoming increasingly important in the pursuit of competitive advantage that determines the final outcome.

Patenting rates and patent propensity have long been used as a well-grounded proxy for measuring technological innovations in the literature (Burhan et. al, 2017). For example, Furman et al. (2002) used patent data to evaluate sources of difference among countries in the production of visible innovative output. Mathews et al. (2005) defined patents as the ability of a country to produce and commercialize a flow of new-to-the-world technologies. There's been a growing consensus that government plays an irreplaceable role in the rate and direction of inventive activity. And the most productive and diverse technological trajectories are likely to build on government-funded inventions (Corredoira et. al, 2018). However, revealing and accurately quantifying how the process of public investment promotes and accelerates innovation remains a difficult undertaking. Proper reporting about the metrics of this paradigm assist the government in determining whether it should exercise its retained rights.

In 1980, government-funded innovation reached a turning point with the US Senate passing the University and Small Business Patent Procedures Act, now known as the Bayh-Dole Act. The key change made by Bayh-Dole was in the ownership of inventions subsidized by federal funding. Before the Bayh-Dole Act, federally-funded research contracts and grants obligated inventors, wherever they worked, to assign the rights for those sponsored inventions over to the federal government (Stevens, 2004). Whereas, with Bayh-Dole, a university, small business, or non-profit institution has the option to pursue ownership of the invention (Loise & Stevens, 2010). If the inventor does claim title, the government still retains a royalty-free license to use the patented technology. Bayh-Dole imposes several requirements, one of which creates a record labelled "Government Interest". The label means that any indication of interest or rights the US government may have on a particular patent is contained in the text of the patent document. This "government interest" statement alerts third parties, who may be negatively affected by improper use of a patent, of their ability to petition the funding agency to exercise these retained rights. Although the number of patents that indicate government interest is small, these patents reveal specific areas of scientific and technological development where the United States government has invested millions of dollars.

Most previous works have focused on tracing the topics highlighted in government-funded research (Huang et. al, 2016) or exploring its socio-economic impacts and public value (Bozeman & Youtie, 2017). Few recent studies have paid attention to assessing the technological importance of patents supported by basic research funding. Rather, they are limited to investments from the private sector (Comins, 2015). This leaves a huge gap in the literature on how government investment activity connects to innovation, particularly from the perspective of bibliometrics. This gap could be filled with evidence that reveals which federal agencies are important to research funding, the dynamics of public-private partnerships, and explains the innovation paradigm. In this paper, we mainly address the following research questions: 1) What are the roles of the federal agencies in supporting utility inventions to accelerate innovation? 2) Which sectors of patent assignees helps to communicate the relationships between funding and invention? 3) Are there variations by agency in terms of technology fields?

Framework and Data

To examine the possible relationship between federal funding and government-funded patents, we first need to collect the patents funded by the federal government. As patent filing is sometimes driven by motives other than seeking protection or the actual granting of a patent, incorporating all patents, including applications, into the corpus may result in a less perfect representation of innovation. So, we used granted patents instead of patents filed in the United States Patent and Trademark Office (USPTO) as a more appropriate data source.

These USPTO granted patent data were collected from Thomson Innovation (https://www.thomsoninnovation.com), which brings together the world's most comprehensive international patent coverage and powerful Intellectual Property analysis tools. Analyses to address the primary research questions included counting the number of inventions, as represented by DWPI patent families in the data. This approach ensured that a single invention was not counted multiple times when included in different patent applications in different jurisdictions. Information about interests by federal government agencies was extracted from the field labeled "Government Interest". The utility patent is a patent that covers the creation of a new or improved product, process, or machine. They are generally treated as the very valuable assets because they give inventors exclusive commercial rights to producing and utilizing the latest technology. Therefore, in this paper, we pay attention to utility patents and exclude the other type of patents, such as Design, Plant, SIR's, Reissue, and Defensive Publication patents. Ultimately, 103,411 utility patents granted by USPTO during 1980-2017 were collected (Retrieved on October 12, 2018). We selected 1980 as the starting point because this was the year the Bayh-Dole Act was adopted. The main government-interest statements patterns are summarized in Table 1.

Table 1. The main government-interest statements patterns in the USPTO patent dataset

Sample	*Description*	*Source*
US4181139	The invention described herein may be manufactured and used by or for the Government for governmental purposes without the payment of any royalty thereon.	The invention was supported by a government department.
US4182158	The invention described herein was made by an employee of the U.S. Government and may be manufactured and used by or for the Government for governmental purposes without the payment of any royalties thereon or therefor.	The invention was produced by a government employee.
US5337603	This invention was made with government support under Contract No. W-7405-ENG-36 awarded by the U.S. Department of Energy. The government has certain rights in the invention.	The invention is licensed under a specific contract.
US6452177	The U.S. Government has certain rights in this invention pursuant to NAS7-1407 awarded by NASA.	The invention was funded by a specific funding program or project.

Taking 1980 as the base year, the growth index of granted utility patents with government-interest statements over time is shown in Figure 1. The trend illustrates that the number of utility patents issued both overall and with government funding generally increased up to 1997. A period of inflation during 1998-2009 temporarily flattened growth, followed by a marked increase in the years after. In fact, government-subsidized patents increase eight-fold from 1980 to 2017 – proportionally higher than the overall five-times growth. However, on the whole, the general growth index was similar for both.

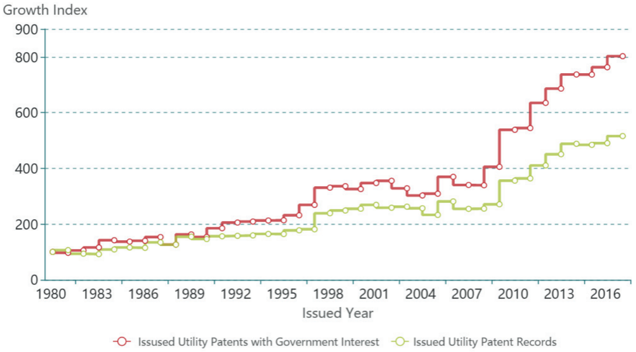


Figure 1. Growth index of issued utility patents overall and with government-interest statements

To explore these statistics in further detail, we charted the two growth rates in terms of percentages over time, as shown in Figure 2. Overall, the percentage of patents with government-interest statements has marginally grown over the past three decades but has tended to hover between 0.8% and 1.4%. As the figure illustrates, the less than 2.0% incidence rate of government-interest statements provides prima facie evidence of under-disclosure. However, this share has shifted substantially over time, rising from 1.2% to 2.0% before 2008 with a steady increase since 2009 to 2.08% in 2016 and 2017.

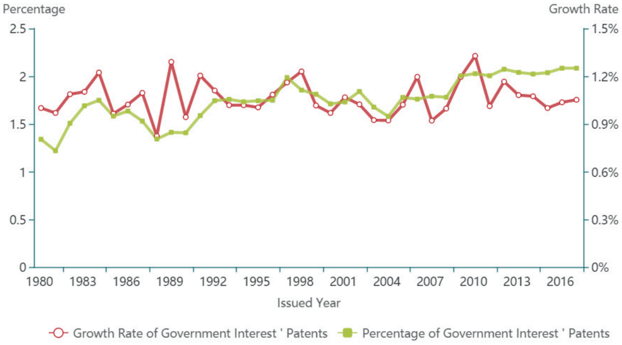


Figure 2. The percentage and growth rate of utility patents with government-interest statements

Results

This section presents the distribution of funded patents by federal agencies, patent assignee sectors by federal agency, and the technology fields funded found in the government-interest records of the dataset.

Distribution of funded patents by federal agencies

Table 2 lists the main federal agencies that supported more than 200 utility patents issued during 1980-2017. The results show the Department of Defense (DoD) is most frequently mentioned, accounting for 25.48% of the granted utility patents supported by the federal government, followed by the Department of Health and Human Services (HHS) and the Department of

Energy (DoE), who participated in 25,796 utility patent families and 20,900 utility patents, respectively. Notably, HHS is the parent agency of the National Institute of Health and the primary government agency responsible for biomedical and public health research. The National Science Foundation (NSF) also plays an important role in government-subsidized patents. The NSF is the agency that supports fundamental research and education in all non-medical fields of science and engineering.

Table 2. Top federal government agencies that supported utility patents during 1980-2017

Federal Agencies	*Records*	*Ratio*	*Federal Agencies*	*Records*	*Ratio*
Dept. of Defense (DoD)	26348	25.48%	Dept. of Agriculture (USDA)	1155	1.12%
Dept. of Health and Human Services (HHS)	25796	24.95%	Dept. of Transportation (DoT)	318	0.31%
Dept. of Energy (DoE)	20900	20.21%	Environmental Protection Agency (EPA)	275	0.27%
National Science Foundation (NSF)	10036	9.70%	Dept. of Homeland Security (DHS)	210	0.20%
National Aeronautics & Space Administration (NASA)	2870	2.78%	Dept. of Education (ED)	209	0.20%
Dept. of Commerce (DoC)	1462	1.41%	Dept. of Veteran's Affairs (VA)	208	0.20%

To further analyze joint funding by multiple federal agencies, we generated a co-funding network map of 12 agencies – see Figure 3. A node indicates an agency, and the size represents the number of patent families funded by that agency. According to the map, the leading four have a relatively strong co-funding relationship with utility patents, especially among DoD, HHS, and the NSF. This could be a signal that these agencies share common interests in certain inventions despite their different agency accountabilities. It is not surprising to see that the National Aeronautics & Space Administration (NASA) has a high co-occurrence with DoD as aeronautics and aerospace research always occupies an important position in defense and military activities.

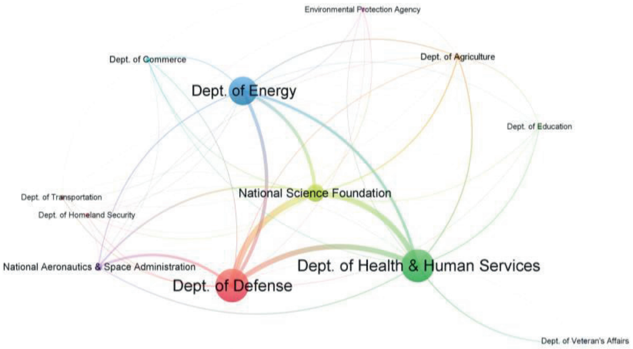

Environmental Protection Agency
Dept. of Commerce
Dept. of Agriculture
Dept. of Energy
Dept. of Education
Dept. of Transportation
National Science Foundation
Dept. of Homeland Security
National Aeronautics & Space Administration
Dept. of Health & Human Services
Dept. of Defense
Dept. of Veteran's Affairs

Figure 3. Joint-funding map for granted utility patents by the top federal government agencies

Figure 4 shows the number of patents funded by the top four federal government agencies (DoD, HHS, DoE, and NSF) during 1980-2017. During the cold war up to the 1980s, technologies related to defense and the military received more attention, so DoD and DoE account for a greater share of the funding than the other agencies. However, with the end of the rivalry between the US and the USSR, more resources were directed toward exploring the potential of science and technology for economic development, and the number of granted patents funded by the four departments gradually began to increase, especially for HHS. The number and, in turn, amount of research funding grew rapidly under the Obama administration for all four departments but, particularly for the NSF, who received a great deal of financial support as part of America's economic stimulus plan.

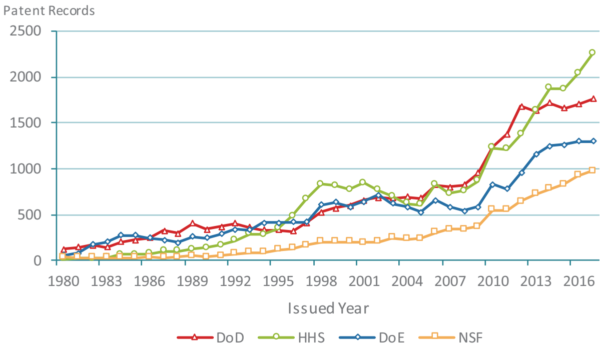


Figure 4. The annual number of issued utility patents funded by leading government agencies

Distribution of patent assignee sectors by federal agencies

Delving deeper into this analysis, we divided the patent assignees into four different sectors: academic & research, corporate, individual, and government. Figure 5 shows the annual activity in patents with government-interest statements by sector. Until the early 1990s, government assignees had the most issued patents. But, around 1994, academic & research assignees took the lead, closely followed by corporate assignees, whose role was taken over by individual assignees after 2003. The landscape shifted again in 2010 when academic & research assignees and corporate assignees jostled for the largest share leaving government-funded patents far behind.

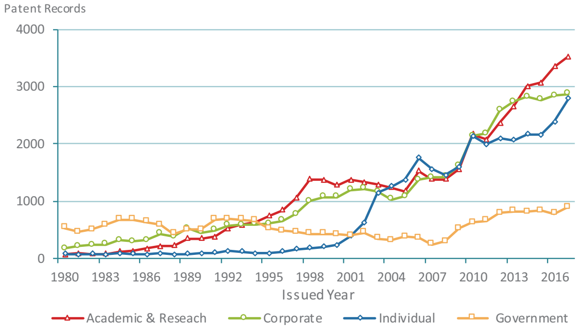


Figure 5. The assignee sectors in issued utility patents with government-interest statements

The top 10 leading assignees for federally-funded patents according to sector are listed in Table 3. Here, we see that distribution of assignees in the academic & research and corporate sectors tends to be more balanced. Whereas, in the government sector, there is a very limited group of assignees. Three subordinate military departments of the DoD – the Army, Navy, and Air Force – along with NASA and the DoE are the most notable.

Table 3. Top 10 most leading assignee in three sector assignees

NO	*Academic & Research Assignee*	*Corporate Assignee*	*Government Assignee*
1	Univ California (5821)	Sandia Corp (2554)	US Sec of Navy (5090)
2	Massachusetts Inst Technology (2653)	Int Business Machines Corp (2251)	US Sec of Army (4569)
3	Univ Leland Stanford Junior (1545)	General Electric Co (1841)	US Sec of Air Force (3198)
4	Univ Texas System (1376)	United Technologies Corp (1178)	NASA US Nat Aero & Space Admin (2612)
5	California Inst of Technology (1199)	Lawrence Livermore Nat Security LLC (2107)	US Dept Energy (2302)
6	Univ Washington (1061)	Honeywell Int Inc (1110)	US Dept Veterans Affairs (668)
7	Univ Johns Hopkins (1048)	Ut-Battelle LLC (985)	US Dept Health & Human Services (527)
8	Univ Florida (1022)	Raytheon Co (936)	US Sec of Agric (121)
9	Univ Michigan (1004)	Boeing Co (844)	US Sec of Commerce (98)
10	Univ New York State Res Found (932)	Westinghouse Electric Corp (709)	US Sec of Interior (60)

Figure 6 shows the number of patents funded by the DoD, HHS, DoE, and the NSF according to sector. The government sector has a relatively low percentage of utility patents supported by these four federal departments, but it is clear that this sector channels more of its funding into the DoE, likely due to its many national laboratories. The corporate sector receives the most support from the DoD and DoE, while universities and research institutions receive the most attention from HHS and the NSF who supports fundamental research and education across all fields of science and engineering.

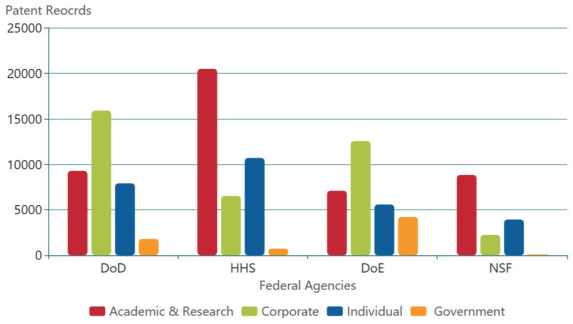


Figure 6. Patent assignee sectors supported by top 4 federal government agencies

Distribution of technology fields by federal agencies

Our third research question concerns which technology fields are most represented in government-funded research. However, appropriately classifying these fields needs careful attention. One solution is to fold the patent subclasses up into the 4-digit umbrella categories provided in the International Patent Classification (IPC) system. However, in 1992, Fraunhofer ISI and the Observatoire des Sciences etdes Technologies, in cooperation with the French patent office (INPI), developed a more systematic technology classification, called the ISI-OST-INPI classification. Their system is loosely based on IPC codes (Schmoch 2008) but has been amended several times to keep pace with the evolutions and revolutions in technological fields, so new codes are added as needed. The latest edition from the WIPO Statistics Database includes 35 technological fields that cover and balance all possible IPCs (an Excel spreadsheet is available at: www.wipo.int/ipstats/en/statistics/patents). The top 20 technology fields are shown in Table 4.

Table 4. Top 20 technology fields reflected in utility patents supported by federal agencies

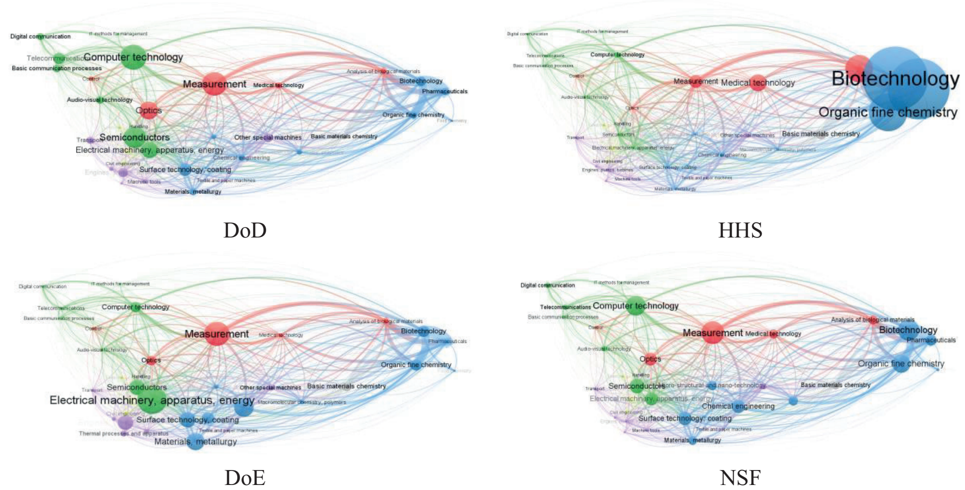
Technology Fields	*Records*	*Percentage*	*Technology Fields*	*Records*	*Percentage*
Biotechnology	20,572	19.89%	Medical technology	6806	6.58%
Pharmaceuticals	17,472	16.90%	Basic materials chemistry	6630	6.41%
Measurement	16,270	15.73%	Chemical engineering	6402	6.19%
Organic fine chemistry	13,836	13.38%	Other special machines	6169	5.97%
Computer technology	10,920	10.56%	Engines, pumps, turbines	5307	5.13%
Electrical machinery, apparatus, energy	10,039	9.71%	Materials, metallurgy	5296	5.12%
Analysis of biological materials	8917	8.62%	Telecommunications	4607	4.46%
Semiconductors	8464	8.18%	Audio-visual technology	3528	3.41%
Optics	7243	7.00%	Macromolecular chemistry, polymers	3521	3.40%
Surface technology, coating	6829	6.60%	Transport	3485	3.37%

Every patent lists at least one IPC, so the co-occurrence of IPCs can be upgraded to the co-occurrence of the 35 technology fields mentioned above. Figure 7 illustrates the structure of funded technology revealed in our analysis. The nodes represent a technology field, and the size of the node reflects co-occurring technology fields. The weight of the links indicates the frequency of co-occurrence, which are drawn in the same color if two fields belong to the same technological sector, i.e., electrical engineering is green, instruments (red), chemistry (blue), mechanical engineering (purple), and “other” (yellow).

From Figure 7, we observe that the NSF has a relatively balanced distribution of funding hot technologies compared to the DoD, DoE, and HHS. Measurement and medical technology in instruments and biotechnology are funded extensively, as is organic fine chemistry in chemistry and computer technology and semiconductors in electrical engineering. The mechanical engineering sector has received less attention from the NSF, which may be due to its inherent nature as an applied discipline.

Given the DoD is largely and directly concerned with national security, its emphasis is on technology relating to the three main types of national security: information, economic, and energy security. Our analysis revealed results consistent with these purviews: the two major sectors DoD supports are electrical engineering and instruments and, specifically, the fields of computer technology, semiconductors and electrical machinery, apparatus, and energy. Moreover, the DoD also focuses on measurement and optics in instruments, which similarly, may be attributed to the strong applicability of this sector to measuring and monitoring security. Electrical machinery, apparatus, and energy in the electrical engineering sector appear to be the focus of the DoE, which is unsurprising since these technologies are the foremost components of a nation's energy. Another obvious technology points to measurement in instruments sector. The most remarkable difference between DoE-funded technology and that of the other agencies is its relatively high attention to mechanical engineering, especially mechanical elements. We reason this is because effective protection of national energy requires sophisticated mechanical engineering technology. Materials, coatings in chemistry and metallurgy, and surface technology form further and greater support by the DoE than the other agencies, mainly due to the compatibility of these types of technologies with the features of the DoE.

Unlike the other institutions, HHS has an evident emphasis on biotechnology, pharmaceuticals, and organic fine chemistry, which fall within the chemistry sector. Biotechnology is a broad area that involves living systems and organisms as the motivation for developing products, such as pharmaceuticals. These issues are highly related to health, medicine, and human services. Therefore, it is reasonable for HHS to give priority to biotechnology and chemistry. As a highly complex subfield of chemical technology, organic fine chemistry also plays a crucial role in biotechnology development. The network map in Figure 7 makes the relationship between organic fine chemistry and biotechnology clear. Instruments is another sector that has received relatively high attention, particularly medical technology and measurement. Like biotechnology, medically-related fields are core to HHS, so it is natural for this agency to invest heavily in medical technology. However, compared to the above three institutes, HHS pays less attention to measurement, possibly because of the differences in HHS's portfolio of responsibilities compared to the other agencies.



DoD

HHS

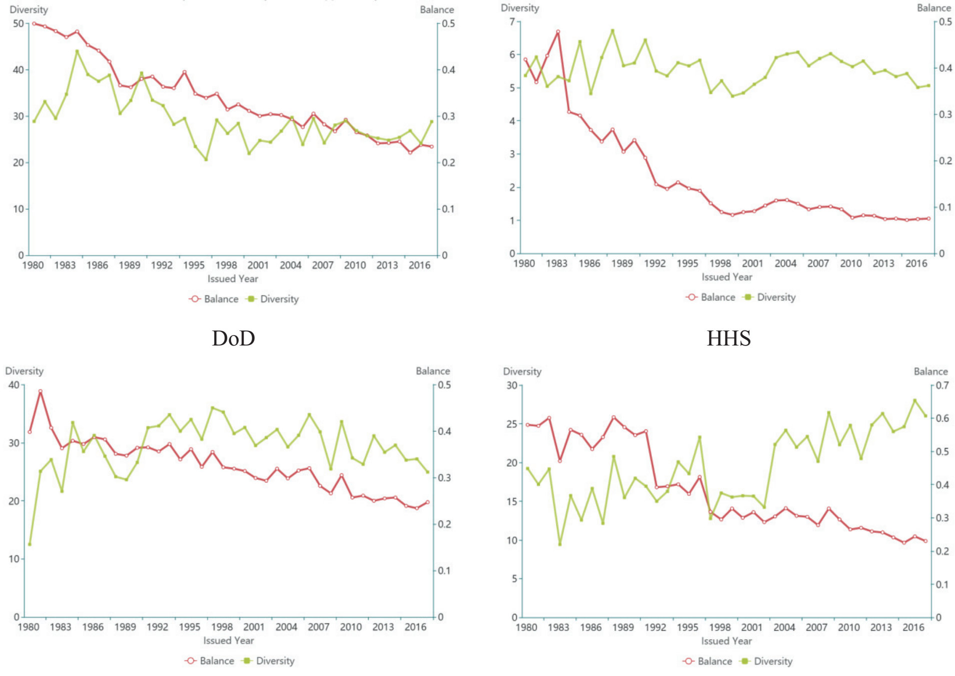
DoE

NSF

Figure 7. The technology overlap map of utility patents supported by four federal government agencies

The above analysis reveals distinct distributions in the technology fields funded by different government agencies. As such, we replaced the 35 broad fields with 4-digit IPC categories to calculate balance and diversity at an aggregate level by patent granted year. The results dissect the annual trends in DoD, HHS, DoE, and NSF funding. Here, "Balance = 1-Geni" corresponds to a measure of evenness or balance in the various 4-digit IPC categories funded by all four government agencies across the entire period of study. The diversity measure, which comprises the components of variety, balance and disparity (Zhang et al. 2016). The results are presented in Figure 8.

Overall, the balance and diversity of the supported patents reveals significant differences over the years. This comparison shows several interesting results. First, the diversity of utility patents supported by the HHS shows much less diversity than the other three agencies, which may be due to HHS's main functions of protecting health and providing essential human services. Medical research consistently remains its clear research focus. Second, almost all of the balance indicators have tended to decrease over the past several decades with few fluctuations. Notably, these values for HHS and the NSF have been relatively stable since 1998. Third, the trend for the DoD reveals a drop in diversity over the full period, but the opposite is true for the NSF. This may tell a story of the DoD restraining its funding support to technological fields that are more directly pertinent to its institutional functions, while the NSF intends to expand its capacity to accelerate fundamental research in the non-medical fields of technology and engineering.



DoD

HHS

DoE

NSF

Figure 8. Balance and diversity indicators for issued utility patents supported by four federal government agencies

Conclusion and Discussion

The aim of this paper was to reveal the paradigm of government activity toward innovations by analyzing government-interest statements in USPTO patents. We arrived at several conclusions throughout our analyses. First, the DoD, HHS, DoE and the NSF were the most prominent federal government agencies to support issued utility patents during 1980-2017, and each has engaged in more and more utility patents in recent years. Second, in the early years of the period studied, most patent assignees fell into the academic & research and corporate sectors. However, individual assignees have shown notable performance in the past ten years. The DoD and the DoE have funded more utility patents in the corporate sector, while universities and research institutions have attracted the most attention from the HHS and the NSF. Third, the technologies that receive the most support by these four agencies, respectively, are generally in line with the agency's primary sphere of responsibilities and government functions. More precisely, the NSF has the most balanced funding structure among the four agencies, and the HHS has the most evident emphasis on medical technologies. The DoD focuses more on technologies relating to various aspects of national security, while the DoE focuses on energy and machinery.

One of the clearest paper trails between public investment offered by federal government agencies and innovation is utility patents. Federally-funded patents are not like those funded by other entities, as government funding is usually conditional on technological sectors and fields. A tentative conclusion points to the US government's interest and desire to stimulate research and development in the specific areas found in our analysis. These insights should deepen our knowledge of possible future directions in the US's national development.

However, although our research sheds some light on the pattern of federal agency support for innovation, issued utility patents are only a proxy for information that could provide a more comprehensive picture. Further study needs to be conducted from more perspectives and in more contexts. For example, do federally-funded utility patents best describe a nation's current capacity for innovation? Do these types of patents have a broad impact on later inventions? Does federal agency support closely correspond to what they care about the most across the board? Do these inventions meet the demand of the strategic development objectives they profess to Congress and the public? Each of these questions and more are worthy of further exploration.

Acknowledgments

This research is supported by the National Natural Science Foundation of China under Grant 71573085 and the Excellence Scholarship in Social Science in HeNan Province (No.2018-YXXZ-10). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the supporters.

References

Bozeman, B. , & Youtie, J. . (2017). Socio-economic impacts and public value of government-funded research: lessons from four us national science foundation initiatives. *Research Policy*, 46(8): 1387-1398.

Burhan, M. , Singh, A. K. , & Jain, S. K. . (2017). Patents as proxy for measuring innovations: a case of changing patent filing behaviour in Indian public funded research organizations. *Technological Forecasting & Social Change*, 123, 181-190.

Comins, & Jordan, A. . (2015). Data-mining the technological importance of government-funded patents in the private sector. *Scientometrics*, 104(2), 425-435.

Corredoira, R. A. , Goldfarb, B. D. , & Yuan, S. . (2018). Federal funding and the rate and direction of inventive activity. *Research Policy*, 47(9), 1777-1800.

Furman, J. L. , Porter, M. E. , & Stern, S. . (2000). The determinants of national innovative capacity. *Research Policy*, 31(6), 899-933.

Huang, Y. , Zhang, Y. , Youtie, J. , Porter, A. , & Wang, X. . (2016). How does national scientific funding support emerging interdisciplinary research: a comparison study of big data research in the us and china. *Plos One*, 11(5), e0154509.

Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics*, 98(3), 1583-1599.

Loise, V. , & Stevens, A. J. . (2010). The bayh-dole act turns 30. *Science Translational Medicine*, 2(52), 52cm27-52cm27.

Mathews, J. A. , & Hu, M. C. . (2005). National innovative capacity in East Asia. *Research Policy*, 34(9), 1322-1349.

Nijssen, D., Rousseau, R., & Hecke, P. V. (1998). The Lorenz curve: A graphical representation of evenness. *Coenoses*, 13(1), 33–38.

Rai, A. K. , & Sampat, B. N. . (2012). Accountability in patenting of federally funded research. *Nature Biotechnology*, 30(10), 953-956.

Ruegg, R. , & Thomas, P. . (2009). Tracing government-funded research in wind energy to commercial renewable power generation. *Research Evaluation*, 18(5), 387-396.

Schmoch, U. (2008). Concept of a technology classification for Country comparison. *Final Report to the World Intellectual Property Organization* (pp. 1-15).

Stevens, A. J. . (2004). The enactment of Bayh-Dole. *The Journal of Technology Transfer*, 29(1). 93-99.

Zhang, L. , Rousseau, R. , & Glänzel, W. . (2016). Diversity of references as an indicator of the interdisciplinarity of journals: taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67(5), 1257-1265.

The HF-rating as a universal complement to the h-index

Yves Fassin[1]

[1] Yves.Fassin@Ugent.be

Ghent University, Department for Marketing, Innovation and Organisation, Tweekerkenstraat, 2, 9000 Gent, (Belgium)

Abstract

Interdisciplinary comparison has been a constant objective of the bibliometric field. The well-known h-index and its alternatives have not achieved this objective. Based on the gh-rating or *gh*ent-rating (Fassin 2018a), a categorization of academic articles into tiers of publications within similar citations ranges, a new ratio is proposed, the *high fame* hf-ratio. This ratio is calculated as the adjusted average of the weighted factors of the researchers' best articles; it leads to an associated hf-rating also designated by the symbols AAA, AA, A, BBB, B, C, D, … etc. comparable to financial ratings such as Moody's and S&P ratings.

The hf-rating provides the average grade of a researcher's best papers benchmarked in their field. This new hfrating induces some qualitative elements in the evaluation of research, includes more selectivity and mitigates between classic h-indices. Adding this hf-rating to the h-index forms a *high fame* HF-rating. This universal HFrating complements the well-known h-index with a relative indication of its influence in its field that also allows inter-field comparison. The methodology is illustrated with examples of researchers with different citations distribution from different disciplines.

Introduction

Indicators are the essence of bibliometrics (Vinkler 2010). Numerous indicators have been proposed in recent years (for a comparative overview see Yan, Wu & Li, 2016). Many bibliometricians have drawn the attention on the difficulty to grasp a complete oeuvre in a single indicator and have proposed additional elements (Costas & Bordons, 2007; Bornmann and Daniel, 2009; Zhang, 2009). Despite the conceptual weakness of any single indicator, the hindex (Hirsch, 2005) has been widely accepted as a simple indicator of a researcher's influence. But it has also been criticized for a number of drawbacks and inconsistencies (Costas & Bordons, 2007; Wendl, 2007; Waltman & Van Eck (2012). Especially its inaptitude for benchmarking researchers from different disciplines (Batista et al. 2006) emphasizes the need for contextualization (Wendl, 2007). An important disadvantage is that the h-index is not sensitive to the increase of number of citations of the most influential (highly-cited) papers (Vinkler, 2010, p. 864). Consequently, a number of variants such as the Kosmulski-index $h^{(2)}$ have been proposed (Kosmulski, 2006). More recently, a '*fame*' index f^2 has been developed, based on a categorization of academic articles, the gh-rating or *gh*ent-rating (Fassin, 2018a).

The *gh*ent-rating and refinements

Radicchi, Fortunato and Castellano (2008) pointed to the universality of citations distributions. In his recent publication, Fassin (2018a) proposed an innovative approach for a categorization of articles in function of their position in the citation distribution rank. This new rating system for academic publications, the gh-rating or *gh*ent-rating, is based on a categorization into tiers of publications within similar citations ranges. These *gh*ent-ratings are comparable to financial ratings such as Moody's and S&P ratings, with categories designated by the symbols AAA, AA, A, BA, BBB, BB, B, CCC, CC, C, D, E, etc.

The categorization makes use of a variable percentile approach based on recently developed htype indexes (Hirsch, 2005; Egghe, 2006). The levels set to categorize articles into the different categories are defined by a mix of standard levels for the higher percentile ranges (for articles with fewer citations), and the h-type percentiles for the lower percentile ranges, i.e., for articles with a large amount of citations. In practice, I opt for a model with three superposing methods to define the thresholds. The basic division rests upon the standard percentiles and two different methods at both end of the distribution ranking: h-percentiles at the top end and fixed thresholds at the lower end, expressed through a minimum number of citations (0, 1 or 2).

The principles behind those ratings lie on an exponentially increase of quality in function of the higher grades. The categories are divided in grades A, B, C, D, E and Z in declining order of citations, each with a corresponding weighted factor defined by a geometric sequence (Fassin, 2018a). The B, C and D categories are delineated through the 10%, 25% and 50% percentile. The A-category is defined through the h-percentile, the percentage of articles within the h-core of the dataset (Rousseau, 2006). The Z-category groups the articles that have not received any citation yet, and the E-category groups those articles with 1 or 2 citations, and those that have not reached the 50% percentile. Figure 1 presents the gh-rating categories on the citation distribution curve.

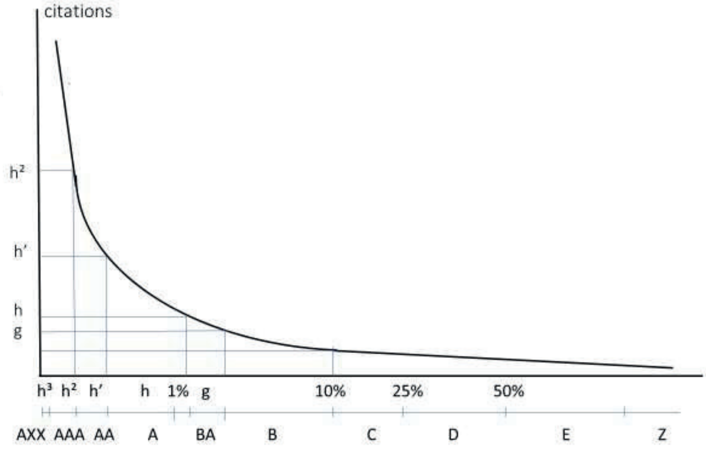


Figure 1 - The citation distribution curve and the gh-rating categories (Fassin, 2018a).

While the lower categories group a larger amount of articles (25% for D and 15% for C), the higher categories comprise around 9% (for B) or 1 % (for A). Those general categories are further subdivided into subcategories with the first letter of the general category, for example CCC, CCD, CC and C for the general category C for the 12.5, 15, 20 and 25%-percentiles.

Further subdivisions are calculated at the top of the distribution on the basis of g, h, h', h^2 and h^3-percentiles, based on their respective indexes. They define the respective AXX, AAA, AA, A and BA categories. The BA-category corresponds to the g-percentile (Fassin, 2018a) based on the g-index of this set of articles defined as the highest rank g such that these g articles together received at least g^2 citations (Egghe, 2006). The $h^{(2)}$-index or Kosmulski-index is equal to h2 if r = h2 is the highest rank such that the first h2 articles each received at least $(h_2)^2$ citations. In analogy, the h^3-index is equal to h3 articles that have at least $(h3)^3$ citations[1], for example an h^3-index of 5 means that this author has 5 articles with at least 5^3 or 125 citations (Fassin, 2018b).

The weighted factors are defined by a geometric sequence: 4, 2, 1, ½ and ¼ with the 'normal' weight of 1 assigned to the 10% percentile band (i.e. B). Sub-categories receive an intermediate weighted factor, as defined in table 1. For each of the publications that fall within the top categories, within the h-core and g-core, a bonus system is constructed: a bonus of 0.25 for the g-core, 0.50 for the h-core, 1 for the h'-core, 2 for the h^2-core, 3 for the h^3-core, is added to the starting weighted factor between 1 for the 10% B category or 2 for the higher cores of the 1% BBB category[2]. An additional bonus of 1 is added for the 0.1% highly-cited articles (for datasets over 500 units). The spread between the weighted factor reaches a factor of 50, with 6 for the most-cited papers and 0.12 for the lowest category without citations. The bonus system for the h-percentiles contributes to mitigate for different h-indexes following different databases. While the h-indexes of Scopus and the Web of Science may differ for 20%, the databases based on Google Scholars attain double of the h-indexes of the Web of Science. In this bonus system, there is still differentiation if articles within the h-core also belong to the 1% percentile category BBB or only to the 10% percentile category B.

Table 3 in appendix presents the successive rating categories with the corresponding percentile and weighted factor (see infra). The sub-categorization allows to make a better differentiation when comparing authors with less cited publications. Table 1 presents a continuum of thresholds based on percentiles and the corresponding rating categories. h-type percentiles can overlap the standard percentiles.

Normalization: the hf-ratio and the HF-rating

A constant objective of the bibliometric field has been the quest for normalization, in order to allow interdisciplinary comparison. The h-index and its alternatives have not succeeded in this endeavor. In order to compare different researcher's contribution, one could select their i best articles, and sum up their corresponding weighted factors[3].

If f_n is the contribution of the n-th paper (the weighted factor of this publication including bonus), then fi is the sum of the weighted factors of the i best cited publications: $fi = \sum f_n$. The top four fi-index f4 is thus $f4 = f_1 + f_2 + f_3 + f_4$.

[1] I will further use the symbols h^2 and h^3, rather than $h^{(2)}$ and $h^{(3)}$.

[2] Which gives 1.25 for that article if in the g-core and also within the 10% (B) or 1.50 if also within the 5% (BB). An article within the 1% (BBB) has a weight factor of 2 points; if also in the h-core it raises to 2.5; if in the h^2core it reaches 4.

[3] In difference with the f^2-index (Fassin, 2018a), that sums up the weighted factors of all the articles in the author's h^2-core.

An alternative for a more balanced comparison is to select a fixed number of the highest cited papers and to calculate the average $hfi = 1/i . \sum f_n$.

Following a common approach in statistical analysis, where often the extreme data are dropped to give a more precise measurement, I propose an adjusted average by dividing the sum of the i papers by (i-1). So, $hfi' = 1 /(i-1) . \sum f_n$ further called the researcher's hf-ratio.

This adjustment allows to avoid disadvantaging younger researchers who have no more than n papers or whose nth paper has not attained the same impact yet.

Reversed conversion of this hf-ratio on basis of same table in appendix (limited to AAA) leads to the categories for researchers, or hf-ratings, AAA, AA, A, BBB, BBC, BB, B, CCC, CCD, CC, C, D, E, etc. with corresponding percentiles.

In practice, I propose to select the researcher's most cited 4 papers (n=4). So, three A papers will give an A-rating to the researcher. A minimum hf-rating of B is obtained with 3 B papers or with 2 B papers and 2 C papers, etc. Some research fields with higher frequency of publications, often with a large number of authors, may chose a larger fixed number than 4. The division by a fixed minimum factor, 3 or (n-1), also allows to mitigate the contribution of authors with only one or two papers in the field but with exceptional amounts of citations: whereas the weighted factor f1 would give them the maximum count of 6, that only paper will give them a hf-ratio of 2. In order to distinguish occasional authors with only one or two papers from researchers with a large body of research in that field, they will receive only the basic categories *A*, *B*, *C* or *D*, put in italics.

In practice, seen the wide dissemination and acceptance of the h-index, I propose to add this hfrating, based on the converted hf-ratio, to the author's h-index to form a HF-rating (*high fame*)[4]. This new HF-rating complements the well-known h-index with a relative indication of its influence in its field that better allows comparison between peers.

The advantages of the hf-rating: Inter-field comparison

The proposed hf-rating offers some advantages, especially a possibility of comparison of the impact of a researcher within his peers. The hf-rating does not aim to rank, but leads to a rating in tiers of articles grouped in comparable categories. This hf-rating allows benchmarking. It gives the average categorization of the researcher's (i or 4) best cited papers benchmarked in his discipline or in her field.

This new hf-rating based on the *gh*ent-rating induces some qualitative elements in the evaluation of research and mitigates between classic h-indices. It includes more selectivity, and allows to single out more influential papers than the traditional h-indexes.

But the great benefit of the hf-rating lies in its universal scope of application. Thanks to its normalization character, the new HF-rating allows, to a certain degree, inter-field comparison. I will illustrate this with a practical example in a few totally different disciplines.

[4] As an alternative for a H^{3}F rating, where the hf-rating would follow the h^2-index.

Table 1 presents a continuum of thresholds based on percentiles and the corresponding rating categories. h-type percentiles can overlap the standard percentiles. The table gives the number of articles (n), the h, h², h³ and g-indexes for datasets for science in general and for each of the 4 disciplines: physics, medicine, entrepreneurship and bibliometrics, selected as 'topic' in the Web of Science search. The disciplines have been chosen on the basis of diversity in size of datasets. The four fields range from around 10.000 articles for bibliometrics over 50.000 for entrepreneurship to half a million publications for medicine or physics. The following columns in table 1 show the highest cited citation count, the required citation for the 0.1%, 1 and 10 % percentiles, as well as the required citations for the g, h, h² and h³-core.

Table 1: Distribution data of different disciplines (Retrieved from Web of Science on 21st May 2018, as 'topic')*.

Discipline	*n*	*h*	*h²*	*h³*	*g °*	*max*	*top 0.1%*	*top 1%*	*top 5%*	*top 10%*	*g*	*h*	*h²*	*h³*
Science	1555926	931	56	17	*2327*	61290	723	221	88	~40	*592*	931	3405	5509
Physics	479105	696	52	16	*1740*	21413	854	220	81	49	*401*	696	2895	4461
Medicine	485321	559	49	15	*1398*	20527	604	160	81	33	*337*	559	2027	3817
Entrepreneurship	48086	278	28	10	*695*	3527	671	199	62	29	*157*	278	822	1331
Bibliometrics	9610	111	16	7	327	795	404	120	48	27	60	111	288	403

° g-index approximately defined as 2.5*h, as total citations is not known

As an illustration of an interdisciplinary benchmarking, I present the calculation of the hf-ratio (table 2) of a selected number of researchers in those different scientific disciplines, physics, AIDS- and entrepreneurship research, and bibliometrics. For each field, I select one of the most influential researchers, and two other authors of which a younger researcher. For the entrepreneurship field, I compare the three most influential authors and a younger author. Hirsch, the founder of the h-index, is positioned in two fields: physics, where he has his largest contribution, and bibliometrics, where he only has a limited number of extremely impactful papers.

The table shows the number of publications and the total number of citations of each researcher, the number of citations of the highest-cited article (c max), the average of citations per paper, and their h and h²-indexes. Then follow the weighted factors of their 4 most cited papers, and the hf-ratio as defined supra. The table is completed with the HF-rating of those researchers.

The comparison illustrates the variety of citation habits and size of different fields that result in higher h-indexes for influential authors of broad large disciplines. In contrast, impactful authors in smaller specialized disciplines have a lower h, but comparable hf-ratios. The resulting HFrating allows – to a certain extent – inter-field comparison. The most influential authors in each discipline obtain an AAA categorization, independent from their largely different h-index: Witten with 137, Montagnier with 63, Shane with 38 and Bornmann with 36; also independent from the huge differences in total citations or number of citations of their best-cited article, about 20 times higher for Witten than for Bornmann.

The use of the hf4' variant for the hf-ratio where the 4 best papers are chosen and divided by 3 allows also to evaluate and benchmark the work of younger authors. In an absolute ranking in order of the number of papers or h-index, the researchers in physics and medicine would rank much higher than their colleagues in smaller disciplines. The comparison of the research oeuvre of Danziger-Isakov in medicine (h-index of 13 for 31 papers) with Rinia in bibliometrics (9 papers with h-index of 5) is nuanced with the A-grade for Rinia and a BB-grade for DanzigerIsakov, the average grade of their best papers benchmarked in their field.

Table 2: h-indexes and hf-ratio's of a selected number of researchers in different disciplines: physics, AIDs and entrepreneurship research, bibliometrics.

Authors	*n*	*tot cit*	*c max*	*avg*	*h*	*h^2*		*f_1*	*f_2*	*f_3*	*f_4*	*hf*		*HF-rating*	
Witten	344	93317	3088	271.3	137	29		6	4	4	4	6		137	AAA
Hirsch	264	14241	1435	53.9	57	14		6	2	1.75	1.75	3.83		57	AA
Rontynen	5	116	48	23.2	5	3		0.75	0.66	0.6		0.67		5	CC
Montagnier	246	5292	5922	99.1	63	16		6	4	2	2	4.67		63	AAA
Tan ET	77	622	166	8.1	11	5		1.5	1.25	1.25	1	1.67		11	BBB
Danziger-Isakov	31	359	42	10.0	13	5		1	1	1	0.75	1.25		13	BB
Wright	141	7692	391	54.5	50	12		2	2	1.75	1.75	2.50		50	A
Shane	61	11280	3371	184.9	38	12		6	6	3	2	5.67		38	AAA
Zahra	73	8930	664	122.3	43	15		3	3	2	2	3.33		43	AA
Minola	15	82	24	5.5	5	3		0.75	0.66	0.6		0.67		5	CC
Bornmann	249	5187	420	20.8	36	9		6	3	3	3	5		36	AAA
Hirsch	3	3695	3205	1231.7	3	3		6	4	1.75		3.92		3	AA
Rinia	9	337	124	37.4	6	5		2	1.75	1.33	1.25	2.11		6	A

Seen the exponential aspect of the categorization of the gh-rating, the categories reflect some proportionate distribution in the researcher's categorization. As a logic result, more researchers in the larger disciplines with more researchers will be able to obtain the higher ratings, than their colleagues from more limited disciplines.

Conclusion

With this extension of the application of the *gh*ent-rating, to a normalized hf-ratio and derived HF-rating, I contribute to a better method for benchmarking researchers among different research disciplines. Indeed, while the h-index and most h-type related indicators depend from the discipline, the normalization offered by the hf-ratio allows identifying tiers of comparable researchers over all fields. The hf-rating provides the average grade of a researcher's best papers benchmarked in their field.

This *high-fame* HF-rating corresponds to the call in bibliometrics for more qualitative indicators and 'responsible metrics' (Editorial Nature, 2015). The additional information provided by the HF-rating adds context to the h-index thanks to the normalization based on the universality of citation distributions.

Like many other indicators, the hf-ratio is only PAC (Probably Approximately Correct) (Rousseau, 2016). The present ratio has limitations, due to the constraints, as based on total citations, that favours established authors, and on full count for multiple authorship. Future developments should tackle these issues of researcher seniority and multiple authorship. Despite these limitations, the hf-ratio and its corresponding HF-rating offer a valuable tool for interdisciplinary benchmarking between researchers. The same methodology and HF-rating can also be applied for the benchmarking of scientific teams or universities.

References

Anonymous (2015), Editorials, *Nature*, July 2015, 523 :127-128.

Batista, P., Campiteli, M., Kinouchi, O. & Martinez, A. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1).179-189.

Bornmann, L., & Daniel, H.-D. (2009). The state of h index research. Is the h index the ideal way to measure research performance? *EMBO Reports*, 10(1), 2–6.

Costas, R. & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1: 193–203.

Egghe, L. (2006). 'Theory and practice of the g-index'. *Scientometrics*, 69(1):131-152.

Fassin, Y. (2018a). A new qualitative rating system for scientific publications and a fame Index for academics. *Journal of the Association for Information Science and Technology*, 69(11):1396–1399.

Fassin, Y. (2018b). A new h^3-index for academic publications. *STI conference,* Leiden, 14 September.

Hirsch, J.E. (2005). 'An index to quantify an individual's scientific research output'. *Proceedings of the National Academy of Sciences USA* 102:16569-16572.

Kosmulski, M. (2006). 'A new Hirsch-type index saves time and works equally well as the original hindex'. *ISSI Newsletter*, 2(3): 4-6.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). 'Turning the tables on citation analysis one more time: Principles for comparing sets of documents'. *Journal of the American Society for Information Science and Technology*, 62(7), 1370-1381.

Radicchi, F. , Fortunato, S. , Castellano C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105 (45) 17268-17272.

Rousseau, R. (2016). Citation data as proxy for quality or scientific influence are at best PAC (Probably Approximately Correct). *Journal of the Association for Information Science and Technology,* 67 (12) 3092-3094.

Vinkler, P. (2010). Indicators are the essence of scientometrics and bibliometrics. *Scientometrics* 85 (3), 861-866.

Waltman, L. and Van Eck, NJ. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*. 63(2007):406–415.

Wendl, (2007). H-index: however ranked, citations need context. Correspondence, *Nature*, 449:403.

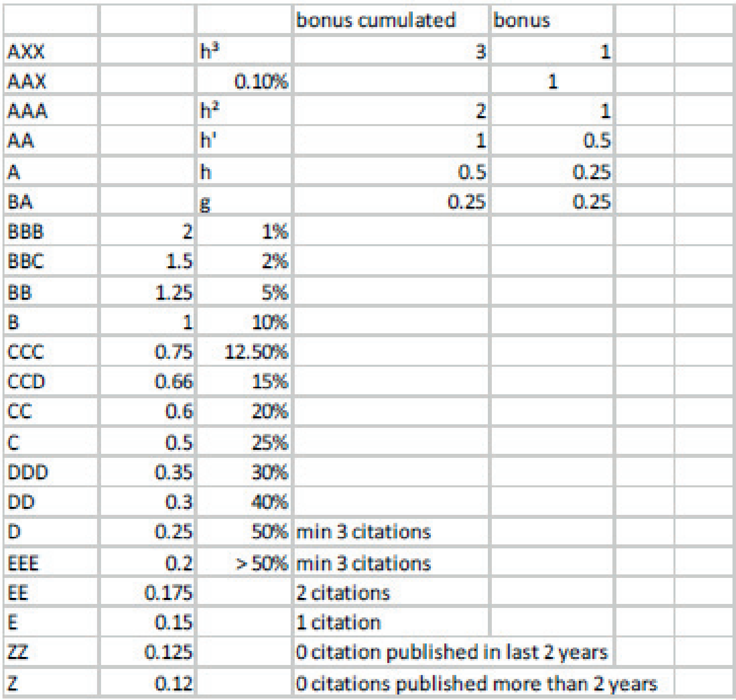
Wilsdon, J. (2015). We need a measures approach to metrics, *Nature*, 523: 129.

Yan, Z. , Wu Q. & Li, X. (2016). Do Hirsch-type indices behave the same in assessing single publications? An empirical study of 29 bibliometric indicators. *Scientometrics,* 109(3).

Zhang, CT, (2009). The e-Index, Complementing the h-Index for Excess Citations. *Plos One*, 4 (5) e5429.

Appendix

Table 3: The *gh*ent-rating percentiles and weighted factors.



			bonus cumulated	bonus		
AXX		h^3	3	1		
AAX		0.10%		1		
AAA		h^2	2	1		
AA		h'	1	0.5		
A		h	0.5	0.25		
BA		g	0.25	0.25		
BBB	2	1%				
BBC	1.5	2%				
BB	1.25	5%				
B	1	10%				
CCC	0.75	12.50%				
CCD	0.66	15%				
CC	0.6	20%				
C	0.5	25%				
DDD	0.35	30%				
DD	0.3	40%				
D	0.25	50%	min 3 citations			
EEE	0.2	> 50%	min 3 citations			
EE	0.175		2 citations			
E	0.15		1 citation			
ZZ	0.125		0 citation published in last 2 years			
Z	0.12		0 citations published more than 2 years			

Quantifying the research preferences of top research universities: why they make a difference?

Barbara S. Lancho-Barrantes[1] and Francisco J. Cantu-Ortiz[2]

[1] b.s.lancho-barrantes@leeds.ac.uk

University of Leeds, LS2 9JT, Leeds, (United Kingdom)

[2] fcantu@tec.mx

Tecnologico de Monterrey, Eugenio Garza Sada 2501, 64849 Monterrey, N.L., (Mexico)

Abstract

Research universities are institutions with a strong vocation and advocacy towards research. They are in the top of the world university rankings because of their excellence in research. This paper analyzes research universities focusing on their research preferences. We have selected the top twelve universities of THE World University Rankings 2019 and Scopus and SciVal as source of data with a five-year publication window. In order to analyze university preferences, we have applied a statistitical technique called cosine similarity. Besides, cluster analysis through VOSviewer showed and classified the terms most used by universities. The results have revealed that research universities have a strong commitment in specific areas. Finally, we have analyzed the scientific production in collaboration between all them and their preferences to cooperate.

Introduction

World leading universities devote to research as a central part of their mission. These institutions focus on the discovery of new scientific knowledge and future researchers training (Mohrman, Ma & Baker, 2008). Research universities make the difference with teaching universities giving more emphasis to research instead of teaching. In fact, pursuing publishing is not something that defines teaching universities. Research universities also commit to teaching as a social role of universities, but their nature is rather shaped by a research infrastructure (Taylor, 2006).

According to League of European Research Universities (LERU) the research universities are the ultimate source of innovation in the economy, society and culture. They train people to think with skepticism, creativity, and high-level capability that society demands (LERU, 2019).

The United States developed the concept of research universities. The Carnegie Classifications defines them as institutions that offer baccalaureate and doctorate programs (Carnegie Foundation, 2001). Taylor (2006) stayed that the key characteristics of leading research universities are: a) Presence of pure and applied research. b) Delivery of research-led teaching. c) Breadth of academic disciplines. d) High proportion of postgraduate research programs. e) High levels of external income. f) An international perspective.

All over the world, countries have recognized that research universities are key to the knowledge economy of the 21st century (Clark, 2004; Etzkowitz & Leydesdorff, 2000). Power & Dusdal (2017a) examined three leading countries in organizational development and scientific innovation (Germany, France and the United Kingdom). They found that global investments or the number of researchers do not influence in countries' productivity. Their findings explained that the institutionalization of research university support high productivity.

Specially, United Kingdom has a group of leading research universities that has attracted the best talent worldwide (Powell & Dusdal 2017b). Research universities are a fundamental element in the connection between research and teaching by giving freedom to teach and to study, autonomy and commitment to science as well as the hosting of future researchers.

During the twentieth century, United States (US) foster a small nucleus of productive "super research universities". This expansion was product of the increase of massive tertiary education in this nation (Fernandez & Baker, 2017). US research universities are research centers increasing the knowledge in all scientific disciplines. They are contributing to the general economy of the country and also to local and regional economies. The US university system is one of the best in the world including the number of Nobel Laureates awarded to their faculty members. Some countries have tried to imitate the model of the US university system, but with limited success. The reason is that most university systems are controlled by governments (Atkinson & Blanpied, 2008).

The Times Higher Education (THE) World University Rankings defines different criteria to include universities, three of which meet the concept of research universities: i: Enough publications – An institution must publish more than 1,000 papers over the previous 5 years, and more than 150 publications in any single year. ii. Thresholds are also applied per subject for the subject rankings. iii: Subject breadth – An institution must not be focused on a single narrow subject area.

Research questions addressed in this work are: What is peculiar about the scientific production of research universities? Is there any research preference in their publications? Do research preferences have any resemblance among them? What are the most used terms in the production of these universities? The most prolific authors, are they national or international? Are these universities collaborating with each other or are they competing among themselves? Which are the preferences to collaborate?

Data and method

We first concentrate on those research universities that state prestigious university brands. We have used the Times Higher Education (THE) Ranking World University rankings 2019 to sort the top twelve institutions. We have used Elsevier's Scopus to extract publications because it is one of the world's most comprehensive bibliometric databases and is employed by THE to calculate rankings. We also retrieved the number of publications of the top universities for each research subject in those years. These information was included in Elsevier's SciVal to apply indicators and in the VOSviewer program to create clusters and maps.

Scientific collaboration is a quality that defines leading world-class institutions since they cannot excel in isolation, as they are purpose-built for cooperation. For this reason, we have also analyzed collaboration among them.

The research universities chosen for the study are as a follows: University of Oxford, University of Cambridge, Stanford University, Massachusetts Institute of Technology, California Institute of Technology, Harvard University, Princeton University, Yale University, Imperial College London, University of Chicago, ETH Zurich, and Johns Hopkins University.

We downloaded 668,204 publications from Scopus which is all the scientific production from the universities in the period of time 2014-2018. The data for each institution was processed using the data management software program, SPSS. Data were retrieved in december 2018.

Similarity measurement among top universities

Two institutions are similar in research preference if their cosine similarity is close. This study compared the research preference using cosine similarity rather than Euclidean distance. Cosine similarity refers to cosine of the angle between two vectors. Generally, the angle between two vectors is used as a measure of divergence between the vectors. Cosine is used as the numeric similarity (where cosine has the property that it is 1.0 for identical vectors and 0.0 for orthogonal vectors) (Singhal, 2001; Zhigang, Gege, & Haiyan, 2017; Lin, Hu, & Hou, 2018)

The cosine similarity of two vectors A and B using a dot product and size as

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

Cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

University performance in Times Higher Education (THE) are grouped into five areas: teaching (the learning environment) (30%); research (volume, income and reputation) (30%); citations (research influence) (30%); international outlook (staff, students and research) (7.5%); and industry income (knowledge transfer) (2.5%).

We selected the metrics: research and citations. They meet 60% of the score and they have a strong influence in the character of a research university.

- Research (volume, income, and reputation) – 30%: the most prominent indicator in this category looks at a university's reputation for research excellence among its peers, based on the responses to our annual Academic Reputation Survey. Therefore: Reputation survey: 18%, Research income: 6% and Research productivity from Scopus: 6%.
- Citations (research influence) – 30%: it examines the research influence by capturing the average number of times a university's published work is cited by scholars globally. The data include more than 25,000 academic journals indexed by Elsevier's Scopus database.

The indicators used in this study (SciVal, 2019)

- Scholarly output which is the total number of publications of an entity.
- % International collaboration indicates the extent to which an entity's publications have international co-authorship.
- Citation count is the total number of citations received by an entity. This is a complement to scholarly output.
- Publications in Top 10 Journal Percentiles, which indicates the extent to which an entity's publications are present in the most-cited journals.
- Outputs in Top Citation Percentiles, which indicates the extent to which an entity's publications are present in the most-cited percentiles.
- Field-Weighted Citation Impact (FWCI) indicates how the number of citations received by an entity's publications compares with the average number of citations received by all other similar publications.

Results

In the following table we can observe the position of the top twelve institutions in THE. We also have applied the bibliometric indicators to scientific production.

Table 1. Indicators applied to top 12 research universities

Research universities	Countries	Overall	Research (volume, income and reputation)	Citations (research influence)	Scholarly Output	Collaboration (%)	Citation Count	Publications in Top 10 Journal Percentiles (%)	Outputs in Top 10 citation percentile (%)	Field- Weighted Citation Impact (FWCI)
		Times Higher Education World University			Scopus- SCIval (2014 - 2018)					
University of Oxford	United Kingdom	96	99.5	99.1	68,034	58.7	855,705	48.7	27.0	2.3
University of Cambridge	United Kingdom	94.8	98.8	97.1	56,726	59.3	699,059	49.5	27.9	2.1
Stanford University	United States	94.7	96.8	99.9	67,662	41.0	966,946	51.6	29.5	2.6
Massachusetts Institute of Technology	United States	94.2	92.7	99.9	44,840	48.9	660,727	55.0	30.7	2.4
California Institute of Technology	United States	94.1	97.2	99.2	22,250	52.0	327,666	47.1	33.6	2.2
Harvard University	United States	93.6	98.4	99.6	148,972	43.6	2,005,555	49.4	29.1	2.3
Princeton University	United States	92.3	93.6	99.4	22,417	46.2	288,466	50.4	28.6	2.4
Yale University	United States	91.3	93.5	97.8	44,497	38.4	537,206	49.2	27.3	2.1
Imperial College London	United Kingdom	90.3	87.7	97.8	57,576	60.3	684,087	49.2	27.1	2.2
University of Chicago	United States	90.2	90.1	99	30,787	35.8	407,384	47.8	27.0	2.2
ETH Zurich	Switzerland	89.3	91.4	93.8	38,806	65.3	445,479	53.1	28.1	2.0
Johns Hopkins University	United States	89	90.5	98.5	65,637	39.4	797,610	44.8	26.3	2.2

We can observe that the two best research universities in the world are from the United Kingdom. They are the corners of the ‘golden triangle’. Golden triangle universities have some of the largest UK university financial endowments. Followed by these are the universities of the United States where most of them are private. Currently, there are more than 250 of these institutions in the United States. In table 2 we can observe the total amount of publication and the percentage with the total of each disciplines. The average of international collaboration is 49%. Percentage of publications in the top 10 journal percentiles is 49.65 and the percentage of outputs in top 10 citation percentiles is 48%.

Table 2. Research preferences in the Top Twelve Research Universities

	University of Oxford		University of Cambridge		Stanford University		Massachusetts Institute of Technology		California Institute of Technology		Harvard university		Princeton University		Yale University		Imperial College London		University of Chicago		ETH Zurich		Johns Hopkins University	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Agricultural and Biological Sciences	4,501	4.65	3,725	4.29	2,997	2.93	1,317	1.92	1,070	2.66	2,670	5.31	1,246	3.39	2,149	3.76	2,518	3.60	1,376	2.92	2,786	4.75	2,155	2.73
Arts and Humanities	4,357	4.50	3,807	4.38	1,848	1.81	616	0.90	128	0.32	2,567	5.11	1,640	4.47	2,035	3.56	242	0.35	1,867	3.96	392	0.67	1,220	1.55
Biochemistry, Genetics and Molecular Biology	10,495	10.85	9,644	11.10	10,828	10.59	5,431	7.93	1,764	4.38	5,052	10.05	1,982	5.40	6,911	12.10	6,775	9.68	5,448	11.57	4,362	7.43	9,434	11.95
Business, Management and Accounting	966	1.00	1,039	1.20	903	0.88	576	0.84	51	0.13	804	1.60	315	0.86	370	0.65	328	0.47	507	1.08	511	0.87	321	0.41
Chemical Engineering	1,588	1.64	1,978	2.28	2,061	2.02	2,600	3.79	883	2.19	1,088	2.16	975	2.66	821	1.44	2,156	3.08	599	1.27	2,032	3.46	872	1.10
Chemistry	3,740	3.87	4,087	4.70	3,533	3.45	4,191	6.12	1,991	4.95	1,991	3.96	1,836	5.00	1,810	3.17	3,581	5.12	1,746	3.71	4,304	7.33	1,738	2.20
Computer Science	4,457	4.61	3,589	4.13	5,983	5.85	5,292	7.72	2,381	5.92	1,984	3.95	2,869	7.82	1,340	2.35	4,709	6.73	1,290	2.74	5,321	9.06	2,582	3.27
Decision Sciences	492	0.51	521	0.60	611	0.60	591	0.86	85	0.21	383	0.76	284	0.77	217	0.38	405	0.58	347	0.74	439	0.75	262	0.33
Dentistry	39	0.04	17	0.02	74	0.07	14	0.02	0	0.00	68	0.14	3	0.01	44	0.08	13	0.02	52	0.11	2	0.00	70	0.09
Earth and Planetary Sciences	3,325	3.44	3,949	4.54	3,324	3.25	3,128	4.57	8,485	21.09	1,721	3.42	2,801	7.63	1,792	3.14	2,136	3.05	1,505	3.19	4,154	7.08	3,061	3.88
Economics, Econometrics and Finance	1,388	1.43	1,076	1.24	1,182	1.16	598	0.87	144	0.36	1,323	2.63	721	1.96	762	1.33	289	0.41	1,042	2.21	697	1.19	435	0.55
Energy	885	0.91	1,145	1.32	1,410	1.38	1,942	2.83	557	1.38	385	0.77	573	1.56	314	0.55	1,721	2.46	173	0.37	1,391	2.37	316	0.40
Engineering	4,476	4.63	6,136	7.06	6,790	6.64	9,936	14.50	4,718	11.73	2,960	5.89	3,020	8.23	1,794	3.14	7,663	10.95	1,090	2.31	6,663	11.35	3,621	4.59
Environmental Science	2,639	2.73	2,251	2.59	2,226	2.18	1,682	2.45	1,156	2.87	1,428	2.84	1,168	3.18	1,361	2.38	2,310	3.30	501	1.06	3,046	5.19	1,177	1.49
Health Professions	488	0.50	365	0.42	794	0.78	159	0.23	18	0.04	164	0.33	47	0.13	370	0.65	385	0.55	232	0.49	195	0.33	757	0.96
Immunology and Microbiology	2,966	3.07	1,813	2.09	2,320	2.27	1,089	1.59	341	0.85	991	1.97	467	1.27	1,469	2.57	2,116	3.02	1,141	2.42	864	1.47	2,386	3.02
Materials Science	3,225	3.33	4,391	5.05	4,284	4.19	6,110	8.92	2,580	6.41	2,294	4.56	2,025	5.52	1,318	2.31	4,108	5.87	1,256	2.67	4,194	7.14	1,936	2.45
Mathematics	3,859	3.99	2,830	3.26	3,418	3.34	3,383	4.94	1,760	4.37	1,802	3.58	2,510	6.84	1,219	2.13	3,158	4.51	1,567	3.33	3,424	5.83	1,671	2.12
Medicine	18,501	19.12	11,516	13.25	25,201	24.64	3,335	4.87	601	1.49	5,754	11.44	1,125	3.06	15,511	27.16	13,965	19.95	12,631	26.81	2,601	4.43	27,493	34.83
Multidisciplinary	1,892	1.96	1,715	1.97	2,267	2.22	1,868	2.73	921	2.29	1,595	3.17	877	2.39	1,150	2.01	1,001	1.43	932	1.98	882	1.50	1,085	1.37
Neuroscience	3,437	3.55	2,847	3.28	3,520	3.44	1,331	1.94	440	1.09	1,550	3.08	640	1.74	2,804	4.91	1,375	1.96	1,351	2.87	915	1.56	3,808	4.82
Nursing	906	0.94	630	0.72	819	0.80	42	0.06	6	0.01	268	0.53	30	0.08	799	1.40	553	0.79	624	1.32	154	0.26	1,489	1.89
Pharmacology, Toxicology and Pharmaceutics	1,106	1.14	899	1.03	974	0.95	509	0.74	107	0.27	350	0.70	73	0.20	1,106	1.94	877	1.25	648	1.38	485	0.83	1,507	1.91
Physics and Astronomy	7,427	7.68	9,277	10.67	8,231	8.05	10,617	15.50	9,683	24.07	4,903	9.75	6,525	17.78	4,064	7.12	6,188	8.84	4,280	9.09	6,967	11.87	5,437	6.89
Psychology	2,108	2.18	1,544	1.78	2,166	2.12	387	0.56	89	0.22	1,661	3.30	493	1.34	2,369	4.15	388	0.55	1,477	3.14	377	0.64	1,486	1.88
Social Sciences	7,311	7.56	5,799	6.67	4,422	4.32	1,705	2.49	270	0.67	4,488	8.93	2,440	6.65	3,151	5.52	948	1.35	3,421	7.26	1,461	2.49	2,516	3.19
Veterinary	170	0.18	328	0.38	79	0.08	65	0.09	6	0.01	39	0.08	22	0.06	55	0.10	99	0.14	2	0.00	85	0.14	100	0.13

The colour green means acceptable values and the red colour might be a room for improvement. Most of research universities have a strong production in Medicine followed by Biochemistry, Genetics and Molecular Biology. However University of Princeton, California Institute of Technology and ETH Zurich have more production in Physics and Astronomy. We must bear in mind that scientific disciplines have different patterns of publication and some disciplines attract more economic funding than others. Both circumstances could affect to scientific production.

Table 3 shows the resulting adjacency matrix, which represents the level of similarity between two pairs of institutions.

Table 3. Adjacency matrix of top research universities' cosine similarity in research areas

	University of Oxford	University of Cambridge	Stanford University	Massachusetts Institute of Technology	California Institute of Technology	Harvard University	Princeton University	Yale University	Imperial College London	University of Chicago	ETH Zurich	Johns Hopkins University
University of Oxford	1											
University of Cambridge	0.835	1										
Stanford University	0.847	0.754	1									
Massachusetts Institute of Technology	0.799	0.826	0.747	1								
California Institute of Technology	0.870	0.772	0.790	0.806	1							
Harvard University	0.810	0.785	0.790	0.824	0.808	1						
Princeton University	0.793	0.866	0.793	0.855	0.782	0.746	1					
Yale University	0.864	0.895	0.828	0.709	0.833	0.780	0.847	1				
Imperial College London	0.784	0.797	0.772	0.713	0.844	0.755	0.812	0.840	1			
University of Chicago	0.851	0.773	0.754	0.767	0.786	0.867	0.739	0.756	0.759	1		
ETH Zurich	0.836	0.794	0.834	0.869	0.903	0.790	0.872	0.809	0.764	0.741	1	
Johns Hopkins University	0.848	0.804	0.854	0.766	0.848	0.863	0.740	0.800	0.839	0.842	0.806	1

Cosine similarity is 0.87 for University of Oxford and California Institute of Technology. This represents a strong similarity between them. However the University of Oxford and Imperial College London has a 0.784 similarity. It means University of Oxford is more like California Institute of Technology than Imperial College London in research preferences.

Next figure analyses the terms used in the scientific production of all universities together. We have chosen a map based on bibliographic data. We used fractional counting method to perform the map.

The minimum number of occurrences of a term is 100. Of 639,854 terms, 1,215 meet the threshold. For each of the 1,215 terms, we calculated a relevance score. Based on this score, the study showed the most relevant terms. The default choice is to select the 60% most relevant terms. Finally, we selected a total of 500 terms.

There are 6 clusters. Red color represents the Cluster 1- Physical sciences with 184 items. Green color represents the Cluster 2- Health sciences with 133 items. The blue color represents the Cluster 3- Social sciences with 85 items. The yellow color is Cluster 4- Life sciences with 79 items. Color pink Cluster 5 with 15 items. Finally the purple color is Cluster 6 with 2 items.

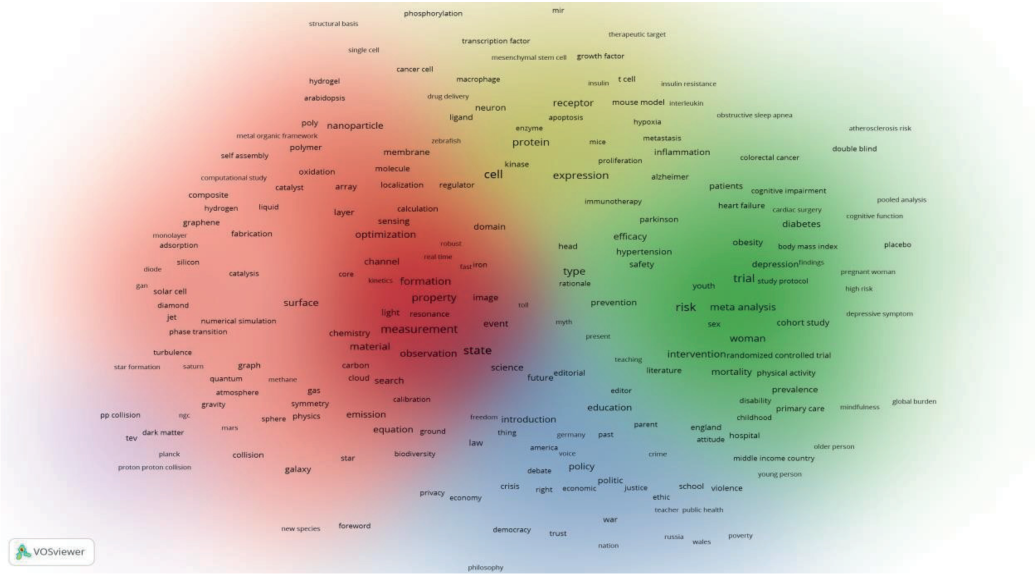


Figure 1. Co-occurrence map based on text data

Risk is the term most used in the production of universities 4,896 occurrences with a relevance of 0.81; Cell is used a number of 4,490 times with a relevance of 0.67; State is used a number of 4,197 with a relevance of 0.29; and Measurement with 3,967 occurrences and 0.82 relevance. Risk, cell, state and measurement could be the same four most common words in the publications of the universities ranked 13-200 in THE.

The next figure analyses the authors in the scientific production of all universities together. We have chosen a map based on bibliographic data for a co-authorship analysis. The fractional counting method was used for this analysis. We have ignored documents with a large number of authors only choosing 10 as the largest number of authors per document. Moreover we selected prolific authors with at least 50 documents and 200 citations. The number of citations of an author equals the total number of citations the documents of the author have received in Scopus.

Of the 450,562 authors, 1,783 meet the thresholds. For each of the 1,783 authors we calculated a total strength of the co-authorship links.

We have applied an author disambiguation process. We checked that the analyzed production belongs to each of the authors and there was no mixed production within the same name.

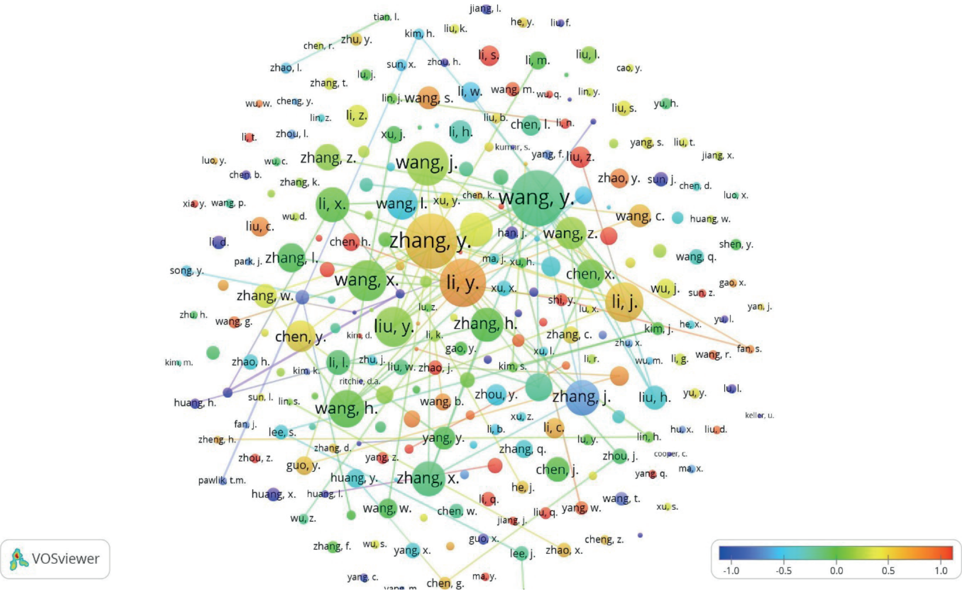


Figure 2. Co-authorship map based on bibliographic data.

There is a total of 16 clusters with 12,219 links and a total link strength of 18,099. We notice a larger amount of Asiatic names appearing the most in the documents of all these institutions.
Cluster 1 obtains the majority of items (38), followed by cluster 2 (28 items), cluster 3 (27 items), cluster 4 (27 items), cluster 5 (26 items), cluster 6 (19 items), cluster 7 (18 items), cluster 8 (17 items), cluster 9 (16 items), cluster 10 (14 items),cluster 11 (7 items), cluster 12 (3 items), cluster 13 (3 items), cluster 14 (2 items), cluster 15 (2 items), cluster 16 (2 items).
Wang, Y is the author with the largest number of documents and links. The second is Zhang, Y.

Are research universities outputs favored by collaboration with these asiatic researchers? or, are they improving the scores of these institutions as a member of staff? These research questions emerged from the results and we will study in future works.

Scientific collaboration among 12 top universities

There are 28,348 documents published in cross-institutional collaboration among those research universities. They published together 13,908 documents in Open Access (articles published in “Gold” OA, including full OA journals, hybrids, open archive and promotional OA). Alternatively, they published 14,440 in other type of access including subscription or green OA.

The next map shows the terms used in the title and abstract of the 28,348 documents in collaboration. We chose binary counting method with a minimum of 20 occurrences. Of the 63,780 terms, 472 meet the threshold. 150 terms selected to appear in the map.

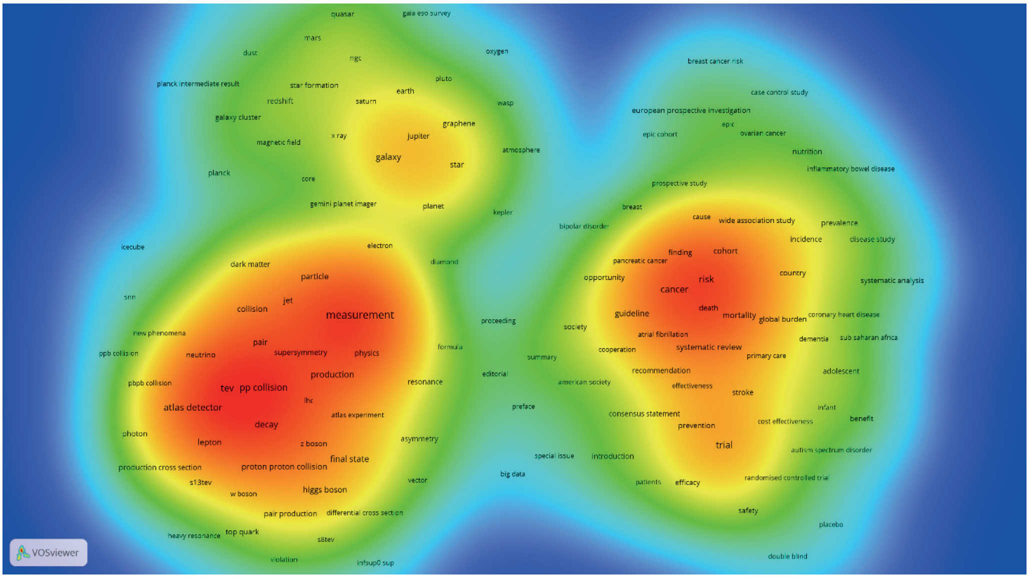


Figure 3. Co-occurrence map based on collaboration production among research universities

There are 7 clusters with 1,408 links and 11,228 link strength. Measurement is the term most repeated with 940 occurrences, a total number of 65 links. Follow by Risk with 542 occurrences and 52 links, Atlas detector with 445 and 45 links and Cancer with 433 occurrences and 45 links.

The fifteen institutions behind these collaborations: University of Oxford (8,071), University of Cambridge (7,300), Massachusetts Institute of Technology (6,963), Harvard University (6,227), Stanford University (6,003), Imperial College London (5,950), California Institute of Technology (4,624), Johns Hopkins University (3,985), University of Chicago (3,974), Princeton University (3,920), Yale University (3,911), ETH Zürich (3,273), CNRS- Centre National de la Recherche Scientifique (2,442), UCL (University College London) (2,435), University of California, Berkeley (2,174).

The funding sponsors that finance these publications: National Science Foundation (NSF) (2,437), National Institutes of Health (NIH) (2,012), U.S. Department of Energy (1,247), European Research Council (1,226), Wellcome Trust (1,097), National Aeronautics and Space Administration (1,071), Science and Technology Facilities Council (973), European Commission (870).

Table 4. Indicators applied to scientific production in collaboration among Research Universities

	Scholarly Output	Citation Count	Citations per Publication	Field-Weighted Citation Impact	Collaboration (%)	Collaboration Impact (%)	Outputs in Top 10 citation percentile (%)	Publications in Top 10 Journal Percentiles (%)
Total	28,348	760,023	27.4	3.82	69.8	30.8	45.4	61.3
Agricultural and Biological Sciences	1,782	37,670	21.1	2.91	74	18.7	41.1	85.4
Arts and Humanities	408	2,604	6.4	2.44	47.3	8.8	13.2	53.3
Biochemistry, Genetics and Molecular Biology	5,157	157,775	30.6	3.5	71.5	32.5	52.4	55.3
Business, Management and Accounting	235	2,907	12.4	3.29	54.5	13.2	28.5	55.7
Chemical Engineering	885	26,982	30.5	3.15	66.9	28.6	53.6	61.4
Chemistry	1,890	49,072	26	2.94	70.2	26.8	54.4	74.7
Computer Science	1,841	28,761	15.6	3.61	64	14.7	21.4	30.7
Decision Sciences	267	3,795	14.2	3.69	55.1	15.5	25.5	46
Dentistry	14	81	5.8	4.47	35.7	8.8	21.4	83.3
Earth and Planetary Sciences	4,313	103,229	23.9	2.79	83.5	26.1	51.5	35.1
Economics, Econometrics and Finance	559	7,519	13.5	3.34	43.8	11.3	33.5	60.8
Energy	417	10,291	24.7	3.35	71.9	27.1	48	58.6
Engineering	2,663	50,225	18.9	3.29	67	20.9	34.5	49.9
Environmental Science	1,149	30,597	26.6	3.87	75.9	23	50	77.5
Health Professions	147	2,666	18.1	3.19	54.4	24.9	37.4	54.2
Immunology and Microbiology	1,186	30,376	25.6	2.94	69.7	26.7	48.4	33.6
Materials Science	2,218	46,981	21.2	3.31	70.7	23.2	41.9	51.9
Mathematics	1,734	17,212	9.9	3.02	62.3	11.4	18.8	33.7
Medicine	6,929	205,172	29.6	5.17	64.6	35.9	45.8	59.3
Multidisciplinary	1,672	103,124	61.7	4.75	73	66.1	67.3	97.9
Neuroscience	1,454	33,846	23.3	2.97	64.7	26.5	49.6	40.6
Nursing	286	5,734	20	3.22	63.6	26	35.3	62.6
Pharmacology, Toxicology and Pharmaceutics	353	5,843	16.6	2.66	64.9	19.9	39.4	50.5
Physics and Astronomy	8,180	219,081	26.8	3.36	81.8	29.1	49.1	42.2
Psychology	608	8,530	14	2.87	50.2	16.4	34.4	52.6
Social Sciences	1,149	14,831	12.9	3.85	48.1	16.5	25.1	60.8
Veterinary	37	299	8.1	2.53	40.5	12.2	24.3	73.5

The scientific production received a total of 760,023 citations. A Field-Weighted Citation Impact of 3.82 and a 61.3% of publications in Top 10 Journal Percentiles. The highest score in citation per publication placed in the field of Biochemistry, Genetics and Molecular Biology (30.6), Chemical Engineering (30.5) and Medicine (29.6). The top Field-Weighted Citation Impact (FWCI) is in the field of Medicine (5.17). The highest percentages of international collaboration are in Earth and Planetary Sciences (83.5), Physics and Astronomy (81.8) and Environmental Science (75.9).

Apart of Multidisciplinary, the fields with highest international collaboration impact are Medicine, Biochemistry, Genetics and Molecular Biology and Physics and Astronomy. The outputs in Top 10 citation percentile are in Chemistry, Chemical Engineering and Biochemistry, Genetics and Molecular Biology. The publications in Top 10 Journal Percentiles are in the field of Agricultural and Biological Sciences, Dentistry and Environmental Science.

Conclusions

The study analyzed research preferences of Top-12 research universities according to the last edition of THE ranking 2019. We have used the Scopus database to extract data and SciVal to apply indicators with a 5-year publication window.
The percentage of international collaboration is 49%. Percentage of publications in the top 10 journal percentiles is 49.65 and the percentage of outputs in top 10 citation percentiles is 48%. Research universities published in the area of Medicine. But others like MIT - Massachusetts Institute of Technology, California Institute of Technology, Princeton University focused their production on Physics and Astronomy.
Cosine revealed the similarity of universities with more analogous research preferences and others. For example California Institute of Technology and ETH Zurich or University of Cambridge with University of Princeton.
Cluster techniques classified outputs into Life Sciences, Physical Sciences, Social Sciences and Health Sciences. The majority of terms from research universities concentrate in the Physical Sciences cluster.
From the co-authorship map have extracted a great number of Asiatic names. This motivates us to investigate if they are responsible of raising scientific production and impact of these universities.

Research universities have 28,348 in common. They are collaborating the most in Physics and Astronomy, Medicine, and Biochemistry. University of Oxford is the most collaborator institution.

Next steps and future works

In future research, we will compare the research patterns of the lower ranked research universities with the Top-12. At the same time, we will analyze the scientific collaboration between the runners-up universities and top universities to know the benefits obtained from this diversity.

We motivate research universities to become more strategic, successful, and quality-oriented in their collaborations.

An emerging recommendation from this study is that governments and private companies such as Fortune 500 companies should do more research investment in universities since it is clear it becomes a profitable asset. The future in rankings will depend more on corporate policies for financing research. It is an imperative need to invest in universities and adopt measures to improve the quality of education and research to become a key player in the higher education sector in the world sphere.

Acknowledgement

The authors acknowledge the kind support given by the University of Leeds and Tecnologico de Monterrey through Intelligent Systems Group to conduct the research reported in this article.

References

Atkinson, R. C., & Blanpied, W. A. (2008). Research universities: Core of the US science and technology system. Technology in Society, 30(1), 30-48.

Carnegie Foundation (2001). The Carnegie Classification of Institutions of Higher Education, Carnegie Foundation, California.

Clark, B. (2004). Sustaining change in universities. Berkshire: SRHE/Open University Press.

Etzkowitz H. & Leydesdorff L. (2001). Universities and the global knowledge economy. Continuum, London, 2001

Fernandez, F., & Baker, D. (2017). Science production in the United States: An unexpected synergy between mass higher education and the super research University. In J. J. W. Powell, D. P. Baker, & F. Fernandez (Eds.), The century of science (international perspectives on education and society) (Vol. 33, pp. 85–111). Bingley: Emerald Publishing Limited.

League of European Research Universities (LERU) (2019) www.leru.org. Accessed 28 January 2019.

Lin, G., Hu, Z., & Hou, H. (2018). Research preferences of the G20 countries: A bibliometrics and visualization analysis. Current Science, 115(8), 1477-1485.

Mohrman, K., Ma, W., & Baker, D. P. (2008). The research university in transition: The emerging global model. Higher Education Policy, 21(1), 5-27.

Powell, J. J., & Dusdal, J. (2017a). Science production in Germany, France, Belgium, and Luxembourg: Comparing the contributions of research universities and institutes to science, technology, engineering, mathematics, and health. Minerva, 55(4), 413–434.

Powell, J. J. W., & Dusdal, J. (2017b). The European Center of science productivity: Research universities and institutes in France, Germany, and the United Kingdom. International Perspectives on Education and Society, 33, 55–83.

SciVal (2019) https://www.scival.com/. Accessed 28 January 2019.

Scopus (2019) https: //www.scopus.com/home.uri. Accessed 28 January 2019.

Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24, 35–43.

Taylor, J (2006). Managing the Unmanageable: The Management of Research in Research-Intensive universities. Higher Education Management and Policy. OECD. 18 (2): 3–4.

University World News (2013). The role of research universities in developing countries. https://www.universityworldnews.com/post.php?story=20130811091502202 Accessed 28 January 2019.

Zhigang, H., Gege, L., & Haiyan, H. (2017). Research preferences of the G20 countries: A bibliometrics and visualization analysis. Paper presented at the ISSI 2017 - 16th International Conference on Scientometrics and Informetrics, Conference Proceedings, 709-720

Measurement variation in bibliometric impact indicators

Stephan Stahlschmidt[1] and Marion Schmidt[2]

[1]*stahlschmidt@dzhw.eu*
Department 2 – Research System and Science Dynamics, German Centre for Higher Education Research and Science Studies (DZHW), Schützenstr. 6a, Berlin, 10117 (Germany)

[2]*schmidt@dzhw.eu*
Department 2 – Research System and Science Dynamics, German Centre for Higher Education Research and Science Studies (DZHW), Schützenstr. 6a, Berlin, 10117 (Germany)

Abstract

The bibliometric measurement process transfers scientific publications and citations into indicators on scientific impact. In defining specific measurement paths researchers hold several degrees of freedom as various methodical decisions are scarcely founded on stringent criteria and their respective implications are not fully understood. These diverse measurement paths result in varying measurements. We propose to compute many possible measurement paths and to analyse the resulting measurement variation. On the one hand, effects of decisions can thus be better understood and on the other hand the resulting measurement variation should be taken into account when using impact values for e.g. funding decisions.

Introduction

Bibliometric indicators result from a measurement process which converts scientific publications and the included references into aggregate values. While the particular selection of citations arises from complex und partially unobserved mechanisms and consequently could be modelled via a stochastic approach, the data generating process of the deterministic bibliometric measurement is perfectly known. The measuring researcher decides upon the process and defines a specific measuring approach, e.g. the choice between Web of Science and Scopus. Figure 1 presents three exemplary measurement decisions, which lead via different measurement paths to eight varying values on the same bibliometric impact indicator.

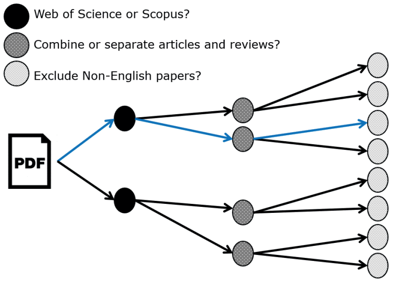

Web of Science or Scopus?
Combine or separate articles and reviews?
Exclude Non-English papers?
PDF

Figure 1: Measurement paths in an exemplified garden of forking paths for a bibliometric analysis.

In defining a particular measurement approach the researcher holds several degrees of freedom, as the implications of each single measurement decision are not fully understood and furthermore the content to be measured, e.g. scientific productivity or impact, consists of latent constructs. The exact extent to which scientific impact is comprehensively covered via citations is disputed (MacRoberts and MacRoberts, 2018), while at the same time citations to an unknown degree occur due to aspects unrelated to scientific impact (Latour & Woolgar, 1986). Citations and scientific impact overlap, but they are not congruent - as any missing Mertonian citation and any existing non-Mertonian citation decrease the signal-to-noise ratio of a citation based measure of scientific impact. Due to the lacking justification of measurement decisions and the lacking clearness about the content to be measured, no – in a statistical sense – true or optimal measurement path might be identified. Instead, several diverging measurement paths through the so-called garden of forking paths (Gelman and Loken, 2014) co-exist. However, most measuring authorities opt for a single measurement path (e.g. CWTS, 2018). In contrast, we propose to embrace this variety and compute several measurement paths. The resulting fuzziness might be understood as a measurement variation. Due to its known and deterministic origin it might not be qualified as a stochastically induced variance, but highlights uncertainness of the measured values caused by the measurement process.

A conceptual model of variation in bibliometrics

The analysis of variation in bibliometrics is predominately discussed in terms of a stochastic variance. In a recent contribution Williams and Bornmann (2016) assume randomness in bibliometric citation counts and propose frequentist statistical inference techniques to quantify its magnitude. For example Bornmann (2017) proposes the estimation of parametric confidence intervals for Journal Impact Factors. Thelwall and Fairclough (2017) extend this approach by proposing a partially randomly determined capacity to produce impactful research observed via publications. Apart from these parametric approaches also non-parametric techniques like bootstrapping (Waltman et al., 2012) and the jackknife (Saḡlam and Friggens, 2018) have been applied. However, these modelling approaches mostly do not account for the data generating process, but rely solely on the observed cross-sectional variation arising in a single measurement path. Furthermore Schneider (2016) argues that frequentist statistical inference on populations is inappropriate. Recently the debate on p-hacking and the reproducible crisis has inspired research to focus on modelling decisions (e.g. Gelman and Loken, 2014; Rohrer et al., 2018). In bibliometrics, the synchronous presentation of different measurement approaches is usually restricted to a limited subset of single parameters, like database coverage (Archambault et al., 2009, Struck et al, 2018), self-citations (Mittermaier et al., 2016) or fractional and whole counting (Waltman et al., 2012).

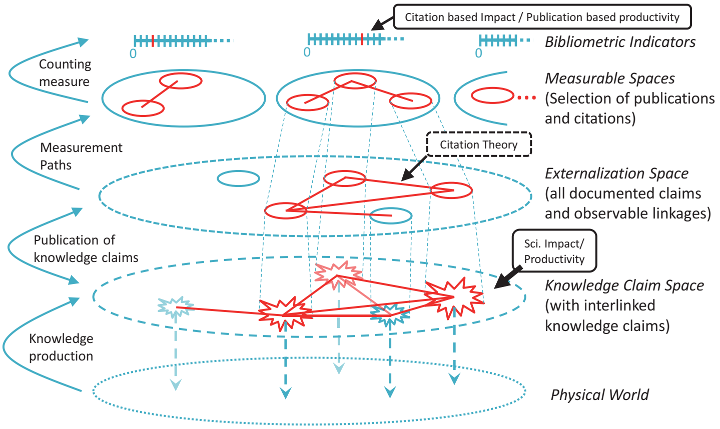

Citation based Impact / Publication based productivity
0
0
0
Bibliometric Indicators
Counting measure
Measurable Spaces (Selection of publications and citations)
Measurement Paths
Citation Theory
Externalization Space (all documented claims and observable linkages)
Publication of knowledge claims
Sci. Impact/ Productivity
Knowledge Claim Space (with interlinked knowledge claims)
Knowledge production
Physical World

Figure 2: Conceptual model of measurement variation

Figure 2 presents a model detailing how such different measurement paths are embedded in the wider mechanism of the science system. We assume several distinct layers which are inter- and intra-related. The bottom layer of the physical world is related to the space of knowledge claims understood as the (debated and extending) knowledge on the physical world. The next layer of the externalization space depicts observable artefacts i.e. documented claims as publications and explicit (e.g. citations) and implicit (e.g. author keywords) links between them.

The topmost layer of the measurable spaces is central to our analysis. The transfer of artefacts from the externalization space to diverging measurable spaces is thought to be governed by diverse measurement paths. As the externalization space is too large to be observed in its entirety, the measurement path extracts a subset of artefacts and consequently renders them accessible to the bibliometric measurement. However, different measurement paths equally co-exist and each path defines a separate although potentially overlapping measurable space. Any such measurable space might be equipped with a counting measure and consequently bibliometric statistics might be calculated. Hence any variation in the measurable spaces consequently causes variation in the values of bibliometric indicators. Given the unknown ground truth of the scientific impact each of these single incarnations of the same indicator maps the respective scientific impact of an entity to a different value constructing en passant separate realities (Desrosières, 1998), upon which funding and policy decision are made.

A quantitative description of measurement variation

For this research-in-progress paper we focus our analysis on eight binary measurement decisions, which result in 256 different measurement paths and values of the same indicator for the same analysed entity. As publications and citations have to be counted separately for every path and are to be supplemented with field-specific expectancy values, we limit our analysis to a computationally feasible random sample of 25% of all potential measurement paths, i.e. we compute 64 parallel bibliometric worlds defined by the respective randomly drawn measurement paths. The following measurement decisions were taken into account: (1) Web of Science or Scopus, (2) include or exclude Non-English publications, (3) combine or separate reviews and articles in normalization, (4) include or exclude self-citations, (5) include or exclude Social Sciences and Humanities (via OECD Fields of Sciences), (6) apply fractional or whole counting to multi-author papers, (7) three-year or five-year citation window and (8) discipline classification by database provider or OECD Disciplines of Sciences.

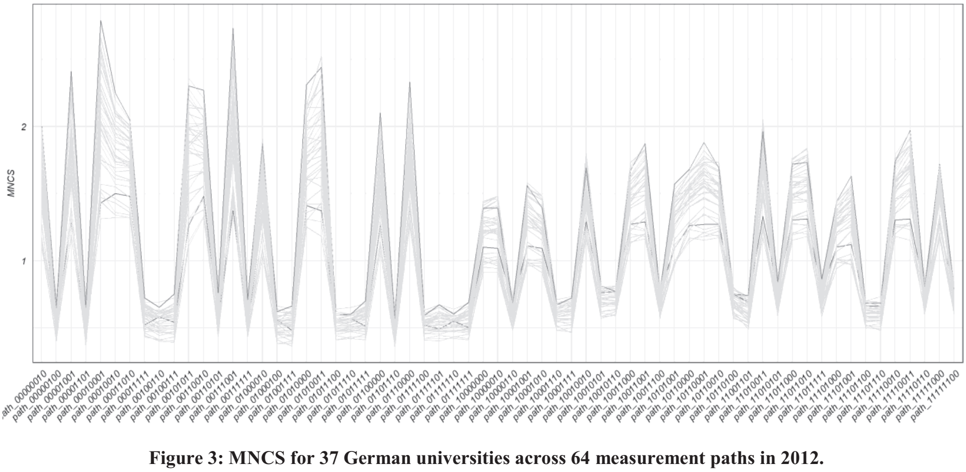


Figure 3: MNCS for 37 German universities across 64 measurement paths in 2012.

For each of these constructed bibliometric realities we compute the *Mean Normalised Citation Score (MNCS)* for every German university. Figure 3 illustrates a preliminary description of the resulting measurement variation. Every line indicates the respective MNCS values (y-axis) across the measurement paths (x-axis) for one of the 37 universities. For example, the

Universität Heidelberg observes a more pronounced variation with values between 0.6 and 2.75, while the equivalent interval for the Freie Universität Berlin starts at 0.5 and ends at 1.5. Given that the value of 1 is commonly understood as the world average, the Universität Heidelberg (Freie Universität Berlin) might find its citation impact in the interval of being 175% (50%) better or 40% (50%) worse than the world average. Consequently the information value on the actual bibliometric impact state seems dubious. Obviously this substantial variation might not be observed if the analysis is limited to a single measurement path. Apart from this institutional perspective on the variation the relative position of these two universities among all other 35 German universities does not vary to a large degree.

In order to gauge the stability across the measurement paths Figure 4 presents the rank-based Spearman correlation matrix between the 37 German universities' MNCS in each of the 64 parallel bibliometric worlds, i.e. every small square colour codes the Spearman correlation in the ranking of the same universities across two measurement paths. Dark blue symbolizes a strong accordance in the ranking of the respective measurement paths, while white denotes no accordance. Negative correlation does not arise.

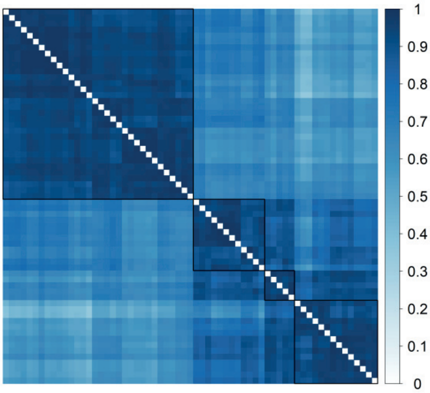


Figure 4: Spearman correlation matrix across the MNCS of 37 German universities resulting from 64 measurement paths.

The correlation matrix has been ordered according to a hierarchical Ward clustering to allow an inspection of the structure causing the aforementioned stability. Four dominant cluster can be identified, the first and biggest one in the upper left corner, the next one in the lower right corner, a third one in the middle and a comparable small one just below on the diagonal, all marked by black lines. The biggest cluster at the upper left corner is constituted by all Spearman correlations between measurement paths based on whole counting, which seem to render greater stability than fractional counting in this preliminary state. Consequently the square to the right (and bottom) of this cluster denotes the rather inconsistent Spearman correlation between whole and fractional counting. On the lower right corner, we find all measurement paths applying fractional counting for the Web of Science, while the equivalent paths for Scopus are subdivided into two clusters. Thus Web of Science seems to be a more stable and probably coherent base less affected by single measurement decisions. The cluster in the middle incorporates measurement paths based on Scopus and fractional counting and the OECD classification, while the smallest cluster holds measurement paths based on Scopus, fractional counting, the Scopus ASJC discipline classification, only English publications and including the social sciences and humanities (SSH). Hence German universities with their rather large corpus of non-English publications in the SSH seem to be uniformly affected by the enlarged database of Scopus in the SSH and Scopus's ASJC discipline classification. These structural observations will be modelled in the next section to infer the direction and size of these effects.

Modelling measurement variation

We model intrinsic values independent of the measurement process by employing a linear mixed model

$$Y_{ij} = \alpha_i + x_{ij}^t\beta + u_{ij}^i\gamma_i + \epsilon_{ij}$$

where

Y_{ij} indicates the *MNCS* of university i corresponding to measurement path j

$i \in [1, \ldots, m]$ denotes the $m = 37$ German universities

$j \in [1, \ldots, n_i]$ states the balanced size of $n_i = 64$ observations per university arising from the diverse measurement paths

α_i denotes the university specific (random) intercept

β describes the vector of fixed effects of the eight binary measurement decisions and

γ_i details the random effects of these measurement decisions on the university i.

Hence the effect of the measurement decision on the universities' MNCS is composed of an overall effect β and an individual effect γ_i allowing for a large degree of flexibility. As we analyse the same 37 German universities in the 64 different measurement paths, we assume the respective MNCS values of the same university to be related throughout all measurement paths and hence obtain a cluster of related MNCS values for every university. The university specific intercept denotes the constant, unaffected part across the observed MNCS values of all measurement paths and accordingly might be understood as the citation-based latent scientific impact irrespective of the variation caused by the measurable spaces of Figure 2.

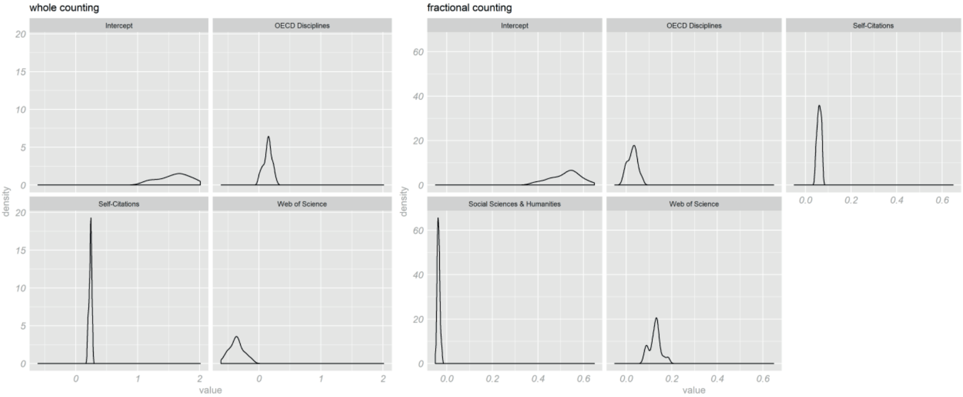


Figure 5: Distribution over German universities of the composed effect of the measurement decisions on the respective MNCS by fractional and whole counting.

The direction and size of the diverse measurement decisions are depicted in Figure 5. Due to fitting issues in this research-in-progress paper we still not present an overall model, but separate models for whole and fractional counting. However, the effect of whole and fractional counting might be inferred from a comparison of the intercepts. Fractional counting reduces the MNCS to a large degree, as observable by the different scales in the x-axis. The application of the broader OECD discipline classification improves the MNCS values slightly, but to a varying degree. Some universities gain more than others from applying the OECD disciplines, possibly due to different disciplinary profiles. The positive effect of including self-citations does not vary to a large degree across universities. However its positive effect allows to infer that

German authors cite themselves more often that the average global author. The inclusion of the SSH reduces the MNCS slightly for fractional counting, while we observe no such effect for whole counting. We also observe no overall effect of non-English papers, separating reviews from articles and the size of the citation window. Although contradictory results for these negligible effects might have been observed in particular measurement paths, they do not hold any overall effect in the multivariate regression framework. The application of Web of Science instead of Scopus has contradictory results, whose reason is still to be investigated.

Outlook

We are currently extending the current analysis by (1) refining the measurement decisions, (2) including further measurement decisions, (3) computing the PP(top10) indicator and (4) drafting a Bayesian model to show how measurement paths in bibliometrics carry considerable consequences for the analysed entities and how "model-based" descriptive statistics might help to alleviate these.

References

Archambault, É. Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the Association for Information Science and Technology*, 60(7), 1320-1326

Bornmann, L. (2017). Confidence intervals for Journal Impact Factors. *Scientometrics*, 111, 1869–1871

CWTS (2018). CWTS Leiden Ranking 2018: Methodology, Leiden: Universiteit Leiden

Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning*. Cambridge: Harvard University Press.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460-465.

Latour, B., & Woolgar, S. (1986). Laboratory Life: The Construction of Scientific Facts (2nd Edition with a New Postscript). London: Sage Publications.

MacRoberts, M.H., MacRoberts, B.R. (2018). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology*, 69, 474–482

Mittermaier, B., Tunger, D., Meier, A., Glänzel, W., Thijs, B. & Chi, P.-S. (2016). Erfassung und Analyse bibliometrischer Indikatoren für den PFI-Monitoringbericht 2017; http://hdl.handle.net/2128/15276

Rohrer, J.M., Egloff, B. & Schmukle, S.C. (2018). Run all the analyses. 51. Kongress der Deutschen Gesellschaft für Psychologie, 15.-20. Sept 2018, Frankfurt.

Sağlam S.Y and Friggens, D. (2018). Cut Your Bootstraps: Use a Jackknife. In: STI 2018 Conference Proceedings (eds. Wouters, P., Costas, R., Franssen, T. and Yegros-Yegros, A.). Leiden: Centre for Science and Technology Studies (CWTS).

Schneider, J. (2016). The imaginarium of statistical inference when data are the population: Comments to Williams and Bornmann, *Journal of Informetrics*, 10, 1243-1248

Struck B., Durning M., Roberge G., & Campbell D. (2018). Modelling the effects of open access, gender and collaboration on citation outcomes: Replicating, expanding and drilling. In: STI 2018 Conference Proceedings (eds. Wouters, P., Costas, R., Franssen, T. and Yegros-Yegros, A.). Leiden: Centre for Science and Technology Studies (CWTS).

Thelwall, M. and Fairclough, R. (2017). The accuracy of confidence intervals for field normalised indicators. *Journal of Informetrics,* 11, 530–540

Waltman, L. , Calero-Medina, C. , Kosten, J. , Noyons, E. C., Tijssen, R. J., Eck, N. J., Leeuwen, T. N., Raan, A. F., Visser, M. S. and Wouters, P. (2012), The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the Association for Information Science and Technology*, 63: 2419-2432.

Williams, R. and Bornmann, L. (2016). Sampling issues in bibliometric analysis. *Journal of Infometrics*, 10(4), 1225-1232.

Changing publication practices in Italy: the case of Social Sciences and Humanities

Antonio Ferrara[1], Carmela Anna Nappi [2] and Francesca Pentassuglio[3]

[1] *antonio.ferrara@anvur.it*
Italian National Agency For The Evaluation Of Universities And Research Institutes
Via Ippolito Nievo, 35 00153 Rome (Italy)

[2] *carmelaanna.nappi@anvur.it*
Italian National Agency For The Evaluation Of Universities And Research Institutes
Via Ippolito Nievo, 35 00153 Rome (Italy)

[3] *francesca.pentassuglio@anvur.it*
Italian National Agency For The Evaluation Of Universities And Research Institutes
Via Ippolito Nievo, 35 00153 Rome (Italy)

Abstract

The paper aims to analyse the major changes in publication practices occurred in Social Sciences and Humanities in the Italian Academia after the implementation and introduction of two new procedures aimed at improving quality of research in Italian academia i.e. a Performance Based Funding System and a new recruitment mechanism for academic staff based on national habilitation. The analysis is carried out by analysing: 1) data on the overall scientific production of professors and researchers holding a permanent position in Italian Universities at a given point in time looking back at their production in the last 15 years (hereinafter "Loginmiur" database); 2) the data collected during the two research assessment exercises carried out by ANVUR in 2012 and 2016 (hereinafter VQR), referred to the period 2004-2014, which represent a specific subset of that production. The main result emerging from the analysis is that the share of journal articles published by scholars in SSH has progressively increased (particularly from 2012 which is the year in which both new procedures were firsly implemented) and, nonetheless, monographs seem to retain a central role in all SSH area. The introduction of the assessment policies seems to have influenced the choice of publishing in those journals defined by ANVUR as "A class". Finally, the score achieved by the research outputs assessed in both VQR exercises prove to be related to the publication's typology.

Introduction

The paper aims to investigate the changes occurred in the publication practices in the SSH after the introduction of two new procedures aimed at fostering research quality in Italian academia i.e. the Research Quality Assessment (hereinafter VQR, Valutazione della Qualità della Ricerca), that is a performance based research funding system (Hicks 2012), and a new recruitment mechanism for academic staff called National Scientific Habilitation (hereinafter ASN) introduced by the Ministry of Education. Our paper wants to understand the response, in terms of publication choices, to the introduction of new rules and incentives in the Italian academia and to the need of meeting new requirements.

Both the measures have been introduced by the Ministry of Education and University (hereinafter MIUR) in 2012.

The Research Quality Assessment exercises (VQR) are aimed at evaluating the research outcomes of state-financed Universities and Research Institutions and at providing an up-to-date assessment of the state of research in the various scientific fields, in order to promote the improvement of research quality in the assessed institutions. Its results are used by MIUR to allocate the merit-based share of the *Fondo di Finanziamento Ordinario* (FFO) – the financing fund of the Italian university system. According to the Law 232/2016, VQR exercises have to be carried out every five years. To the time we are writing, ANVUR has already completed two rounds of assessment, one covering the publication years 2004-2010 (hereinafter VQR1) and the second covering the publication years 2011-2014 (hereinafter VQR2). A new exercise, covering the years 2015-2019, is going to be issued by MIUR by March 2020.

The two VQR exercises adhered to broadly similar rules and methodologies. All research outputs submitted in the SSH areas[i] were assessed using peer review: each research output was assigned to two members of the panel of experts (GEV member). Panel members had to choose independently two external experts (even three in case of disagreement) who assessed the paper following three criteria (relevance, originality and internationalization in the first VQR; originality, methodological rigor and attested or potential impact in the second VQR).

Almost simultaneously to the introduction of PRFS, the Ministry of Education introduced a new hiring procedure for academic professors in which the National Scientific Habilitation (ASN) is a necessary requirement to apply for permanent positions of Full and Associate Professor in Italian Universities.

In order to get the ASN, candidates have to reach minimum values of the indicators of scientific qualification defined by the Ministry of Education (MIUR) and proposed by ANVUR. Indicators of scientific qualification in Social Sciences and Humanites: number of journal articles and book chapters published in the last 10 years; number of journal articles published in A-class journals in the last 15 years; number of books published over the last 15 years of career. In order to calculate such indicators in humanities and social sciences, ANVUR rates journals in two lists (A-class journals and scientific journals).

The hypothesis behind our analysis is that the introduction of Performance Based Funding Systems (PRFS) and in general of procedures that push academics towards research of higher quality, such as ASN procedure, may induce changes (including unintended ones) in the publication practices of the affected scientific communities (Moed, 2008, Jiménez-Contreras, De Moya Anegón, & López-Cózar, 2003). The impact on publication practices may depend (at least in part) on the characteristics of the scientific communities involved.

In the SSH, scholars are often more affected by the national context, even in the selection of research topics; moreover, they tend to publish – besides journal articles – book chapters, books and works belonging to niche typologies (Lariviere *et al*. 2006, Nederhof 2006, Bolton & Kuteeva 2012, Kuteeva & Airey 2013).

Past experience shows that the introduction of research assessment processes has frequently influenced publication practices, particularly by fostering publications in indexed journals (Leite, Mugnaini, & Leta, 2011; Paji, 2015), in addition to increasing the use of English (Li and Flowerdew 2009; Li 2014). In other words, research evaluation tends to steer researchers in SSH towards modes of dissemination (if not production) of knowledge closer to those usually considered as more typical of STEM disciplines.

What emerges from the literature (Aagard 2015) is that the incentives introduced by research evaluation systems have a considerable impact even when the effects on funding are limited, as it happens in Norway (where evaluation results determine 2% of the funding). Therefore, we may expect a greater impact in those countries where these effects are stronger, such as in Italy (where evaluation results determined, in 2018, 19% of the overall ordinary funding of universities).

The existing literature suggests that research evaluation systems actually impact publication practices: Kulczycki *et al*. (2018) highlighted the sharp reduction of the number of monographs in Poland between 2011 and 2014, as a result of the introduction of an evaluation system which discouraged from publishing research outputs other than articles in indexed journals. By comparing the situation of eight EU countries, they found that in the same period the share of monographs and journal articles remained stable in Denmark, Finland, Norway and Flanders (the share of journal articles being much higher than in Italy).

In order to understand the changes occurred in SSH publication practices in Italy, we use two sources of data: 1) the dataset of publications of Italian scholars in service at the beginning of 2018, looking back at their production in the last 15 years (2003-2017); 2) database of research poutputs submitted to the two VQR exercises. The analysis focuses on the publication channels favoured by scholars in SSH, and on the presence (or absence) of the research outputs submitted for evaluation in the main international databases (Web of Science and Scopus). We also try to understand whether the introduction of these new procedures have actually achieved the intended effects i.e. fostering research quality looking at the scores obtained in both VQRs.

Scientific production in SSH area: 2003-2017

The first question we try to address is whether the introduction of research assessment in the Italian university system has had unintended effects, such as that of increasing the overall number of publications, as happened in UK after the introduction of RAEs (Moed 2008, Moya et al. 2015) , and especially whether it has influenced the choices of publication channels: our expectation is that specific publication typologies have become more or less frequent depending on whether or not they have been considered eligible for the evaluation procedures

We present and analyse data regarding research outputs published between 2003 and 2017 by all professors holding a permanent position in Italian universities in February 2018, looking back at the last 15 years of their career. Those data have been extracted from the national database containing all the publications of Italian scholars (henceforth denominated Loginmiur).

Looking at the total number of publications by year, we note that the effect of boosting the number of publication do not happen in Italy. In fact research production have an hump shape that describes the typiucal research life cycle of academics: the number of publuication grows at the beginning of the career, reaches a peak and hence decreases (Figure 1).

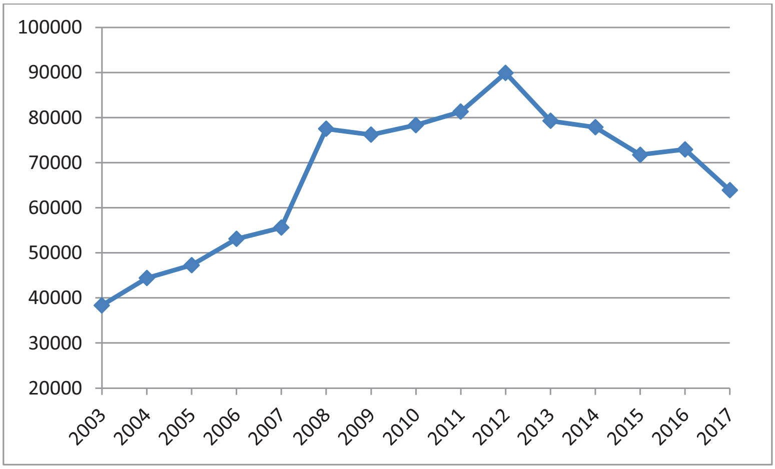


Figure 1. Number of publications by SSH Italian academics in service as of February 2018 per year (2003-2017).

In order to understand the changes occurred in publication practices regarding the research output typologies, we calculated for each year the share of publications in the following typologies: monographs, journal articles, book chapters, conference proceedings and a residual category containing minor and niche publication typologies.

At aggregate level (i.e. for all SSH disciplines, last image of Figure2) data show a significant increase of the percentage of journal articles after 2012, a minor decrease of both monographs and proceedings. The increase of journal articles is particularly marked in the area of Economics and Statistics, where the percentage almost doubled, but is still significant in the other areas (except Architecture and Law). On the other hand, the share of monographs decreased in the 15 years-period in all areas (its percentage decreases from 8% in 2003 to 5% in 2017), and shows a fluctuating trend. A peak in the share of monographs is reached in 2012 in all areas, probably as a result of the first ASN round. Comparing the percentage variation in 2003-2011 with that of 2012-2017, we find a slight reduction of the share of monographs (from 6,2 to 5,6%).

The percentage of book chapters grows in almost all areas (the aggregate percentage goes from 32% to 34%); the highest increases occur in Law and Philology, Literary studies, Art History, while in Economics and Statistics we register a steep decrease from 29% to 21% and in Political and Social Sciences a slight decrease (see Figure 1).

At aggregate level and in all single SSH disciplines, we also see clearly that the share of publications in outlets not relevant for ASN (i.e. not computed in the indicator valid for ASN) sharply decline after 2012.

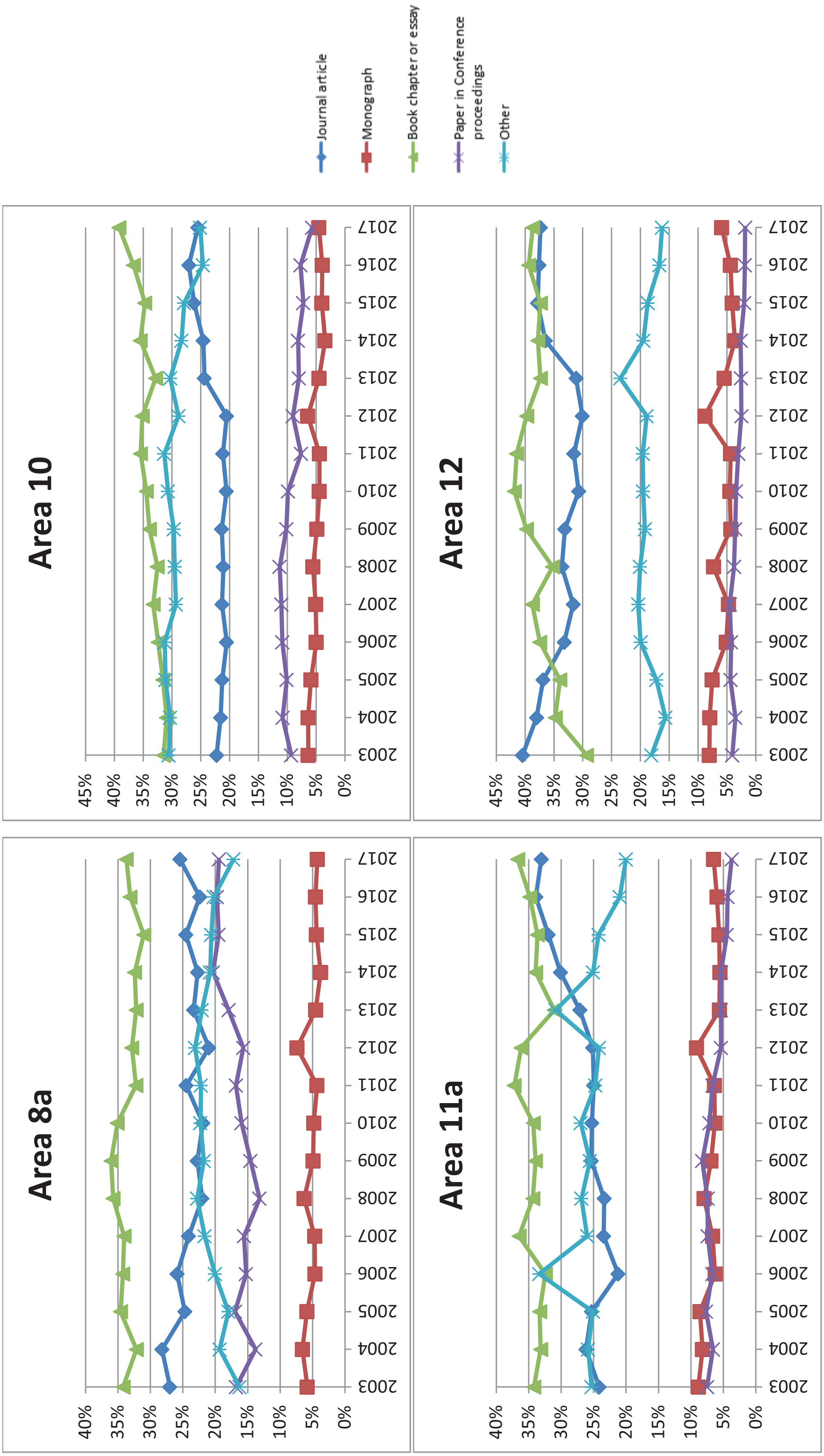
We analyse publication data as aggregated according to the indicators required by the Ministry to access the Habilitation procedure (detailed in the Ministerial Decree n. 120/2016). The habilitation is granted on the basis of an evaluation of the candidate’s titles and publications; in order to have access to this evaluation phase, applicants are required to reach fixed thresholds for at least two out of three indicators of scientific qualification.

The first indicator defined by the Ministry refers to journal articles (published in the so-called “scientific journals”, according to ANVUR classification) and book chapters[ii]. The second indicator refers to articles published in “A class” journals (representing a subset of scientific journals)[iii]. The third indicator refers to books and includes (besides monographs) niche typologies such as critical editions, critical editions of excavations, publication of unedited sources, and so on.

Figure 3 shows the distribution of publication typologies valid for the ASN indicators between 2003 and 2017. In this period, a sharp increase of the share of journal articles and book chapters can be observed: at aggregate level, it grows from 58.2 to 75.9% of the total scientific production, the same trend being reflected in all areas.

If we compare the percentage increase between 2003 and 2011 with that between 2012 and 2017, we find that neither the introduction of the ASN (firstly launched in 2012) nor the announcement of the VQR (whose public Call had been issued in November 2011) significantly accelerated a process which was, as a matter of fact, already ongoing. Considering all areas, indeed, the share of articles and book chapters increases of 11 percentage points in the decade before 2011, with respect to an increase of 6 percentage points in the six-year period after 2012.

However, the introduction of ASN, although not affecting the choice of publishing a journal article, seems to affect the choice of which journal researchers publish in. In fact, looking at the evolution of publication in the so-called “A-class journals” (see again Figure 2), we find a percentage increase (at aggregate level) of 6 percentage points of the share of those kind of publications between 2012 and 2017, with respect to a constant share of publications in class A journals in the period before 2012.This increase is widespread in all SSH area, although it is not homogenous across areas: the increase of the share of A-class journal articles is particularly remarkable in the areas of History, Philosophy and Education (from 8,6 to 14,5%) and Economics and Statistics (from 9,8 to 22,2%), and less so in Architecture (from 5,5% to 8%), Antiquities, philology, literary studies, art history (from 9,7 to 12,3%), and Law (from 17, 9 to 22,4%).


Area 10
Area 12
Area 8a
Area 11a
Journal article
Monograph
Book chapter or essay
Paper in Conference proceedings
Other
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

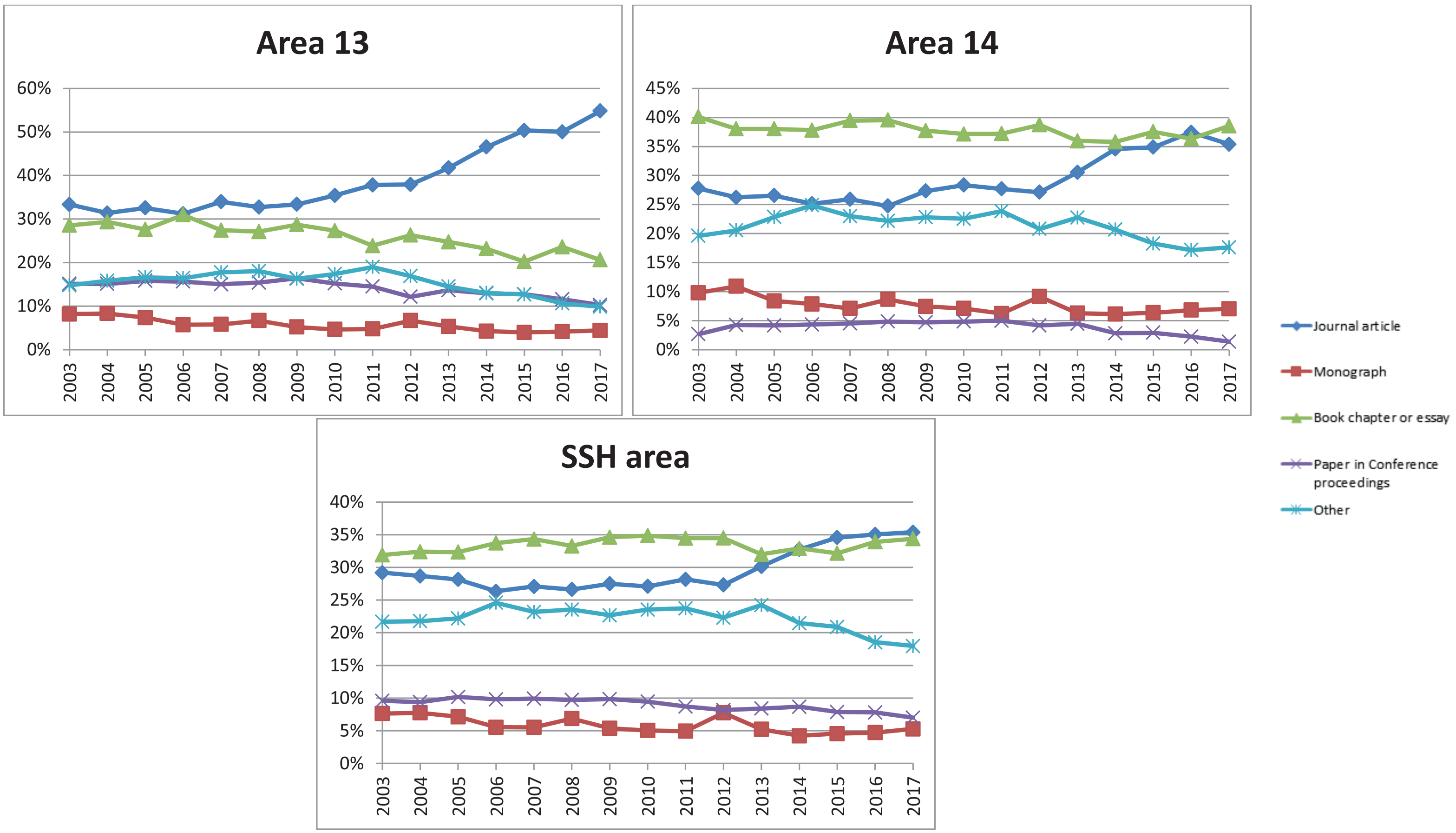


Figure 2. Percentage of publications in SSH (per year, publication typology and subject area)

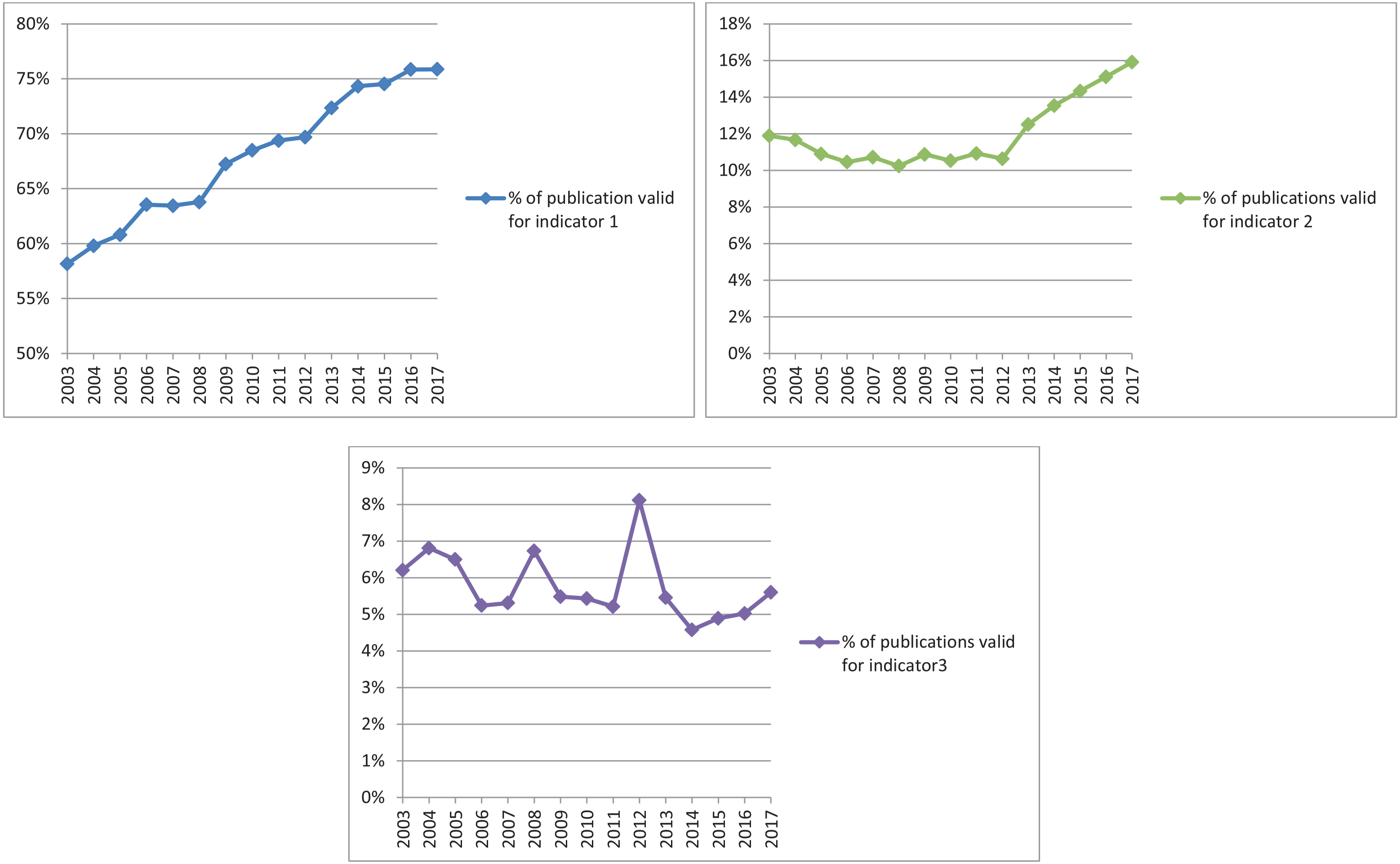


Figure 3. Share of publications in SSH disciplines valid for the three ASN indicators

Typologies of research outputs submitted for VQR1 and VQR2

In this section we look at the changes in publication outlets using VQR data. VQR database represent a subset of the publications taken into account in the previous paragraph. In each assessment exercise, institutions are in fact asked to submit for each affiliate researcher a selection of research outputs (three in VQR1 and two in VQR2) deemed as the best in terms of scientific quality.

We replicate the previous analysis on this subset in order to understand whether the patterns elicited for the entire set of publications persist (or not) once we consider only the best ones, in order to detect possibly new phenomena.

Both in VQR1 and in VQR2, monographs represented a large share of the research outputs submitted by Italian scholars, the percentage being 25% and 20% for VQR1 and VQR2 respectively (see Table 1). The higher percentages emerging from VQR data with respect to Loginmiur data (the share of monographs out of the total scientific production in the same period equal to 6%) may be explained by the fact that, according to the literature, scholarly monograph is deemed to be the most high-profile research output in almost all the SSH disciplines (Thompson 2002; Verleysen-Ossenblok 2017): hence, it is predictable that monographs are overrepresented among outputs selected as the best ones in a given timeframe.

Scholarly monographs, therefore, retain a central role in all SSH areas, except in that of Economics and Statistics, whose publication practices are closer to those of STEM disciplines. According to VQR2 data, the highest share of monographs is to be found in the area of History, Philosophy and Education (27,1%).

Despite the large share of monographs submitted for evaluation, a comparison between VQR1 and VQR2 data shows a decreasing percentage in all SSH areas. The reduction is more significant in the areas of Political and Social Sciences (-10,8%), and of History, Philosophy and Education (-7,1%), faced to Law (-1,9%) and Architecture (-2,5%). The share of book chapters decreased in almost all areas except Philology, Literary Studies, Art History and Political and Social Sciences, where it remains stable.

The share of journal articles submitted for evaluation in both VQRs is much lower in SSH disciplines than in STEM disciplines. In STEM areas, indeed, the percentage of journal articles out of the total of submitted publications is around 90% on average, much higher than in SSH areas where the percentage, at aggregate level, ranges between 34% in VQR 1 and 44% in VQR2. Factors such as the language and the orientation towards a “local” scientific debate may explain this difference. The share of journal articles submitted for evaluation increases between the two evaluations in all areas (+10% on average). Namely, in VQR1 monographs and book chapters are the most recurring types of publication in all SSH areas, except Economics and Statistics; in VQR2 journal articles are, on the contrary, the most widely used publication channel. The highest percentage increases are found in the areas of History, Philosophy and Education (+11,8%) and Political and Social Sciences (+12,7%) and Economics and Statistics (+10,9%). Indeed, the latter area is exceptional as its researchers submitted the highest share of journal articles in both VQR; in VQR1, in particular, this share reaches 62,5%, and is thus much higher than in any other SSH discipline. This is due to the peculiar characteristics of the area, that is closer to the STEM disciplines in terms of publication behaviour. It is very likely that also the assessment criteria chosen by the panel of experts, relying on bibliometric indicators, have been a contributing factor.

Table 1. Sample statistics of publication typologies in VQR1 and VQR2: number, percentage, mean scores and score standard deviations.

Type	*Stats*	*Architecture*		*Philology, Lit. Studies, Art History*		*History, Philosophy, Education*		*Law*		*Economics and Statistics*		*Political and Social Sciences*	
		VQR1	*VQR2*	*VQR1*	*VQR2*	*VQR1*	*VQR2*	*VQR1*	*VQR2*	*VQR1*	*VQR2*	*VQR1*	*VQR2*
journal article	*Nr.*	931	863	3581	2722	2116	2012	3841	3194	7281	5887	1180	1138
	%	17.7	26.3	26.2	32.6	23.3	35.1	32.8	39.4	62.5	73.4	28.3	41
	score	0.52	0.51	0.70	0.61	0.62	0.60	0.59	0.54	0.52	0.62	0.52	0.51
	st. dev.	0.37	0.29	0.32	0.26	0.36	0.27	0.34	0.25	0.43	0.35	0.36	0.28
monograph	*Nr.*	1319	743	3163	1572	3106	1552	3055	1958	1504	635	1447	663
	%	25.1	22.7	23.2	18.8	34.2	27.1	26.1	24.2	12.9	7.9	34.7	23.9
	score	0.58	0.58	0.72	0.71	0.64	0.66	0.63	0.64	0.11	0.23	0.51	0.51
	st. dev.	0.34	0.35	0.32	0.36	0.36	0.32	0.35	0.36	0.28	0.28	0.35	0.31
book chapter	*Nr.*	2022	973	4512	3072	3000	1764	4205	2532	2303	1197	1334	895
	%	38.5	29.7	33.1	36.8	33.1	30.8	36	31.3	19.8	14.9	32	32.3
	score	0.49	0.45	0.69	0.60	0.58	0.56	0.56	0.49	0.14	0.17	0.41	0.42
	st. dev.	0.36	0.27	0.33	0.26	0.35	0.27	0.35	0.25	0.27	0.20	0.35	0.27
conf. proceeding	*Nr.*	594	420	1869	538	594	203	356	121	444	206	77	4
	%	11.3	12.8	13.7	6.4	6.5	3.5	3	1.5	3.8	2.6	1.8	0.1
	score	0.47	0.42	0.67	0.58	0.58	0.50	0.56	0.50	0.08	0.15	0.41	0.63
	st. dev.	0.35	0.25	0.32	0.26	0.37	0.26	0.38	0.27	0.23	0.18	0.45	0.15
other	*Nr.*	388	281	519	450	257	196	239	295	117	100	137	75
	%	7.4	8.6	3.8	5.4	2.8	3.4	2	3.6	1	1.2	3.3	2.7
	score	0.53	0.55	0.63	0.65	0.55	0.61	0.30	0.41	-0.03	0.11	0.47	0.44
	st. dev.	0.37	0.29	0.45	0.38	0.44	0.31	0.56	0.27	0.43	0.24	0.39	0.33

Source: authors' elaboration on ANVUR data

Monographs obtain an higher average score than all other publication typologies (statistically significant at 1%[iv]), in both VQR exercises in the following areas: Architecture; Philology, Literary Studies and Art History; History, Philosophy and Education; Law (see Table 1). Economics and Statistics is again an outlier, as journal articles get on average scores much higher than monographs in both VQRs (0.52 vs 0.11 in VQR1; 0.62 vs 0.23 in VQR2). In Political and Social Sciences the scores obtained by monographs and journal articles are similar (the difference is not statistically significant), yet they are higher than those assigned on average to other publication typologies. We can hence conclude that in almost all the "pure Humanities disciplines" monographs continue to represent the outlet to which researchers entrust the more relevant contributions/findings. Differently, both in Economics and Statistics and in Political and Social Sciences, researchers seem to behave in this respect more similarly to STEM disciplines, choosing journal articles as an important publication outlet for their best research.

Table 2. Quota of journal articles indexed in bibliometric databases submitted for evaluation in VQR1 and VQR2 by area.

Area	*VQR 1*				*VQR 2*			
	Total nr. output	*% journal articles*	*% WoS*	*% Scopus*	*Total nr. output*	*% journal articles*	*% WoS*	*% Scopus*
Architecture	5,254	3.0	1.8	2.8	3,280	10.1	4.9	8.8
Philology, Literary Studies, Art History	13,644	3.3	2.0	2.7	8,354	4.8	2.1	4.2
Histor, Philosophy, Education	9,073	6.0	3.8	5.3	5,727	10.9	7.2	9.9
Law	11,696	0.4	0.2	0.4	8,100	1.0	0.4	0.8
Economics and Statistics	11,649	37.1	30.8	34.5	8,025	58.3	40.3	56.3
Political and Social Sciences	4,175	6.3	8.1	9.6	2,775	15.6	11.6	17.1

Source: authors' elaboration on ANVUR data

As for journal articles, a specific analysis can be carried out with regard to those published in journals indexed in the main international databases (i.e. Clarivate WoS and Scopus). By comparing the data of VQR1 and VQR2 (see Table 2), there is evidence of a considerable increase of the share of publications indexed in WoS and Scopus in all SSH areas. The highest increases are registered in Architecture (+7%), Economics and Statistics (+21%) and in Political and Social Sciences (+9%), while the lowest is reported for Law (+0.5%). This is in line with two general trends observed also elsewhere: a trend to favour journals (including international journals) as publication channel, which determines an higher degree of coverage of publications belonging to SSH in the main databases and a trend of increasing coverage of SSH disciplines journals in the international databases. Therefore, it is not easy to separate the impact of these trends from that produced by the introduction of research evaluation practices, which may have affected the habits of Italian scholars, as happened in other countries (Leite, Mugnaini, & Leta, 2011, Paji, 2015).

VQR data also allow to examine the coverage in the two main databases of the subset of publications selected to be submitted for evaluation in SSH. The degree of coverage varies of course depending on the research field, but the journals where SSH research is published are more likely found in Scopus than in Clarivate WoS (much as it occurs internationally: see Norris & Oppenheim 2007).

The WoS coverage is widest for the area of Economics and Statistics (40,3%), followed by History, Philosophy and Education (7,2%). We find higher percentages of publications submitted for evaluation and indexed in Scopus; again, their share peaks in the areas of Economics and Statistics, Political and Social Sciences , History, Philosophy and Education. In all areas, however, we find an increase of the share of publications indexed in international databases between VQR1 and VQR2.

Conclusions

Our analysis, carried out using two unique databases of research output in SSH, comes to the following conclusions. Italian academics did react to the new procedures in order to meet new requirements. The introduction of the research assessment policies and new hiring procedure turns out to have influenced the choice of publication outlets; more particularly, it increased the share of articles appearing in journals defined by ANVUR as "A-class". Nonetheless, monographs retain a central role in all SSH disciplines except for Economics and Statistics. As a further result of the analysis, it can be noted that the score achieved by the monographs and journal articles in both VQR exercises are higher than those of any other publication channel. Henceforth, we may conclude that researchers in SSH entrust their best research findings to these two types of publication.

References

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Science and Public Polic*y, 42(5), 725-737

Archambault, É. & Lariviere, V. (2010). The limits of bibliometrics for the analysis of the social sciences and humanities literature. In World Social Science Report Knowledge, Knowledge Divides, 251-254. Retrieved from http://unesdoc.unesco.org/images/0018/001883/188333e.pdf.

Bolton, K., & Kuteeva, M. (2012). English as an academic language at a Swedish university: Parallel language use and the 'threat' of English. *Journal of Multilingual and Multicultural Development*, 33(5), 429-447.

Hicks, D. (2004). The four literatures of social science. In H. F.Moed, W. Glänzel, U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 473-496), Dordrecht: Kluwer Academic Publishers.

Hicks, D. (2012). Performance-based university research funding systems. *Research policy*, 41(2), 251-261.

Jiménez-Contreras, E., De Moya Anegón, F., & López-Cózar, E. D. (2003). The evolution of research activity in Spain: The impact of the National Commission for the Evaluation of Research Activity (CNEAI). *Research Policy*, 32, 123-142.

Kulczycki, E., Engels, T. C., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Istenič Starčič, A., & Zuccala, A. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*, 1-24.

Kuteeva, M., & Airey, J. (2014). Disciplinary differences in the use of English in higher education: Reflections on recent language policy developments. *Higher Education*, 67(5), 533-549.

Larivière, V., Archambault, É., Gingras, Y. & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8), 997-1004.

Moya, S., Prior, D., & Rodríguez-Pérez, G. (2015). Performance-based incentives and the behavior of accounting academics: Responding to changes. Accounting Education, 24(3), 208-232.

Moed, H. F. (2008). UK Research Assessment Exercises: Informed judgments on research quality or quantity, *Scientometrics*, 74(1), 153-161.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1), 81-100.

Leite, P., Mugnaini, R. & Leta, J. (2011). A new indicator for international visibility: exploring Brazilian scientific community. *Scientometrics*, 88(1), 311-319.

Li, Y. & Flowerdew, J. (2009). International engagement versus local commitment: Hong Kong academics in the humanities and social sciences writing for publication. *Journal of English for Academic Purposes*, 8(4), 279-293.

Li, Y. (2014). Seeking entry to the North American market: Chinese management academics publishing internationally. *Journal of English for Academic Purposes*, 13(1), 41–52.

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of informetrics*, 1(2), 161-169.

Paji, D. (2015). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice,. *Scientometrics*, 102(3), 2131–2150.

Thompson, J. W. (2002). The death of the scholarly monograph in the humanities, Citation patterns in literary scholarship. *Libri*, 52(3), 121-136.

Verleysen, F. T., & Ossenblok, T. L. (2017). Profiles of monograph authors in the social sciences and humanities: an analysis of productivity, career stage, co-authorship, disciplinary affiliation and gender, based on a regional bibliographic database. *Scientometrics*, 111(3), 1673-1686.

[i] In our analysis we use the definition used in VQR and SSH areas comprises: Architecture (area 8.1 in VQR); Philology, Literature Studies and Art History (area 10 in VQR); History, Philosophy, and Education (area 11.a in VQR); Law (area 12 in VQR); Economics and Statistics (Area 13 in VQR); Political and Social Sciences (area 14in VQR). In VQR, a panel for each of the mentioned areas was appointed.

[ii] The first indicator includes further publication typologies such as: law case notes (provided they are published in journals classified as "scientific"), entries in dictionaries or encyclopaedias, forewords, afterwords, papers conference proceedings (provided they are published in journals classified as "scientific" or associated to an ISBN code).

[iii] The second indicator also includes law case notes (provided they are published in "A-class" journals).

[iv] A T-Tests for pairs of typology shows that monographs always got a statistically significant higher score than any other typology. T-tests results are available on request.

Measuring changes in country scientific profiles: the inertia issue

Wilfriedo Mesheba[1], Egidio Luis Miotti[2], Frédérique Sachwald[3]

[1] wilfriedo.mescheba@hceres.fr
OST-Hcéres, 2 rue Albert Einstein, 75013 Paris (France)
[2] egidio.miotti@hceres.fr
University Sorbonne Paris Cité & OST-Hcéres, 2 rue Albert Einstein, 75013 Paris (France)
[3] frederique.sachwald@hceres.fr
OST-Hcéres, 2 rue Albert Einstein, 75013 Paris (France)

Abstract

The literature comparing the scientific profiles of high-income research-intensive countries and those of developing and emerging countries find a higher degree of inertia in the case of the former. This diagnosis may however be biased by the size of the country and the maturity of the research system. This contribution shows that the volume of publications has a decisive influence on the inertia that is measured. The demonstration is based on the computation of distances between the specialization indexes overtime at the finest level of the Web of Science classification (175 categories, excluding Social Sciences and Humanities) and for 80 countries representing over 95% of world publications. These distances, which measure changes in specialization, prove to be highly sensitive to publication volumes. We therefore propose to cancel out the volume effect, at least partially, by computing a residual inertia. Once the size effect has been neutralized, the evolution measured by the residual distance reveals very different dynamics of the scientific profiles among countries.

Introduction

The scientific profile of a country results from the interactions between the global generation and diffusion of knowledge and the domestic research system. Changes in this profile are driven both by the dynamics of science and by public policies, including the allocation of resources to different disciplines and research issues. Public support for certain specialized institutions or the financing of research programs influence the disciplinary profile of countries.

Over the long term, as an economy develops and investments in research grow, the scientific system matures and diversifies, becoming less strongly specialized in a small number of research fields and disciplines (Archibugi and Pianta, 1991). Emerging countries that have been investing heavily in their scientific systems for several decades are therefore undergoing a process of diversification. Their scientific profile does, however, still remain much more specialized than that of the high-income, research-intensive countries. The evolution of the scientific profile of China over the last few decades illustrates these dynamics. The country has increased its scientific publications very rapidly and has at the same time somewhat reduced its specialization in chemistry. A number of European Union members in a catching-up phase since the 1990s have also experienced a substantial evolution of their scientific profile (Peter and Bruno, 2010; Campbell et al., 2013). Medium to long term bibliometric analyses have pointed to the much more dynamic scientific profiles of emerging countries as opposed to research-intensive, high-income countries (Yang et al. 2012, Radosevic and Yoruk, 2014).

The opposition between dynamic scientific profiles in emerging countries and inert profiles in research intensive countries is at odds with the idea that the latter strive to advance the scientific frontier. Moreover, since the beginning of the century, the development of the scientific system has been recognized as a major policy area in order to promote knowledge-based economies. In this perspective, in many countries research funding has been increasingly based on the prospective impact of research on invention, innovation and the economy more generally. This trend has been reinforced by the emphasis public policies have put on the need for scientific research to contribute to solving societal challenges. Indeed, this objective may lead to specific funding schemes and research programs to tackle multidisciplinary issues.

Scientometrics may not be able yet to clearly evaluate societal impact as the impact of scientific contributions on various sectors of society (Bornmann and Haunschild, 2017), which entails generating new indicators from new data sources. Scientometrics can however measure the capacity of different countries to start new endeavours and refocus their research system on promising fields. In its study of the French innovation system, for example, the OECD (2014) stressed that the scientific profile of France did not seem to respond to the priorities set by public policies. The OECD report measured the "inertia" of the French scientific profile with the cosine distance between 2000 and 2011, on the basis of the Web of Science subject categories. This type of analysis does not take account of the fact that the development of a scientific system generates a certain degree of inertia that is at least partly due to the growth in the volume of publications. The volume of publications is only partially endogenous to the implementation of a scientific development strategy and is also a manifestation of the size of the country (population, GDP, etc.). In other words, the size of a country has a mechanical impact on the apparent inertia of its research system. The role of size has already been noted in studies dealing with countries' scientific profiles on their dynamics (Peter & Bruno 2010; Almeida et al 2009). The impact of size has however not been quantified and there does not exist a measure of the dynamics of scientific profiles correcting for the influence of size.

The contribution of this paper to the literature on the dynamics of scientific profiles is twofold. Based on an aggregate measure of the evolution of country scientific profiles, it first shows how the volume of a country's scientific production impacts the measure of changes in those profiles. It then designs a measure of the dynamics of scientific profiles that corrects for the size bias.

The rest of the paper is organised as follows. Part two goes over the usual indicators for measuring scientific specialization and presents the data. Part three shows the ranking of countries in terms of the dynamism of their scientific profile. Part four demonstrates the influence of the size and complexity of the scientific system on its ability to change over time. It also designs a measure of the dynamics of specialization that corrects for the identified size bias. The conclusion summarises the results and discusses policy implications.

Method and data

The question appears simple at first sight. We need only to quantify the specialization of a country and measure its change between two years that are quite far apart from each other in order to ascertain its degree of inertia or structural change concerning this specialization. In fact, however, the analysis encounters difficulties in providing a precise definition and measuring specialization and its change.

The disciplinary structure of the publications may constitute a first measurement of specialization.

$$PD_{i,t}^{j} = \frac{\#\ of\ publications\ country\ i\ in\ scientific\ area\ j}{\#\ of\ publications\ of\ country\ i}$$

The disciplinary composition of a country's publications does not allow a comparative analysis to be made between countries or with a reference geographic area, however. Some disciplines might represent a large share of publications in all countries, or in certain groups of countries. In this respect, the indicator of this share needs to be normalized,

The specialization index (Activity Index, AI) provides this normalization.

$$AI_{i,t}^{j} = \frac{PD_{i,t}^{j}}{PD_{ref,t}^{j}}$$ [i]

AI is a double ratio and analysing the change in specialization therefore comes up against two problems. First, the fact that we cannot determine the reasons for a variation in a simple manner, as it may be due to variation in one or several components of the formula. Also, changes in the AI are sensitive to the volume of publications of the countries: the higher the world share of a country, the more that country influences the world structure of publications, which mechanically tends to make that country appear more inert than smaller countries.

Measurement of the change in scientific profile

To measure the change in the scientific profile, we need to assess how the distribution (specialization) vectors vary over time. This consists in putting together the information from the initial and final values in order to obtain an indicator of the change between the two vectors. These measurements are broken down into two main groups: measurements of association or similarity and measurements of distance (metric or not) or dissimilarity.

An appropriate compound measurement must possess at least two properties.

1) Invariance to the aggregation level. Dividing a category in two must not modify the value of the measurement when the change in each of the two sub-categories is identical to the change in the initial category.
2) Invariance to the addition of a category of negligible weight. The addition or deletion of a category of a small weight should modify the value of the measurement only marginally.

As well as being easy to use, the Manhattan distance satisfies these requirements due to its linear structure on positive values:

$$D_{i,X}^{Manhattan} = \sum_{j=1}^{n} |X_{i,t}^{j} - X_{i,\tau}^{j}| \qquad with\ t > \tau\ ,\ X_{i,t}^{j} = PD_{i,t}^{j}\ or\ X_{i,t}^{j} = AI_{i,t}^{j}$$

In the specific case of the use of the AI for a country F:

$$D_{F,AI}^{Manhattan} = \sum_{j=1}^{n} |\frac{PD_{F,t}^{j}}{PD_{W,t}^{j}} - \frac{PD_{F,\tau}^{j}}{PD_{W,\tau}^{j}}| \qquad with\ t > \tau\ \ (1)$$

The formula for the final year of the analysis period becomes:

$$\frac{PD_{F,t}^{j}}{PD_{W,t}^{j}} = \frac{PD_{F,\tau}^{j}}{PD_{W,\tau}^{j}} \frac{(1+g_F^j)}{(1+g_W^j)}\ (2)$$

with g_F^j the rate of growth of the share of discipline j for the country and g_W^j the rate of growth of the share of discipline j for the world, we can write

$$D_{F,AI}^{Manhattan} = \sum_{j=1}^{n} |\frac{PD_{F,\tau}^{j}}{PD_{W,\tau}^{j}} \frac{(1+g_F^j)}{(1+g_W^j)} - \frac{PD_{F,\tau}^{j}}{PD_{W,\tau}^{j}}|$$

After simplification:

$$D_{F,AI}^{Manhattan} = \sum_{j=1}^{n} |\frac{PD_{F,\tau}^{j}}{PD_{W,\tau}^{j}} \frac{(g_F^j - g_W^j)}{(1+g_W^j)}|$$

We define:

$$\alpha^j = \frac{(g_F^j - g_W^j)}{(1+g_W^j)}$$

where α^j is the dynamism of the share of discipline j for country F relative to the world.

The distance of the specialization between two dates can therefore be expressed as the product of the AI of the initial year with the relative dynamism coefficient.

$$D_{F,AI}^{Manhattan} = \sum_{j=1}^{n} |\frac{PD_{F,\tau}^{j}}{PD_{W,\tau}^{j}} \alpha^j|\ (3)$$

Computation requires a stable classification of research fields as the appearance of new disciplines distorts the analysis in terms of specialization. This is all the truer when the subject of the new discipline was previously divided between several disciplines in the classification prior to the change. There will therefore be an automatic decrease in the activity in those fields and the appearance of a new specialization without a counterpart in time, with the result being a false structural change in specialization[ii].

Data

Computations were performed on 176 WoS subject categories (SSH are not included). The data correspond to the years 1999/2001 and 2011/2013[iii] for 80 countries representing over 90% of world publications. Fractional counts were used. Complementary sources were used for variables such as the per capita income (purchasing power parity) and population (World Bank and Eurostat for the data for EU-28,). See Appendix 1.

Ranking of countries based on the dynamics of their scientific profile

Table 1 shows that countries with a mature scientific system experience little change in their profile, contrary to countries producing a smaller volume of publications.

Table 1: Manhattan distance for the Activity index between 1999-2001 and 2011-2013

Country	Manhattan Distance	Rank Manhattan
EU28 +	30.21	1
Japan	42.22	2
France	44.17	3
United States	44.84	4
Sweden	45.31	5
Germany	45.72	6
Italy	48.77	7
Switzerland	53.22	8
Belgium	53.45	9
United Kingdom	55.35	10
Israel	57.58	11
Canada	59.41	12
Austria	60.25	13
Russian Federation	61.23	14
Finland	61.66	15
Greece	62.17	16
Netherlands	62.97	17
Australia	64.97	18
China	65.37	19
Hungary	68.91	20
Spain	69.14	21
Denmark	71.38	22
Ireland	72.07	23
Taiwan, Province of China	73.10	24
Mexico	74.88	25
India	75.11	26
Norway	75.74	27
Korea, Republic of	77.21	28
Portugal	78.63	29
Turkey	80.09	30
Czech Republic	88.63	31
Poland	89.94	32
Argentina	89.98	33
Iran (Islamic Republic of)	90.60	34
Egypt	91.19	35
New Zealand	93.55	36
South Africa	94.76	37
Ukraine	98.17	38
Brazil	111.27	39
Algeria	112.70	40
Bulgaria	114.57	41
Slovenia	117.69	42
Thailand	120.15	43
Slovakia	120.88	44
Morocco	122.25	45
Singapore	122.95	46
Saudi Arabia	123.18	47
Tunisia	124.02	48
Chile	127.05	49
Romania	128.23	50
Belarus	128.68	51
Viet Nam	137.86	52
Pakistan	141.85	53
Estonia	152.75	54
Uruguay	156.64	55
Croatia	159.67	56
Cuba	163.96	57
Malaysia	164.55	58
Colombia	166.27	59
Lebanon	170.93	60
Jordan	171.18	61
Iceland	171.36	62
Kenya	180.82	63
Bangladesh	182.82	64
Nigeria	187.54	65
Venezuela, Bolivarian Republic of	190.10	66
Philippines	218.41	67
Senegal	227.47	68
United Arab Emirates	245.50	69
Latvia	247.08	70
Lithuania	248.31	71
Costa Rica	251.38	72
Cyprus	255.83	73
Cameroon	264.30	74
Burkina Faso	272.91	75
Indonesia	293.02	76
Peru	299.06	77
Ivory Coast	301.76	78
Luxembourg	307.02	79
Ecuador	308.48	80
Bolivia, Plurinational State of	510.13	81

High-income countries thus seem to have stable scientific profiles, while emerging countries exhibit more dynamic profiles. This result is similar to the conclusions reached by Radosevic and Yoruk (2014). In order to conclude whether this stability is actually inertia, our next step is to measure the potential impact of size and related characteristics of national scientific systems.

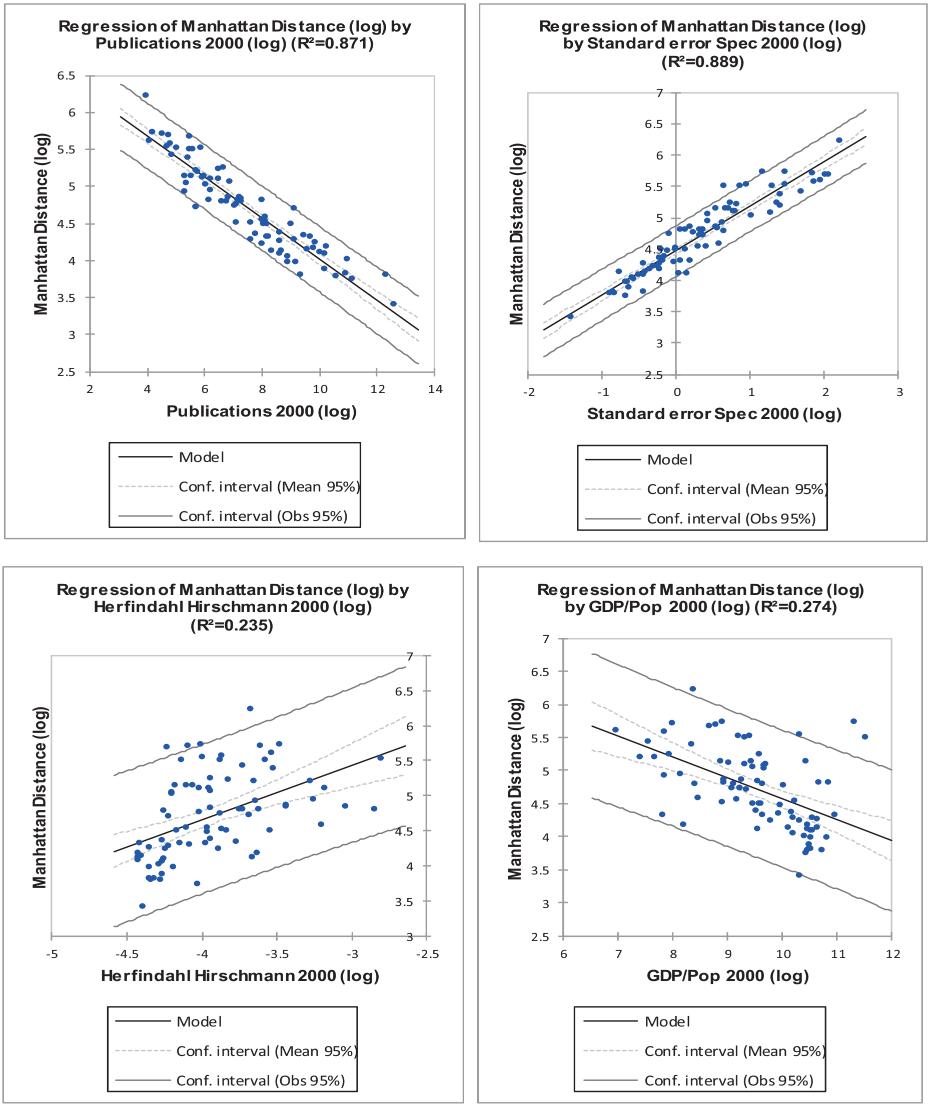


Figure 1: Manhattan distance and characteristics of the scientific systems

Size in terms of volume of publications can impact the Manhattan distance through two channels. The first one is direct: a country that produces a large volume of publications, all other things being equal, requires very significant changes in order to modify its profile substantially. Conversely, in a country with a small volume of publications, even marginal increases in the number of publications will modify its specialization substantially. The second channel is indirect: a country that produces a very large volume of publications covers a broad spectrum of disciplines and has a balanced specialization, whereas a country with a limited number of publications can cover only a restricted range of research fields. In the former case, an increase in the publications in a particular research field changes the specialization only marginally, whereas in the latter case, it might be changed radically.

Moreover, size influences two other characteristics of national scientific systems: the degree of specialization and disciplinary concentration. The degree of overall focus or specialization of countries is measured by the standard deviation of the specialization indicator over the 176 subject categories. High dispersion means that the scientific system is focused on a number of fields, while low dispersion means that the country has a less contrasted profile with similar research capacities in many fields. In high-dispersion systems, a small variation in the number of publications in a non-specialization field can cause substantial change in specialization.

Disciplinary concentration of scientific systems may also be measured. We calculate the Herfindahl-Hirschmann index based on the share of each subject category in total publications in 2000. High concentration is considered as an indicator of contrasted specialization and low concentration as an indicator of diversification.

Finally, the volume of publications itself is not totally exogenous since it tends to increase with the development and maturation of the scientific system. For an approximation of the level of development of the scientific system, we use the level of income per capita on a purchasing power parity basis. Obviously, using this indicator as a proxy of the development of the scientific system is not enough. Per capita income depends, among other things, on the productive structure of the country. For example, the oil-producing countries have high per capita incomes (depending on oil price cycles), but their scientific systems are far from having comparable depth and level of diversification to those of the developed countries.

The impact of the volume and disciplinary diversity of publications on the dynamics of specialization is tested through an analysis of the correlations between the Manhattan distance and four variables: volume of publications in 2000; dispersion of specialization by subject category; disciplinary concentration; GDP per head.

To obtain an overview and taking account of the collinearity between the variables that were used, we performed a Principal Component Analysis (PCA) followed by a typology based on Agglomerative Hierarchical Clustering (AHC).

PCA results

The PCA shows a single highly significant axis (67.7% of the total variability), thereby demonstrating strong collinearity between the variables chosen (Table 2 and Graph 2).

Table 2: PCA, Eigenvalues and correlations factors and variables

	F1	F2	F3	F4
Eigenvalue	2.705	0.701	0.465	0.129
Variability (%)	67.632	17.523	11.618	3.227
Cumulative %	67.632	85.155	96.773	100.000
Correlations				
Publications 2000 (log)	-0.875	0.387	-0.167	0.239
Standard error Spec 2000 (log)	0.904	-0.335	0.027	0.263
GDP/Pop 2000 (log)	-0.790	-0.273	0.547	0.050
Herfindahl Hirschmann 2000 (log)	0.706	0.604	0.370	0.015

The F1 axis opposes countries with high-income and a significant number of publications, and those with a very marked specialization profile (high specialization standard deviations) and disciplinary concentration.

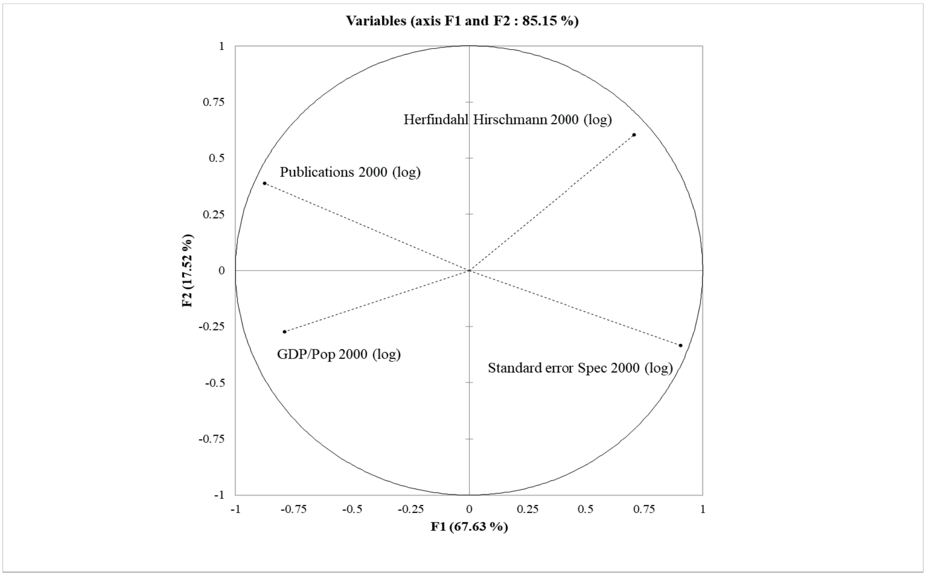


Figure 2: Circle of correlations

AHC results

The AHC produces four clearly differentiated groups (Table 3)

Table 3: AHC, a four-group typology

1	2	3	4
Germany	Argentina	South Africa	Saudi Arabia
Australia	Brazil	Algeria	Bangladesh
Austria	China	Bulgaria	Belarus
Belgium	Korea, Republic of	Cyprus	Bolivia, Plurinational State of
Canada	Hungary	Colombia	Burkina Faso
Denmark	India	Croatia	Cameroon
Spain	Iran (Islamic Republic of)	Egypt	Chile
United States	Mexico	United Arab Emirates	Costa Rica
Finland	New Zealand	Estonia	Ivory Coast
France	Poland	Iceland	Cuba
Greece	Russian Federation	Jordan	Ecuador
Ireland	Singapore	Lebanon	Indonesia
Israel	Slovenia	Lithuania	Kenya
Italy	Taiwan, Province of China	Luxembourg	Latvia
Japan	Czech Republic	Malaysia	Morocco
Norway	Turkey	Slovakia	Nigeria
Netherlands		Thailand	Pakistan
Portugal		Tunisia	Peru
United Kingdom		Uruguay	Philippines
Sweden		Venezuela, Bolivarian Republic of	Romania
Switzerland			Senegal
EU28 +			Ukraine
			Viet Nam

Table 4 presents the median values of the variables used in the PCA, according to the typology groups from the AHC. We note that Group 1 is made up of high-income economies and presents the highest median volume of publications, the lowest standard deviation of specialization, the lowest disciplinary concentration and the highest level of per capita GDP. At the other extreme, Group 4 is made up of developing economies which produce a small number of publications, have a very high standard deviation of specialization and disciplinary concentration, and a much lower level of development. The median Manhattan distance increases from group 1 to group

4: countries endowed with a diversified scientific system are those in which structural change measured by the Manhattan distance is the smallest.

Table 4: Medians* of the population by AHC group

	All	Group 1	Group 2	Group 3	Group 4
Manhattan (Distance)	**114.57**	58.49	84.36	162.11	187.54
# publications in 2000	**1 403.79**	12 867.19	4 653.92	406.03	264.31
Standart Error Specialization 2000	**1.33**	0.58	0.97	1.93	3.90
Herfindahl-Hirchmann Index	**0.01871**	0.01320	0.01921	0.01866	0.02930
GDP/Pop 2000	**14 732.48**	36 650.19	15 291.34	12 703.29	4 483.81

**: Preference was given to the median over the mean in order to avoid any bias due to extreme values. For example, Saudi Arabia is in group 4 despite a GDP comparable to that of developed economies.*

A normalisation of structural change

We build on the correlation observed in Table 4 to "normalise" the measure of structural change. We use the residuals from an equation that expresses the Manhattan Distance (MD) as a function of the volume of publications or characteristics of the scientific system as a structural change stripped of the volume effect.

$$MD_j = \alpha + \beta X_j + \mu_j$$
$$\mu_j = MD_j - \alpha - \beta X_j$$

The result of the estimation is presented in Table 5.

Table 5: Robust OLS of the Manhattan distance on volume

	Manhattan distance (log)
# Publications (fract) in 2000 (log)	-0.278***
	(0.0129)
Constant	6.795***
	(0.0999)
Observations	81
R-squared	0.871

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6 presents the ranks of the countries by increasing Manhattan distance and their ranks calculated using the residuals of equation in Table 5. Table 6 also classifies countries according to the extent of their change in specialization based on a univariate partitioning into 5 groups using the Fisher algorithm. It shows that the correction for size has a major impact on the evaluation of the dynamism of country specialization over the period.

Structural change is the greatest in Group 5. The United States stands out: it was initially ranked 4th (very strong inertia), while it ranks 79th after correction of the volume effect (strong structural change). The impact of the normalization is lower for the EU28: it was initially ranked 1st and comes in 58th after correction, joining the Moderate inertia group. Evolution is similar for Japan: it ranked 2nd and is 47th after correction, also in the moderate inertia group.

The disciplinary profile of a number of small high-income countries appears less dynamic after correction for size: Ireland, Norway. Other small high-income countries have quite inert profiles by both measures: Sweden, Belgium, Austria. France was initially ranked 3rd and is ranked 29th after correction, in the Strong inertia group. Italy experiences a similar change. Germany, which was initially close to France (rank 6), experiences a much larger change to rank 50 in the

Moderate inertia group. Spain and the UK are in the Moderate change group, like China. The UK experiences a similar very substantial change (from rank 10 to 75) but less dramatic than that of the US. At the opposite, after correction for size, a number of developing countries appear among the most inert ones, like Algeria, Vietnam or Iran.

Table 6: Manhattan Distance Calculated and Corrected.

Country	Rank Manhattan before correction	Rank Manhattan after correction	Group
Algeria	40	1	1 Very strong inertia
Ireland	23	2	
Viet Nam	52	3	
Sweden	5	4	
Greece	16	5	
Belgium	9	6	
Hungary	20	7	
Iran (Islamic Republic of)	34	8	
Austria	13	9	
Israel	11	10	
Finland	15	11	
Tunisia	48	12	
Switzerland	8	13	
Portugal	29	14	
Uruguay	55	15	
Norway	27	16	
Mexico	25	17	
Iceland	62	18	
Egypt	35	19	
Morocco	45	20	
Denmark	22	21	2 Strong inertia
Thailand	43	22	
Lebanon	60	23	
Pakistan	53	24	
Bulgaria	41	25	
Estonia	54	26	
Italy	7	27	
Czech Republic	31	28	
France	3	29	
Burkina Faso	75	30	
Slovenia	42	31	
Belarus	51	32	
Colombia	59	33	
Argentina	33	34	
Turkey	30	35	
South Africa	37	36	
Senegal	68	37	
Bangladesh	64	38	
Saudi Arabia	47	39	
Kenya	63	40	
New Zealand	36	41	
Netherlands	17	42	
Jordan	61	43	
Slovakia	44	44	
Cuba	57	45	
Taiwan, Province of China	24	46	3 Moderate inertia
Japan	2	47	
Ukraine	38	48	
Cyprus	73	49	
Germany	6	50	
Chile	49	51	
Romania	50	52	
Australia	18	53	
Philippines	67	54	
Luxembourg	79	55	
Russian Federation	14	56	
Ecuador	80	57	
EU28 +	1	58	
Canada	12	59	
Cameroon	74	60	
Malaysia	58	61	
Costa Rica	72	62	
Ivory Coast	78	63	4 Moderate change
Korea, Republic of	28	64	
Spain	21	65	
Croatia	56	66	
Poland	32	67	
India	26	68	
Peru	77	69	
China	19	70	
United Arab Emirates	69	71	
Nigeria	65	72	
Singapore	46	73	
Latvia	70	74	
United Kingdom	10	75	
Venezuela, Bolivarian Republic of	66	76	
Lithuania	71	77	5 Strong structural change
Indonesia	76	78	
United States	4	79	
Brazil	39	80	
Bolivia, Plurinational State of	81	81	

Robustness

A change in the distance measure marginally modifies the correlations with the volume effect and our result. The Pearson correlation between the Manhattan distance and other distance measures (Canberra, Euclidean, Bray and Curtis, Khi^2, Cosines), particularly with the Euclidean distance, is very high and significant. Although correlation is less strong with the Cosines measure (similarity), the Spearman rank correlation is also very high. As a result, modifying the distance measure does not change the ranks of countries significantly. Concerning the volume, the Pearson correlation is very high and significant in all cases. A sensitivity analysis has been conducted (not reproduced); all correlations are significant at 1% level[iv].

Conclusions and further research

Designing an indicator of the evolution of national scientific profiles comes up against methodological difficulties. The indicator of changes in specialization is highly influenced by the structural characteristics of scientific systems, such as the volume of publications and disciplinary diversification.

Measuring changes in the specialization index in order to evaluate the dynamism of a research system or the impact of specific policies or research strategies is therefore not appropriate. This contribution has shown that it is necessary to implement a normalization procedure. The normalization that is proposed constitutes a reasonable approximation of the measurement of the medium to long term evolution of specialization by eliminating the size effect.

Country rankings are quite different when the specialization index is normalized. In particular, a number of research-intensive high-income countries show a substantially more flexible scientific profile. The normalized measure does on the contrary confirm that some high-income countries have a quite inert profile. After correction, developing countries have quite inert scientific profiles, while large emerging countries appear substantially more dynamic. Further research will be necessary to explain these differences in the impact of normalization. One objective would be to discuss the ability of countries to implement scientific strategies or to be able to develop emerging research areas.

References

Addinsoft (2019). XLSTAT statistical and data analysis solution. Long Island, NY, USA. https://www.xlstat.com.

Almeida, J. A. S., Pais, A. A. C. C., & Formosinho, S. J. (2009). Science indicators and science patterns in Europe. Journal of Informetrics, 3(2), 134-142.

Balassa, B. (1965). Trade liberalization and ‘revealed’ comparative advantage. The Manchester School of Economics and Social Studies, 32(2), 99–123

Bongioanni, I., Daraio, C., & Ruocco, G. (2014). A quantitative measure to compare the disciplinary profiles of research systems and their evolution over time. Journal of Informetrics, 8(3), 710-727.

Bornmann, L. & R. Haunschild (2017). Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? Scientometrics, 110:937–943

Frame, J. (1977), Mainstream research in Latin America and the Caribbean, Interciencia, 2: 143–148

Glänzel W., Debackere K. & Meyer M. (2008). ‘Triad’ or ‘tetrad’? On global changes in a dynamic world. Scientometrics, 74, 71-88.

Liang, L. M., Havemann, F., Heinz, M., & Wagner-Dobler, R. (2006). Structural similarities between science growth dynamics in China and in western countries. Scientometrics, 66(2), 311–325.

Peter, V. & N. Bruno, (2010). International Science & Technology Specialisation: Where does Europe stand? European Union Studies and reports. EUR 24198. ISBN 978-92-79-14285-7.

OECD (2014), Examens de l’OCDE des politiques d’innovation – France.

Radosevic S. & Yoruk E. (2014). Are there global shifts in the world science base? Analysing the catching up and falling behind of world regions. Scientometrics, 101:1897–1924.

Pianta, M. & Archibugi, D. (1991). Specialization and size of scientific activities: A bibliometric analysis of advanced countries, Scientometrics, 22, 341-358.

Rousseau R. (2018). The F-measure for Research Priority. Journal of Data and Information Science Vol. 3 No. 1, 2018 pp 1–18. DOI: 10.2478/jdis-2018-0001.

Yang, L. Y., Yue, T., Ding, J. L., & Han, T. (2012). A comparison of disciplinary structure in science between the G7 and the BRIC countries by bibliometric methods. Scientometrics, 93, 497–516.

Appendix 1: Countries and data

Country	Manhattan distance	# Publications in 2000	Standard Error Specialisation 2000	Herfindahl-Hirschmann Index 2000	Level of income	GDP/Pop (PPP) 2000
South Africa	94.762	3015	1.495	0.01900	Upper middle income	9539
Algeria	112.701	294	1.418	0.02514	Upper middle income	10211
Germany	45.718	54269	0.423	0.01336	High income	36765
Saudi Arabia	123.179	1251	1.370	0.05764	High income	43071
Argentina	89.982	3540	1.123	0.01551	Upper middle income	14900
Australia	64.968	17467	0.697	0.01195	High income	35378
Austria	60.246	5492	0.599	0.01419	High income	38844
Bangladesh	182.818	292	2.256	0.02597	Lower middle & Low income	1642
Belarus	128.680	903	1.687	0.03212	Upper middle income	7563
Belgium	53.452	7238	0.512	0.01291	High income	37189
Bolivia, Plurinational State of	510.125	52	9.004	0.02539	Lower middle & Low income	4412
Brazil	111.271	9021	1.361	0.01472	Upper middle income	11371
Bulgaria	114.567	1127	0.914	0.02077	Upper middle income	8833
Burkina Faso	272.910	58	7.020	0.02930	Lower middle & Low income	1075
Cameroon	264.303	122	6.392	0.02100	Lower middle & Low income	2604
Canada	59.412	25962	0.645	0.01202	High income	37432
Chile	127.054	1404	1.761	0.03217	High income	14315
China	65.365	27513	0.800	0.02638	Upper middle income	3701
Cyprus	255.834	109	2.553	0.01833	High income	30086
Colombia	166.265	385	2.186	0.01808	Upper middle income	8308
Korea, Republic of	77.213	12543	0.837	0.02295	High income	20757
Costa Rica	251.383	153	4.368	0.06032	Upper middle income	9878
Ivory Coast	301.759	92	6.343	0.02705	Lower middle & Low income	2948
Croatia	159.667	961	3.597	0.01937	High income	15742
Cuba	163.961	484	2.201	0.04140	Upper middle income	11154
Denmark	71.384	5643	0.636	0.01289	High income	42338
Egypt	91.191	2001	1.001	0.02093	Lower middle & Low income	7388
United Arab Emirates	245.498	242	3.667	0.01595	High income	102635
Ecuador	308.483	66	4.361	0.03062	Upper middle income	7388
Spain	69.143	19156	0.779	0.01431	High income	29967
Estonia	152.755	413	2.732	0.01505	High income	15703
United States	44.842	220396	0.436	0.01305	High income	45986
Finland	61.660	5782	1.034	0.01198	High income	34887
France	44.171	39354	0.403	0.01384	High income	34881
Greece	62.173	4059	0.657	0.01217	High income	24839
Hungary	68.910	3010	0.727	0.02054	High income	17952
India	75.112	16192	1.048	0.01871	Lower middle & Low income	2495
Indonesia	293.020	242	7.811	0.01457	Lower middle & Low income	5806
Iran (Islamic Republic of)	90.595	1196	0.986	0.02875	Upper middle income	13135
Ireland	72.068	2055	0.791	0.01468	High income	38850
Iceland	171.364	198	1.723	0.01651	High income	34045
Israel	57.575	7332	0.547	0.01404	High income	26825
Italy	48.769	27110	0.531	0.01405	High income	36536
Japan	42.216	68116	0.510	0.01773	High income	33872
Jordan	171.177	399	2.048	0.01729	Upper middle income	7235
Kenya	180.819	319	4.032	0.03754	Lower middle & Low income	2132
Latvia	247.078	264	2.380	0.02796	High income	11175
Lebanon	170.926	252	1.951	0.01527	Upper middle income	12532
Lithuania	248.310	359	1.914	0.02077	High income	12190
Luxembourg	307.024	67	3.186	0.01824	High income	81690
Malaysia	164.554	675	2.152	0.01912	Upper middle income	16310
Morocco	122.248	742	1.123	0.02402	Lower middle & Low income	4484
Mexico	74.882	3833	0.841	0.01589	Upper middle income	15683
Nigeria	187.544	652	3.898	0.02178	Lower middle & Low income	2848
Norway	75.743	3576	1.204	0.01204	High income	58045
New Zealand	93.547	3428	1.334	0.01656	High income	27620
Pakistan	141.852	502	1.521	0.03867	Lower middle & Low income	3495
Netherlands	62.970	14227	0.464	0.01211	High income	41722
Peru	299.056	115	7.447	0.01672	Upper middle income	6563
Philippines	218.411	227	4.109	0.02946	Lower middle & Low income	4224
Poland	89.938	8096	0.810	0.02157	High income	14732
Portugal	78.630	2427	0.801	0.01402	High income	25999
Romania	128.230	1398	1.192	0.04771	Upper middle income	10471
United Kingdom	55.352	57994	0.561	0.01370	High income	33266
Russian Federation	61.226	22253	1.146	0.02573	Upper middle income	14051
Senegal	227.466	132	5.432	0.02408	Lower middle & Low income	1911
Singapore	122.946	3036	1.436	0.02667	High income	51706
Slovakia	120.885	1409	1.417	0.02353	High income	15605
Slovenia	117.695	1201	1.269	0.01793	High income	22723
Sweden	45.314	11507	0.639	0.01296	High income	36855
Switzerland	53.216	9870	0.509	0.01509	High income	50776
Taiwan, Province of China	73.099	9230	0.961	0.01695	High income	27172
Czech Republic	88.631	3217	0.888	0.01907	High income	21194
Thailand	120.148	890	1.908	0.01415	Upper middle income	9189
Tunisia	124.017	471	1.053	0.01933	Lower middle & Low income	7574
Turkey	80.086	5474	0.846	0.01936	Upper middle income	13862
EU28 +	30.211	291197	0.239	0.01231	High income	30326
Ukraine	98.170	3323	1.782	0.04083	Lower middle & Low income	4797
Uruguay	156.644	210	1.527	0.01506	High income	12875
Venezuela, Bolivarian Republic of	190.096	760	2.088	0.01933	Upper middle income	14413
Viet Nam	137.863	208	1.847	0.02608	Lower middle & Low income	2562

[i] The Activity Index (AI), originally introduced in bibliometrics by Frame (1977) was constructed by Balassa (1965) for the analysis of the trade specialization (also known as Revealed Comparative Advantage indicator).

[ii] The most striking example is Nanoscience & Nanotechnology subject category in the WoS. It did not exist in 2000 and represented 0.69% of world publications in 2012. To eliminate the bias introduced by the appearance of this discipline, it would have been necessary to retropolate the nomenclature. We decided to accept this bias that particularly affects large specialized countries in 2012.

[iii] When it is mentioned 2000 (2012) in the text it is necessary to understand the average of the years 1999-2001 (2011-2013).

[iv] Rousseau (2018) proposed an alternative indicator, the F-measure. However, it actually presents similar problems with respect to the influence of publication volume (Pearson correlation is strongly significant, at 0.83, with $p<0.001$).

Highly cited references in PLOS ONE and their in-text usage over time

Wolfgang Otto[1], Behnam Ghavimi[1], Philipp Mayr[1], Rajesh Piryani[2], Vivek Kumar Singh[3]

[1] *{firstname.lastname}@gesis.org*
GESIS - Leibniz Institute for the Social Sciences, 50667 Cologne (Germany)

[2] *rajesh.piryani@gmail.com*
Department of Computer Science, South Asian University, New Delhi (India)

[3] *vivekks12@gmail.com*
Department of Computer Science, Banaras Hindu University, Varanasi (India)

Abstract

In this article, we describe highly cited publications in a PLOS ONE full-text corpus. For these publications, we analyse the citation contexts concerning their position in the text and their age at the time of citing. By selecting the perspective of highly cited papers, we can distinguish them based on the context during citation even if we do not have any other information source or metrics. We describe the top cited references based on how, when and in which context they are cited. The focus of this study is on a time perspective to explain the nature of the reception of highly cited papers. We have found that these references are distinguishable by the IMRaD sections of their citation. And further, we can show that the section usage of highly cited papers is time-dependent: the longer the citation interval, the higher the probability that a reference is cited in a method section.

Introduction

Scientific publications are highly structured texts which incorporate specific properties related to their references. The accessibility of full-text publications has broadened up the possibilities for analysing citation behavior and the usage of referenced publications in bibliometrics. The objectives of this research-in-progress paper are highly cited PubMed papers in a corpus of the open access journal PLOS ONE[1] during the period between 2006 and 2017. As a highly cited PubMed paper, all papers (n=666) are taken into account which are cited in more than 100 PLOS ONE papers of our corpus. We call these highly cited papers *top-666*.

The main goal is to describe the highly cited papers based on extracted information from our corpus; this includes metadata of citing publications and citation contexts. So the information we gather is not from the full-text of the referenced publication, but the citing publications. On the paper level, we use the year of publication of the citing publication. On citation context level we use information about the sections based on the IMRaD scheme[2] which is "the most used standard of today's scientific discourse" (Sollaci & Pereira, 2004). As a second distinguishing feature, we examine the co-citation count on the citation context level. With the latter information, it is possible to investigate how a paper is perceived over time based on the citation interval. The citation interval is the time distance of the citing paper to the time of publication of the reference. With this information, we are able to describe the top cited references based on how, when and in which context they are cited. We use this information to describe the "citation history" of a specific cited object over time.

Having a deeper understanding and methodology for the usage of gathered information from citations based on full text is, on the one hand, helpful to understand the temporal citation

[1] https://journals.plos.org/plosone/

[2] i.e. Introduction, Method, Results, and Discussion

patterns of references, on the other hand, the information can be used to build better tools for information retrieval systems for suggesting related research literature (e.g. Huang et al., 2015).

In our study we will address the following research questions:

- RQ 1: In which IMRaD section are the top-666 references cited?
- RQ 2: How is the proportion of concrete IMRaD sections evolving over time?
- RQ 3: Is the number of co-citations on citation context level declining when the citation interval is becoming longer?

Related Work

Citations are an important parameter of connectivity of related research works. A lot of studies have focused on analysing citations for different purposes ranging from assessment of the quality of an article to tracing the flow of ideas on a topic. Sugiyama et al. (2010) have suggested that there could be two kinds of citation analysis: (1) Citation Counts and (2) Citation Context Analysis. They argue that citation context analysis could be a better approach to determine the influence of a research article. Unlike simple counts, citation context analysis identifies the contextual relationship between citing research articles and referenced articles by applying various NLP and Machine Learning approaches (Hernández-Alvarez & Gómez, 2016). It processes the text of articles, particularly that portion where it cites another article. This is called citation context, i.e. the sentence where a specific reference is cited. Relatively few studies have been carried out on citation context analysis.

Some researchers have performed a sentiment analysis by incorporating the citation context with the subjectivity analysis of citations (Athar & Teufel, 2012; Abu-Jbara et al., 2013; Athar, 2014). In a recent work, Bertin et al. (2016) have used the linguistic patterns to analyse the citation context and its location in the IMRaD structure of a research article to determine the recognition of citer motivation (Teufel et al., 2006). Another work by Small (2018) has studied the phenomenon of highly cited research articles based on citation context. They have examined citation context for linguistic patterns which are associated with different types of referenced research articles. As types, they consider method and non-method publications.

Boyack et al. (2018) have analysed the in-text citation characteristics in the larger dataset of Elsevier full-text journal articles and PubMed Open Access Subset articles. They have identified that all fields such as references, sentences, in-text citation numbers per article have increased over time. They also found that highly cited publications are often cited only once per publication. An et al. (2017) have carried out a study to identify the characteristics of highly cited authors on the basis of citation location and contexts using NLP techniques. They have worked on the ACL Anthology dataset (Bird et al., 2008). Atanassova & Bertin (2016) investigated positions of in-text citation in IMRaD structure regarding the age of cited papers.

Similar to Boyack et al. (2018) and Bertin et al. (2016), we search for patterns in the position and the linguistic context references are cited. But in opposite to them, we change the perspective from the citing publication to the referenced publication. Analogue to Small (2018), we are interested in retrieving information for specific highly cited references from their citation context. But unlike Small, our goal is not the classification of references based on a predefined schema. We want to provide a basis for a comprehensive analysis of the references based on the citation contexts. Additionally, we introduce time aspects in respect to citation interval to the analysis of highly cited references.

Methods

Description of the Dataset

Our research object is a corpus of citation contexts of highly cited references. Each context consists of the sentence in which the citation occurs. For all sentences, we add information about the citing publication and basic information (incl. publication year) of the referenced publication. This information originates from the reference sections of the citing publications. No additional information source is used. To create the dataset, we start with a corpus of 176,856 papers from PLOS ONE published between 2006 and 2017. While we are interested in citation contexts of highly cited papers we selected all references which are cited in more than 100 PLOS ONE articles. To be able to get additional metadata and keep the problem of deduplication away, we choose only references with PubMed ids in the reference part of the citing publications. An example of a citation context with related metadata is shown in Table 1. In total, the number of references cited in more than 100 publications is 666. 127 publications are discarded because of missing PubMed id. In the next step, we filter all citation contexts which do not reference one of the top-666 publications. Because not every publication in our corpus is referencing one of these publications, the number of citing papers used in our study shrinks down to 62,127.

Corpus statistics

The number of relevant citation contexts for our analysis is 173,630. This number reflects the fact, that only in 0.5 percent of the citation contexts (total: 31,746,769) at least one top reference is cited. These top-referenced publications are published between 1951 and 2015. Only 69 of them are published before 2010. We have to keep in mind that just for cited articles published after 2007 it is possible to examine statistics for short citation intervals. The distribution of the number of citation contexts per top-666 follows a power-law-like shape. The most cited reference is mentioned in 3,363 citation contexts. This reference is a method paper by Livak and Schmittgen[3] titled "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method". The lowest number of citation contexts is 75, where the median reference is mentioned in 184 contexts.

Table 1. Example of extracted data for one citation context.

DOI of citing paper	10.1371/journal.pone.0013678
Citation context	*"The molecular mechanisms of nuclear reprogramming are still unsolved although recent reports have shown that reprogramming of human somatic cells can be achieved in vitro by retroviral expression of four transcription factors creating induced pluripotent stem (iPS) cells, which are comparable to ES cells [1], [2], [3], [4]."*
Pubmed-ID of reference	18035408
Section title	Introduction
Pub. year of the citing paper	2010
Pub. year of the reference	2007
Citation interval	3 years (2007-2010)
No. of co-citations	4

[3] This paper is also the top 1 cited paper in a similar study by Small (2018).

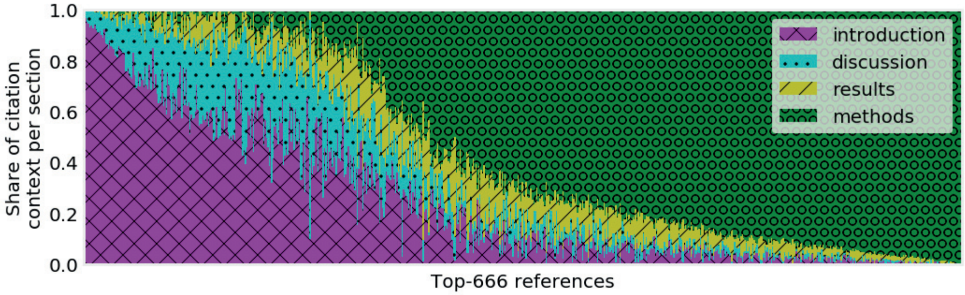


Figure 1. Share of section where all top-666 references are cited.

Results

Distribution over sections

The first question we want to answer with the data from our corpus for the top-666 references is: In which section are the top-666 cited (RQ 1)? Small (2018) used the share of citation contexts located in the method section of a citing publication to predict what type of reference it is. The usefulness of this feature to predict the type of the referenced publication provides a hint, that their usage in a specific section could distinguish highly cited references. Figure 1 shows for each of our top-666 references on the x-axis the proportion of sections in which the citation contexts were found. For more than half of our references the most used section for citing is the method section. But on the other hand, there is a larger group of references which are mainly used in introduction and discussion. We tried to correlate the section usage with the mean citation interval of the top-666. This did not show a significant result. In general, this means, that the section usage of a reference is not dependent on the age of a reference while citing.

Sections of citation contexts over time

To be able to delineate the influence of age of referenced literature and the type of the section in the citing papers we measure the citation interval (RQ 2). The citation interval is the time distance between the publication date of the citing paper and the publication date of the reference paper. We accumulate all citation contexts in groups for each possible citation interval to the corresponding reference. This grouping results in a minimum of zero years to a maximum of 65 years. Of course, the amount of citation contexts in each group varies. Eighty percent of the citation intervals are between 5 to 20 years. For citation intervals larger than 25 years, we have not much account on absolute numbers of citation contexts as well as number of different referenced objects. For each group, we calculated the share of sections and visualized the results in Figure 2.

The result of this analysis is that a reference with a high citation interval, i.e. an older publication at the time of citing, is likely to be used in a method part. The second result is that there is a change for the most frequently used section based on the citation interval at the beginning of a reference lifetime (0-3 years). In the first two years, a top-cited publication is more likely to be used in the introduction part, later it is more often used in the method part. This usage is a hint that the function of the citation is changing over time and needs further examination in future work.

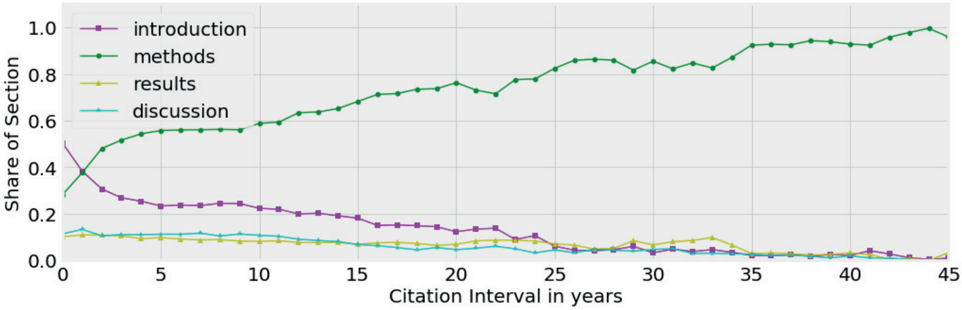


Figure 2. The proportion of IMRaD sections over time.

Number of co-citations on citation context level

The third introspection (RQ 3) of the usage of highly cited references concerns their co-citation on the citation context level. For each relevant context based on citing a top-666 reference we count the number of all references. This approach is defined by Liu & Chen (2012) as sentence-level co-citation. The fact that a publication is referenced alone in one citation context could be a sign that the reference stands on its own and there seems to be no need of citing similar publications. The larger the citation interval is, the higher could be the probability, that there is a low number of co-cited publications in the same sentence.

To investigate the number of co-citations depending on the citation interval, we binned all citation contexts for each citation interval (0-40 years) separately. For each year we have calculated the mean value from the number of quotations per citation context. The mean value is between 1 and 2. Figure 3 (a) shows that the mean value is declining. We calculated the coefficients of a linear regression describing this correlation which is *-0.54* and stated to describe a significant relation. But probably we have to include the knowledge that the higher the citation interval, the greater the probability that the citation context is in a method part of a paper. This fact raises the question whether this phenomenon is also able to describe the declared co-citation time effect.

Here we can come up with a derived research question: Is the high proportion of "method"-citation contexts within longer citation intervals explaining the change of co-citation mean? To answer this question, we divide the citation context into four groups based on location in one of the sections *Introduction, Methods, Results, and Discussion*. After this grouping, we applied the same procedure as for the first calculation. Figure 3 (b) reflects the mean co-citation number in the context given a citation interval for different sections. For all the curves we figured out that there are no significant relations between citation interval and the number of co-citations. We combine the results of this non-significance with the knowledge that the proportion of the citation context of the method increases due to the length of the citation interval (Figure 2). Then, the fact that the mean number of co-citations in the method-contexts is the lowest (Figure 3 (b)) explains the decreasing curve of Figure 3 (a).

Future Work

Future studies aim to replicate results by concerning a larger data corpus, for example by considering the whole PubMed corpus. Also, future research should consider the potential effects of various citation intervals more carefully. For example, analysis of word usage (e.g., adjective, verb, and noun) in citation contexts of highly cited papers and changes over time might be addressed. Besides, the categorization of the cited works into different types or disciplines by their citation contexts could be an essential field for future research.

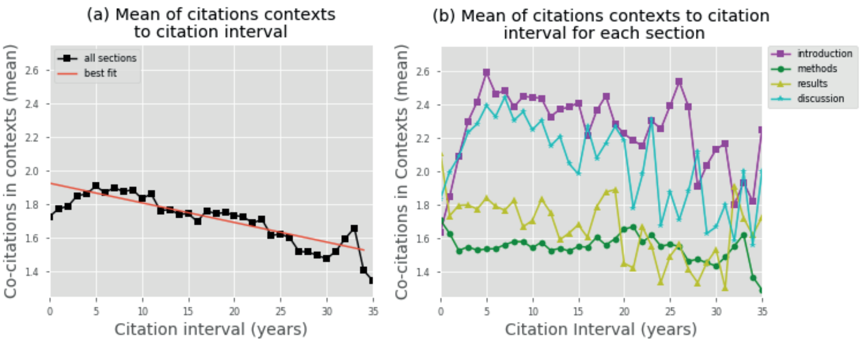


Figure 3. The curves describe the decrease of the mean number of co-citation in contexts for longer citation intervals. (a) Describes the overall decline incl. the best-fit line. (b) Reflects the means for each section individually.

Acknowledgments

The authors acknowledge the enabling support provided by the Indo-German Joint Research Project titled 'Design of a Sciento-text Computational Framework for Retrieval and Contextual Recommendations of High-Quality Scholarly Articles' (Grant No. DST/INT/FRG/DAAD/P-28/2017) for this work.

References

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards nlp-based bibliometrics. In Proceedings of ACL 2013 (pp. 596-606).

An, J., et al. (2017). Exploring characteristics of highly cited authors according to citation location and content. JASIST, 68(8), 1975-1988.

Atanassova, I., & Bertin, M. (2016). Temporal properties of recurring in-text references. D-lib Magazine, 22(9/10).

Athar, A., & Teufel, S. (2012). Context-enhanced citation sentiment detection. In Proceedings of ACL 2012 (pp. 597-601). ACL.

Athar, A. (2014). Sentiment analysis of scientific citations (No. UCAM-CL-TR-856). University of Cambridge, Computer Laboratory.

Bertin, M., et al. (2016). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. Scientometrics, 109(3), 1417-1434.

Bird, S., et al. (2008) The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics. In Proceedings of LREC 2008, (pp. 1–5).

Boyack, K. W., et al. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. Journal of Informetrics, 12(1), 59-73.

Hernández-Alvarez, M., & Gómez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering, 22(3), 327-349.

Huang, W., et al. (2015). A Neural Probabilistic Model for Context Based Citation Recommendation. In AAAI (pp. 2404-2410).

Liu, S & Chen C. (2012). The proximity of co-citation. Scientometrics, 91(2), 495-511.

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty, Journal of Informetrics, 12(2), 461-480.

Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. Journal of the medical library association, 92(3), 364.

Sugiyama, K., et al. (2010). Identifying citing sentences in research papers using supervised learning. 2010 International Conference on Information Retrieval & Knowledge Management (CAMP).

Teufel, S. et al. (2006). Automatic classification of citation function. Proceedings of EMNLP 2006.

A bibliometric perspective on the roles of government funding and international collaboration in scientific research

Ping Zhou[1], Xiaojing Cai[1], Wenjing Xiong[1], and Xiaozan Lyu[1,2]

pingzhou@zju.edu.cn; cai_xj@zju.edu.cn;wenjing_xiong@zju.edu.cn; lvxz1991@zju.edu.cn;

[1]Zhejiang University, School of Public Affairs, Dept of information Resources Management, No. 866 Yuhangtang Road, 310059 Hangzhou (China)

[2]Leiden University, Central for Science and Technology Studies (CWTS), Faculty of Social and Behavioral Sciences, Leiden University, 2300 AX Leiden (The Netherlands)

Abstract

Government funding and international collaboration play significant roles in science development. Both may play roles in publication citation impact. The current study focused on funded international collaboration in six countries (i.e., China, Brazil, South Africa, Germany, The Netherlands and US). Different countries vary in arrangement of government funding sources in support of competitive research projects. China and Brazil are centralized to a unique agency (i.e., NSFC and CNPq), whereas the rest four countries are relatively decentralized. The six focal agencies of the current study (i.e., NSFC, NSF, DFG, NWO, NRF and CNPq) are more efficient than non-focal agencies in general in raising citation impact, with NSF of US performs best. Not all countries get benefit from international collaboration in terms of raising citation impact of publications: developing countries (i.e., China and Brazil in current study) benefit more than developed countries. Collaborating with developed countries especially the US can be a first option for choosing foreign partners. With regard to interaction between funding and international collaboration, DFG performs best in raising citation impact of international collaboration.

Introduction

Government funding plays a significant role in the development of science, and have attracted extensive interests of academic community. Positive effects of government funding in raising citation impact of publications (e.g., Yan, Wu, & Song, 2018) and breakthrough inventions (Corredoira, Goldfarb, & Shi, 2018) have been found. Variations in terms of funding efficiency may exist between different funding systems. Among many factors, national research evaluation systems (Sandström & Van Besselaar, 2016), academic freedom, and university stratification play significant roles in affecting funding efficiency (Sandström & Van den Besselaar, 2018). In current era that scientific discovery increasingly rely on wide-spread collaboration (Bozeman & Corley, 2004; Choi, Yang, & Park, 2015; Cimini, Zaccaria, & Gabrielli, 2016) including international collaboration (Frame & Carpenter, 1979; Wagner, Park, & Leydesdorff, 2015). Studies on international collaborative research have kept expanding so rapidly that some even wonder it as an emerging field (Chen, Zhang, & Fu, 2018; D'Ippolito & Rüling, 2019). In addition to analysing the roles of international collaboration in publication productivity and citation impact (Bozeman, Fay, & Slade, 2013; Lee & Bozeman, 2005; Van Raan, 1998), network position of specific countries (Adams, 2012; Guan, Yan, & Zhang, 2017; Todeva & Knoke, 2005; Wang, Wang, & Philipsen, 2017; Zhao & Guan, 2011), as well as collaboration patterns (Leydesdorff & Wagner, 2008; Todeva & Knoke, 2005) are also important topics. The effect of international collaboration in raising citation impact of publications of different countries vary significantly and is very much dependant on collaborating partners (Bote, Olmeda-Gómez, & de Moya-Anegón, 2013; Lancho-Barrantes, Guerrero-Bote, & de Moya-Anegón, 2013; Sud & Thelwall, 2016). Collaborating with the US especially with US researchers as corresponding authors would benefit most (Bote et al., 2013;

De Moya-Anegon, Guerrero-Bote, Lopez-Illescas, & Moed, 2018; Moya-Anegón, Guerrero-Bote, Bornmann, & Moed, 2013).

The above studies imply that both funding and international collaboration play significant roles in scientific research, which leads to perspectives combining the two factors together. One typical perspective is on the role of funding in facilitating international collaboration, and with no exception, positive effect has been confirmed (Cimini et al., 2016; Clark & Llorens, 2012; Liu, Liang, Tuuli, & Chan, 2018; Ubfal & Maffioli, 2011). Another popular perspective is to investigate if international collaboration facilitates access to funding supports (Zhou & Tian, 2014). Some studies explore the effect of funding and international collaboration separately on citation impact without investigating if mutual effect exists between the two factors together (Leydesdorff, Wagner, & Bornmann, 2019; Zhou, Zhong, & Yu, 2013). International collaboration plays positive role in raising citation impact but the effect of government funding tends to be a small adverse (Leydesdorff et al., 2019).

In fact, funding and international collaboration may interact with each other, and thus affect citation impact comprehensively. The current study will contribute in this regard. Considering that the effect of the two factors (i.e., funding support and international collaboration) on citation impact may also be dependent on other factors such as collaboration size, funding sources, research fields, and so on, the current study will take most of the possible factors into account by focusing on publications acknowledging government funding supports of six different countries, so as to clarify the following questions: (1) What is the contribution of funded research to science? Does country variation exist? (2) Is there difference in terms of citation impact between publications acknowledging support of national funding agencies of different countries? (3) What is the role of international collaboration in different countries? Is the effect of first and corresponding authorship different? Does effect of collaboration with developed countries the same as that with developing countries? (4) How do funding and international collaboration interact with each other? Upon answering the four questions, discussion will be carried out.

Data and methods

Bibliometric data in 2009-2016 are extracted from the CWTS-licensed version of the Web of Science (WoS) database of Clarivate. Six countries, namely, China, Brazil, South Africa, Germany, The Netherlands, and the US are included to illustrate the situations in both developing and developed countries. The targeted funding agencies (will be called *focal agencies* latter) supporting basic research of the six countries are the National Natural Science Foundation of China (NSFC), National Council for Scientific and Technological Development of Brazil (CNPq), National Research Foundation of South Africa (NRF), German Research Foundation (DFG), The Netherlands Organization for Scientific Research (NWO), and the US National Science Foundation (NSF). Funded publications are harvested from CWTS funding organization database originated from the WoS index fields FO and FT.

Field classification is based on CWTS' 35 subject categories. Because not all scientific fields are supported by the focal agencies of the current study, we only cover subject categories that are main supporting areas of the six funding agencies. Thus, 22 subject categories[1] remained. Since the current study focuses on journal publications, only those with journal papers as main form of research outcome will be covered. Although highly productive in journal publications, the area *Clinical Medicine* is not covered because it is not the major area supported by the National Science Foundation (NSF). In the end, six subject categories including *Basic Life Sciences, Biological Sciences*, *Chemistry & Chemical Engineering*, *Environmental Sciences & Technology*, *Mathematics* as well as *Physics & Materials Science* remain.

Two sets of indicators measuring publication productivity and citation impact are used. For productivity measurement, we apply percentage of funded publications and top-1% highly cited publications. The former measures the extent of funded research, and the latter measures high-quality research. The *Mean Normalized Citation Scores* (MNCS) of CWTS is used to measure average citation impact of a publication set under variable citation window.

International collaboration is defined as two or more countries appear in author addresses of a publication. Single-author publications are considered non-international collaboration even if different country affiliations appear. Publications with two authors in which both share the same country and one of them has a second country affiliation would be treated as international collaboration. In co-authored publications, the roles of different authors vary with that of first and corresponding authors most significant. Investigating country affiliations of first or corresponding authors may provide deeper insights on the role of a country in international collaboration. The positive roles of US corresponding authors have been found in different studies (Bote et al., 2013; De Moya-Anegon et al., 2018; Lancho-Barrantes et al., 2013). The current study will explore both first- and corresponding-authored collaboration of a target country. Because the effect of collaboration with scientific leading and following countries may differ, we classify countries into two sets - Group 7 countries (i.e., Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) and non-G7 countries. Collaboration with US will be analyzed specially because of its absolute leading position in science.

To analyze the effect of different factors on citation impact, publications are classified into different types as illustrated in Table 1.

Table 1. Publication classification.

Publication set	*Definition*
All publication	All publications regardless of funded or not.
PFA	Publications acknowledging any types of funding support.
PFAs*	Publications acknowledging funding support of a focal agency (i.e., NSFC, NSF, DFG, NWO, CNPq, NRF).
PFA without PFAs	Publications not including financial support of the focal agencies.
PnotFA	Publications without funding acknowledgement.
PFAs and IntNatl	Internationally collaborated publications acknowledging funding support of a focal agency.
(PFA without PFAs) and IntNatl	Internationally collaborated publications with funding not including the focal agencies.
First country	First-author publications of a target country in international collaboration.
Reprint country	Corresponding-author publications of a target country in international collaboration.
G7 countries	Canada, France, Germany, Italy, Japan, the UK, and the US

* There may exist publications claiming to be funded by a focal agency but no authors from that country, which may interfere with a country-based result. To avoid such situation happening, we further define a publication set funded by a focal agency in the following way: A publication acknowledging funding support of a focal agency should have author(s) from the country of the focal agency. For example, a publication acknowledging NSF grant should have author(s) from the USA.

Results

Both descriptive and regression results will be illustrated. Ordinary Least Square (OLS) regression with robust standard errors is used to measure effect of funding and/or international collaboration on citation impact of different types of publication sets.

Descriptive analysis

Funding supports play important roles in national publication production. Around 67% of publications of the six countries have acknowledged grant support. With over 83% of funded publications, China is most prominent. Situations in the rest five countries are similar (62% ~ 67%). In terms of percentage of publications acknowledging grant of the focal agencies in the total national funded publications (i.e., % (PFAs/PFA)), that of NSFC is the highest (69%). In other words, about 69% of China's funded publications are supported by NSFC. The role of CNPq is similar, it contributes about 65% of Brazil's funded publications. The NSF's percentage contribution to US funded publications is the least (18%), whereas that of DFG, NRF and NWO are respectively 36%, 27% and 23% (Figure 1a).

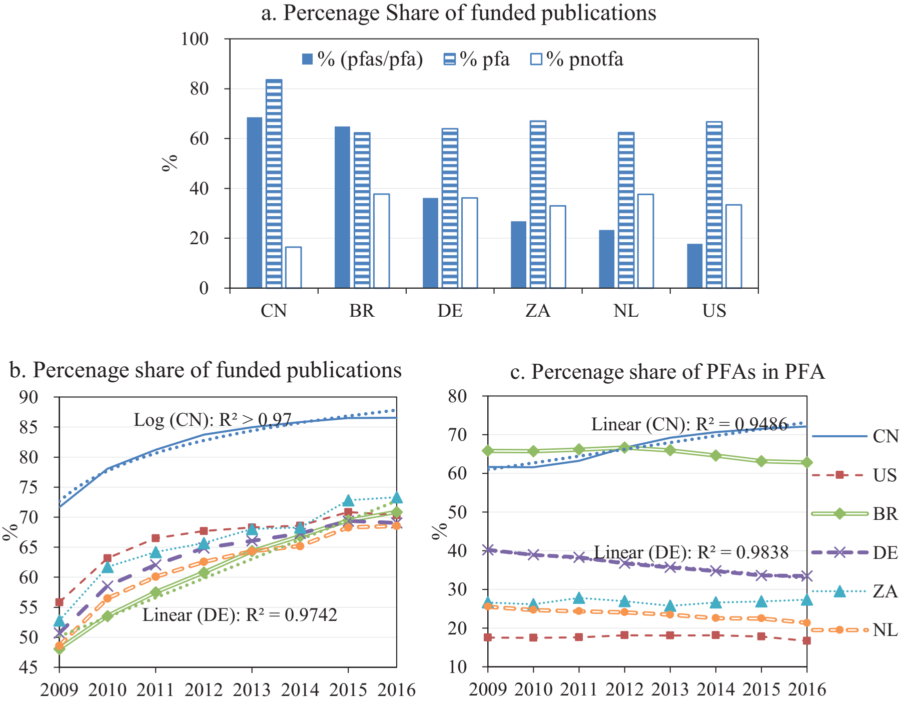


Figure 1. Output and annual trend in 2009-2016

The percentage share of funded publications in national total has kept growing, although each country has its own trend. China has the highest percentage and grows logarithmically. The growth trend of Brazil has kept strong. The rest four countries share similar trends with growth styles changed in 2014 and converged at around 68% in 2016 (Figure 1b). With respect to percentage of publications funded by the focal agencies (PFAs in PFA), that of NSFC has kept growing. There is no significant change for that of NRF. Percentage decline is also seen for the rest four agencies, especially DFG and NWO (Figure 1c).

Citation impact of all publication sets measured by MNCS is, however, a different landscape. The Netherlands performs best in the three types of publications (i.e., PFA, PFAs, and PnotFA), followed by the US and Germany. Publications with funding support perform, in general, better than unfunded publications. With significantly higher citation impact than all-funded publications of US, NSF-supported publications perform best, followed by those funded by NWO and NSFC. In contrast to the US situation, citation impacts of all funded publications of

Germany, South Africa and Brazil are higher than those funded by the focal agencies of the corresponding countries (i.e., DFG, NRF, and CNPq). NSFC funding is critical to Chinese publications in raising citation impact. NSFC-funded publications have higher citation impact than those funded by NRF, although unfunded publications of China have lower citation impact than unfunded publications of South Africa (Figure 2a). As expected, similar situation happens in producing highly-cited publications. The Netherlands performs best, followed by the US and Germany.

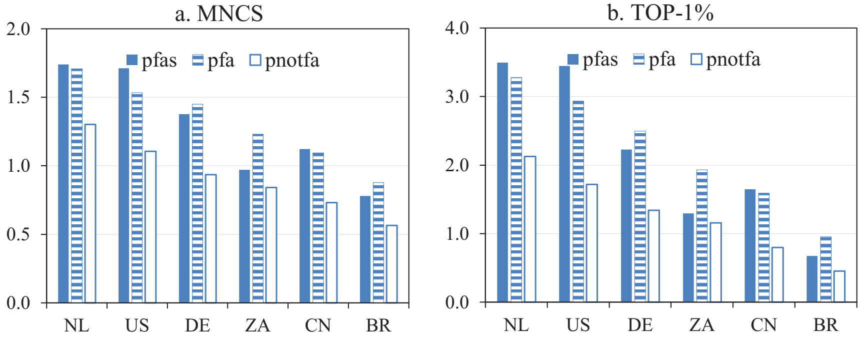


Figure 2. MNCS and percentage of top-1% highly cited publications (2009-2016).

Regression analysis

The effect of different independent variables on citation impact measured by *NCS* (i.e., the dependent variable) is obtained by applying OLS regression (Table 2). The independent variables include funding types (i.e., *PnotFA, PFA without PFAs*, *PFAs*), international collaboration (i.e., *IntNatl*), internationally collaborated publications acknowledging funding support not from the focal agencies (i.e., *PFA without PFAs* and *IntNatl*), internationally collaborated publications acknowledging support of the focal agencies (i.e., *PFAs and IntNatl*), first-author publications, reprint (or corresponding) publications, as well as publications collaborated with US, G6 or non-G7 countries. Interaction between funding support and international collaboration is also included to investigate synergy between funding support and international collaboration. Other factors including publication year, number of references, number of authors, number of institutions, number of countries, publication types (i.e., review, article), as well as scientific fields may also affect citation impact, and thus are set as controlled variables.

Funding support, regardless of focal- or non-focal-agency sources, has positive effect on citation impact of publications, although such contribution varies among countries and depends on the variables. Except DFG and NRF, the effects of NSFC, NSF, NWO and CNPq are higher than non-focal-agency funding of the same country. Citation impact of Chinese publications with non-NSFC funding support is 7.3% higher, whereas those with NSFC funding is 10.1% higher than those with no funding support ($e^{0.071}$ and $e^{0.096}$). For US publications, non-NSF funding support may raise citation impact by 6.9% whereas NSF support has higher effect of 10.8% ($e^{0.067}$ and $e^{0.103}$). The Dutch NWO raised citation impact by 10.4%, which is higher than 7.1% of those with non-NWO funding ($e^{0.069}$ and $e^{0.099}$). Similar situation applies to Brazilian CNPq and non-CNPq funding. The opposite is true in Germany: citation-impact-raising effect (by 9.4%) of non-DFG funding is higher than DFG funding (by 4.8%) ($e^{0.090}$ and $e^{0.047}$). The situation in South Africa is most unique – there is no significant change of citation impact of publications funded by NRF whereas non-NRF funding can raise citation impact by 1.2%.

The effect of international collaboration varies between countries. With 12.9% increase of citation impact by collaborating with foreign partners, China benefits most among the six countries. Brazil is also a beneficiary with 7.9% increase of citation impact. On the contrary, the rest four countries including the US, The Netherlands, South Africa and Germany have citation impact been reduced respectively by 8.1%, 6.7%, 4.6% and 4.0% in collaborating with other countries. When the interaction between funding and international collaboration is considered, the situation is a little bit complex. The effect of non-focal-agency (*PFA without PFAs*) support on raising citation impact of internationally collaborated publications is lower than that of non-internationally collaborated publications of countries like China, Germany and The Netherlands, but slightly higher than that of non-internationally collaborated publications of a country like the US and South Africa. There is no significant difference in internationally collaborated publications of Brazil whether with or without non-focal-agency support. The effect of the focal agencies on raising citation impact of internationally collaborated publications is lower than that on publications of non-international collaboration in China, Brazil, and Netherlands, and is higher in Germany, but no significant difference in US and South Africa. In general, funding support, regardless with or without the focal agencies (i.e., NSFC or NWO) of China and The Netherlands, may be less effective in raising citation impact of internationally collaborated publications than in raising citation impact of non-internationally collaborated publications. In contrast to the rest five focal agencies (i.e., NSFC, NSF, CNPq, NRF, NWO), German DFG performs better in raising citation impact of internationally collaborated publications than that of publications of non-international collaboration (Table 2).

Table 2. Results of OLS regression.

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Variables	*CN*	*US*	*BR*	*DE*	*ZA*	*NL*
PFA without PFAs	0.071***	0.067***	0.076***	0.090***	0.012**	0.069***
	(0.002)	(0.001)	(0.003)	(0.003)	(0.006)	(0.006)
PFAs	0.096***	0.103***	0.096***	0.047***		0.099***
	(0.001)	(0.002)	(0.003)	(0.003)		(0.007)
IntNatl	0.121***	-0.078***	0.076***	-0.039***	-0.045***	-0.065***
	(0.005)	(0.003)	(0.008)	(0.003)	(0.012)	(0.008)
(PFA without PFAs) and IntNatl	-0.011***	0.004*		-0.013***	0.040***	-0.013*
	(0.004)	(0.002)		(0.004)	(0.009)	(0.008)
PFAs and IntNatl	-0.025***		-0.050***	0.021***		-0.033***
	(0.004)		(0.005)	(0.004)		(0.010)
First country	-0.051***	0.048***	-0.048***	0.014***	-0.035***	0.036***
	(0.004)	(0.003)	(0.009)	(0.002)	(0.007)	(0.005)
Reprint Country	-0.013***	0.032***	-0.028***			
	(0.004)	(0.003)	(0.009)			
Collaborate with USA	0.111***		0.113***	0.130***	0.137***	0.137***
	(0.005)		(0.008)	(0.003)	(0.012)	(0.007)
Collaborate with G6	0.041***	0.072***	0.045***	0.067***	0.087***	0.072***
	(0.005)	(0.003)	(0.008)	(0.002)	(0.011)	(0.007)
Collaborate with Non-G7	0.072***	0.027***	0.028***		0.061***	0.014**
	(0.005)	(0.004)	(0.008)		(0.012)	(0.007)
Length of years	0.012***	0.014***	0.013***	0.012***	0.013***	0.014***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
log_refs	0.285***	0.219***	0.173***	0.228***	0.210***	0.244***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.004)	(0.003)
log_au	0.116***	0.120***	0.034***	0.058***	0.015**	0.023***
	(0.001)	(0.001)	(0.003)	(0.002)	(0.007)	(0.004)
log_ins	-0.059***		-0.020***		0.052***	0.014*
	(0.002)		(0.004)		(0.010)	(0.007)
log_country	-0.085***	-0.034***	0.097***		0.026	0.030**
	(0.009)	(0.006)	(0.013)		(0.022)	(0.014)

Variables	(1) CN	(2) US	(3) BR	(4) DE	(5) ZA	(6) NL
doc_review	0.189***	0.310***	0.259***	0.300***	0.267***	0.285***
	(0.005)	(0.003)	(0.009)	(0.005)	(0.015)	(0.009)
doc_letter	0.580***	0.405***	0.407***	0.534***	0.410***	0.461***
	(0.015)	(0.007)	(0.021)	(0.018)	(0.043)	(0.023)
basiclife	0.004**	-0.047***	-0.033***	-0.045***		-0.057***
	(0.002)	(0.002)	(0.004)	(0.003)		(0.004)
biosci	0.067***	-0.030***	0.007*	0.015***		-0.032***
	(0.002)	(0.002)	(0.004)	(0.003)		(0.005)
chemengn	0.167***	0.057***	0.058***	0.013***	0.060***	0.016***
	(0.001)	(0.002)	(0.004)	(0.002)	(0.006)	(0.005)
envitech	0.159***	0.016***	0.051***	0.057***	0.022***	0.022***
	(0.002)	(0.002)	(0.004)	(0.003)	(0.006)	(0.005)
math	0.256***	0.131***	0.127***	0.125***	0.113***	-0.029***
	(0.003)	(0.003)	(0.007)	(0.004)	(0.011)	(0.009)
phymat	0.145***	0.056***	0.010**	0.047***	0.047***	
	(0.002)	(0.002)	(0.004)	(0.003)	(0.006)	
Constant	-0.861***	-0.499***	-0.520***	-0.504***	-0.549***	-0.458***
	(0.007)	(0.005)	(0.011)	(0.006)	(0.019)	(0.015)
Observations	1,085,537	1,168,590	143,000	387,207	38,781	100,467
F	7203.171	6954.223	836.380	2656.112	353.347	599.659
r2	0.155	0.125	0.161	0.142	0.195	0.148

Robust standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

[a] According to F test, there are significant differences between coefficients of *Pfa without Pfas* and *Pfas* in models 1-4 and 6.

Different roles and partners with different scientific positions may also affect citation impact. When developed countries (i.e., US, Germany, and The Netherlands) act as first authors, citation impact would be higher than non-first authorship of these countries. On the contrary, when developing countries (i.e., China, Brazil and South Africa) act as first authors, citation impact of publications would be lower than those of non-first authorship of the three countries. When corresponding authorship is considered, similar results happen to China and Brazil: International collaboration with Chinese or Brazilian researchers acting as corresponding authors, citation impact would be lower than those of non-corresponding authorship of the two countries. The positive role of US happens again when corresponding authorship is concerned. The effect of corresponding authorship of the rest countries (i.e., Germany, Netherlands, and South Africa) is insignificant. The outstanding role of US in raising citation impact of internationally collaborated publications is further confirmed by the standardized coefficients of OLS regression. Collaborating with G7 countries excluding the US is also a way of raising citation impact, although such effect is less than that collaborating with US. Non-G7 countries may also help raising citation impact of the six target countries except Germany, but with the least effect among the three types of partner countries (i.e., US, G6, and non-G7). The situation of China is exceptional: collaborating with non-G7 countries may have higher citation impact than collaborating with G6 countries (Table 3).

Table 3. Standardized coefficients of OLS regression.

Collaboration partners	*(1)* *CN*	*(2)* *US*	*(3)* *BR*	*(4)* *DE*	*(5)* *ZA*	*(6)* *NL*
US	0.062***		0.080***	0.083***	0.097***	0.090***
	(0.005)		(0.008)	(0.003)	(0.012)	(0.007)
G6	0.021***	0.047***	0.036***	0.049***	0.073***	0.060***

Collaboration	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
partners	*CN*	*US*	*BR*	*DE*	*ZA*	*NL*
	(0.005)	(0.003)	(0.008)	(0.002)	(0.011)	(0.007)
Non-G7	0.038***	0.022***	0.025***		0.060***	0.012**
	(0.005)	(0.004)	(0.008)		(0.012)	(0.007)

Standardized beta coefficients; Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Conclusions and discussion

With absolute high percentage of publications supported by the focal agencies, China and Brazil have more centralized government resources supporting basic research than the rest four countries. The NSFC of China supports basic research in almost all areas, while in the US, the responsibility for supporting basic research is spread across the National Science Foundation (NSF), National Institute of Health (NIH), and Department of Energy (DOE). The percentage share of funded publications in all publications of a country has kept growing, and each country has its own way of growth. China has the highest percentage of funded publications, and the percentage of NSFC in the funded total has kept growing, which is in stark contrast to the focal agencies in the other five countries, especially DFG and NWO with a decreasing share. In other words, Chinese researchers rely heavily on NSFC funding whereas US, Germany, Dutch and South African scholars have more options in seeking national funding sources.

With regard to citation impact and top-1% highly cited publications, developed countries (i.e., The Netherlands, US and Germany) perform better than developing countries (i.e., South Africa, China and Brazil), and Netherlands performs best. Funding supports are important in raising citation impact, although variations exist among countries. In supporting high-quality research in terms of citation impact, some focal agencies like NSF are more efficient than non-NSF funding support, some (i.e., NWO and NSFC) do not make much difference from non-focal agencies, and some (i.e., DFG, NRF, and CNPq) even are less efficient.

The positive effect of funding support in raising citation impact of publications is further confirmed in OLS regression analysis. Citation-raising effect of most of the focal agencies (i.e., NSFC, NSF, NWO and CNPq) are usually higher than that of non-focal agencies. Nonetheless, it is not true in German and South African situation. The situation in South Africa is most unique – there is no significant change of citation impact of publications funded by NRF whereas non-NRF funding can raise citation impact.

Not all countries get benefit from international collaboration in terms of raising citation impact of a nation's publications. In most cases, developing countries benefit more than developed countries. Collaborating with developed countries especially with the US can be a first option in choosing international partners. Although citation impact might be lowered in collaborating with developing countries, such collaboration should still be encouraged for developed countries, because it may complement shortage of human resources and help young scholars from developing countries grow up.

When interaction of funding support and international collaboration is concerned, the situation is very much dependent on countries and funding sources. The citation-raising effect of non-focal-agency funding on international collaboration is slightly lower than on non-international collaboration of China, Germany and The Netherlands, but slightly higher on international collaboration of US and South Africa. To Brazilian publications, effect of non-focal-agency support in raising citation impact of publications with or without international collaboration does not show significant difference. The impact-raising effect of the focal agencies in publications with international collaboration over non-international collaboration also varies – some (i.e., NSFC or NWO) with less efficiency, some (i.e., NSF, NRF) with insignificant difference; the German DFG is the only one that is more efficient. An important reason, among others that may lead to the different effect of the six focal agencies on citation impact, can be evaluation principles for research projects. The DFG practice can be an excellent example for other agencies supporting research in basic science.

Notes

[1] Agriculture and Food Science, Astronomy and Astrophysics, Basic Life Sciences, Basic Medical Sciences, Biological Sciences, Biomedical sciences, Chemistry and Chemical Engineering, Civil Engineering and Construction, Clinical Medicine, Computer Sciences, Earth Sciences and Technology, Electrical Engineering and Telecommunication, Energy Science and Technology, Environmental Sciences and Technology, General and Industrial Engineering, Health Sciences, Instruments and Instrumentation, Mathematics, Mechanical Engineering and Aerospace, Multidisciplinary Journals, Physics and Materials Science, Statistical Sciences.

Acknowledgments

This study is supported by National Natural Science Foundation of China (grant number: 71473219, 71843012). Thank Dr. Rodrigo Costas, a senior researcher of the Centre for Science and Technology Studies (CWTS) in Leiden University for suggestions of selecting target countries and indicators. Dr. Clara Calero-Medina, a researcher of CWTS offered help in collecting funded data of CWTS-licensed version of the Web of Science database (WoS) of Clarivate.

References

Abramo, G., D'Angelo, A. C., & Murgia, G. (2017). The relationship among research productivity, research collaboration, and their determinants. *Journal of Informetrics*, *11*(4), 1016–1030. https://doi.org/10.1016/j.joi.2017.09.007

Adams, J. (2012). Collaborations: The rise of research networks. *Nature*. https://doi.org/10.1038/490335a

Bote, V. P. G., Olmeda-Gómez, C., & de Moya-Anegón, F. (2013). Quantifying the Benefits of International Scientific Collaboration. *Journal of the American Society for Information Science*, *64*(2013), 2353–2361. https://doi.org/10.1002/asi

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy*. https://doi.org/10.1016/j.respol.2004.01.008

Bozeman, B., Fay, D., & Slade, C. P. (2013). Research collaboration in universities and academic entrepreneurship: The-state-of-the-art. *Journal of Technology Transfer*. https://doi.org/10.1007/s10961-012-9281-8

Chen, K., Zhang, Y., & Fu, X. (2018). International research collaboration: An emerging domain of innovation studies? *Research Policy*, *48*(1), 149–168. https://doi.org/10.1016/j.respol.2018.08.005

Choi, S., Yang, J. S. W., & Park, H. W. (2015). The triple helix and international collaboration in science. *Journal of the Association for Information Science and Technology*, *66*(1), 201–212. https://doi.org/10.1002/asi.23165

Cimini, G., Zaccaria, A., & Gabrielli, A. (2016). Investigating the interplay between fundamentals of national research systems: Performance, investments and international collaborations. *Journal of Informetrics*, *10*(1), 200–211. https://doi.org/10.1016/j.joi.2016.01.002

Clark, B. Y., & Llorens, J. J. (2012). Investments in Scientific Research: Examining the Funding Threshold Effects on Scientific Collaboration and Variation by Academic Discipline. *Policy Studies Journal*, *40*(4), 698–729. https://doi.org/10.1111/j.1541-0072.2012.00470.x

Corredoira, R. A., Goldfarb, B. D., & Shi, Y. (2018). Federal funding and the rate and direction of inventive activity. *Research Policy*. https://doi.org/10.1016/j.respol.2018.06.009

D'Ippolito, B., & Rüling, C. C. (2019). Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications. *Research Policy*, (January), 1–15. https://doi.org/10.1016/j.respol.2019.01.011

De Moya-Anegon, F., Guerrero-Bote, V. P., Lopez-Illescas, C., & Moed, H. F. (2018). Statistical relationships between corresponding authorship, international co-authorship and citation impact of national research systems. *Journal of Informetrics*, *12*(4), 1251–1262. https://doi.org/10.1016/j.joi.2018.10.004

Frame, D., & Carpenter, M. P. (1979). International Research Collaboration Published by : Sage Publications , Ltd . Stable URL : https://www.jstor.org/stable/284574 REFERENCES Linked references are available on JSTOR for this article : Y. *Social Studies of Science*, *9*(4), 481–497.

Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics*, *11*(2). https://doi.org/10.1016/j.joi.2017.02.007

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & de Moya-Anegón, F. (2013). Citation increments between collaborating countries. *Scientometrics*, *94*(3), 817–831. https://doi.org/10.1007/s11192-012-0797-3

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*. https://doi.org/10.1177/0306312705052359

Leydesdorff, L., & Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2008.07.003

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, *13*(1). https://doi.org/10.1016/j.joi.2018.12.006

Liu, A. M. M., Liang, O. X., Tuuli, M., & Chan, I. (2018). Role of government funding in fostering collaboration between knowledge-based organizations: Evidence from the solar PV industry in China. *ENERGY EXPLORATION & EXPLOITATION*, *36*(3), 509–534. https://doi.org/10.1177/0144598717742968

Moya-Anegón, F., Guerrero-Bote, V. P., Bornmann, L., & Moed, H. F. (2013). The research guarantors of scientific papers and the output counting: A promising new approach. *Scientometrics*, *97*(2), 421–434. https://doi.org/10.1007/s11192-013-1046-0

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics*. https://doi.org/10.1007/BF02093973

Sandström, U., & Van Besselaar, P. Den. (2016). Quantity and/or quality? The importance of publishing many papers. *PLoS ONE*, *11*(11), 1–16. https://doi.org/10.1371/journal.pone.0166149

Sandström, U., & Van den Besselaar, P. (2018). Funding, evaluation, and the performance of national research systems. *Journal of Informetrics*, *12*(1). https://doi.org/10.1016/j.joi.2018.01.007

Sud, P., & Thelwall, M. (2016). Not all international collaboration is beneficial: The Mendeley readership and citation impact of biochemical research collaboration. *Journal of the Association for Information Science and Technology*, *67*(8), 1849–1857. https://doi.org/10.1002/asi.23515

Todeva, E., & Knoke, D. (2005). Strategic alliances and models of collaboration. *Management Decision*. https://doi.org/10.1108/00251740510572533

Ubfal, D., & Maffioli, A. (2011). The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, *40*(9), 1269–1279. https://doi.org/10.1016/j.respol.2011.05.023

Van Raan, A. F. J. (1998). The influence of international collaboration on the impact of research results - Some simple mathematical considerations concerning the role of self-citations. *SCIENTOMETRICS*, *42*(3), 423–428. https://doi.org/10.1007/BF02458380

Wagner, C. S., Park, H. W., & Leydesdorff, L. (2015). The continuing growth of global cooperation networks in research: A conundrum for national governments. *PLoS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0131816

Wang, L., Wang, X., & Philipsen, N. J. (2017). Network structure of scientific collaborations between China and the EU member states. *SCIENTOMETRICS*, *113*(2), 765–781. https://doi.org/10.1007/s11192-017-2488-6

Yan, E., Wu, C., & Song, M. (2018). The funding factor: a cross-disciplinary examination of the association between research funding and citation impact. *Scientometrics*, *115*(1), 369–384. https://doi.org/10.1007/s11192-017-2583-8

Zhao, Q., & Guan, J. (2011). International collaboration of three `giants' with the G7 countries in emerging nanobiopharmaceuticals. *SCIENTOMETRICS*, *87*(1), 159–170. https://doi.org/10.1007/s11192-010-0311-8

Zhou, P., & Tian, H. (2014). Funded collaboration research in mathematics in China. *Scientometrics*, *99*(3), 695–715. https://doi.org/10.1007/s11192-013-1212-4

Zhou, P., Zhong, Y., & Yu, M. (2013). A bibliometric investigation on China-UK collaboration in food and agriculture. *Scientometrics*, *97*(2), 267–285. https://doi.org/10.1007/s11192-012-0947-7

Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches

Alexander Tekles[1] and Lutz Bornmann[2]

[1] *alexander.tekles.extern@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society,
Hofgartenstr. 8, 80539 Munich (Germany)
Ludwig-Maximilians-University Munich, Department of Sociology, Konradstr. 6, 80801 Munich, Germany

[2] *bornmann@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society,
Hofgartenstr. 8, 80539 Munich (Germany)

Abstract

Adequately disambiguating author names in bibliometric databases is a precondition for conducting reliable analyses at the author level. In the case of bibliometric studies that include many researchers, it is not possible to disambiguate each single researcher manually. Several approaches have been proposed for author name disambiguation but there has not yet been a comparison of them under controlled conditions. In this study, we compare a set of unsupervised disambiguation approaches. Unsupervised approaches specify a model to assess the similarity of author mentions a priori instead of training a model with labelled data. In order to evaluate the approaches, we applied them to a set of author mentions annotated with a ResearcherID, this being an author identifier maintained by the researchers themselves. Apart from comparing the overall performance, we take a more detailed look at the role of the parametrization of the approaches and analyse the dependence of the results on the complexity of the disambiguation task. It could be shown that all of the evaluated approaches produce better results than those that can be obtained by using only author names. In the context of this study, the approach proposed by Caron and van Eck (2014) produced the best results.

Introduction

Bibliometric analyses at an individual level depend on the adequate identification of the authors' oeuvres. At best, all of an author's papers should be considered without fail, while other papers should not be falsely assigned to that author. Getting as close as possible to this ideal situation is especially important since poorly disambiguated data may distort the results of analyses at an author level (Kim and Diesner 2016; Kim 2019). Some identifiers that uniquely represent authors are available in bibliometric databases. These, however, are maintained by the researchers themselves (e.g. ResearcherID, ORCID) – implying a low coverage and the possibility of deliberate false assignments – or are based on an undisclosed automatic assignment (e.g. Scopus Author ID) – which does not allow an assessment of the quality of the algorithm (the algorithm is not publicly available). Automatic approaches that try to solve the task of disambiguating author names have thus been proposed in bibliometrics. This task presents a non-trivial challenge since different authors may have the same name (homonyms) and one author may publish under different names (synonyms).

Table 1. Examples for homonyms and synonyms in bibliometric databases

Publication title	Author name	Author ID
Social theory and social structure	R. Merton	1
The Matthew effect in science	Robert Merton	1
Allocating Shareholder Capital to Pension Plans	Robert Merton	2

Table 1 shows the titles, the author names and an author identifier for three publications, including both homonyms and synonyms. The author names of the first two publications are synonyms since they refer to the same person but differ in terms of the name. The author names of the last two publications are an example of homonyms since they refer to different persons but share the same name.

In this study, we compare four unsupervised disambiguation approaches. In order to evaluate the approaches, we applied them to a set of author mentions annotated with a ResearcherID, this being an author identifier maintained by the researchers themselves. Apart from comparing the overall performance, we take a more detailed look at the role of the parametrization of the approaches and analyse the dependence of the results on the complexity of the disambiguation task.

Related work

In order to find sets of publications corresponding to real-world authors, approaches for disambiguating author names try to assess the similarity between author mentions by exploiting metadata such as co-authors, subject category, journal, etc. In order to reduce runtime complexity and exclude a high number of obvious false links between author mentions, most approaches reduce the search space by blocking the data in a first step (On et al. 2005). The idea is to generate disjunctive blocks so that author mentions in different blocks are very likely to refer to different identities, and therefore the comparisons can be limited to pairs of author mentions within the same block (Newcombe 1967; Levin et al. 2012). A widely used blocking strategy for disambiguating author names in bibliometric databases is to group together all author mentions with an identical canonical representation of the author name, consisting of the first name initial and the surname (On et al. 2005).

The algorithms to disambiguate author names that have been proposed up to now differ in several respects (Ferreira, Gonçalves, and Laender 2012). One way to distinguish between different approaches is to classify them as either unsupervised or supervised (Smalheiser and Torvik 2009). Supervised approaches try to train the parameters of a specified model with the help of certain training data (e.g., Torvik and Smalheiser 2009; Ferreira et al. 2010; Levin et al. 2012; Ferreira et al. 2014). The training data contains explicit information as to which author mentions belong to the same identity and which do not. The model trained on the basis of this data is then used to detect relevant patterns in the rest of the data. Unsupervised approaches, on the other hand, try to assess the similarity of author mentions by explicitly specifying a similarity function based on the author mentions' characteristics. We will focus on unsupervised approaches in the following. Supervised approaches entail several problems, especially the challenge of providing adequate, reliable and representative training data (Smalheiser and Torvik 2009).

The unsupervised approaches for disambiguating author names that have been proposed so far vary in several ways. First, every approach specifies a set of attributes and how these are combined to provide a similarity measure between author mentions. In order to determine which similarities are high enough to consider two author mentions or two groups of author mentions as referring to the same author, some form of threshold for the similarity measure is necessary. This threshold can be determined globally for all pairs of author mentions being compared, or it can vary depending on the number of author mentions within a block that refers to a single name representation. Block size dependent thresholds try to reduce the problem of an increasing number of false links for a higher number of comparisons between author mentions, i.e. for larger name blocks (Caron and van Eck 2014; Backes 2018).

Another way in which the approaches differ is the clustering strategy that is applied, i.e. how similar author mentions are grouped together. All clustering strategies used so far in the context of author name disambiguation can be regarded as agglomerative clustering algorithms (Ferreira, Gonçalves, and Laender 2012), especially in the form of single-link or average-link clustering. More specifically, single-link approaches define the similarity of two clusters of author mentions as the maximum similarity of all pairs of author mentions belonging to the different clusters. The idea behind this technique is that each of an author's publications is similar to at least one of his or her other publications. In average-link approaches, on the other hand, the two clusters with the highest overall cohesion are merged in each step, i.e. all objects in the clusters are considered (in contrast to just one from each cluster in single-link approaches). This rests on the assumption that an author's publications form a cohesive entity. As a consequence, it is easier to distinguish between two authors with slightly different oeuvres compared to single-link approaches, but heterogeneous oeuvres by a single author are more likely to be split.

Previous author name disambiguation approaches have usually been evaluated in terms of their quality. This evaluation is always based on measuring how pure the detected clusters are with respect to real-world authors (precision) and how well the author mentions of real-world authors are merged in the detected clusters (recall). However, different metrics have been applied when assessing these properties. Furthermore, different datasets have been used to evaluate author name disambiguation approaches (Kim 2018). It is therefore difficult to compare different approaches based on their previous evaluations.

Approaches compared

We focused on unsupervised disambiguation approaches in our analyses (see above). Since these approaches require no training data to be provided a priori, they are more convenient for use with real-world applications. Furthermore, narrowing the set of approaches down to unsupervised ones facilitates their comparison, whereas more aspects have to be considered if they are compared with supervised approaches (e.g., the quality of the training data, which type of supervised model is chosen), making this kind of a comparison more incomprehensible. We chose four approaches in addition to a naïve approach, which only considers the canonical representation of author names. These were selected to cover a wide variety of features that characterize unsupervised approaches for disambiguating author names. We applied the approaches to data from the Web of Science (WoS, Clarivate Analytics) that had already been pre-processed according to a blocking strategy, as described above. More precisely, all author mentions that share the author name representation specified by surname and first initial of the first name have been assigned to the same block. Therefore, all author mentions referring to one real-world author should be in one of these blocks, but there may be several authors represented by one name block. However, there were already some splitting errors in the blocking step (e.g. spelling errors, errors due to name changes).

Implementation of the four selected disambiguation approaches

(1) Cota, Gonçalves and Laender (2007) proposed a two-step approach that considers the names of co-authors, publication titles and journal titles. In a first step, all pairs of author mentions that share a co-author name are linked. The linked author mentions are then clustered by finding connected components with regard to this matching. The second step iteratively merges these clusters if they are sufficiently similar with respect to their publication titles or journal titles. The similarity of two clusters (one for publication titles, one for journal titles) is defined as the cosine similarity of the two TF-IDFs (term frequency-inverse document frequency) for the clusters' publication titles (or journal titles). Two clusters are merged if one of their similarities (either with regard to publication titles or to journal titles) exceeds a predefined threshold. This process continues until there are no more sufficiently similar clusters to merge, or until all author mentions are merged into one cluster.

(2) Schulz et al. (2014) proposed a three-step approach based on the following metric for the similarity s_{ij} between two author mentions i and j:

$$s_{ij} = \alpha_A \left(\frac{|A_i \cap A_j|}{\min(|A_i|,|A_j|)}\right) + \alpha_S\left(|p_i \cap R_j| + |p_j \cap R_i|\right) +$$
$$\alpha_R\left(|R_i \cap R_j|\right) + \alpha_C \left(\frac{|C_i \cap C_j|}{\min(|C_i|,|C_j|)}\right) \quad \text{(I)}$$

Here, A_i denotes the co-author list of paper i, R_i its reference list and C_i its set of citing papers. The first step links all pairs of author mentions with a similarity (determined by formula (I)) exceeding a threshold β_1 and a set of clusters is determined by finding the corresponding connected components. In the second step, these clusters are merged in a very similar way. In order to determine the similarity $S_{\gamma\kappa}$ of two clusters γ and κ, the similarities between author mentions within these clusters are combined by means of the following formula:

$$S_{\gamma\kappa} = \sum_{i\in\gamma j\in\kappa} \frac{s_{ij}\Theta(s_{ij})}{|\gamma||\kappa|}, \; \Theta(s_{ij}) = \begin{cases} 1 \; if \; s_{ij} > \beta_2 \\ 0 \; if \; s_{ij} \leq \beta_2 \end{cases} \quad \text{(II)}$$

Here, |γ| denotes the number of author mentions in cluster γ (similarly for cluster κ). As the formula shows, only those similarities between author mentions that exceed a threshold β_2 are considered when calculating the similarity between two clusters. As in the first step, this cluster similarity is used to link clusters if they exceed another threshold β_3 in order to find the corresponding connected components. The third step of this approach finally adds single author mentions that have not been merged to a cluster in either of the first two steps, provided its similarity with one of the cluster's author mentions exceeds a threshold β_4.

(3) Caron and van Eck (2014) proposed measuring the similarity between two author mentions based on a set of rules that rely on several paper-level and author-level characteristics. More precisely, a score is specified for each rule, and all of the scores for matching rules are added up to an overall similarity score for the two author mentions (see Table 2). If two author mentions are sufficiently similar with regard to this similarity score, they are linked and the corresponding connected components are considered oeuvres of real-world authors. The threshold for determining whether two author mentions are sufficiently similar depends on the size of the corresponding name block. The idea behind this approach is to take into account the higher risk of false links in larger blocks. Higher thresholds are therefore used for larger blocks to reduce the risk of incorrectly linked author mentions.

Table 2. Rules for rule-based scoring proposed by Caron and van Eck (2014)

Field	Criterion	Score
Email	exact match	100
All initials, more than one	exactly two matching initials	5
	more than two matching initials	10
	conflicting initials	-10
First name	matching general name	3
	matching non-general name	6
Address (linked to author)	matching country and city	4
Co-authors	one shared co-author	4
	two shared co-authors	7
	more than two shared co-authors	10
Grant number	at least one shared grant number	10
Address (linked to publication, but not linked to author)	matching country and city	2
Subject category	matching subject category	3

Journal	matching journal	6
Self-citation	at least one publication citing the other	10
Bibliographic coupling	exactly one shared cited reference	2
	exactly two shared cited references	4
	exactly three shared cited references	6
	exactly four shared cited references	8
	more than four shared cited references	10
Co-citation	exactly one shared citing reference	2
	exactly two shared citing references	3
	exactly three shared citing references	4
	exactly four shared citing references	5
	more than four shared citing references	6

(4) Backes (2018) proposed an approach that starts by considering each author mention as one cluster. An agglomerative clustering algorithm is then employed that iteratively merges clusters if they are sufficiently similar, i.e. two clusters are connected if their similarity exceeds a quality limit l. The similarity metric indicating how similar two clusters are takes into account the specificity of the author mentions' metadata. For example, if two author mentions share a very rare subject category this might be a stronger indicator of the author mentions for the same author compared to a very common subject category. This strategy is applied to compute a similarity score for each characteristic under consideration. When using this approach in our study, we considered the following characteristics: titles, abstracts, affiliations, subject categories, keywords, co-author names, author names of cited references, and email addresses. Backes (2018) proposed several variants to combine these scores into a final similarity score of two clusters. In the variant implemented in this study, the scores are combined in the form of a linear combination with equal weights for all characteristics' scores. Each iteration of the clustering process merges all pairs of current clusters whose similarity exceeds l. The quality limit l is designed to have a linear dependence on the block size $|author\ mentions|$, whereby the parameter λ specifies this relationship (see formula (III)).

$$l = \lambda \cdot |author\ mentions| \qquad \text{(III)}$$

Parameter specification

Some form of threshold (or a set of thresholds) has to be specified for each of the four approaches. Since such thresholds have not been proposed for all approaches by the authors, and some of the proposed thresholds produce poor results for our dataset, we fitted them with regard to our data. This allows a better comparability since the thresholds are matched to the particular datasets they are applied to. Our procedures for specifying the thresholds maximize the evaluation metrics $F1_{pair}$ and $F1_{best}$ (see below).

We specified such a procedure for each of the approaches that allowed an efficient consideration of a wide range of thresholds. A set of thresholds uniformly distributed over the complete parameter space was chosen as candidate set for the approach of Cota, Gonçalves and Laender (2007). We also specified the thresholds for the approach of Schulz et al. (2014) by evaluating a candidate set of parameters; in this case, the candidate set of thresholds was chosen on the basis of the parameters proposed in the original paper. The parametrization of this approach was further optimized by fitting β_1, β_2 and β_3 independently from β_4. β_4 was subsequently chosen based only on the best combination of the other thresholds, which substantially reduces the search space. We believe this to be an adequate procedure for finding the thresholds since the last step of this disambiguation approach (which is based on β_4) has only a minor influence on the final result. For the approach proposed by Caron and van Eck (2014) we initially had to define the block size classes that divide the blocks into several classes with regard to the number of author mentions in them. Similar to Caron and van Eck (2014), we defined six block size classes. Our specification of the classes aims at reducing the variance of optimal thresholds within a class.

For the approach of Backes (2018), we had to modify the approach slightly in order to define a feasible procedure for fitting the parameter λ, which determines the quality limit l for a given block. Instead of linking all pairs of clusters whose similarity exceeds a given l in each iteration, we iteratively merged only those pairs of clusters whose similarity equals the maximum similarity of all current pairs of clusters (the clusters are recomputed after each merger). These similarities were taken as estimates for the quality limit that would yield the clustering of the corresponding merger step. This modification may produce results that are different to the original approach, since the order in which the author mentions are merged may change and the similarities between clusters depend on the previous mergers. However, we assume that these changes would produce only minor differences that do not influence any general conclusions on the approach. Our implementation merges the most similar clusters in each iteration, i.e. the most reliable mergers are applied iteratively until the quality limit is reached. Correspondingly, the original approach follows the idea that all cluster similarities exceeding a certain quality limit indicate reliable links between the corresponding clusters.

Data

We collected metadata for a subset of author mentions from the WoS for our analyses. In order to provide a gold standard that represents sets of author mentions corresponding to real-world authors, we only took author mentions with a ResearcherID linked to them in the WoS into account. More specifically, all person records that are marked as authors and that have a ResearcherID linked to at least one paper published in 2015 or later have been considered. It is very likely that this procedure excludes all author mentions with ResearcherIDs referring to non-author entities (e.g. organizations) and takes into account only such ResearcherIDs that have been maintained recently. We applied the same standardization for all name-based metadata as was used to block author mentions, i.e. a canonical name representation is used consisting of first name initial and surname. We only considered name blocks comprising at least five real-world authors. This selection allowed us to focus on rather difficult cases where the author mentions in a block actually have to be disambiguated across several authors. All in all, this data collection procedure results in 1,057,978 author mentions distributed over 2,484 name blocks and 29,244 distinct ResearcherIDs. The largest name block ("y. wang") comprises 7,296 author mentions.

Results

Evaluation metrics

The evaluation of author name disambiguation approaches is generally based on assessing their ability to discriminate between the author mentions of different real-world authors (precision) and their ability to merge the author mentions of one real-world author (recall). Even though these concepts are widely accepted and referenced, different specific evaluation metrics have been used in the past. In the following, we focus on two types of evaluation metrics. First, we calculate the pairwise precision (P_{pair}), pairwise recall (R_{pair}) and pairwise F1 ($F1_{pair}$) (Levin et al. 2012; Caron and van Eck 2014; Backes 2018) for each of the approaches. Whereas the pairwise precision measures how many of the links between author mentions in the detected clusters are correct, the pairwise recall measures how many of the links between author mentions of real-world authors are correctly detected. Pairwise F1 is the harmonic mean of these two metrics. In formulae (IV)-(VI), $pairs_{author}$ denotes the set of pairs of author mentions where both of the author mentions refer to the same author, and $pairs_{cluster}$ denotes the set of pairs of author mentions where both author mentions are assigned in the same cluster by the disambiguation algorithm. Each of the pairwise evaluation metrics can take values between 0 (no true links between author mentions detected) and 1 (all true links between author mentions detected).

$$P_{pair} = \frac{|pairs_{author} \cap pairs_{cluster}|}{|pairs_{cluster}|} \tag{IV}$$

$$R_{pair} = \frac{|pairs_{author} \cap pairs_{cluster}|}{|pairs_{author}|} \tag{V}$$

$$F1_{pair} = \frac{2P_{pair}R_{pair}}{P_{pair}+ R_{pair}} \tag{VI}$$

Second, we calculate metrics to measure how reliably a cluster can be attributed to one specific author (best precision P_{best}) and how well an author can be attributed to one specific cluster (best recall R_{best}).

More specifically, the best precision represents the fraction of author mentions that refer to the most represented author in the corresponding cluster. The most represented author of a cluster is defined as the author with the largest group of author mentions in this cluster. Accordingly, the best recall represents the fraction of author mentions that are assigned to the cluster with the most author mentions of the corresponding author. Similar to the pairwise F1, the best F1 $F1_{best}$ combines best precision and best recall in the form of their harmonic mean. In formulae (VII)-(IX), $author\ mentions_{best\ author}$ denotes the set of author mentions referring to the author most of the corresponding cluster's author mentions refer to, $author\ mentions_{best\ cluster}$ denotes the set of author mentions assigned to the cluster with the most author mentions of the corresponding author and $author\ mentions$ denotes the set of all author mentions. Technically speaking, P_{best}, R_{best} and $F1_{best}$ can also take values between 0 and 1. However, $author\ mentions_{best\ author}$ and $author\ mentions_{best\ cluster}$ will always contain at least one author mention. Actually, these evaluation metrics will thus always be greater than 0.

$$P_{best} = \frac{|author\ mentions_{best\ author}|}{|author\ mentions|} \tag{VII}$$

$$R_{best} = \frac{|author\ mentions_{best\ cluster}|}{|author\ mentions|} \tag{VIII}$$

$$F1_{best} = \frac{2P_{best}R_{best}}{P_{best}+ R_{best}} \tag{IX}$$

Each of these formulae can either be applied to the complete dataset or to a subset of author mentions. For example, the results of one name block can be evaluated by only considering author mentions within this block when computing the evaluation metrics.

Overall results

The results for the approaches described above are summarized in Table 3. The table shows the evaluation metrics described in the previous section for all of the approaches. All of the approaches produce better results than the naïve baseline disambiguation. The approach proposed by Caron and van Eck (2014) performs best among the examined approaches with regard to both $F1_{pair}$ and $F1_{best}$. If one compares the approaches of Schulz et al. (2014) and

Backes (2018), the two evaluation metrics yield different rankings. Whereas the latter approach performs better with regard to $F1_{pair}$, the first performs better with regard to $F1_{best}$. Finally, the approach of Cota, Gonçalves, and Laender (2007) performs only slightly better than the baseline disambiguation. The precision in particular is very low in this case, due mainly to a high number of false links between author mentions in the first step (merging author mentions with shared co-authors).

Table 3. Overall results for all approaches

Approach	P_{pair}	R_{pair}	$F1_{pair}$	P_{best}	R_{best}	$F1_{best}$
Baseline	0.095	1.000	0.173	0.322	1.000	0.487
Cota, Gonçalves, and Laender (2007)	0.111	0.858	0.196	0.442	0.913	0.596
Schulz et al. (2014)	0.453	0.457	0.455	0.799	0.750	0.773
Caron and van Eck (2014)	0.831	0.787	0.808	0.916	0.885	0.900
Backes (2018)	0.674	0.622	0.647	0.761	0.699	0.729

Figure 1 shows the distribution of the disambiguation quality over block sizes (the mean of all blocks of a specific size is plotted). This distribution is shown for the case where the thresholds are specified as described above ("original") and for the case where the optimal thresholds for each single block are used ("flexible"). The results reveal that the disambiguation quality varies strongly across name blocks. The quality generally worsens for large blocks. Therefore, the disambiguation process may produce biases with regard to the frequency of the corresponding name representation. One reason for the disambiguation quality's dependence on the size of the name block is the larger search space to find clusters of author mentions. This increases the search complexity in general, implying a greater potential for false links between author mentions. Some approaches try to reduce this problem by allowing for block size dependent thresholds (see next section). Even though the negative relationship between block size and disambiguation quality can be observed for all approaches, the decline in quality is not equal in all of them. Especially for the approach of Caron and van Eck (2014), the influence of the block size is relatively small.

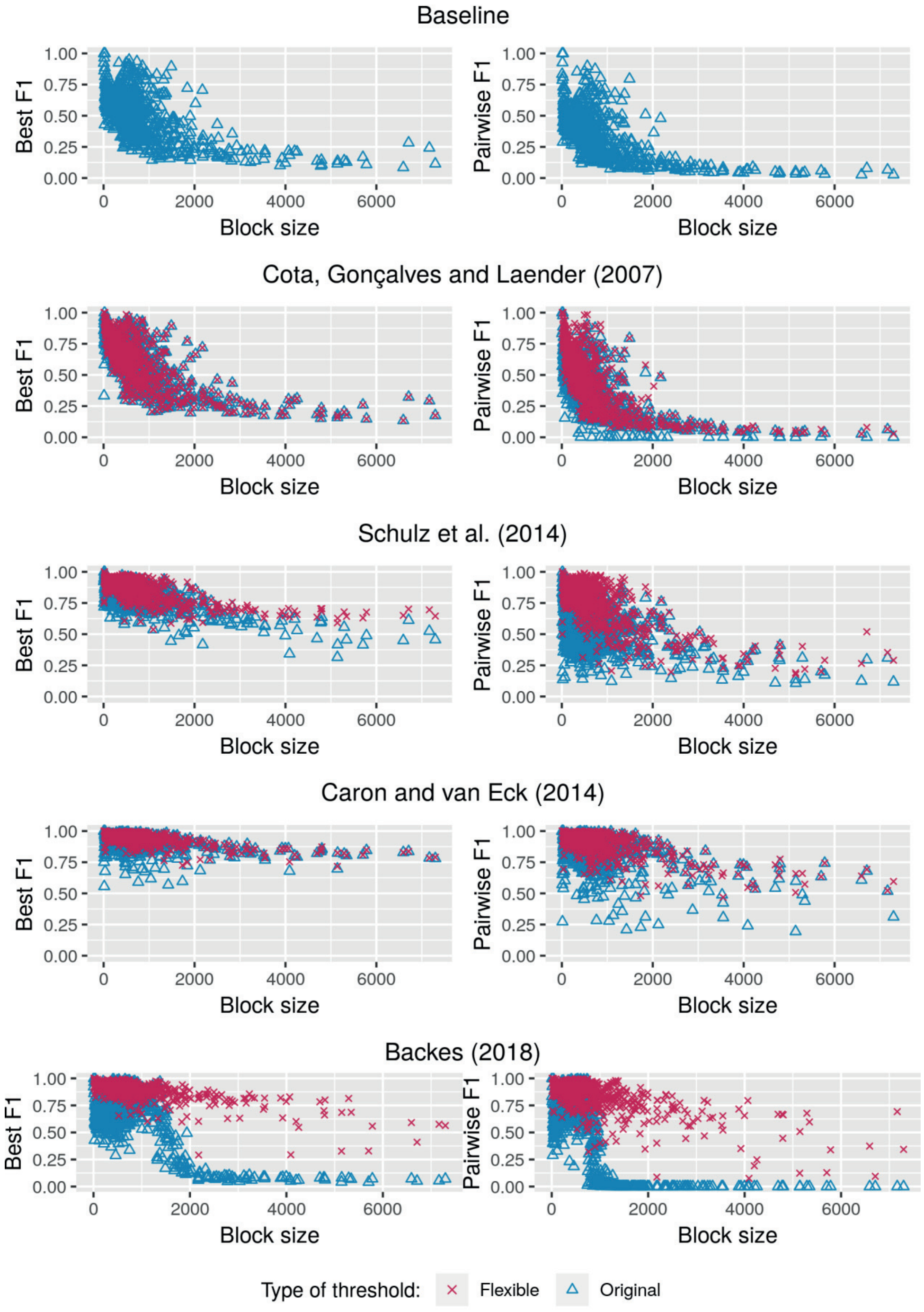
Influence of parametrization on the disambiguation quality

Among the approaches included in our comparison, Caron and van Eck (2014) and Backes (2018) used block size dependent thresholds. As described above, the first approach is based on defining one threshold for each of six block size classes, whereas the threshold is linearly dependent on the block size in the second approach. Table 4 shows the block size classes and corresponding thresholds used by our implementation for the approach of Caron and van Eck (2014). In contrast, the approaches of both Cota, Gonçalves, and Laender (2007) and Schulz et al. (2014) use global thresholds for all block sizes.

Table 4. Block size classes and thresholds for Caron and van Eck (2014)

Block size	Threshold ($\mathbf{F1_{pair}}$)	Threshold ($\mathbf{F1_{best}}$)
1-500	21	19
501-1000	22	21
1001-2000	25	23
2001-3000	27	25
3001-4500	29	25
>4500	29	27

Figure 1. Distribution of disambiguation quality over block sizes



In Figure 1, the results based on optimal thresholds for each single block (flexible thresholds) represent an upper bound for the quality over all possible thresholds. Flexible thresholds would not greatly improve the quality of the approach of Cota, Gonçalves, and Laender (2007) since the results based on global thresholds are very close to the results based on completely flexible thresholds. The reason for this is that the quality is dominated by the first step of the approach, which does not employ any threshold at all. The second step, on the other hand, does not change the results significantly, so that the effect of the thresholds is rather small. In contrast, the approach of Schulz et al. (2014) could benefit from using flexible thresholds, especially for large blocks.

Similar to the approach of Cota, Gonçalves, and Laender (2007), the difference between the original implementation and the one with flexible thresholds is rather small for the approach of Caron and van Eck (2014). However, the choice of thresholds does affect the result in this case, as shown by the comparison with an implementation based on a constant threshold for all block sizes. Table 5 shows the evaluation results for the approach of Caron and van Eck (2014) with three different types of thresholds: a constant threshold for all blocks ("Constant"), the thresholds of the block size classes shown in Table 4 ("Block size classes"), and the optimal threshold for each single block ("Flexible"). These results show that the original implementation produces better results than those obtained using a constant threshold. This means that the somewhat rough partitioning between six block size classes already allows for an adequate differentiation with regard to the threshold, and that this strategy improves the disambiguation result compared to a constant threshold over all block sizes. In contrast, the strategy of specifying a threshold which is linearly dependent on the block size, as employed by the approach of Backes (2018), is unable to find good thresholds over the complete range of block sizes. This is due mainly to a drop in the recall (together with an increasing precision) for large blocks. The thresholds chosen by the algorithm are thus too high for large blocks. Hence, a linear relationship between block size and threshold would not appear to be an adequate strategy for large blocks. The fitted thresholds for the approach of Caron and van Eck (2014) also confirm that a nonlinear relationship between block size and threshold may be more suitable.

Table 5. Results for different types of thresholds for Caron and van Eck (2014)

Type of threshold	P_{pair}	R_{pair}	$F1_{pair}$	P_{best}	R_{best}	$F1_{best}$
Constant	0.690	0.741	0.714	0.879	0.880	0.880
Block size classes	0.831	0.787	0.808	0.916	0.885	0.900
Flexible	0.907	0.850	0.878	0.954	0.897	0.924

The results in Figure 1 and Table 5 demonstrate that the disambiguation quality can be improved if flexible thresholds dependent on the block size are specified. However, the specification of adequate thresholds is generally a non-trivial task since it depends on the data at hand. Likewise, the thresholds proposed previously for the approaches examined in this paper do not correspond to the thresholds fitted with regard to our dataset.

Discussion

The disambiguation of units (researchers, research groups, institutions etc.) for bibliometric analyses is an important topic in research evaluation. The results of evaluation studies can only be as good as the underlying data. For example, Clarivate Analytics annually publishes the names of highly cited researchers who have published the most papers belonging to the 1% most highly cited in their subject categories (see https://hcr.clarivate.com). The reliable attribution of papers to corresponding researchers is an absolute necessity for publishing this list of researchers. Although different disambiguation approaches have been developed and implemented in local bibliometric databases (e.g., Caron and van Eck 2014), there is hardly any comparison of the approaches. However, this comparison is necessary to obtain indicators of the best approaches, or those conditions on which the performance of the approaches depends. In this paper, we compared different author name disambiguation approaches based on a dataset containing author identifiers in the form of ResearcherIDs. This allows a better comparison of different approaches than if previous evaluations are used since these are generally based on different databases. Our results show that all of the approaches included in the comparison perform better than a baseline that only uses a canonical name representation of the authors for disambiguation. Although the comparison does not point to the recommendation of one approach for all disambiguation tasks, it does provide evidence of when which approach can produce good results – especially with regard to the size of corresponding name block sizes. As our analyses show, the parametrization of the approaches can have a significant effect, which depends largely on the data at hand. Therefore, the context of the disambiguation task has to be taken into account for a proper implementation of an algorithm. In the context of this study, the approach proposed by Caron and van Eck (2014) produced the best results.

Future research should further examine how different author name disambiguation approaches behave and how certain features affect the disambiguation results. For example, the set of characteristics used by the approaches may play an important role. Since the approaches included in our comparison use different sets of characteristics, differences in the results may be due in part to the choice of the characteristics used. A more detailed analysis of this choice in future studies may shed more light on which set of characteristics is most suitable for which context.

Understanding how author name disambiguation approaches behave is important in order to improve the algorithms and to assess the effect they have on analyses building on the disambiguated data. A good understanding of the behaviour is the basis for reliable analyses at the individual level.

Acknowledgments

The bibliometric data used in this paper are from an in-house database developed and maintained in cooperation with the Max Planck Digital Library (MPDL, Munich) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics, formerly the IP & Science business of Thomson Reuters (Philadelphia, Pennsylvania, USA). We would like to thank Robin Haunschild and Thomas Scheidsteger from the Central Information Service for the institutes of the Chemical Physical Technical (CPT) Section of the Max Planck Society (IVS-CPT) for providing the computational infrastructure for conducting our analyses.

References

Backes, Tobias. 2018. "Effective Unsupervised Author Disambiguation with Relative Frequencies." In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18*, 203-12.

Caron, Emiel, and Nees Jan van Eck. 2014. 'Large scale author name disambiguation using rule-based scoring and clustering', *19th International Conference on Science and Technology Indicators*.

Cota, Ricardo G., Marcos André Gonçalves, and Alberto H. F. Laender. 2007. "A Heuristic-based Hierarchical Clustering Method for Author Name Disambiguation in Digital Libraries." In *XXII Simpósio Brasileiro de Banco de Dados*.

Ferreira, Anderson A., Marcos André Gonçalves, and Alberto H. F. Laender. 2012. 'A Brief Survey of Automatic Methods for Author Name Disambiguation', *ACM SIGMOD Record*, 41.

Ferreira, Anderson A., Adriano Veloso, Marcos André Gonçalves, and Alberto H. F. Laender. 2014. 'Self-training author name disambiguation for information scarce scenarios', *Journal of the Association for Information Science and Technology*, 65: 1257-78.

Ferreira, Anderson A., Adriano Veloso, Marcos André Gonçalves, and Alberto H.F. Laender. 2010. "Effective self-training author name disambiguation in scholarly digital libraries." In *Proceedings of the 10th annual joint conference on Digital libraries*. Gold Coast, Queensland, Australia.

Kim, Jinseok. 2018. 'Evaluating author name disambiguation for digital libraries: a case of DBLP', *Scientometrics*, 116: 1867-86.

Kim, Jinseok. 2019. 'Scale-free collaboration networks: An author name disambiguation perspective', *Journal of the Association for Information Science and Technology*.

Kim, Jinseok, and Jana Diesner. 2016. 'Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks', *Journal of the Association for Information Science and Technology*, 67: 1446-61.

Levin, Michael, Stefan Krawczyk, Steven Bethard, and Dan Jurafsky. 2012. 'Citation-based bootstrapping for large-scale author disambiguation', *Journal of the American Society for Information Science and Technology*, 63: 1030-47.

Newcombe, Howard B. 1967. 'Record linking: The design of efficient systems for linking records into individual and family histories', *American Journal of Human Genetics*, 19.

On, Byung-Won, Dongwon Lee, Jaewoo Kang, and Prasenjit Mitra. 2005. 'Comparative Study of Name Disambiguation Problem using a Scalable Blocking-based Framework', *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*.

Schulz, Christian, Amin Mazloumian, Alexander M. Petersen, Orion Penner, and Dirk Helbing. 2014. 'Exploiting citation networks for large-scale author name disambiguation', *EPJ Data Science*, 3: 11.

Smalheiser, Neil R., and Vetle. I. Torvik. 2009. 'Author Name Disambiguation', *Annual Review of Information Science and Technology*, 43.

Torvik, Vetle I., and Neil R. Smalheiser. 2009. 'Author name disambiguation in MEDLINE', *ACM Transactions on Knowledge Discovery from Data*, 3: 1-29.

Open Access uptake of universities in the R-Quest countries under various OA mandates

Thed van Leeuwen [1] and Jesper Schneider [2]

[1] *leeuwen@cwts.nl*
Leiden University, Centre for Science & Technology Studies (CWTS), Kolffpad 1, Leiden (Netherlands)

[2] *jws@ps.au.dk*
University of Aaurhus, Department of Political Science - Danish Centre for Studies in Research and Research Policy, Bartholins Allé 7, building 1331, 029, 8000 Aarhus C, (Denmark)

Abstract

This paper presents an analysis of the uptake of OA publishing by universities in the Leiden ranking, distributed over five countries. The countries we study are part of a large scale supra-national study on research quality, in which a variety of methods is used. Linking policy initiatives to bibliometric results on OA publishing, based on Web of Science and enriched by Unpaywall data, leads to perspectives on the way the universities across the five countries have different foci on OA publishing. Finally, we will show which are the consequences of the launch of Plan S, an international funder driven policy initiative focused on a certain preferential way of OA publishing.

Introduction

In this paper we present the results of research conducted in the R-Quest project. R-Quest is a project funded by the Research Council Norway (RCN), in which the focus is on research quality as an element of research management and evaluation. The R-Quest project is in its third year, of a 5/8 year stretch. In this particular analysis, we will focus on the aspect of openness of research communication, in other words the uptake of Open Access publishing by the universities as represented in the Leiden Ranking, for the five countries in the R-Quest study (Norway, Denmark, Sweden, the Netherlands, and the UK). We will compare closed with OA published outputs, as well as make comparison between Gold, Green, Hybrid and Bronze OA. We will position these against the national mandates at play, in order to see to what extent universities behave mandate compliant. All of this has been complicated by last September's launch of Plan S, a policy initiative that intents to roll back the relevance of Hybrid OA publishing, as it does not seem to function as transitioning from Closed to Gold OA, as most journals (and thus their publishers) stick to the double dipping situation of toll access, with additional APC based OA publishing in these journals.

Rationale for this analysis

Given where we stand now, one can expect that notions of Open Science, and hence Open Access publishing, will further change the way scholarly communication patterns develop over the next 10-15 years. This means a re-orientation on scholarly work, involving doing research, data sharing, collaboration, publishing and reviewing, etc., but also might change the publishing world (mega-journals and platforms, instead of regular publishers and their journals, as well as the ways research and scholarly work is being funded., and this might all heavily affect the academic reward system. Given all of this, one might expect that Open Science and Open Access will affect notions of research quality, as studied in the project.

OA Policies: mandates of open access publishing

In order to be able to monitor and compare between the five countries in the study, and the universities in the Leiden Ranking, we have to contextualize the developments against the national policy background. In the case of analysis of OA uptake, that means introducing the national OA mandates. OA mandates can be defined as the policies adopted by research institutions, research funders, or government, requiring research communities (which is usually university faculty or research staff and/or research grant recipients) to make their published, peer-reviewed journal articles and conference papers available as Open Access (Harnad et al, 2004). Most Open Access (or nowadays Open Science) mandates distinguish various forms of OA publishing. Green OA publishing is a first variation, which consists of the (self-)archiving of final, peer-reviewed drafts in a freely accessible institutional or disciplinary repository. A second variation is Gold OA, which means publishing your manuscripts in fully open access journals, with or without paying APCs (Article Production Costs). A third variation, mostly not mentioned in OA mandates, is the hybrid form of publishing your manuscripts, in an open access format in an otherwise toll-access journal (this is called the hybrid format of OA publishing). A fourth type of OA publishing, which will be considered here is Bronze OA, a type first reported on by Piwowar et al (Piwowar et al, 2017), and further analysed on a large scale by Martin et al (Martin-Martin et al, 2018).

The five countries in the study have each their own, specific OA mandates. The Norwegian OA mandate is focused on Green OA, preferably via institutional repositories, with no explicitly defined goals in time to reach a full 100% OA Norwegian output. In the Netherlands, pressure is somewhat higher, as there the focus is on Gold OA, with science policy explicitly stating terms when the full 100% has to be reached: initially the planning spoke of 60% in 2018, and 100% in 2024, which is shortened to setting the 100% goal to be reached in 2020. This would apply on journal publishing only, as it is recognized that book publishing is somewhat more complex. In Denmark, like in Norway, the focus is mostly on Green OA publishing, irrespective in what type of repositories, with no hard goals set in time to reach a full 100% OA uptake. In Sweden, the national OA mandate is quite free, as there is no explicit preference for either Green or Gold OA, the only requirement relates to having a CC-BY license, which allows re-use of the information. There are no goals set in time, to reach a full 100% OA uptake by the Swedish research communities. However, books and chapters are exempted from this policy, as there is a strong realization that the change in business models with respect to that type of scholarly communication and hence the publishing of those forms of communication might take more time. Finally, in the UK the focus is on Gold OA, with the explicit requirement that in the next Research Excellence Framework (REF) exercise only (Gold) OA published material is to be accepted as underlying the assessment procedure. So here we clearly distinguish various perspectives on both the type of OA publishing, as well as the time path involved in which this has to materialize.

Recent policy developments involving OA publishing: Plan S

In September 2018, Science Europe launched Plan S, a policy initiative to further stimulate the publishing in open access format of results of scientific/scholarly work. The initiative was taken by a coalition of supranational and national funders, from a number of European countries. The plan basically prescribes the researchers that profit form a grant from their agencies, to publish the resulting outcomes in open repositories, platforms, or journals that are open and stay open, all to be realized by 2020. The plan mentions a number of principles, with the major, overarching principle being that research outcomes from public grants must be published in open access journals or platforms. An important incentive for this development was the dissatisfaction of policy with the pace of the OA uptake and development. The type of OA publishing that was intended to signal a transition period ("hybrid" OA), is seen as an attempt by the publishing industry to stall the process of transition. Therefore, hybrid OA has to further change into either Gold or Green OA. Consequently, any reactions from the publishing industry, to create mirror journals for the toll access journals is considered as non OA, and non-Plan S compliant, unless there are clear transformative agreements with the respective publishers for the future development of the respective journals into a further opening up.

Creating OA tags in the Web of Science

Until recently, the Web of Science (WoS) disclosed only in a limited sense OA labels related to indexed publications (van Leeuwen et al, 2018). Of a more recent nature is the implementation of the Unpaywall dataset on WoS. Before this current study, we have developed a methodology developed within CWTS, based upon freely accessible/no costs involved data sources, as we reported on in previous work (van Leeuwen et al, 2017)[1]. This methodology has a clear drawback, namely that it is not possible to distinguish within the OA published material, the Green OA published material from the Hybrid OA published material. However, an important advantage is the high level of control on how OA tagging of WoS indexed publications occurs. In the development of this method, we defined two criteria, namely that of ***sustainability*** (data are in the public domain, without immediate and direct risk of disappearing behind a pay-wall), and ***legality*** (inclusion in the data source should not be based on 'illegal acts' by individual researchers, so no copy right breaches involved).

The source we used for this analysis is Unpaywall. The inclusion of Unpaywall in our database of WoS indexed publications, will also lead to analyses to compare both methodologies, and the degree of complementarity in both directions, that is the direction of the methodology to tag OA developed by CWTS (however, that is beyond the goals of the current study and will be reported elsewhere). The implementation of Unpaywall would allow us to identify hybrid OA, which is extremely important in the light of discussing the consequences of Plan S, a recently launched policy initiative to take out hybrid publishing from the OA landscape, as it does not promote a further transition of academic publishing into a full OA situation, as well as Bronze OA. This type of OA publishing, in which the publishing industry initiated OA availability of published material, is only partially compliant to the criteria defined, and described above, namely compliant with the legality criteria (as no copy rights are breached, these remain with the publisher instead of the authors), but not compliant with the sustainability criteria (as the publishers can decide at any moment what contents will be available OA). It is important to note that the other types of OA publishing, Gold, Green and Hybrid all involve a certain degree of engagement or involvement of the academic actors in the science system (e.g., researchers/authors, librarians). A description of the methodology on tagging the publications in the Leiden Ranking can be found at the CWTS Blog (van Leeuwen et al, 2019).

Data on Leiden Ranking of universities for the R-Quest countries and the universities

Within CWTS, we can access the list of universities that are presented via the Leiden Ranking. In this case, we selected for the five countries in this study all publications related to these universities, and could do a re-run of the basic indicators available in the CWTS tool box of research metrics. Analyses focusing on the OA uptake within the institutional setting

[1] The sources we used in the study are the DOAJ and the ROAD lists of Gold OA journals, which together span the universe of Gold OA publishing, while the CrossRef, PubMedCentral, and OpenAIRE databases functioned as sources to determine Green OA in the WoS database. In linking sources to the database, we used various identifiers, such ISSNs for the Gold OA journals, DOIs for the other three databases, PubMedIDs for PubMedCentral data, while we also applied a fuzzy matching algorithm to link OpenAIRE to WoS. These five sources complied with both our criteria of sustainability and legality. Two sources that did not comply with our second required criteria are SciHub and ReserachGate, as these sources consistently show breaches of copy right.

have not been conducted in similar comparative fashion before until recently (Bosman & Kramer, 2018). In our analysis we will use the national mandated situation as a proxy for the institutional settings, rather than compare the institutional attitude towards OA for every single university in our analysis with their OA publishing uptake. We worked with three indicators: the number of publications, which is actually the number of normal articles, reviews, and letters as processed for journals covered in the Web of Science database, in the period 2014-2017/2018. On the basis of this set of selected publications, we calculated mncs and mnjs scores (Waltman et al, 2012). MNCS stands for the comparison of the real impact of a set of publications with expected citation scores, based upon output similarity in the exact same fields, years, and documents, while mnjs stands for the comparison of the impact of journals selected for publishing with the expected field citation impact, based upon output similarity in the exact same fields, years, and documents. We selected mnjs as well, not being a standard indicator in the Leiden Ranking, as this indicator informs us about the choice of journals with a certain prestige/reputation/standing in the fields to which these journal belong.

Results

To start the description of results, we first need to describe how the level of analysis is constructed. In Table 1, the second column the number of universities from the respective countries in the Quest study, as represented in the 2018 Leiden Ranking (LR 2018). We refer to these in an aggregated manner, by using the label "national/institutional" as the variable that represents at the national level in each of the five countries, for only these LR 2018 universities.

The next columns in Table 1 contains the aggregate numbers of publications for the universities in the study, in the various formats in which the output can appear in the journals as processed for the WoS. So we first present the Closed format output, followed by Gold OA, Green OA, Hybrid OA, and bronze OA. As stated before, we do not consider Bronze as a sustainable type of OA publishing, we report it though for reasons of completeness.

Table 1: Overview of numbers of universities in the study, and the national/institutional distribution of output over types of OA, 2014-2017/2018

	Nr Leiden Ranking universities	*Closed*	*Gold*	*Green*	*Hybrid*	*Bronze*
Denmark	*5*	*27108*	*8971*	*27634*	*5064*	*6839*
Netherlands	*13*	*64740*	*19946*	*61418*	*18623*	*15475*
Norway	*5*	*16254*	*6126*	*15305*	*3801*	*3286*
Sweden	*11*	*44418*	*15094*	*41375*	*11814*	*9653*
United Kingdom	*48*	*106599*	*45711*	*193623*	*55493*	*30178*

In Table 2, we present the relative situation, the shares per type of publishing for the five countries LR 2018 covered universities. Please note that due to choosing for the full picture, whereby overlapping forms of OA publishing are made visible (see van Leeuwen et al, 2019), percentages do not sum up to 100%, due to the double counting of publications mostly under Green OA. The share of OA published or available OA output for the UK is clearly higher as compared to the other four countries.

Table 2: Overview of the national/institutional distribution of shares of the output over types of OA, 2014-2017/2018

	Closed	*Gold*	*Green*	*Hybrid*	*Bronze*
Denmark	*46%*	*15%*	*46%*	*9%*	*11%*
Netherlands	*47%*	*14%*	*44%*	*13%*	*11%*
Norway	*47%*	*18%*	*44%*	*11%*	*10%*
Sweden	*47%*	*16%*	*44%*	*12%*	*10%*
United Kingdom	*33%*	*14%*	*60%*	*17%*	*9%*

If we look at Figure 1, which presents the contents of Table 2 in a graphical manner, we clearly observe that all five countries have roughly 50% or more of the output in OA format published or available in 2014-2017. We carefully mention here ’available’, as we want to make clear by choosing this wording to indicate that with respect to Bronze OA, this is not authors’ choice, but rather publishers’ choice, while the other OA format types are clearly an effect of authors’ choice. Given the fact that this involves some 10% of the national/institutional outputs of the countries involved, this is not a neglible part of the national/institutional outputs in OA format.

Among the five countries in this study, the largest share of OA published/available output is observed for the UK. Some 60% of all output in 2013-2016 are in some form of OA format, compared to some 50% for the other countries and their LR 2018 covered universities.

A rather surprising outcome is the low share of output published via the Gold OA route: all five countries cover about 10/15 % of the national/institutional output in this OA type, which is all the more remarkable since two of the five countries have made this explicit target in their national OS/OA mandates, namely the UK and the Netherlands. And although Green OA is not the explicit preference for the UK, we observe for the UK the largest share of their national/institutional output as published via the Green OA route (60%). A more or less equal situation exists for Denmark, although here an explicit reference was made to Green OA in the national Danish OS/OA mandate (18%). For the other three countries, Green OA is equally developed (about 45 %). For all five countries and their universities studied, we observe a strong preference for Hybrid OA publishing, a situation we will delve into more deeply in the next section. Finally, what is made available via the publishers as Bronze OA varies around 10% for the five countries in the study.

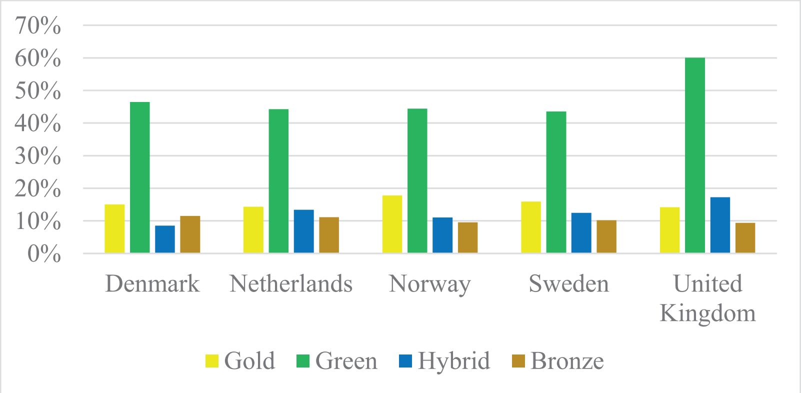


Figure 1: National/institutional distribution of output over types of OA publishing, 2014-2017/2018

If we shift our attention towards the way the five countries in the study have taken up the so called CWTS compliant OA forms, as displayed in Figure 2), we clearly see that for four countries (Denmark, the Netherlands, Norway, and Sweden), the share of the total publication output is varying around 45%, while for the UK this is around 60% in the period 2014-2016/17. Please note that this consists of Gold OA, Green OA and Hybrid OA only.

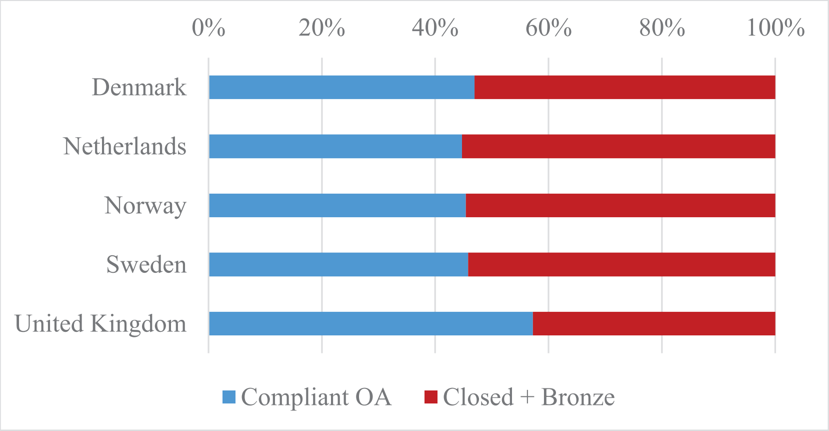


Figure 2: National/institutional distribution of output over OA vs non OA publishing, 2014-2017/2018

In Figure 3, we apply the philosophy of Plan S, thereby taking out Hybrid OA format published materials, as not compliant with Plan S, and thereby no longer policy relevant when in comes to understanding the uptake of OA publishing in the universities in the five countries in the study. We then notice a slight decrease in shares of OA published outputs from the country/institutional setting, a situation that sets back the shares of OA publishing to varying around 40%, with the Netherlands, Norway and Sweden slightly above that 40%, and Denmark around 45%. The UK then ends up with some 55% of the national/institutional output being published in Plan S compliant journals.

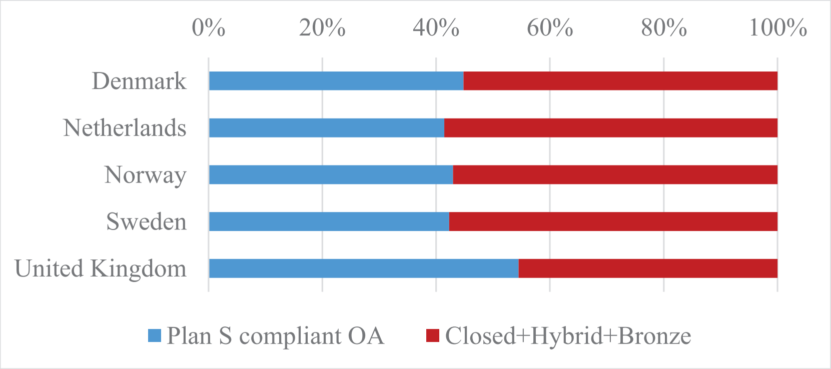


Figure 3: National/institutional distribution of output over Plan S compliant OA vs non OA publishing, 2014-2017/2018

What is important to note here, is the relative small difference between the two types analyses, as we clearly observe that the proportion of Hybrid and Bronze OA is decreasing, while on the other hand Green OA is increasing. Green OA overlaps strongly with these two types, so in that respect, the effect of the implementation of Plan S will have little effect on the overall uptake of OA publishing, while still keep in mind that these figures are mainly influenced by the increase of Green OA, rather than Gold OA (the main goal of Plan S, full and immediate access via Gold OA journals).

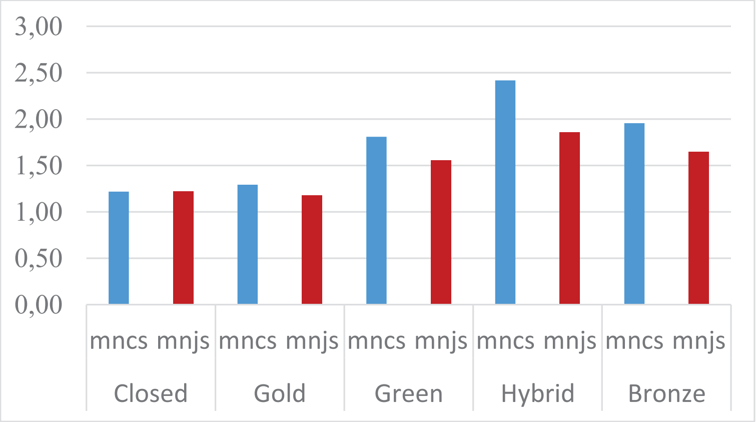


Figure 4a: Impact scores (mncs, mnjs) related to national/institutional distribution over various OA output types for Denmark, 2014-2017/2018

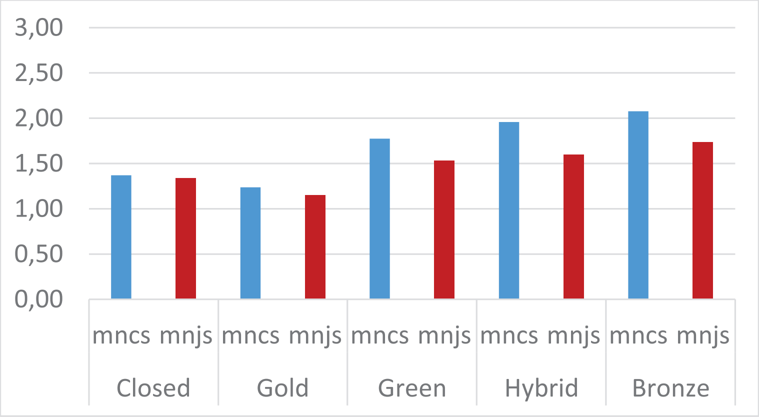


Figure 4b: Impact scores (mncs, mnjs) related to national/institutional distribution over various OA output types for the Netherlands, 2014-2017/2018

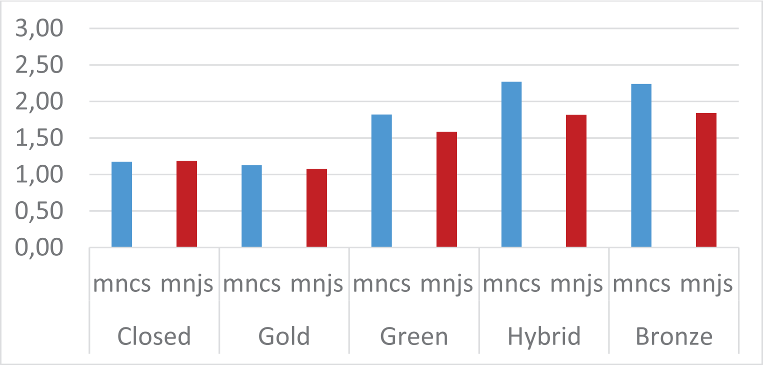


Figure 4c: Impact scores (mncs, mnjs) related to national/institutional distribution over various OA output types for Norway, 2014-2017/2018

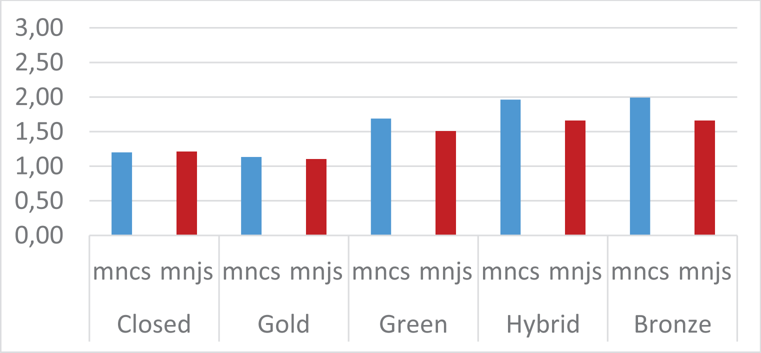


Figure 4d: Impact scores (mncs, mnjs) related to national/institutional distribution over various OA output types for Sweden, 2014-2017/2018

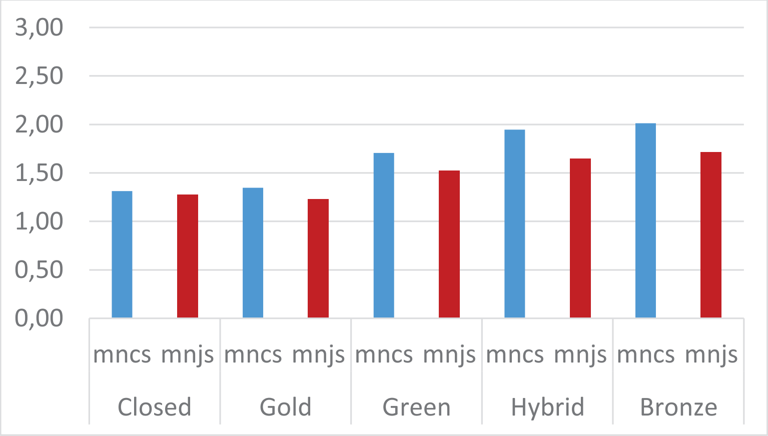


Figure 4e: Impact scores (mncs, mnjs) related to national/institutional distribution over various OA output types for the UK, 2014-2017/2018

If we then add citation impact measures to the analysis, we can create overviews as displayed in Figure 4a-4e. In these figures 4a-4e we present for the five countries the national/institutional citation impact analyses, by type of publishing. In each graph we present the output that was published in non OA format ("Closed"), as well as the Bronze OA output, the Gold, Green and Hybrid OA format output. The impacts score mncs (blue bar) expresses the actual normalized impact over the whole body of publications per type per country/institutional set of publications, while mnjs (red bar) depicts the level of the impact of the journals which were chosen for publishing, compared to the field(s) into which these journals belong.

A first observation relates to the impression of overall impacts. The five graphs clearly show that outputs from the universities as covered by the LR2018 all produce impact levels that are, even considered across all types of publishing, as above worldwide average impact level (with the value one representing worldwide average impact level). A next observation is that in all five countries, Closed published outputs compete with Gold OA for retrieving lowest impact values. This clearly indicates that both the actual impact, as well as the journals chosen for this set publications, is not having the highest standard, with the Netherlands and Sweden as cases in which the impact of Gold OA published outputs is even somewhat lower as compared to the Closed published output. A third observation is that Hybrid and Bronze OA published outputs are reaching high impact scores on mncs. Denmark is the country where Hybrid OA published output reaches the highest impact level, both overall as for the journals selected to publish in. For Sweden, Norway and the UK, the impact levels on mncs are equally high for Hybrid and Bronze OA, while for the Netherlands we observe that Bronze OA output reaches highest impact level. A fourth observation relates to the impact levels of the journals in which the Green and Hybrid OA format output was published: for the Netherlands, Sweden and the UK these reach mnjs values varying around the value of 1.5 for both Green and Hybrid OA. For Denmark and Norway, the differences between mnjs values of Green and Hybrid OA outputs are somewhat larger. These high impact levels for these two types of OA publishing are important, as this is indicative of high impact level journals in the fields to which these belong, an observation which would probably correspond to these journals being published in journals with relative high JIF values (van Leeuwen and Moed, 2002). This is important, as the JIF plays an important role in the discussion in Plan S about perceived journal quality (expressed by high JIF values), and the way that might influence academic careers. A final observation relates to the Bronze OA type of publishing. Here we clearly see that this publisher initiative based OA form is having high impact scores. Clearly, publishers use these publications to showcase their journals, and the strength these journals have in the respective fields these journals belong to. In all five cases, the displayed outputs in Bronze OA format have the highest impact scores, compared to the other four types of academic outputs.

Discussion

This study shows the uptake of OA publishing in the five countries under study in the RCN funded research project R-Quest. This particular analysis focuses on the level of universities, as represented by the Leiden Ranking 2018, and the way these universities (in an aggregate fashion) perform on the various types of OA publishing. As the notion of research quality is central, this study evolves around the way Plan S effects the OA uptake, and the potential effects this policy initiative has on the way OA publishing is taken up in the universities in the five countries in the study. As Plan S aims at diminishing the hybrid form of OA as a way to publish findings in an open access format, this study tries to unravel the various forms of OA publishing, versus impact measures connected to the various forms of publishing.

The tense reactions by academic communities in the countries connected to Coalition S, the funding agencies that joined forces to implement Plan S, shows the complex relationship between perceived quality of journals, as reflected by high JIF values, and in this study represented by mnjs values, and notions of research quality. The disappearing or diminishing of possibilities to publish hybrid OA, means limiting the opportunities to publish in journals with a relative high level of prestige or reputations, as reflected by high JIF values. This indicates that in the transition phase we are currently, in academia, towards a system of full open science based working environments, notions of research quality are potentially repeatedly re-defined, and might be highly volatile for a period of time.

This study also signals some methodological issues. A first issue to mention in this respect is the fact that more sources become available that backwardly declare outputs as Open Access (journals being included in the DOAJ, and thereby being treated and counted as OA, although at the moment of publishing not being OA, or vice versa, journals that previously be on the DOAJ list, and have now disappeared from that list). A second issue is the fact that datasets such as WoS and Scopus lack certain unique identifiers for early years, thereby creating issues related to 'gaps' in what was actually OA, or could have been counted as OA (the missing DOIs in earlier years of our WoS subscription caused this type of problems, only recently resolved by Clarivate Analytics).

A conceptual issue that keep returning is the definition of the various types of OA publishing, and the way these various forms are being treated as policy relevant or not. The Unpaywall database has solved many of the problems in defining OA publishing types, while the way OA publishing forms are considered as policy relevant is much less stabilized or clearly defined. In future research we will try to come up with potential solutions to this problems as well.

Acknowledgments

Funding for the R-Quest project comes from the Research Council of Norway, grant number 256223 (the R-QUEST centre).

References

Bosman, J. & B. Kramer, Open access levels: a quantitative exploration using Web of Science and oaDOI data. Peer J, January 11, 2018. (https://peerj.com/preprints/3520/)

Harnad, S et al (2004).The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials Review. 30 (4): 310–314

Martin-Martin, A, Costas R, van Leeuwen TN & Delgado-Lopez, E. Evidence of Open Access of scientific publications in Google Scholar: a large-scale analysis. Journal of Informetrics, 2018, 12 (3), 819-841

Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. (2017) The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ Preprints 5:e3119v1

van Leeuwen, T.N. and H. F. Moed, Development and application of journal impact measures in the Dutch science system, Scientometrics 53, 2 (2002), pp. 249-266.

van Leeuwen T.N. Meijer, I., Yegros-Yegros, A., and Costas, R. Developing indicators on Open Access by combining evidence from diverse data sources. 2017. Retrieved December 7, 2001 from: https://arxiv.org/abs/1802.02827v1

van Leeuwen, T.N., C. Tatum & P.F. Wouters. Exploring Possibilities to Use Bibliometric Data to Monitor Gold Open Access Publishing at the National Level Journal of the American Society of Information Science & Technology, 2018, 69 , 1161-1173

Van Leeuwen, T.N. R. Costas, and N. Robinson Garcia, Indicators of open access publishing in the CWTS Leiden Ranking 2019, CWTS Blog. May 15th 2019 (https://www.cwts.nl/blog?article=n-r2w2a4&title=indicators-of-open-access-publishing-in-the-cwts-leiden-ranking-2019)

Waltman, L., N.J. van Eck, T.N. van Leeuwen, M.S. Visser, and A.F.J. van Raan, Towards a new crown indicator: Some theoretical considerations, Journal of Informetrics, 2011, 5 (1), 37-47

Using Pat2Vec Model to Discover the Technology Structure

Xiaomei Wang[1] Ting Chen[2] and Guopen Li[3]

[1] wangxm@casisd.cn [2] chenting@casisd.cn [3] liguopeng@casisd.cn
Institutes of Science and Development, Chinese Academy of Sciences, 100190 Beijing (China)

Abstract

The characteristics of patents are different from academic papers, classical citation analysis methods may not be suitable for patent analysing. Identifying the textual features of patents accurately is still challenging job. Recent deep learning techniques brought human-competitive performances in various tasks of fields, including image recognition, natural language understanding etc. In this work, we build a set of processes for feature extraction, clustering and visualization of technology structure discovery using machine learning model. First, we trained an unsupervised patent features extraction model (Pat2Vec) which learns the properties from millions of patent texts in high-dimensional vector space. Then we clustered 291,493 Triadic Patent Families from 2012 to 2017 in vector space to classify the technology structure. Finally, we visualised the large-scale relationship of the patents and clusters of patents using the LargeVis algorithm. The analysis results show that the feature extraction accuracy of Pat2Vec is higher than the traditional text feature extraction model. The technology structure map obtained by clustering patents provides the possibility for further analysis.

Introduction

A World Intellectual Property Organization (WIPO) report points out that almost 90-95% of the world's R&D outcomes are covered in patent publications with the remaining 5-10% reported in the technical literature (essays and publications) (Liu & Yang, 2008). In order to understand the world's technological innovation layout, discover the emerging or hot technologies, it is necessary to discover the structure of technology, which mostly represented by patents.

It is different from the patent citation and the citation of the academic paper, the motivations of citation by the applicants and the examiners are also different. In addition, the citation of the patent is very sparse if we compare them with the papers. According to our statistics, less than 30% of the triadic patents in the past six years have direct citations, and 37% have co-citation relationships. Therefore, the classic citation analysis method in patent analysis may lack some important patents. The commonly used patent IPC, USPC classification is mainly based on functionality, sometimes difficult to correspond to the industrial technology classification. Besides the classification system is usually updated very slowly, unable to reflect the key technologies' latest changes. Content-based patent analysis can avoid the defects of citation relations, but most of the content analysis is based on the keywords or key phrase, which may have the problem of polysemy, it makes the accuracy and usability of text analysis not very efficient. Therefore, how to improve the accuracy of patent content analysis is still a challenge. Recent deep learning techniques brought human-competitive performances in various tasks of fields, including image recognition, natural language understanding etc. This study tries to use Doc2vec model to train the patent feature vector model Pat2Vec and get technology structure more accurately and usefully.

Related work

Existing studies using patent maps to describe the global technical structure mainly include: 1) SCITech Strategies garnered a scientific map containing 20 million papers and 2 million patents (https://www.scitech-strategies.com/). The map contains the largest number of patents, and the patent relationship is mainly established by the citation relationship between the patent and the paper. The practical application of the patent map has not been found yet. The main problem of the patent map is missing some patents because of using patents and papers citations relationship; 2) Japan's JST uses Top1% high-cited patents for co-citation clustering to generate technology front (Jibu, 2014). The problem with this map is that valuable front-end patents are not necessarily to be the highly cited patents, and similar patents do not necessarily have a citation relationship between them. 3) Clarivate Analytics' Derwent Innovation platform produced a patent map using key term co-occurrence relationships, the main drawback of term co-occurrence relationships which are described above.

Several other studies have been tested to automatically extract topic classification using latent Dirichlet allocation (LDA) (Blei, Ng & Jordan, 2003). In essence, the topic is usually a cluster of keywords. It is sometimes difficult to determine the specific topic represented by the group of keywords because the intersection of similar keywords between the topics.

In 2013, Word2vec was created by a team of researchers led by Mikolov et al. at Google(2013a, 2013b), it takes as its input a large corpus of text and embed them into an vector space, typically of several hundred dimensions, embedding vectors created using the Word2vec algorithm have many advantages compared to earlier algorithms such as latent semantic analysis.

On top of Word2vec, Doc2vec or Paragraph2vec (Le & Mikolov, 2014), Sentence Embeddings, is an unsupervised algorithm that can obtain vector representations of sentences, paragraphs, or entire text, and is an extension of Word2vec. Compared with the Word2vec model, Doc2vec adds a paragraph matrix in the process of training the word vector, which is more suitable for the text length of the patent text title abstract.

Data

According to the "World Intellectual Property Indicators 2018" published by WIPO, in 2017 global innovators submitted a total of 3.17 million patent applications, which achieved growth for the eighth consecutive year. However, the value distribution of patents is not balanced, and many patents have lower value. In order to avoid excessive patent noise to impact analysis results, this study selected "Triadic" patents for analysing.

Triadic patents are a series of corresponding patents filed at the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO) and the Japan Patent Office (JPO), for the same invention, by the same applicant or inventor (OECD, 2005). Since the United States, Japan, and the European Union are the three most important markets in the world, and the patent application and maintenance costs are high, triadic patents are generally considered to have high economic value, it may be able to reflect the value of a country's technological invention and competitiveness in the international market.

This study selected 291,493 triadic patents from 2012 to 2017 as the analytical data set, with the patent family as the smallest analysis unit. 3 million patents from US Patent Office were used as training data for model.

Methods

The research flow of this study is shown below. In order to explore the law of patent texts, a doc2vec model start learn text representation of patent from 3 million USPTO patents, then the triadic patents text is placed in the semantic model to extract the text vector features of each

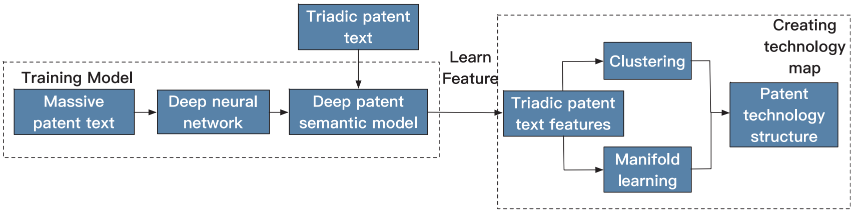


Figure 1 Method flow chart

triadic patent. Finally, the patented technical structure is discovered through clustering, and the relationship between patents is visualized through the large-scale manifold learning model.

Patent feature extraction model based on massive patent text training

The patent's feature extraction model mainly uses the Doc2vec model based on neural network embedding model. We use more than 3 million patents from USPTO in 10 years, mainly using the title and abstract to train the model, we call this model Pat2Vec. Before we train our model, we applied standard NLP pre-processing, such as removed punctuation and stop words, then applied lemmatization and Stemming to patent text. The training algorithm of the model is distributed memory (PV-DM), it acts as a memory that remembers what the topic of the paragraph is. While the word vectors represent the concept of a word, the document vector intends to represent the concept of a document. The PV-DM model is superior and usually performed better than another model Distributed Bag (PV-DBOW).

Clustering based on patent feature vectors

The triadic Patent text feature produced by Pat2vec model is a 100-dimensional feature vectors. In this study, we directly cluster patents in the feature vector space instead of converting the feature vector back into the similarity matrix and then do the community detection. Because compared to community discovery in the network, the technology is more mature and accurate to use traditional machine learning clustering in the feature space(Goyal& Ferrara, 2018; Gunasekaran et al., 2017). We tried hierarchical clustering and spatial partitioning clustering algorithms. Since the quantity of triadic patents is close to 300,000, we choose the fastest K-means clustering algorithm.

Visualization method

When the high-dimensional feature vector of the patent is converted into a distance-based network relationship, there is a loss of information. Based on our visualization study (Chen et al., 2018), the manifold learning based dimensionality reduction visualization model can visualize high-dimensional data better than convert data into the distance-based graph.

The map of massive data usually shows the layout on the topic level. Many articles in one topic are displayed as a point in the map. Since the topic may be large or small, it may affect the visualization effect. This study attempts to create a map with single patent as the basic unit to show the patent layout from the micro level, reflecting the more details of the technology layout. Based the patent level map, construct clustering level map by the location of patents in the subject. The two maps with hierarchical are consistency.

Nearly 300 thousand patents exceed the t-SNE (Maaten & Hinton, 2008) algorithm's computing limit, so this study uses the big data visualization algorithm LargeVis (Tang et al., 2016) for creating a large map based on patent level visualization. LargeVis improved the operation speed of t-SNE algorithm, reduced the complexity of the algorithm from nonlinear to linear, and greatly improves the speed of dimensionality reduction.

Experiments and evaluation

Patent text feature extraction

The test data set was constructed based on the patent classification IPC code. We selected two patents that belong to same subgroup of IPC as similar patents, and then randomly add a patent outside the subgroup but belonging to the same larger group, we treat the third patent as noise. After that, different feature extraction models were used to determine which two patents in the three patents are closer, and whether the noise can be accurately identified. We randomly constructed 4000 sets of triplets to verify the patent feature extraction effect of the Doc2vec model. Experiments show that the feature extraction effect of the Pat2vec model is much better than the traditional models.

Table 1 the accuracy of feature extraction models

Model	Accuracy
Pat2vec	**86.5%**
Tf-idf	73.3%
LSA	76.9%
LDA	78.1%

Clustering parameter selection

The K-means clustering algorithm needs to specify the number of clusters K before clustering, but determining the number of K has always been a challenge. In this study, Sum of squared error, Calinski-Harabasz Index, Davies-Bouldin Index, and Silhouette Coefficient were used to optimize the number of K. The curves of the first three evaluation algorithms start to converge when K=800 to 1200. If the number of patent topics is only 1200, this does not meet the expert's understanding of the patent topics. The Silhouette Coefficient curve increases with the number of clusters and exhibits multiple peaks at K=700, 6400, and 8100. Combining expert judgment with subsequent analysis needs, we choose K=6400.

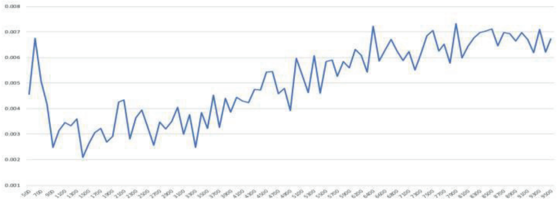


Figure 2 Silhouette Coefficient curve

The clustering results have been interpreted by experts from several fields. The experts believe that the patents in the topic are relatively consistent.

Comparison of patent topics visualization

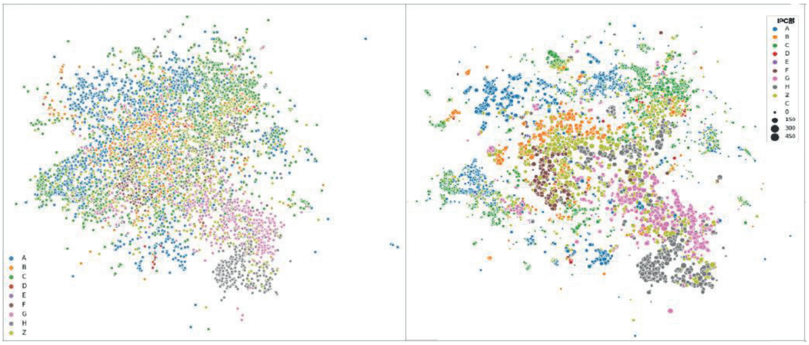


Figure 3 Two clustering representation point algorithms, the left is the intra-cluster coordinate averaging algorithm, and the right is the intra-cluster coordinate density maximum area averaging algorithm.

To construct a coordinate position of a technology topics based on each patent location. This study has tried two methods, first way is using the average coordinates of the patents in the technical subject clustering. Second way is using the average position of the highest density coordinates, as shown in Figure 3. The first method is affected by the uneven coordinate point distribution, and the distribution of the topics in the map is uniform, the aggregation effect is poor. In the second method, the topics has better aggregation and a clearer outline.

Discover the Technology Structure

We map the structure of technology by the relationship of nearly 300 thousand patents (Figure 4). The colour in the left figure is the IPC classification. It can be seen that there are clear separations between different patent IPC classes at the global structure. At the local structure, the sub-topics inside the same IPC classes are also relatively clear.

The colour on the right is the fields obtained by combining the technical classifications of WIPO's 35 fields. As can be seen from the figure, the map can discover the local structure accurately. For example, the class of "Electrical machinery, apparatus, energy" is divided into four large groups: lithium-ion battery/fuel cell, electric vehicles/electric power, electric motor/permanent magnet, and electrical connector.

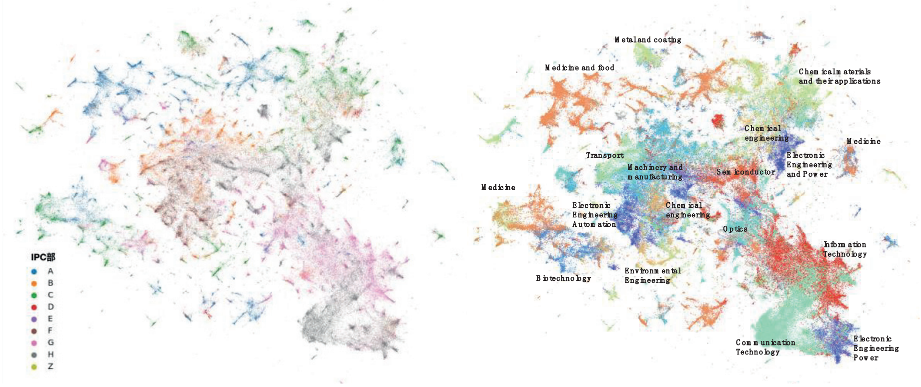


Figure 4 World-wide triadic patent structure map 2012-2017

We further created a heatmap based on patent topics level, the darker area represents higher density of patents in unit area (Figure 5). From the figure, communication and information technology are the areas with the highest density, and the most competitive field in all countries. The number of topics is large, and the average number of patents in the topics is over 70. The next highest dense areas are machinery and manufacturing, optical semiconductors. The areas related to biomedicine and organic chemistry have the lowest density. Although the number of topics in these areas is the largest, but the average number of patents within the topics is the least, only about 20 items in each topic.

In future, we will further analyse the patent layout of countries and institutions based on the patent technology map.

Discussion and future work

We used the machine learning technology to detect and visualize the technical structure of nearly 3 million patent families. A large number of patent texts were used for training the patent text feature extraction model Pat2Vec, to explore the hidden deep relationship and between patented technologies, which improves the accuracy of traditional patent feature extraction model. In order to reduce the loss of information, this study no longer uses the traditional community detection based on distance between patents but clustering directly in high-dimensional vector space; A direct dimensionality reduction method based on manifold learning is adopted to display patent topics level and patents level.

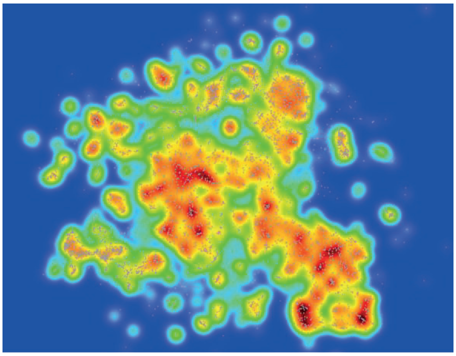


Figure 5 Technology structure heat map

The Pat2Vec model has good versatility and can be applied to other patent data sets and other types of data such as funds, standards, etc. Furthermore, we will build more test datasets for accurate testing and find more accurate patent feature extraction models, clustering models and visualization methods by combining multiple heterogeneous features. After that, we will start researches on the technology hotspots and emerging technology detection based on the technical structure map.

Acknowledgments

This work is funded by the strategic research project of the Development Planning Bureau of the Chinese Academy of Sciences.

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine* Learning *Research,* 3, 993-1022.

Chen, T. Li, G. Wang, X. (2018).Mapping Research Funding in networking and embedding. *8th Global TechMing Conference*, Leiden, Netherlands

Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151, 78-94.

Gunasekaran, K., Muralikumar, J., Sudarshan, S., Srinivasan, B., & Malliaros, F. (2017). NetGloVe: Learning Node Representations for Community Detection. *In 6th International Conference on Complex Networks and Their Applications*.

Jibu, M (2014) Mapping of scientific patenting: toward the development of 'J-GLOBAL foresight', *Technology Analysis & Strategic Management*, 26:4, 485-498

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *In International Conference on Machine Learning* (pp. 1188-1196).

Liu, C. Yang, J. (2008). Decoding Patent Information Using Patent Maps. *Data Science Journal*, 7(0), 14-22.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2579-2605.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at the International Conference on Learning Representations.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 3111-3119.

OECD (2005), Main Science and Technology Indicators, Vol. 1.

Tang, J., Liu, J., Zhang, M., & Mei, Q. (2016). Visualizing large-scale and high-dimensional data. *In Proceedings of the 25th International Conference on World Wide Web (pp. 287-297).*

Public Policy and the Evolution of Technology Transfer in France

Nicolas Carayol[1] and Elodie Carpentier[2]

[1] *nicolas.carayol@u-bordeaux.fr*
GREThA, CNRS and Department of Economics, University of Bordeaux, Pessac 33608 (France)

[2] *elodie.carpentier@u-bordeaux.fr*
GREThA, CNRS and Department of Economics, University of Bordeaux, Pessac 33608 (France)

Abstract

The direct contribution of professors and researchers employed by Universities and Public Research Organizations has been emphasized as increasingly important sustain nations' technological edge. We quantify this phenomenon according to a new methodology in a systematic nation-wide empirical investigation over an eighteen-year period. We find that academia (professors and researchers at universities and PRO) account for more than ten percent of all the nation's patents. We estimate that nearly twenty percent of the sixty-one thousand French faculty members, researchers and engineers whose patenting behavior we investigate have invented at least one patent. As a series of policy reforms of the academic sphere have been introduced in France toward implementing the post Bayh-Dole US style of IP management, we appreciate how this translates into modified behaviors of the different players (professors and researchers and universities). In particular, we find evidence for a transitory learning phase of universities after the introduction of the 1999 Innovation Act.

Introduction

In 1980, the Bayh-Dole Act is introduced in the US with the intention of fostering the transfer of technology from universities to the private sector. Through this legislation, the federal government waives to the universities the intellectual property rights (IPR) over federally funded inventions. It contributed to an increase in academic patenting (Mowery et al., 2001) and despite the initial concern (Trajtenberg, Henderson & Jaffe, 1997), Mowery & Ziedonis (2002) conclude that the overall importance of these inventions did not decrease.

The commercial success of the US stimulated several European governments to introduce reforms in order to converge toward the university ownership regime. In some countries as Germany (Czarnitzki et al., 2015), Norway (Hvide & Jones, 2018) or Finland (Ejermo & Toivanen, 2018), the end of the professor privilege had a negative impact on academic invention. Thus, the apparently same ownership model provides contrasted evidence as regards its effect on academic invention on both sides of the Atlantic.

In order to reconcile this contrasted evidence, Geuna and Rossi (2011) suggest that comparing the US reform with the European ones is misleading. Whereas the Bayh-Dole Act transferred the IPR from the federal government to universities, hence making the management of IPR *closer* to the inventor, the end of the professor privilege in European countries made the IPR management *further away* from her.

Another natural explanation is that the IPR management is a complex activity that requires time, financial support and capabilities to be performed efficiently (Grimaldi et al., 2011). Universities and PRO may need to learn how to manage IPR. In the US, Mowery, Sampat & Ziedonis (2002) observe that it took 15 years to US universities without previous IPR management experience to produce inventions of similar importance to those by experienced universities. On the reverse, Hvide & Jones (2018) observe that in Norway the gap in terms of importance between academic and non-academic patents is widening after the reform. Hence, they do not find supportive evidence for such learning effect. It may be because they observe patenting activity only seven years after the reform. It could also come from an insufficient financial support from the government to public universities.

We use the French case to examine the existence of a learning effect after a Bayh-Dole type reform is introduced and meanwhile how universities cope with the reform requirements in practice. France is a particular case as the universities ownership model has always prevailed but in practice universities used to waive a large part of their IPR to companies. In 1999, on the belief that academic invention was weak as compared to the level of investment in fundamental research, the government voted the Innovation Act. Its goal was to incentivize universities to increasingly retain the ownership of its inventions.

We build a dataset on all French professors and researchers (approximately 65,000 persons) and their patenting activity from 1995 to 2012. We observe that 15% of them invented at least once over the period. The overall number of academic patents is around 56,000 which is 10% of all patents invented in France. Even before the Innovation Act, this share was 5.5%.

We find supportive evidence of a transitory learning phase after the Innovation Act. In the first six years post-reform, universities and PRO indeed increase the share of IPR retained and made important use of the co-ownership regime with companies. When distinguishing institutions with a previous IPR management experience (incumbents) from those without such experience (entrants), we observe that entrants are co-patenting even more and longer than incumbents. We interpret such evidence as a middle ground strategy allowing universities and PRO to comply with the new model while still in a learning phase.

The next section exposes the data collection. The following one portrays the academic patenting in France. Then, the third section pictures the evolution of the legal framework of universities and the recent trends in academic patenting while the fourth one investigates the shift in technology transfer strategies of universities. The last section concludes.

Data collection: Identifying academic inventions

An academic patent is a patent invented by at least one university researcher, independently of the type of applicant: it can be a university, a public research organization, the scientist herself, or a private company either exclusively; or jointly with other types.

Professors and researchers

In order to identify these patents, we need to determine which inventors are employed by universities. We have lists of professors and researchers associated with laboratories recognized by the Ministry of Higher Education and Research which contain, among other information, the first and last names, status (researcher or teacher-researcher), research and teaching fields, research unit, date of birth and gender. It is important to understand that all the French research system is organized in research laboratories (see Carayol & Matt, 2004). This covers in fact very different sorts of institutions, which size for instance can vary from a few professors and researchers to several hundreds. These data are compiled every four- to-five years. Each year approximately, a fourth of all labs in France provide such lists. As our lists have been constituted since year 2005 to year 2016, most labs in France have been surveyed this way at least twice. Some may have been surveyed only once because they have been created or have been terminated over the period. Another reason why a lab could have been surveyed only once is because the coverage of the data is increasing over the years so that some labs were not considered in the first years. As professors and researchers are likely to be listed several times (either in the same lab or by different labs if they moved for instance), a systematic disambiguation has been performed through various automatic and manual procedures. All in all, the data concern 61,223 faculty members, researchers and some engineers.

One important challenge of our work will be to identify whether the probability of those professors and researchers to invent is evolving over time. We therefore create a panel dataset.

Inventors

Patent data are extracted from the "EPO Worldwide Patent Statistical Database" (PATSTAT). We restrict the data to all inventors having a home address in France and having participated in the invention of a patent filed at the EPO, the USPTO or the INPI (French national office). That dataset does not contain any identification code for inventors. Therefore, a preliminary step is to define a reliable identifier for inventors. To do this, we use the algorithm proposed by Carayol, Cassi & Roux (2015). This algorithm basically relies on a Bayesian approach to compute similarity scores which are the probability that several identities correspond to the same person given a series of observables (i.e. the applicant, technological field, address, and backward citations to other patents). The algorithm uses a set of nearly five thousand verified (positive and negative) matches to benchmark itself. The error rate remains very limited (less than 2%).

Filtering academic patents

Once the inventor disambiguation step is completed, we use a statistical model to estimate the probability that all possible matches between academic researchers and professors on the one side and inventors on the other side are correct.

This is in fact a three-stage procedure. In a first stage, we estimate a logit model on a set of validated and unvalidated couples. This benchmark was constituted on the basis of experts (mainly professionals of technology transfer employed in universities) identifying professors as potential inventors. The benchmark is constituted of nearly seven hundred professor-inventor pairs having similar family names. We found out that regressions per office are more efficient. As explaining variables, we include the Jaccard similarity between family names, the inventor name frequency (in log), the distance between the patent technological classification and the professors' scientific disciplines as defined by Magerman, Callaert & Van Looy (2017) (in log), as well as weighted dummies signaling consistency between the professor age and the patent application year and between the assignee name and the professor employing institution. The regressions are performed by office as a Hausman test shows logit coefficients are significantly different across offices. The regression results for each office are presented in Table 1.

Table 1: Logistic regressions on the benchmark, per office.

	EP	EP	FR	US	US	US	US
name similarity	22.43***	21.83***	12.74***	38.79*	38.10***	39.11***	38.02***
inventor's name frequency	1.23***	1.24***	0.64***	1.65**	1.68***	1.68***	1.71***
assignee/employer consistency	6.24***	6.23***	1.27***				
tech/discipline consistency	0.57**	0.59***	0.43**	0.64	0.72		
age/year consistency	0.67		1.25***	-2.57		-2.65	
intercept	-32.19***	-29.19***	-21.84***	-38.77*	-47.58***	-38.62**	-47.23***
N	682	682	2829	115	115	115	115

The second step uses the estimated coefficients to predict the probability that potential matches are correct or incorrect on the whole reference population.

In the third step, we consider various thresholds of the probability for accepting of rejecting matches that we note p.

Let $TP(p)$ denote the number of true positives in the benchmark for a given threshold probability value p, $FP(p)$ is the number of false positives, and $FN(p)$ the number of false negatives. We can then compute the standard precision indicator: $P(p) = \frac{TP(p)}{FP(p)+TP(p)}$, as well as the recall indicator $R(p) = \frac{TP(p)}{TP(p)+FN(p)}$. As precision and recall vary in opposite directions with the threshold p value, we compute a synthetic indicator that takes both into account:

$$F_{\beta}(p) = (1 + \beta^2) \times \frac{P(p)\times R(p)}{\beta^2 \times P(p)+R(p)}, \quad (1)$$

with $\beta > 0$. In that expression, the parameter β weights precision and recall. If $\beta < 1$, precision is more weighted than recall, and the reverse holds when $\beta > 1$. As we would like our results to not be sensitive to a particular value of β, all our statistics will be computed for $\beta = 0.5, \beta = 1$ and $\beta = 2$.

We display, in Figure 1 the computed values of those indicators for the different threshold probability values p.

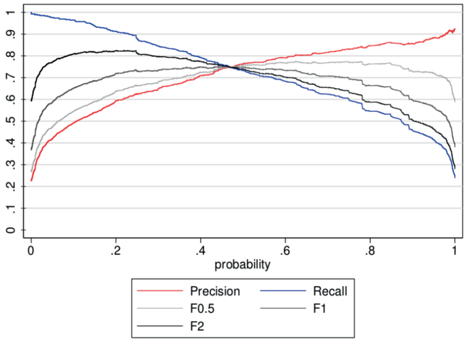


Figure 1: Precision, recall et F_{β} (when $\beta = 0.5$, $\beta = 1$ or $\beta = 2$) for different threshold probability values.

Our goal is to find and "optimal" p value, for a given β and patent office i. That is, we want to find: $p^*_{\beta,i} = argmax_p\{F_{\beta,i}(p)\}$, for each office i, with $F_{\beta,i}(p)$ the indicator defined in Equation 1, calculated on the patents of office i only. Given that we consider three offices and the F-measure is computed for three different values of β, we end out with a series of 9 = 3×3 thresholds. Optimal thresholds are sensibly different for each considered office (See Table 2).

Table 2: Optimal thresholds for each office and different β values.

office	Threshold	F-measure
EPO	$p^*_2 = 0.14$	$F_2 = 0.82$
	$p^*_1 = 0.44$	$F_1 = 0.77$
	$p^*_{0.5} = 0.46$	$F_{0.5} = 0.82$
INPI	$p^*_2 = 0.20$	$F_2 = 0.78$
	$p^*_1 = 0.45$	$F_1 = 0.69$
	$p^*_{0.5} = 0.74$	$F_{0.5} = 0.71$
USPTO	$p^*_2 = 0.24$	$F_2 = 0.86$
	$p^*_1 = 0.24$	$F_1 = 0.71$
	$p^*_{0.5} = 0.24$	$F_{0.5} = 0.60$

Calculating the "expected" number of academic patents

Let $N^1_{\beta,i}$ be the set of patents that are validated in office i and for a probability threshold $p^1_{\beta,i}$. The cardinal of that set is $n^1_{\beta,i}$ and $n^0_{\beta,i}$ is the number of non-validated patents.

The patents in $N^1_{\beta,i}$ are the validated ones for some criterion captured by β. As such those patents are the one that we will keep as our set of academic patents on which we will calculate many things. But assuming that $n^1_{\beta,i}$ reflects the expected number of academic patents would be slightly misleading. Indeed, patents counted in the underlying set have been misreported as such (false positives) while patents in the complement are also misallocated (false negatives). The expected number of patents can be calculated as follows:

$$\hat{x}_\beta = \sum_i \left[\frac{TP(p^*_{\beta,i})}{FP(p^*_{\beta,i})+TP(p^*_{\beta,i})} \times n^1_{\beta,i} + \frac{FN(p^*_{\beta,i})}{FN(p^*_{\beta,i})+TN(p^*_{\beta,i})} \times n^0_{\beta,i}\right],$$

for a given threshold $p^*_{\beta,i}$ for each office i ∈ {EPO, INPI, USPTO}.

The different modes of calculation provide very similar numbers as the estimated number of patents invented by our population of professors goes from 25,624 to 26,503. Moreover, 4.6% of French-invented patents originate from Academia. For the period 1995-2001 and only patents filed at the EPO, Lissoni et al. (2008) provide a lower estimation with 3,4% of the 48973 French-invented being academic. For the same period, we obtain a comparable number of patents invented in France, but a greater share (6%) of academic ones.

Table 3: Expected number of academic patents for several β values (from 1995 to 2012).

Office	$\hat{x}_2$		$\hat{x}_1$		$\hat{x}_{0.5}$		All French-invented patents
EPO	11072	(6.1%)	12320	(6.9%)	12325	(7%)	177286
INPI	11173	(4.5%)	10804	(4.3%)	10473	(4.1%)	250605
USPTO	3379	(2.5%)	3379	(2.5%)	3379	(2.5%)	134315
Total	25624	(4.5%)	26502	(4.6%)	26177	(4.6%)	562206

Academic patenting in France

The starting point of our methodology to collect all academic patents is the list of professors and researchers in France. However, the survey started in 2005 (until 2016) and covers only universities and PROs recognized by the French Ministry of Higher Education and Research (MHER). Hence, we know as a fact that we miss all patents invented by professors and researchers that were no longer active in 2006 or belong to a university or PRO not recognized by the MHER. We recover some of these missing academic patents by adding all patents that did not match on inventor's name but are owned by at least one French university or PRO to our academic patents pool. This leads us to Table 4, according to which at least 10% of all inventions in France between 1995 and 2012 come from Academia, and this is a floor value as academic patents owned exclusively by the private sector remain missing for the reasons mentioned above.

Table 4: Expected number of academic patents for several β values (from 1995 to 2012) – Alternative method.

Office	$\hat{x}'_2$		$\hat{x}'_1$		$\hat{x}'_{0.5}$		All French-invented patents
EPO	19786	(11.1%)	21034	(11.8%)	21039	(11.8%)	177286
INPI	24973	(10%)	24604	(9.8%)	24273	(9.6%)	250605
USPTO	10963	(8.1%)	10963	(8.1%)	10963	(8.1%)	134315
Total	55722	(9.8%)	56600	(10.1%)	56275	(10%)	562206

One of the motivation for the policy reforms introduced in Europe in order to converge to the Anglo-Saxon model was the common belief that European countries were lagging the US in terms of technology transfer. According to Lissoni (2012), the inventions made in universities and research institutions in the USA between 1994 and 2002 represented 6% of all US-invented patents. In Italy, the Netherlands and Sweden this share is 4%, 4.3% and 6.2% respectively.

The importance of academic patenting is France is estimated at 3.4%, which we can reevaluate with our data at rather 8%. Hence, the underperformance of France in terms of academic patenting appears unverified. Even before the introduction of the Innovation Act in 1999, France was doing at least as well as any other advanced economy.

We breakdown inventions by technological field. In Table 5, the first column of percentages indicates the repartition of academic patents across technological fields, while the second column of percentages point out the weight of academic patents among all inventions in the specific technological field. The leading fields in Academia are Chemistry (42.7%), followed by Electrical Engineering and Instruments (20.6% and 20.2% respectively), these shares being higher than in the overall distribution of patents for Chemistry and Instruments (30.2% and 12.8% respectively for all French-invented patents). Moreover, in the Instruments field, academic patents account for 15.1% of all patents, and in Chemistry 13.6%. While Mechanical Engineering is the second most important technological sector at the national level, academic inventions are clearly underrepresented in this field.

Table 5: Distribution of academic patents across technological fields for several β values (from 1995 to 2012).

Technological sector	Academic patents only ($\hat{x}'_1$)			All French-invented patents	
Chemistry	35549	(42.7%)	(13.6%)	259599	(30.2%)
Electrical Engineering	17175	(20.6%)	(9%)	191244	(22.2%)
Instruments	16805	(20.2%)	(15.1%)	110607	(12.8%)
Mechanical Engineering	11082	(13.3%)	(4.9%)	224255	(26.2%)
Other fields	2699	(3.2%)	(3.7%)	70176	(8.1%)
Total	83309	(100%)	(9.6%)	855881	(100%)

Finally, Table 6 provides the participation level of professors and researchers to the academic invention phenomenon by scientific discipline (for medical and hard sciences only). Overall, one professor over five has filed at least one patent between 1995 and 2012. In line with the previous findings, professors in Chemistry are particularly involved, with 32.5% of them being academic inventors. In all other disciplines the share of professors that are also inventors is very homogeneous at approximately 22%, except for Universe Science (12.5%).

Table 6: Repartition of professors and researchers involved in academic patenting by scientific discipline (1995 – 2012)

Scientific field	professors-inventors		All professors
Chemistry	2485	(32.5%)	7653
Applied Bio. Ecology	2119	(22.3%)	9497
Fundamental Biology	3426	(23%)	14890
Medicine	3315	(24.3%)	13624
Engineering Sciences	2806	(24%)	11714
Math	1601	(20.9%)	7674
Physics	2186	(24.4%)	8965
Universe Science	453	(12.48%)	3630
Total	8367	(21.4%)	39069

Notes:

– 23670 professors and researchers in Human and Social Sciences are not represented in this table. 1101 of them have invented at least one patent over the period, but 300 are already accounted for in one the other engineering and hard sciences disciplines. The net share of inventors in Human and Social Sciences is 3.4%. If these SHS inventors are included in the full sample (61223 researchers), the global share of academic inventors goes down to 15.2%.

Science policy and recent trends in technology transfer in France

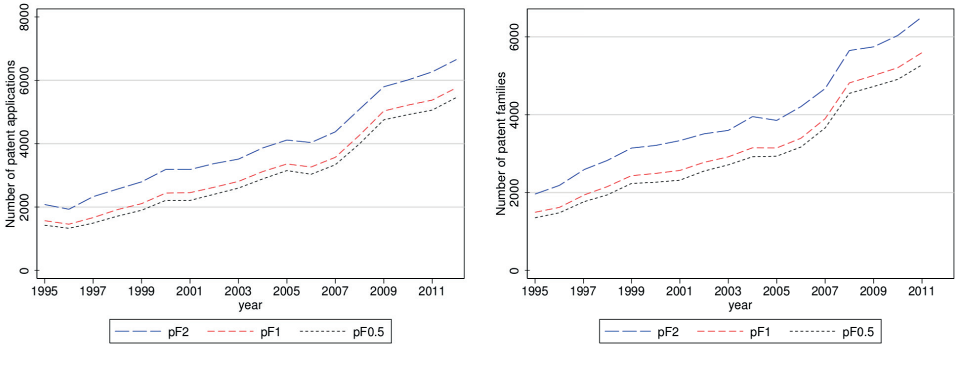
Science policy and technology transfer

Before exposing the recent evolution of technology transfer in France, we need to first depict the numerous policy initiatives affecting technology transfer introduced in the last fifteen years. In France, universities and PROs had the rights over the research of their professors and researchers and thus a legislation equivalent to the Bayh-Dole Act was not necessary. However, these institutions did not manage this function historically. Therefore, national policy aiming at encouraging technology transfer essentially consisted in sustaining the development of technology transfer in these public institutions.

The most important piece of policy is the 1999 law on Innovation and Research (Loi Allègre) allowing universities to create internal services for managing contracts and transfer (SAIC). It created a series of public incubators. The law also modified the status of professors and researchers (who are civil servants in France), favoring mobility to (temporary) positions offered by companies, allowing professors to perform consulting activity for companies (to a certain extent), and take equity positions in start-ups capital.

In 2005, the Agence Nationale de la Recherche (ANR) has been created to implement project-based funding in France. Therefore it funds numerous collaborative research projects between companies and academic labs. That law also created the Instituts Carnot (somehow designed under the Fraunhofer Institutes model) which were created to support collaborative research with companies. In 2010, a very large financial plan called "Programme des Investissements d'Avenir" (PIA) was launched to sustain the emergence of strong research players (laboratories and university sites) in the French research system. It supported the creation of 14 SATTs ("Sociétés d'Accélération de Transfert de Technologie"). The goal was to increase the professionalization of transfer and to favor other economies of scale, in particular in research universities, by increasing their budget thanks to substantial subsidizes (more than 0.8 billion euros altogether).

The left graph of Figure 2 shows that the number of academic inventions has dramatically increased in France over the last two decades. The right graph concerns patent families. That variation corresponds to an increase by a factor 3.



Note: the year of family applications are earlier or equal to the year of applications. Hence, we have consistent estimations until 2011 for families and until 2012 for applications.

Figure 2: The evolution of academic patenting in France: number of patent applications in the three offices (left graph) and number of patent families (right graph).

Policy and different actors' behaviors

Propensity calculation

The above calculations may be misleading as in fact the underlying population of professors and researchers that we consider is likely to be increasing over the period. Indeed, our population identification corresponds to years 2005-2016. Most of them are likely to be active for those years, but less in the preceding subperiods. Therefore, we should now calculate which of those professors and researchers are to be considered each year and exclude all the inventions attached to them outside these individualized ranges of time. This leads to Figure 3. The number of patents is larger than in the previous figure, we now consider inventions in patents, and those co-invented by several of our professors and researchers are counted several times. We can see in this Figure 3 that the number of professors and researchers is increasing very rapidly as well, at a slope which is comparable or slightly higher than the one of their patents. Therefore, it seems that the propensity to invent (expected number of patents per year per capita) remains pretty stable over the considered period.

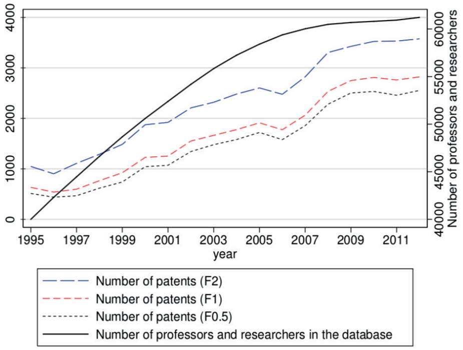


Figure 3: The evolution of academic patenting in France, with respect to the reference population.

More generally, it is then likely that other variables may affect the probability to invent over time. As these factors are not controlled for in the previous simple analysis, our estimates may end out being biased. We run negative binomial regressions (as the dependent variable is a count variable, which is overdispersed), to estimate the probability to invent each calendar year controlling for confounding factors such as age, age squared, academic position and university of affiliation and it appears that the introduction of such control variables only marginally reduces the coefficient associated with the years.

Furthermore, there exists a natural propensity to invent outside of Academia. Our strategy is then to deduce from the academic propensities to invent the natural (or non-academic) ones in order to observe a potential effect of public policies on the propensity to invent in Academia every year. The following regressions will not include anymore the control variables as we do not have personal information for the non-academic inventors. Thus, we run negative binomial regressions to estimate the probability to invent each calendar year. The incidence rate ratios of each year dummy on the number of academic patents (red curve) calculated with threshold F1($\hat{x}_1$) as well as the number of non-academic patents (blue curve) are synthesized in Figure 4. The propensity to invent in Academia is clearly departing from the one in the non-academic sector from 2006 ongoing (Figure 4). The largest year dummy incidence observed is for year 2010 which is up to 3.5 times more than 1995, or 2.5 when deducing the natural propensity to invent. At this point, we cannot say with certainty whether it is specifically due to the Innovation Act of 1999 or to the creation of the ANR, the national institution dedicated to project funding on a nation-wide basis, or even of the Instituts Carnot favoring the technology transfer, or a combination of all these public policy implementations. However, it appears clearly that the public policy in its globality triggered the post 2007 increase in propensity to invent in academia.

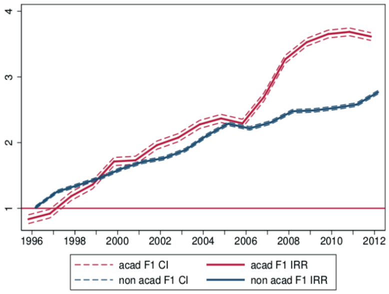
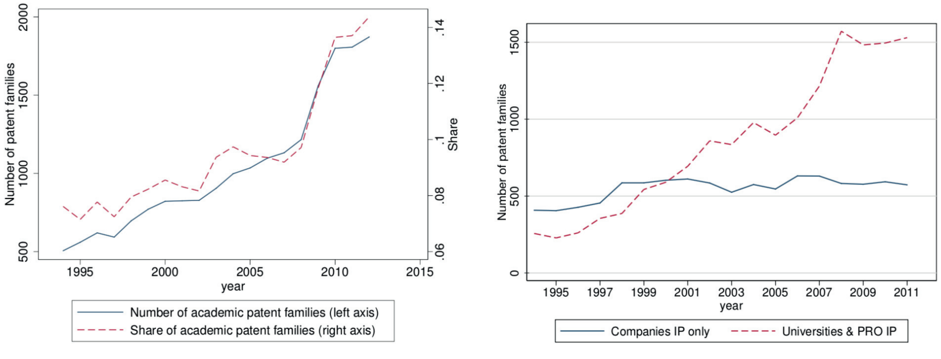


Figure 4: How the propensity for invent is affected over the years for non-academic (left panel) and academic (right panel) inventions (dependent variable xˆt1).

The adjustment of technology transfer strategies

If policies affected the propensity to invent, they may also have affected the propensity to assign the rights of those patents to universities. We investigate now how the propensities to invent vary according to the type of assignees.

We have extracted all patents (from the same offices as before) and identified which ones are owned by universities, government labs and schools, it will be our universities and PROs category, and those owned by companies will be the category of the same name. The left graph of Figure 5 shows that academic institutions have taken an increasing part of patenting activity in France over the last 18-year period, from less than 8% to more than 14% of all patent families in 2012. That variation corresponds to an increase in the number of patent families by a factor 2.5. If we focus on the most recent period, we observe an impressive shift over the 2007-2010 period: the number of patent families owned by universities and PRO raised from less than 10% to nearly 14% of all patent families.



Note: In the left graph, the identification of university-owned (and PRO) patent families (in the EPO, INPI and USPTO) is based on the identification of at least one public player among the applicants (a university, school or PRO). We restrict to the patents one inventor of which at least is localized in France. The right panel is based on patent families invented by at least one of our professors and researchers, patent co-owned by universities and PRO and companies are included into the universities and PRO category.

Figure 5: The evolution of university owned and non-university owned patents in France.

Therefore, we have identified in the academic patents how they distribute according to their type of assignee. The yearly volume of patents owned by companies only remains very stable.

Meanwhile, the volume owned by universities and PRO exclusively or with companies has remarkably increased. If at this point it is difficult to attribute with certainty this evolution to one piece of policy or the other, this may indicate an effect of the policies rather on the ownership structure of academic patents than on the invention phenomenon itself.

We therefore make the same calculations as in the previous subsection, relying on a controlled population of professors and researchers over the years. Figure 6 shows our results. The academic patents owned by companies only remains very stable over the whole period while those owned by university only or in shared property between university or PRO and companies is in rapid growth. The universities seem to increasingly retain ownership over their inventions. Next, we run negative binomial regressions to estimate how the probability to invent varies over time, for each different type of assignee.

Figure 7 shows how the incidence ratio rates vary according to the year dummy, for each type of assignee. The propensities to invent are clearly following divergent paths. Even though the propensity to assign right to the business sector are higher from 1998 ongoing as compared to 1995, it remains particularly constant from then until 2012. As a comparison, the propensities are up to 6 times higher when it comes to assign rights to the universities and PRO only, and up to 20 times to co-assign rights between the business and universities and PRO sectors. This points out a change in the universities technology transfer strategies.

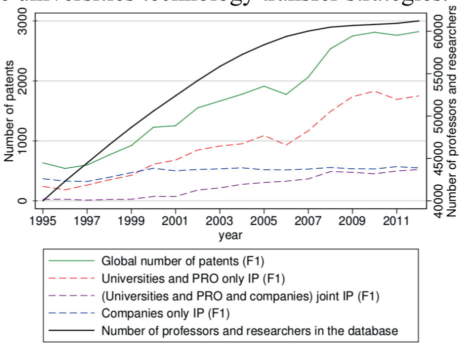
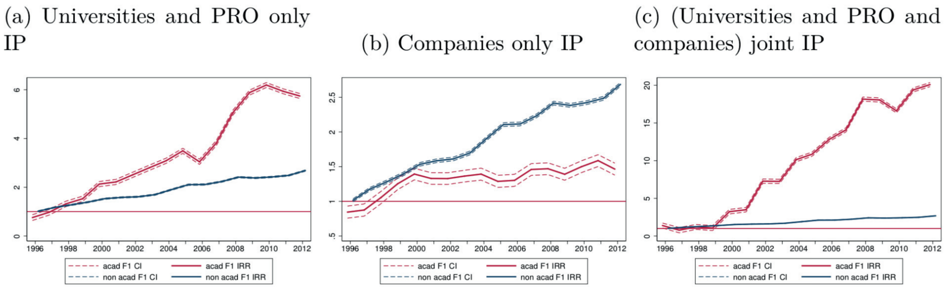


Figure 6: The evolution of academic patenting in France, with respect to the reference population.

However, our database evidences a great heterogeneity in the practices of universities concerning the allocation of IPR. For example, while the University of Grenoble Alpes historically retained ownership of more than 80% of its inventions over the period 1995-2012, the University of Lorraine uses to wave more than 40% of its IPR to companies. Moreover, when most institutions increasingly retain the ownership, some did not change their TT strategy. Public research organizations such as CNRS, CEA and INSERM waive their property rights for a small share of patents only (usually between 20% and 30%) whereas the main universities usually leave between 40% and 60% of their patents. Increasingly, patents are co-owned by universities (and PROs) and companies, as proved by some large universities such as Lyon 1, Bordeaux 1, or Rennes 1 which share the property of a quarter to a third of their patents. If we focus more specifically on the five main universities (Aix-Marseille, Paris 11, Paris 6, Grenoble 1 and Lyon 1), we notice that Grenoble 1 keeps the full property of 53% of its inventions, while Aix-Marseille abandons almost the same share to companies and Lyon 1 co-owns 29%.


(a) Universities and PRO only IP

(b) Companies only IP

(c) (Universities and PRO and companies) joint IP

Figure 7: How the propensity for invent a patent for different types of IP owners is affected over the years (dependent variable $\hat{x}_t^1$ per type).

As evidenced previously, we observe a progressive shift in technology transfer strategies of universities and PROs in France. Indeed, these public entities are increasingly retaining property over their inventions. Companies lose the full ownership of academic inventions to the benefit of co-ownership with the universities and PROs until 2006, and then to the universities and PROs full property.

A one size-fits all reform may be inappropriate as regard to the diversity of the French institutional landscape in terms of IPR management experience. More than 50% of universities and PRO never applied for a patent during the 18 years under study. During the pre-reform period (1995-1999), 72 institutions applied for 2 patents or less (the entrants) and 17 institutions filed 10 patents of more (the incumbents). This last group includes the universities Paris 7, Strasbourg, Montpellier, Grenoble Alpes and Paris 6 and they applied for 51 to 88 patents in the 5 pre-reform years. Finally, the level of experience of 28 institutions which applied for 3 to 9 patents is unclear so we exclude them from the sample. For each group separately, we compute the growth rate of their share of patents co-owned with companies (Figure 8). It appears clearly that universities without pre-reform IP management experience used the co-ownership regime 6 times more often in 2000-2002 as compared to 1997-1999. They used it more and longer than incumbents.

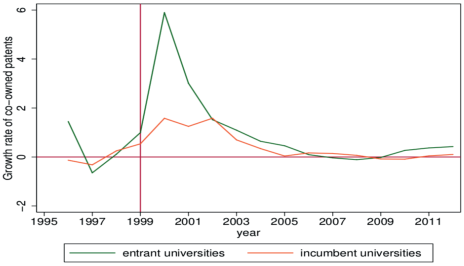


Figure 8: The growth rate of co-ownership share by entrants vs. incumbents.

Conclusion

After the implementation of the Bayh-Dole Act in the 80's in the USA, numerous studies followed in order to assess the real impact of this law and its efficiency (Mowery et al. (2001), Mowery et al. (2015), Mowery & Ziedonis (2002), Baldini, Grimaldi & Sobrero (2007) to mention just a few). In France, similar reforms since 1999 have been implemented in order to progressively reshape the research context of academic professors and researchers. However, still little is known on the impact of these policy initiatives on the universities technology transfer. First of all, our study aims at evaluating the amplitude of the academic patenting phenomenon in France over a large period. Next, we assess the impact of the different policies on the technology transfer strategies of universities and public research organizations.

According to our results, academic patenting represents up to 10% of all inventions in France over 1995-2012 (5.5% before the reform), that is more than 56,000 patents. Controlling for the evolution of the population and the confounding effect of different individual characteristics, we observe that the propensity to invent is increasing over time. After the Innovation Act was introduced, university professors and PRO researchers increasingly patented with a 6-year lag. Then, we explore the ownership structure. The share of patents owned by universities and PRO only or co-owned with a company is clearly increasing over time, while the share of patents owned by companies only remains constant. The universities are thus increasingly retaining ownership over their inventions. Looking at universities and PRO IP management strategies, we observe they adopted a transitory co-ownership strategy before starting to significantly apply on their own (in particular entrants), which we interpret as corresponding to a "learning phase".

References

Baldini, N., R. Grimaldi & M. Sobrero (2007). To patent or not to patent? A survey of Italian inventors on motivations, incentives, and obstacles to university patenting. *Scientometrics* 70(2):333–354.

Carayol, N., L. Cassi & P. Roux (2015). Unintended closure in social networks. The strategic formation of research collaborations among French inventors (1983-2005). *mimeo GRETHA working Paper*.

Carayol, N. and M. Matt (2004). Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy* 33:1081–1102.

Czarnitzki, D., K. Hussinger, P. Schliessler & A. Toole (2015). Individual Versus Institutional Ownership of University-Discovered Inventions. *ZEW Working Paper* 15-007.

Ejermo, O. & H. Toivanen (2018). University invention and the abolishment of the professor's privilege in Finland. *Research Policy* 47(4):814–825.

Geuna, A. & F. Rossi (2011). Changes to university IPR regulations in Europe and the impact on academic patenting. *Research Policy* 40(8):1068–1076.

Grimaldi, R., M. Kenney, D. Siegel & M. Wright (2011). 30 years after Bayh–Dole: Reassessing academic entrepreneurship. *Research Policy* 40(8):1045–1057.

Hvide, H. & B. Jones (2018). University Innovation and the Professor's Privilege. *American Economic Review* 108(7):1860–98.

Lissoni, F., P. Llerena, M. McKelvey & B. Sanditov (2008). Academic patenting in Europe: new evidence from the KEINS database. *Research Evaluation* 17(2):87–102.

Lissoni, F. (2012). Academic patenting in Europe: An overview of recent research and new perspectives. *World Patent Information* 34(3):197–205.

Magerman, T., J. Callaert & B. Van Looy (2017). Science Informing Technology: Towards a Concordance Table Using Large Scale NPR-WOS Matching. *Technical report KU Leuven*.

Mowery, D. & A. Ziedonis (2002). Academic patent quality and quantity before and after the Bayh–Dole act in the United States. *Research Policy* 31(3):399–418.

Mowery, D., B. Sampat & A. Ziedonis (2002). Learning to patent: Institutional experience, learning, and the characteristics of US university patents after the Bayh-Dole Act. *Management Science* 48(1):73–89.

Mowery, D., R. Nelson, B. Sampat & A. Ziedonis (2001). The growth of patenting and licensing by US universities: an assessment of the effects of the Bayh-Dole act of 1980. *Research policy* 30:99–119.

Mowery, D., R. Nelson, B. Sampat & A. Ziedonis (2015). *Ivory tower and industrial innovation: University-industry technology transfer before and after the Bayh-Dole Act*. Stanford University Press.

Trajtenberg, M., R. Henderson & A. Jaffe (1997). University versus corporate patents: a window on the basicness of invention. *Economics of Innovation and New Technology* 5(1):19–50.

The Diffusion of Zebrafish in Latin American Biomedical Research

A Study of Internationalisation Based on Bibliometric Dynamic Network Data

Rodrigo Liscovsky Barrera

rliscovs@ed.ac.uk

University of Edinburgh, Institute for the Study of Science, Technology and Innovation (ISSTI), Old Surgeon's Hall, High School Yards, Edinburgh (UK)

Abstract

Zebrafish (*Danio rerio*) has become an attractive experimental animal model in contemporary biomedical research. This study analyses the international dimension of the diffusion process of this popular model from a multi-level perspective by looking at the collaborative links forged by research institutions and the patterns of mobility displayed by individual researchers across time. To do so, first, it takes insights from diffusion of innovations theory and applies novel statistical techniques to measure network dynamic exposures using bibliometric data. Second, it builds on recent methodologies for developing scientific mobility indicators also based on bibliometric data. The analysis compares network diffusion patterns in selected countries with those displayed in Latin America. The growth of zebrafish research in this region is unprecedented and constitutes an interesting case to answer, from a new angle, long-standing questions on internationalisation dynamics put forward by STS scholars. Results show that a slow and progressive diffusion process has driven the use of zebrafish where high levels of network exposure (resorting to others with prior experience in the use of the model) play an important role. In the case of Latin America, however, expertise-based collaborations are not predominantly international yet international mobility is a common characteristic among early adopters.

Background: zebrafish research in Latin America

Although the use of zebrafish in biomedical research has been growing steadily on the world-stage after the successful completion of the first two large-scale random mutagenesis screens of zebrafish embryos in 1996 (see Nüsslein-Volhard, 2012), in Latin America this growth has been unprecedented (Buske, 2012). The promise of zebrafish in this region is mainly due to the economy of the model allowing average laboratories which often operate in national scientific systems shaped by budget constraints and with less well-developed science infrastructure, to conduct word class research (Allende et al, 2011). This has translated into a remarkable increase of publications from Latin American researchers, which has outpaced the growth of publications in other model organisms. Moreover, international collaboration seems to play a key role as expressed by the remarkable growth of internationally co-authored publications (see *figures 1a* and *1b*).

An important source that contributed to the spread of zebrafish in the scientific community are stock centres. However, the formation of a regional stock centre, as those existing in the U.S., Europe and China (Friedmann et al, 2015), was considered out reach for this region given the budgetary efforts this would have required (Allende et al, 2011: 31). Instead, in December 2010 a group of principal investigators from Argentina, Brazil, Chile and Uruguay decided to create the Latin American Zebrafish Network (LAZEN) with the aim of "enabling resource sharing, starting collaborative research, identify funding opportunities and to enhance training" (ibid: 31). In this sense, one of the central discussions of the regional meetings, which later included research groups from Colombia, Ecuador, Mexico and Peru, revolved around how to disseminate and standardize regional zebrafish work. This has proven to be especially useful because the region is characterised by a lack of access to commercial holding systems and many of the researchers are self-taught in the art of fish husbandry (ibid: 33).

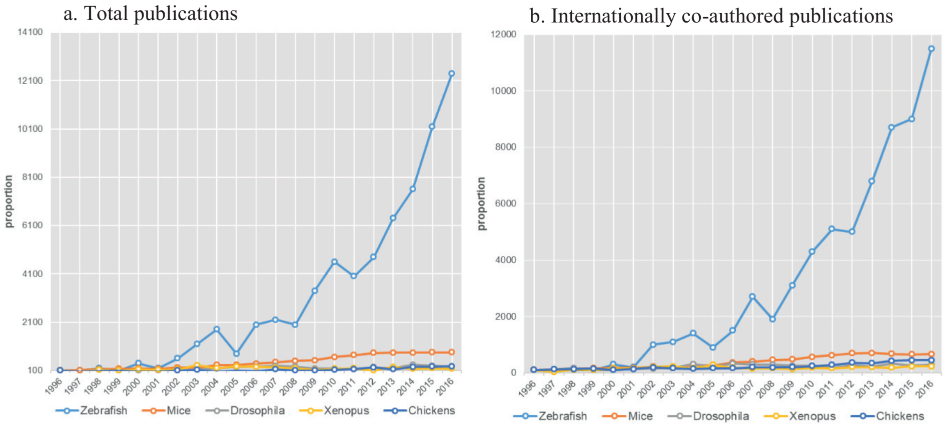


Figure 1. Evolution of publications in Latin America by type of model organism.
1996 = base 100. Source: CWTS Scopus XML internal database. Oct. 2018.

The reasons for the growing use of zebrafish in biomedical research are manifold including high genetic homology to humans, being a sufficient physiological complex in-vivo model that reproduces quickly and abundantly, having an external and transparent development, the fact that it is space/cost-efficient and ease for experimental manipulation among other factors (Kalueff et al, 2014a; 2014b). In most cases, the choice for this model organism is said to vary according to the research questions. In this sense, zebrafish is considered an effective third-path between simple multicellular organisms and complex and expensive vertebrate models such as mice and rats (Bateson Centre, 2014).

However, are scientific and practical factors the only factors that explain the rapid diffusion of zebrafish as an attractive novel organism in biomedical research? Do social factors including interpersonal links, shared expertise and mobility patterns influence this growth as well? How do these factors help explain the introduction of the model in Latin American biomedical research?

Science studies focused on Latin America have noted the prevalence of centre-periphery dynamics to explain how researchers from the region are incorporated and contribute to the global production and circulation of knowledge. Adopting insights from dependency theory and world-system approach, some scholars have noted how internationalisation played a key role in key disciplines such as biomedicine in determining the orientation of local research agendas (Kreimer 2006). In this sense, the personal relationships that local scientists forged with research leaders from central countries often characterised by a ‘subordinated integration’ where the latter retain a ‘cognitive control’ of local research (Kreimer, 2013: 443) via the exportation of techniques and research choices. Internationalisation and network dynamics are thus key elements to examine in the diffusion of zebrafish as a model organism in Latin American biomedical research.

Approach: diffusion theory, dynamic networks and international mobility

This study relies, on the one hand, on contributions from diffusion of innovations theory. Having its roots in anthropology, economics, sociology and other disciplines, diffusion theory seeks to explain how new ideas and practices spread within and between communities (Valente, 2010). The premise, confirmed by empirical research, is that new ideas and practices spread through interpersonal links (Ibid). Hence, social factors rather than economic ones are key factors influencing adoption behaviour (Valente and Rogers 1995; Valente, 2010).

The emphasis of the role of interpersonal links implies adopting a social network approach to study diffusion processes. Diffusion network studies provide useful empirical data for measuring network influences on diffusion (Valente, 2010). However, most of these studies are based on static networks (Valente, 2015) and fail to consider the dynamic nature of diffusion processes. Therefore, in order to study dependency dynamics, this study is based on previous dynamic network studies which investigate how exposure to prior adopters is related with adoption (Valente, 1995; 2005; 2015).

Second, this paper considers the key role that scientific mobility plays in knowledge diffusion and exchange processes (Robinson-Garcia et al, 2018). The study of scientific mobility is closely related to the study of internationalisation dynamics where researcher's international trajectories together with international collaboration proved to be highly correlated as measures of international engagement (Wagner & Jonkers, 2017: 32).

Data

Database 1: diffusion through collaboration

The data was retrieved in October 2018 from CWTS' Scopus custom XML database, which covers publications from 1996 onwards. In order to select publications on zebrafish in biomedical research, a SQL query was designed that used zebrafish's descriptors as defined in the Medical Subject Headings (MeSH) developed by the NIH. MeSH provides hierarchically-organised terminology for indexing and cataloguing of biomedical information such as MEDLINE/PubMed. The designed query parameters helped searching for zebrafish's MeSH descriptors[1] in the abstracts of publications that resulted in an initial sample of 28,973 papers published from 1996 to 2016.

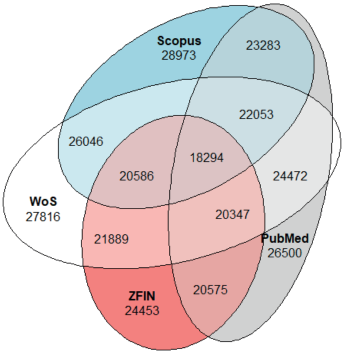


Figure 2. Venn Diagram of SQL query results depicting unique and overlapping publications identified in each bibliometric database.

To validate the representativeness of collected papers (that is, papers reflecting the use of the model in biomedical research) a two-step process was followed. First, the same query was

[1] MeSH descriptors: *B. rerio, Brachydanio rerio, D. rerio, Danio rerio, Zebra Fish, Zebra Fishes, Zebra danio, Zebrafishes.* Source: MeSH Descriptor Data 2018.

replicated in alternative bibliometric databases including the Web of Science (WoS), PubMed and the online repository of publications developed by the Zebrafish Information Network (ZFIN) (see ***figure 2***). Publications overlaps across datasets were then analysed and the results showed that with WoS publications, there is a match of 49.4% whereas with ZFIN there is a 59.1% overlap. Overlaps with ZFIN are largely explained by the fact that the online repository includes a large number of PhD and master theses and non-peer-reviewed publications. A close inspection of journals in the WoS sample showed a variety of journals not belonging to the Life Sciences. However, with publications from PubMed there is match a 94.7% suggesting that a wide majority of retrieved articles are medical research articles. Second, to ensure retrieved publications refer to research that made explicit use of zebrafish as an experimental model (be that as the main model or as a complementary one), both the query and a small random sample of articles were reviewed and validated by an expert in biomedical research and zebrafish.

Sampling:

Unique research institutions were identified using their affiliation ID in Scopus. Furthermore, as the focus is set on the analysis of diffusion via collaborative links, papers with a large number of contributors were dismissed in order to select papers where institutions are thought to have made a substantial contribution. Papers with no more than 15 institutions were chosen by inspection of the distribution of the number of institutions per paper and the cumulative density function. This approach is consistent with previous studies (Deville et al, 2014; Martin et al, 2013). As seen in ***figure 3***, a deviation from the power-law fitting line is observed for those papers containing more than 15 institutions in the sample. Applying this filter results in a set of 28,624 unique publications, which means that only 1.20% of all collected papers are missing from the initial query.

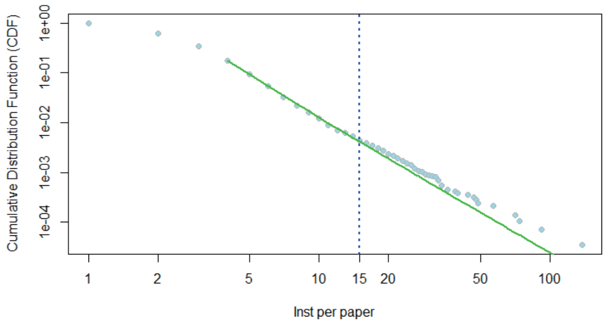


Figure 3. Cumulative Density Function of number of institutions per paper. The vertical line falls at 15 institutions corresponding roughly to the point where the distribution deviates from the power law

Data structuring:

'netdiffuseR' is an R package developed by Young et al (2018) that allows conducting empirical statistical analysis, visualization and simulation of diffusion and contagion processes on networks via so-called '*diffnet*' objects. These objects are lists that hold a series of other objects such as a graph, *toa* (an integer vector of size *n* that holds information about the time of adoption), a matrix of cumulative adopters and vertex's both static and dynamic attributes among other. To build a diffnet object, the package can read data in various formats including cross-sectional and longitudinal surveys as well as edge-lists. In this case, data was structured as a longitudinal co-authorship edge-list hence the resulting graph is dynamic and contains both static and dynamic attributes. To assure that institutions listed in the edge-list are institutions that used the model consistently, only those who published in more than one year were included in the sample. The application of this filter returns a sample of 3,771 unique institutions of which 124 are from LAZEN countries[2].

The *time of adoption* (toa) of each institution was computed by taking the year of first publication on zebrafish. Other attributes include country of affiliation (static), the size of the network at each time point (dynamic), the publication year (dynamic) and the number of collaborative countries at each time point (dynamic).

The resulting dynamic graph is composed of 7,220 nodes (including adopters and non-adopters) and 21 time points (1996 - 2016) recording 64,419 co-authorship events in total.

Database 2: mobility trajectories of zebrafish researchers

In order to analyse patterns of mobility among zebrafish researchers, an additional dataset was built comprising all the records of publications by individual authors included in the zebrafish dataset (database 1). Individual scientists were identified using their unique author ID in Scopus which combines all publication records from an author and his/her possible name variants. Recent studies have reported that the use of the Scopus author ID allows conducting precise and reliable analyses (Moed et al., 2013; Aman, 2017). The whole record of publications was also extracted from CWTS' Scopus custom XML database although in this case the constructed debased also includes publications that are not about zebrafish. In total, the mobility database contains 2,197,571 publication records.

Sampling:

Only authors that published more than once on zebrafish were selected. The application of this filter returned a total sample of 25,511 unique researchers which represents 31.49% of the total number of researchers included in the zebrafish dataset (n = 81,012). In addition, in order to identify authors from LAZEN countries, the previous cut-off parameter was further restricted to include only authors that published a minimum of two years under an affiliation from one of the countries that take part in the regional network. In total of 170 unique LAZEN authors were identified. The affiliations histories of each LAZEN author identified were then manually inspected in order to identify and exclude authors with multiple affiliations that do not necessarily reflect a Latin American researcher.

Data structuring:

To build the mobility dataset, institutional affiliation changes from each individual's history of publications were analysed following the classification developed by Robinson-Garcia et al (2018), which distinguishes between migrants and travellers.

In a researcher mobility trajectory (see ***table 1***), '*Directionality*' indicates whether it is possible to reliably establish if an author has been chronologically affiliated first to his/her country of first affiliation and then to any other country, which is different from the country of origin. '*Rupture*' refers to a mobility of event where a researcher's country(ies) at t_n (t=0) are not found among the affiliations of the researcher at t_n+1 (Ibid).

[2] Argentina, Brazil, Chile, Colombia, Ecuador, Mexico, Peru and Venezuela.

Table 1. Sample of a researcher's mobility trajectory. Directionality implies a researcher gains additional affiliations while maintaining affiliation with his/her country of origin.

Au-id	*pub year*	*country*	*t*	*event type*	*mobility class*
60001812493	2003	Argentina	0	Origin	Migrant
60001812493	2007	Argentina	4	Origin	Migrant
60001812493	2007	United Kingdom	4	Directionality	Migrant
60001812493	2007	Argentina	4	Origin	Migrant
60001812493	2008	United Kingdom	5	Rupture	Migrant
60001812493	2009	United Kingdom	6	Rupture	Migrant
60001812493	2010	United Kingdom	7	Rupture	Migrant
60001812493	2011	United Kingdom	8	Rupture	Migrant
60001812493	2012	Argentina	9	Origin	Migrant

Consequently, researchers can be classified as:

- *Migrant*, if they display a directional mobility event and a point of rupture with their country of origin (t=0) at any point in time;
- *Directional Travellers*, if they display a directionality event but no rupture throughout their publication history;
- *Non-Directional Travellers*, if they have at least one mobility event but no directionality and no rupture with their country of origin.
- *Not mobile*, if they lack any mobility event (Ibid).

The resulting database contains 214,905 mobility events expanding from 1996 to 2016. Furthermore, mobility events for the 170 Latin American researchers identified were further classified as:

- *LAZEN Origin*, if researchers' country of origin belongs to any of the LAZEN countries;
- *International Origin*, if researchers' country of origin is different from any of the LAZEN countries;
- *International Migrant*, if they have had a rupture event at a country that is not from the LAZEN region;
- *LAZEN Migrant*, if they have had a rupture event at any of the LAZEN countries;
- *International Directionality*, if they added affiliations from countries outside the LAZEN region;
- *LAZEN Directionality*, if they added affiliations from any of the LAZEN countries.

Methods:

Diffusion through collaboration

Adoption of an innovation does not occur immediately and automatically. Rather, adoption is a process that involves different stages such as becoming aware of the innovation, learning more about it, trying it and eventually adopting it (Valente, 2005). To observe this process, we can examine the evolution of the number of (cumulative) adopters across time. Trends of adoption can be further compared between communities of adopters – in our case, research institutions from different countries. The adoption process for Latin American research institutions is therefore compared with those observed in the U.S., the U.K, Germany and China, which are the main centres of scientific production on zebrafish.

However, despite its robustness, this measure – and other standard diffusion measures such as the *Rate of Adoption* or the *Hazard Rate*[3] – do not help explain dynamics at the individual level and at each point in time for the duration of the diffusion process. Fortunately, more dynamic models have been developed which treat time more explicitly and measure how an actor's social network influences the adoption and diffusion of innovations across time. To do this, these models – and diffusion research in general – rely on a key measure referred as *network exposure* (Valente, 2005). Network exposure is measured with the following equation:

$$E_t = (S_t \text{ X } [x_t \text{ o } A_t])$$

Where S_t is the network in time *t*, x_t is an attribute vector of size *n* at time *t*, A_t is the t-th column of the cumulative adopters matrix (a vector of length *n* with a_{ti} = 1if *i* has adopted at or prior to *t*), *o* is the kronecker product (element-wise), and X is the matrix product (Young et al, 2018). This means that for each individual institution, we can analyse whether it has adopted zebrafish as an experimental research model and how many of its immediate collaborators (co-authors) have adopted it at each point in time (outdegree), at or prior to ego's adoption. Moreover, thresholds levels can be calculated which are each vertexes' exposure at the *time of adoption* (toa).

In line with the previous measure, national and international exposure and threshold levels can be calculated for any given research institution in the social system. This is done by creating sub-graphs partitions of national collaborations and then calculating the proportion of collaborators who are adopters but from a country that is different to ego's country and then considering the whole size of ego's social network at each point in time. Therefore, *international network exposure* (E_i) is calculated as follows:

$$E_i = \frac{(E_o - E_n)}{Wi}$$

Where E_o is ego's simple exposure matrix at time *t*, E_n is on Ego's matrix exposure to national collaborators at time *t* (the network exposure matrix calculated from the sub-graph) and Wi is ego's social weight matrix at time *t*.

Considering this, we could expect a decrease in the effect of external expertise over time; a statement that is consistent with the general diffusion model, which argues that early adopters are influenced by sources external to the community (Valente, 2015: 90). That is, because there are no or few adopters in the community, early adopters tend to rely on information external to the community and, as diffusion develops, external dependency will tend to decrease and thresholds levels increase. Contrary to this, based on existing analyses on the influence of scientific internationalisation dynamics in Latin America, among early adopters of this region we should expect high levels international exposure and thresholds. In other words, Latin American research institutions will tend to rely on research institutions from central countries as expertise on the use of the model is expected to diffuse from the centre to the periphery.

[3] The Rate of Adoption is commonly referred as the relative speed in which actors adopt a given innovation (Rogers, 2003). The Hazard Rate refers to the instantaneous probability of adoption at each time representing the likelihood members will adopt at that time (Allison 1984).

Diffusion through mobility

The final part of the analysis is devoted to analyse the role of internationalisation in the diffusion process via bibliometric indicators of mobility. The interaction of researchers' mobility trajectories with their respective time of adoption (toa) is analysed. Mobility events and classes are computed for each researcher in the social system and grouped by distinct years of first adoption. This allows observing the extent to which an individual' mobility experiences have an influence on his or her timing for adoption. In addition, for each Latin American researcher, his/her geographical location at *toa* is recorded.

Taken together, these two methodologies provide a more dynamic and detailed account on the weight of internationalisation dynamics in the diffusion of zebrafish as a model organism in biomedical research.

Findings and discussion

Adoption process:

Instead of an exponential and rapid growth as documented elsewhere, a rather slow and progressive diffusion process has driven the use of zebrafish in biomedical research. As seen in ***figure 4***, the proportion of cumulative adopters (research institutions) increases gradually in all of the selected countries. In the case of China, this slow trend is even more noticeable although adoption among Chinese research institutions seems to accelerate from 2007 onwards. In both Latin America and China, however, adoption among research institutions displays a clear upward trend in the least years of the period of reference.

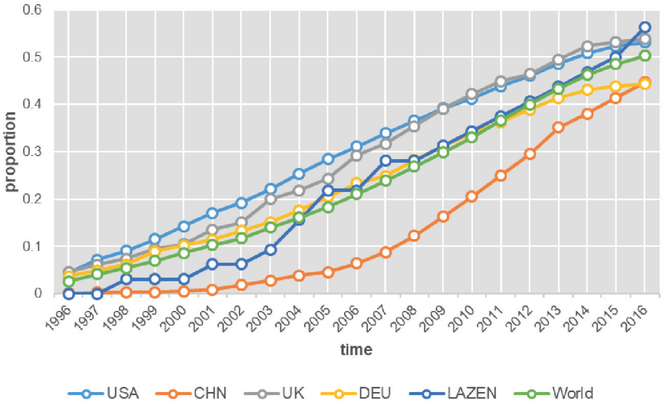


Figure 4. Evolution of cumulative Adopters Compared (%)

Network exposure and thresholds:

When looking at the network exposures at time of adoption (threshold) we can see that in both the global and Latin American communities, high network exposure levels are a common characteristic with a majority of adoptions taking place in the last years of the period of reference (***figure 5***). Particularly, research institutions in Latin America largely adopted zebrafish as an experimental model organism when a majority of their institutional collaborators (threshold >= 60%) had already adopted it.

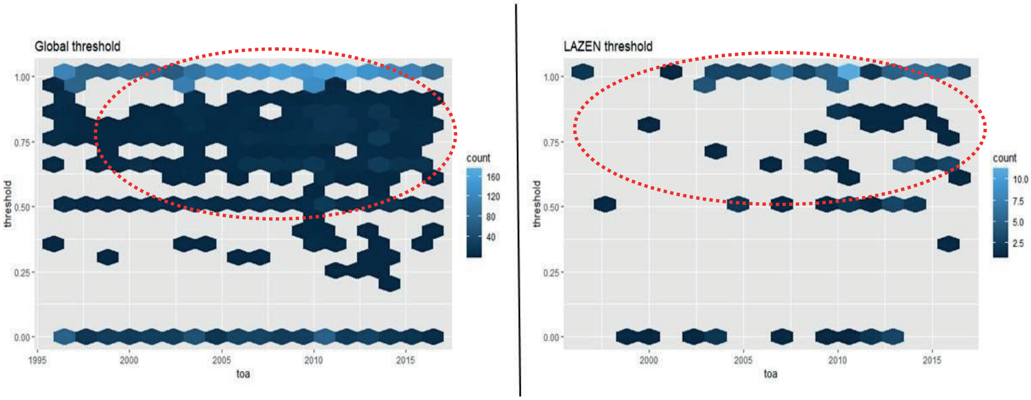


Figure 5. Research institutions' network exposure at adoption. Thresholds are each vertexes' exposure at the time of adoption (toa). *n* = 3,771 (global community) / *n* = 124 (LAZEN community)

This could indicate a prevalence of expert-based dependency dynamics in the adoption behaviour of Latin American research institutions but a closer look at the origin of such expertise reveals a different picture. As seen in ***figure 6***, expertise-based collaborations among Latin American research institutions are not predominantly international. The evolution of international exposures levels in the LAZEN countries shows that access to international expertise has been growing over the last years yet the average percentage of international collaborators with experience in using the model never exceed 30%.

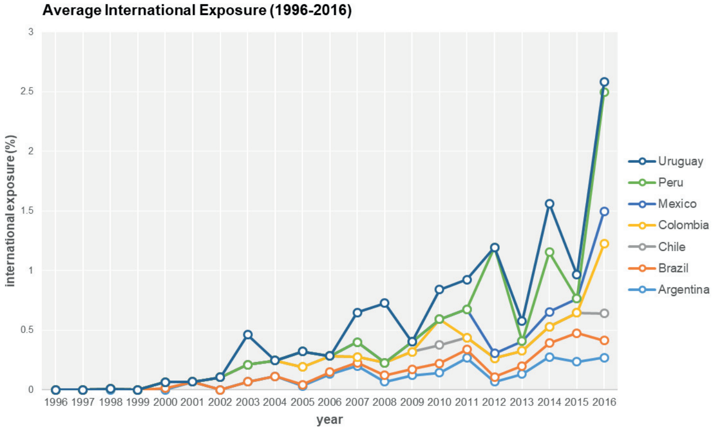


Figure 6. Evolution of average international network exposure levels among LAZEN research institutes (users).

Moreover, at *time of adoption* (toa), international exposure is even less determinant. When adopting the model for the first time, the large majority of Latin American institutions resort to national expertise or no external expertise at all. As seen in ***figure*** *7*, in most cases international threshold levels – resorting to foreign research institutions with experience in the use of the model – are equal or below 50%.

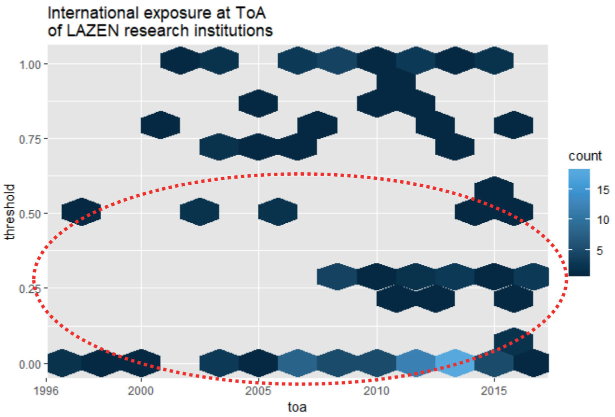
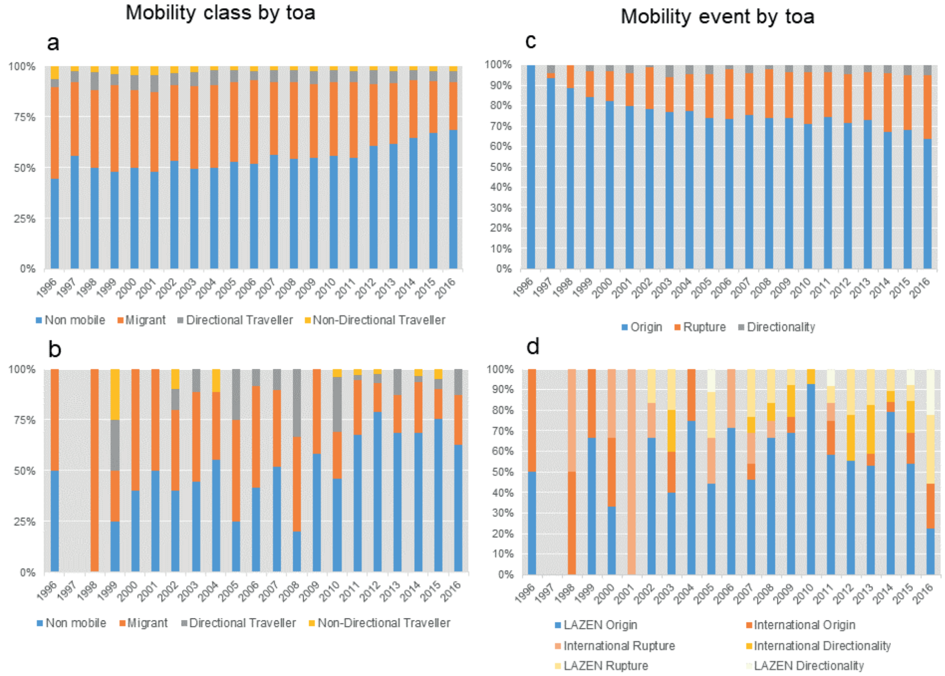


Figure 7. International network exposure at adoption for LAZEN institutions.
Thresholds are each vertexes' international network exposure at the time of adoption (toa).

International mobility
At the micro level, the distribution of international mobility patterns offers a more contrasting picture in terms of internationalisation dynamics (***figures 8a-d***). Among early Latin American adopters, the proportion of international mobile researchers – migrants, directional and non-directional travellers – (***figure 8b***) is substantially higher than that of non-mobile researchers, whereas in the global community the share of international mobile researchers has been decreasing over the years (***figure 8a***). This shows that international mobility is a key factor in the early adoption of the model, and this is especially true for Latin America researchers. The reduction of the share of Latin American mobile researchers from 2010 onwards, on the other hand, may suggest that returnees could have contributed to the diffusion of the model among local colleagues. However, the distribution of mobility classes does not consider the type of mobility event that took place at the time of adoption for each individual researcher. ***Figures 8c*** and ***8d*** provide a more detailed snapshot in this regard. In the global community of adopters, the large majority of adoptions (researchers' first publication on zebrafish) took place at their country of origin (***figure 8c***).

Among early Latin American researchers, however, adoption seems to have taken place when they were working abroad. Many early Latin American researchers either initiated their zebrafish research abroad (international origin) or started working with this model after travelling to another country (international rupture) (***figure 8d***) – mostly the U.S., the United Kingdom or Germany. In this sense, migration seems to be a common pattern among Latin American zebrafish pioneers. On the other hand, international directionality and regional mobility events (rupture or directionality) are more frequent among lagers although adoption at LAZEN countries becomes more prevalent over the last years of the period of reference – except for 2016.



Figures 8a-d. International mobility trajectories by time of adoption. Figures 8a and 8c display mobility class and mobility events at *toa* for the global community of adopters. Figures 8b and 8d shows mobility class and mobility rupture events of the researchers from LAZEN countries.

Conclusions

Zebrafish has become an attractive model organism in biomedical research ever since the publication in 1996 of the results of the first two large screens for zebrafish mutants performed in Tübingen (Germany) and Boston (U.S.). Using bibliometric data, the present paper analysed the diffusion process of this particular model while paying special attention to the Latin American region where its growth has been unprecedented. Results show that a slow and progressive diffusion process has driven the use of zebrafish in the wider community of researchers where high levels of network exposure seemed to play an important role. However, in the case of Latin America, international-based expertise does not seem to have a crucial influence in the adoption of zebrafish as a model organism among research institutions. At the micro level, the influence of internationalisation dynamics is clearer. International mobility is a common pattern among Latin American zebrafish pioneers, whereas gaining additional country affiliations (either from countries from outside or within the region) is more frequent among lagers.

Overall, this paper presented a multi-level and more dynamic account on the weight of internationalisation dynamics in knowledge diffusion processes by combining network diffusion measures with international mobility indicators. In particular, the explicit treatment of time in the presented measures allowed developing a more nuanced analysis of internationalisation dynamics. At the meso level, this was done by observing how institutions' international co-authorship networks influence the adoption and diffusion of innovations across time. At the micro level, it relied on novel methodologies for measuring the movement of scientists and examined how proposed mobility taxonomies interact with the timing of adoption of zebrafish as experimental animal model in biomedical research.

References

Allende, M.L., Calcaterra, N.B, Vianna, M.R and Zolessi F.R. (2011). First Meeting of the Latin American Zebrafish Network. *Zebrafish*, 8(1), pp.31-33.

Allison, P. (1984). *Event history analysis regression for longitudinal event data*. Beverly Hills: Sage Publications.

Aman, V. (2017). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Paper presented at the 22th Conference on Science, Technology & Innovation Indicators (STI 2017)*, ESIEE, Paris.

Bateson Centre (2014). *History of the zebrafish as a model organism*. [online] Fish For Science. Available at: http://www.fishforscience.org/zebrafish/origins/ [Accessed 15 Jun. 2018].

Buske, C. (2012). *Zebrafish research: growing demands in South America*. [online] Noldus.com. Available at: https://www.noldus.com/blog/zebrafish-research-growing-demands-in-south-america [Accessed 7 Jun. 2018].

Deville, P., Wang, D., Sinatra, R., Song C., Blondel, V.D and and Barabási A.L. (2014) Career on the Move: Geography, Stratification, and Scientific Impact. *Scientific Reports*. Vol.4. pp.1-7.

Friedmann, T., Dunlap, J.C. and Goodwin, S.F (eds.). (2015) *Advances in Genetics*, Vol.92. Academic Press.

Kalueff A.V., Echevarria, D.J. and Stewart, A.M. (2014a). Gaining translational momentum: More zebrafish models for neuroscience research. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*. Vol. 55. pp-1-6.

Kalueff A.V, Stewart A.M., Gerlai R.(2014b) Zebrafish as an emerging model for studying complex brain disorders. *Trends Pharmacol Sci*. 35(2). pp-63–75.

Kreimer, P. (2006). ¿Dependientes o integrados?: La ciencia latinoamericana y la nueva división internacional del trabajo. *Nómadas*, 24, pp 199-212.

Kreimer, P. (2013). Internacionalización y tensiones para un uso social de la ciencia latinoamericana. Del siglo XIX al XXI, In: Restrepo Forero, O. (ed.). *Ensamblando Estados*, Bogota: Universidad Nacional de Colombia. pp 437-452.

Martin, T., Ball, B., Karrer, B. & Newman, M. E. J. (2013). Coauthorship and citation patterns in the physical review. *Phys. Rev*. 88.

Moed, H.F., Aisati, M., and Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, 94. pp.929–942

Nüsslein-Volhard, C. (2012). The zebrafish issue of Development. *Development* 139, pp.4099-4103.

Robinson-García, N., Sugimoto, C.R, Murray, D. Yegros-Yegros, A., Larivière, C. and Costas, R. (2018). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, 13, pp.50–63

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York, NY: The Free Press.

Valente, T. W. (1995). *Network models of the diffusion of innovations*. Cresskill, NJ: Hampto Press.

Valente, T.W. (2010). Diffusion of Innovations. In Valente, T.W (ed.) *Social Networks and Health: Models, Methods, and Applications*. University Press Scholarship Online. pp-1-31.

Valente, T. W., (2015). Diffusion of innovations theory applied to global tobacco control treaty ratification. *Social Science & Medicine*. 145. pp-89-97.

Valente, T. W., & Rogers, E. M. (1995). The origins and development of the diffusion of innovations paradigm as an example of scientific growth. *Science Communication: An Interdisciplinary Social Science Journal*, 16. pp.238-269.

Wagner, C. S., & Jonkers, K. (2017). Open countries have strong science. *Nature*, 550(7674), 32.

Young, P. H., Dyal, S., Hayes, T., Valente, T. W. (2018). *netdiffuseR: Analysis of Diffusion and Contagion Processes on Networks*. Accessed: https://github.com/USCCANA/netdiffuseR

The use of Gold Open Access in four European countries: An analysis at the level of articles

Gunnar Sivertsen[1], Raf Guns[2], Emanuel Kulczycki[3], and Janne Pölönen[4]

[1] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education, Toyen, Oslo, Norway

[2] *raf.guns@uantwerpen.be*
Centre for R&D Monitoring, Faculty of Social Sciences, University of Antwerp, Antwerp, Belgium

[3] *emek@amu.edu.pl*
Scholarly Communication Research Group, Faculty of Social Sciences, Adam Mickiewicz University, Poland

[4] *janne.polonen@tsv.fi*
Federation of Finnish Learned Societies, Helsinki, Finland

Abstract

We assess the use and potential of Gold Open Access (OA) in Finland, Flanders (Belgium), Norway, and Poland by comparing data at the level of articles from full-coverage databases in each country. The inclusion of the journals in the Directory of Open Access Journals (DOAJ) is used as a reference to determine Gold Open Access. Gold OA is on the rise in all four countries and across fields, but some countries, especially Norway, and some fields have a substantially larger proportion of OA publications than others, with the overall share of Gold OA ranging from 5.7% to 17.3%. Especially in the SSH, a mixture of local and international journals can be found, many of which are not indexed in databases like Web of Science. As such, our results indicate that an overview of the state of Gold OA is preferably obtained by comparing DOAJ to a full-coverage database.

Introduction

Open Access (OA) to research has been one of the major topics of discussion in the area of scholarly communication for over a decade. Traditionally, a distinction is made between author self-archiving – Green OA – and publishing in an OA journal – Gold OA. A more refined model has been proposed by Martín-Martín et al. (2018). Using the terminology of these authors, we focus on *libre*, *immediate* and *permanent* access to the *accepted peer-reviewed* text of journal articles. For the sake of brevity, we will use the 'Gold OA' terminology.

The Directory of Open Access Journals (DOAJ) has emerged as one of the major sources of information on OA journals (Piwowar et al., 2018), although it does not cover all Gold OA (Björk, 2019). Basic requirements for inclusion in the DOAJ include immediate access (no embargo) to all content in the journal; having a registered ISSN; and displaying clear information on editor, editorial board, author guidelines, and article processing charges (APCs). In March 2014, DOAJ launched a new and more stringent set of criteria for inclusion (Van Noorden, 2014), leading to rejection of many journals that were previously included. In January 2019, the DOAJ covers 12,420 OA journals.

In this paper, we examine and compare to what extent researchers in four European countries/regions – Finland, Flanders (Belgium), Norway, and Poland – make use of journals that are in the DOAJ. These countries have been chosen because each maintains a full-coverage database (Sīle et al., 2018). also covers journals that are not indexed in international databases like Web of Science (WoS) as well as journals that do not register DOIs. This sets our study apart from most other studies, which rely on WoS or Scopus (Archambault et al., 2014; Bosman & Kramer, 2018; European Commission, 2019), and is especially relevant for the social sciences and humanities. As such, the study provides a complete picture of how widespread Gold OA is among peer-reviewed journal articles in these countries.

Data and methods

For each country, we take into account all peer-reviewed journal articles published between 2011 and 2017 by authors at the country's research institutions. However, the temporal and/or disciplinary scope of the Flemish and Polish data is smaller due to limitations of the data sources in these countries. Table 1 provides an overview.

The metadata of journal, conference and book publications from fourteen Finnish universities is stored in the VIRTA Publication Information Service for the period 2011-2017 (Pölönen, 2018). In case of scientific publications, it is indicated if they are peer-reviewed or not. For this study we selected peer-reviewed journal articles published in 2011-2017. For the year 2017, the data collection is not complete. Each publication is also assigned a cognitive field classification according to OECD Fields Of Science (FOS; OECD, 2015). Finnish universities' co-publications appear as duplicates, and they may have different field classification. We use deduplicated publication counts but one article can be counted in several fields. A small number of publications is assigned to category 'Other', and so can be counted toward the total for all fields but is excluded from the field-specific analyses.

Table 1. Overview of data per country

Country	Time period	Fields	Number of articles	Number of journals
Finland	2011-2017	All fields	169,231	15,434
Flanders	2011-2016	All fields	114,134	12,214
Norway	2011-2017	All fields	123,865	14,173
Poland	2013-2016	SSH	120,111	8,577

The Flemish PRFS (Engels & Guns, 2018) consists of multiple parameters, two of which count scientific publications in, respectively, the WoS and the VABB-SHW. The VABB-SHW is a database that was constructed to alleviate the shortcomings of WoS in covering the social sciences and humanities. We consider all journal articles published in 2011-2016 that are counted in the Flemish PRFS, both in WoS (n=81,936) and in VABB-SHW (n=12,635). The analysis at disciplinary level is carried out using a cognitive classification (Guns et al., 2018) based on OECD FOS; 4 publications that could not be assigned to a discipline were discarded.

Data for Norway are derived from the Norwegian Science Index (NSI), a subset of the Current Research Information System in Norway (Cristin), with complete coverage since 2011 of all peer-reviewed scientific and scholarly publications from most research organizations in the country. The bibliographic data in NSI represent books, journal articles, articles in edited volumes, and articles in peer-reviewed conference series (Sivertsen, 2016). Only journal articles are included in this study, and they are counted only once even if several institutions have contributed to them. Field classifications are mapped against OECD FOS.

The data from Poland are limited to the years 2013–2016 and to the social sciences and humanities (SSH). In these years, Polish SSH scholars published 120,111 articles (deduplicated at the national level). Disciplines or fields are assigned according to a qualification-based classification (typically based on the author's PhD). 9,147 co-authored articles involve authors from both social sciences and humanities and are assigned to both fields.

An overview of DOAJ-covered journals, obtained from the DOAJ website, is matched against each national database by comparing the ISSN(s) recorded per publication to the print and online ISSNs registered in DOAJ. Our analysis includes all journals in DOAJ, whether or not they have been accepted after March 2014. If a journal has only started providing OA content in a given year, only publications from that year or later are considered to be OA. In addition to a general overview, we also present the results for four broad fields: Natural sciences & technology, Medical & health sciences, Social sciences, and Humanities.

Results

The overall share of Gold OA articles varies considerably by country and by field, ranging from 5.7% (Social sciences, Flanders) to 17.3% (Medical & health sciences, Norway). In each of the four fields, Norway has the largest share of Gold OA articles (Figure 1). North- and West-European countries tend to exhibit similar publication patterns, while Eastern European countries sometimes behave somewhat differently (Kulczycki et al., 2018). This does not appear to carry over to Gold OA publishing, at least not in the SSH: the share of Flemish OA publications is lower in both social sciences and humanities than any of the other three countries. This suggests that national context and incentives may play an important role.

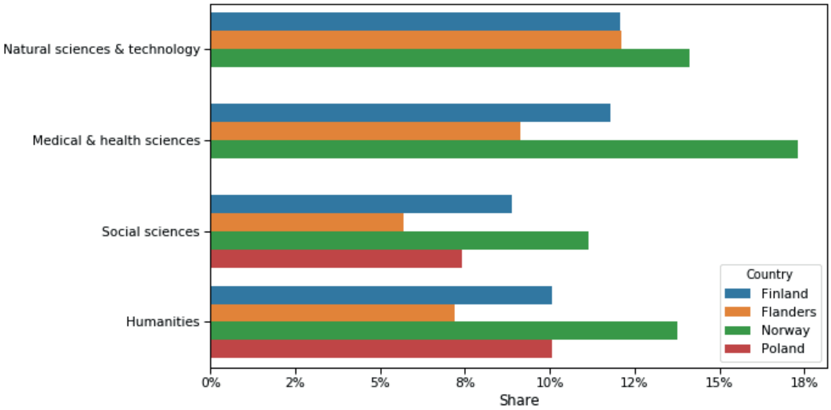


Figure 1. Share of Gold OA articles per field and country

The differences between countries and fields notwithstanding, the overall trend is clear: the share of Gold OA articles is linearly increasing (Figure 2). This increase may be due to multiple factors: the establishment of new Gold OA journals, changes to the business models of existing journals, and changes in journal choice of researchers. Figure 2 suggests that the ratios between the four countries are mostly stable, with Norway having the largest share of OA, followed by Finland and Poland, and finally by Flanders. The recent steep increase for Norway in the SSH is partly due to the establishment of a national OA platform for the most central journals published in the Norwegian language in SSH disciplines (Sivertsen, 2018).

Table 2 displays the 5 most used OA journals in Finland, Flanders and Norway. The top-5 tends to be dominated by international journals that are mostly multidisciplinary or from the natural sciences. Only the large multidisciplinary journals *PLOS One* and *Scientific reports*, as well as *Journal of High Energy Physics*, figure among the most used OA journals in all three countries. Because the Polish data is limited to the SSH, the Polish top-5 is completely different and does not contain any WoS-indexed journals.

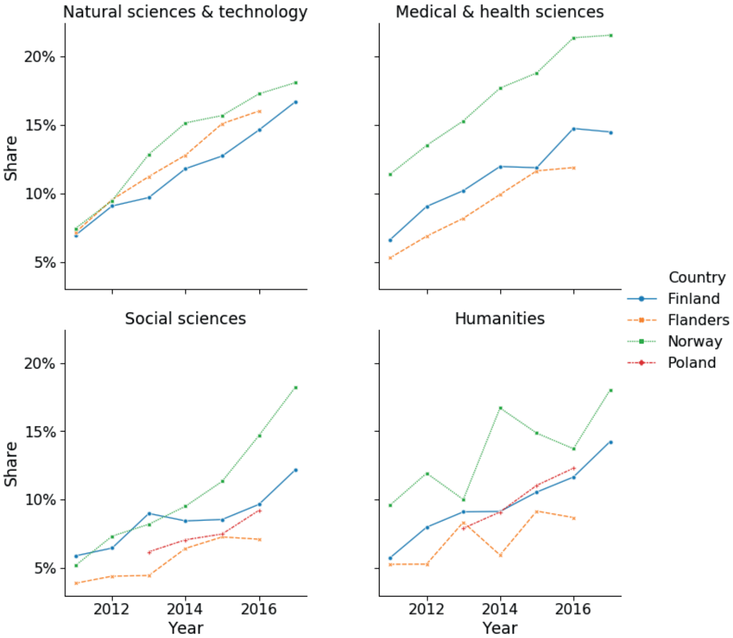


Figure 2. Evolution of share of Gold OA articles per field and country

Table 2. Top-5 most used OA journals per country

Finland	Flanders	Norway
PLoS ONE	PLoS ONE	PLos ONE
Scientific reports	Scientific Reports	Scientific Reports
Atmospheric chemistry and physics	Optics Express	BMC Public Health
Nature communications	Journal of High Energy Physics	BMJ open
PLoS genetics	BMC Public Health	Atmospheric chemistry and physics

For each of the four fields, we investigate one discipline in more detail: Biological sciences (Natural sciences & technology), Clinical medicine (Medical & health sciences), Educational sciences (Social sciences), and Languages and literature (Humanities). As can be seen from Table 3, the variability between disciplines and countries is, again, substantial. First, some disciplines are an order of magnitude larger than others in terms of number of articles. These size differences are not similar across countries, e.g., Educational sciences appears to be much larger (relatively speaking in terms of the number of articles) in Finland than in Flanders and Poland. Second, the share of OA publications of a discipline seems to be dependent on local circumstances.

Table 3. Number of publications and OA share per discipline and country

		Biological sciences	Clinical medicine	Educational sciences	Languages and literature
Finland	Total	12,375	32,291	4,086	2,656

	OA	2,075	3,088	583	275
	Share of OA	16.8%	9.6%	14.3%	10.4%
Flanders	Total	12,608	18,021	1,444	2,608
	OA	1,450	1,139	104	276
	Share of OA	11.5%	6.3%	7.2%	10.6%
Norway	Total	14,148	32,755	4,899	3,575
	OA	1,544	5,545	808	646
	Share of OA	10.9%	16.9%	16.5%	18.1%
Poland	Total	-	-	6,985	17,917
	OA	-	-	617	1,280
	Share of OA	-	-	8.8%	7.1%

We also investigate the most used OA journals per discipline per country. The distribution of papers per OA journal tends to be highly skewed, with the top-10 journals typically accounting for 50% or more of all OA publications in a given discipline. It is noteworthy that the two most important OA journals in Finland for both Biological and Medical & health sciences are the large multidisciplinary journals *PLoS ONE* and *Scientific reports*. Since disciplines in Flanders are currently assigned at the journal level, publications from either journal are treated as multidisciplinary, even if they may be about, e.g., biology.

Table 4. Number of non-English or multilingual journals among 10 most used OA journals

	Educational sciences	Languages and literature
Finland	4/10	6/10
Flanders	3/10	4/10
Norway	9/10	5/10
Poland	9/10	6/10

All top-10 journals for Biological sciences and Clinical medicine are English language, mostly published in the UK, US, Switzerland (Frontiers) and the Netherlands (Elsevier). Exceptions include Bulgaria (*ZooKeys*), Sweden (*Acta dermato-venereologica*), and Italy (*Haematologica*). The situation is rather different in the SSH, where we also find journals published in other languages (Table 4). These may target a local audience through use of the local language, but there are also examples of non-English journals that reach a broad international audience (e.g., *Zeitschrift fur interkulturellen Fremdsprachenunterricht* or *Teoría de la Educación*; cf. Sivertsen, 2018). In addition, there are several instances of multilingual journals, which accept articles written in two or more different languages. As for country of publication, we observe that in some cases the top-10 is largely international, albeit with greater geographical variation than for the natural and medical sciences (e.g., Educational sciences in Finland and Flanders). Other cases exhibit much more concentration in one or a few countries. In Poland, the ten most used journals of both SSH disciplines are all published in one of three Central and Eastern European countries (Poland, Lithuania, Ukraine), including the multilingual and English-language journals.

Discussion and conclusions

By comparing the contents of full-coverage databases to DOAJ, we are able to make an accurate assessment of the current state of Gold OA to peer-reviewed articles in four European countries. The same type of analysis can be used to monitor the further development towards Gold OA.

The overall share of Gold OA differs substantially between countries as well as between fields, and ranges from 5.7 to 17.3%. This finding suggests that the share of Gold OA depends not only on the number of possible OA publishing outlets in a given discipline, but also on more local and contextual factors, such as incentives and perceived quality level. Gold OA is on the rise in Finland, Flanders, Norway and Poland.

A closer investigation into four specific disciplines shows that the most important journals in Biological sciences and Clinical medicine tend to be English-language journals, mostly published by large international publishers. Note, however, that the results from Flanders for these two disciplines may be biased in favour of English-language journals, since the data for Natural sciences & technology and Medical & health sciences derive from WoS. In the SSH disciplines, we find both local and international journals. The latter group can be published in English or another international language, or in multiple languages. All in all, the results demonstrate that, especially for the SSH, the state of Gold OA can only be fully assessed by comparing to a full-coverage database.

References

Archambault, É., et al. (2014). *Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013*. European Commission.

Björk, B.-C. (2019). Open access journal publishing in the Nordic countries. *Learned Publishing*. https://doi.org/10.1002/leap.1231

Bosman, J., & Kramer, B. (2018). Open access levels: a quantitative exploration using Web of Science and oaDOI data. *PeerJ*, e3520v1. https://doi.org/10.7287/peerj.preprints.3520v1

Engels, T. C. E., & Guns, R. (2018). The Flemish performance-based research funding system: A unique variant of the Norwegian model. *Journal of Data and Information Science*, *3*(4), 45–60. https://doi.org/10.2478/jdis-2018-0020

European Commission (2019). *Open science monitor*. https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/trends-open-access-publications_en

Guns, R., et al. (2018). A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics*, *116*(2), 1093–1111. https://doi.org/10.1007/s11192-018-2775-x

Kulczycki, E., et al. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*, *116*(1), 463–486. https://doi.org/10.1007/s11192-018-2711-0

Martín-Martín, A., et al. (2018). Unbundling Open Access dimensions: a conceptual discussion to reduce terminology inconsistencies. https://doi.org/10.17605/OSF.IO/7B4AJ

OECD. (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264239012-en

Piwowar, H., et al. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

Pölönen, J. (2018). Applications of, and Experiences with, the Norwegian Model in Finland. *Journal of Data and Information Science*, *3*(4), 31–44. https://doi.org/10.2478/jdis-2018-0019

Sīle, L., et al. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, *27*(4), 310–322. https://doi.org/10.1093/reseval/rvy016

Sivertsen, G. (2016). Publication-Based Funding: The Norwegian Model. In M. Ochsner, S.E. Hug, H.D. Daniel (Eds.), *Research Assessment in the Humanities. Towards Criteria and Procedures* (pp. 79-90). Zürich: Springer Open.

Sivertsen, G. (2018). Balanced multilingualism in science. *BiD: Textos Universitaris de Biblioteconomia i Documentació*, (40). https://doi.org/10.1344/BiD2018.40.25

Van Noorden, R. (2014). Open-access website gets tough. *Nature News*, *512*(7512), 17. https://doi.org/10.1038/512017a

Evolution of Topics and Novelty in Science

Omar Ballester[1] and Orion Penner[2]

[1] *omar.ballester@epfl.ch*
École Polytechnique Fédérale de Lausanne, Lausanne (Switzerland)

[2] *orion.penner@epfl.ch*
École Polytechnique Fédérale de Lausanne, Lausanne (Switzerland)

Abstract

Methods of estimating the similarity between individual publications is an area of long-standing interest in the scientometrics community. Traditional methods have generally relied on references and other metadata, while text mining approaches based on title and abstract text have appeared more frequently in recent years. In principle, Topic Models have great potential in this domain. But in practice, they are often difficult to successfully employ and, in particular, are notoriously inconsistent as latent space dimension grows. That is, running the same model, with the same parameters, on the same data, but with a different random seed produces radically different similarity estimates as the number of topics increase. In this manuscript we develop a simple, but novel, methodology for evaluating the robustness of topic models. Employing that methodology, we find that the neural network based Doc2Vec approach seems capable of providing (statistically) robust estimates of document-document similarities, even for topic spaces far larger than prudent for the most common topic model approach: Latent Dirichlet Allocation. As this is a work in progress, we do not venture deeply into the question of whether these estimates also reflect reality, but do provide some preliminary evidence and future directions for those efforts.

Introduction

Methods for understanding the topics and concepts of individual publications are a matter of long-standing interest within the scientometric and informetric communities. Indeed, going back to some of Garfield's earliest thinking on citation indexes (1955), he identified a goal of an “association-of-ideas” index. In those thoughts he further developed the role such an index would play in the literature-search process and highlighted the value of a “sub-micro” or “molecular” level approach over one focused on “classification”.

Today document similarity and clustering is a vibrant area of research within the scientometric and informetric community and appearing in many contexts, including information retrieval, the mapping of science, and as an input to rich studies of the individuals and institutions engaged in the research production process. Much of today's work, in line with Garfield's early vision, finds citation and co-citation at the centre of their formulation of contextual or contextual similarity. Although that relationship may be more tenuous than generally accepted, see (Borner, Chen, & Boyack, 2003) for an in-depth exploration.

Increases in computational capacity and the availability of (electronic) data have opened many new avenues for estimating document similarity and carrying out clustering. While the range of options and ideas is vast, in this manuscript we focus on “Topic Models” –a group of techniques arising largely from the computer science literature. As the input to these techniques is textual data (specifically, a collection of text documents) they offer an interesting twist on traditional approaches for understanding the topics and concepts that make up individual publications and, in turn, estimating document similarities and clustering. As discussed below, these techniques are certainly not without their flaws[i] (Velden et al., 2017) but are also well positioned to exploit the rapidly growing body of textual, and perhaps even full text, data.

In this manuscript we develop a robust approach for calculating pair-wise similarities between documents based on state-of-the-art topic modelling techniques. We compute the similarity between researchers which, in turn, allows us to obtain the topical overlap (or proximity) between them.

With this text-only approach, we obtain a continuous knowledge domain space from which we can cluster and delineate topics as narrowly as desired, estimate interdisciplinarity, and observe the evolution and direction of research.

Background

Topic models are statistical models designed to extract from a set of documents the relevant "topics", and in turn, provide a representation of each document within that "topic" or latent space. More pragmatically, it is to infer from a set of document-term vectors a set of document-topic vectors (establishing the extent to which each topic pertains to each document) and a set of topic-term vectors (establishing the extent to which each topic is associated with each term). In this task, a topic model will exploit hidden semantic structure within and across the documents. As it is the case that each document is treated as a bag-of-words topic models cannot exploit local structure (*i.e.* grammar or the specific order of words within a sentence). But rather they exploit structure that emerge at the document level. For example, that the word "table" in the context of a document also containing the words "wood" and "legs" conveys a different meaning than "table" in a document containing "row" and "column". It is ultimately through the exploitation of high-level correlations in the co-occurrence of individual terms, as well as groups of terms, that the topic model produces its document-term and document-topic vectors.

In this manuscript we will test topic models in terms of their capacity to robustly estimate pairwise document similarities. Specifically, we have chosen Non-negative Matrix Factorisation (NMF) (Lee & Seung, 1999), Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) and doc2vec (Le & Mikolov, 2014). NMF decomposes the document-term matrix into a product of two matrices, which by design may have only non-negative entries. LDA is based on a probabilistic model of language in which decomposition stochastically produces two matrices. doc2vec is a relatively new neural network-based approach built upon the similarly new word2vec (Mikolov, Chen, Corrado, & Dean, 2013) word-embedding algorithm. doc2vec formulates the problem as one of predicting an omitted word within a short (3 to 15) contiguous sequence of text.[ii] Treating the neural network's hidden layer as the latent space, one can infer document-topic couplings from the model's parameters. Although it should be noted that, strictly speaking, doc2vec may not be considered a "true" topic model as the topic-term couplings are not easily inferred. But this is acceptable as the fundamental elements of the statistical analysis presented herein are document-document similarity scores, which require only the document-topic vectors.

A specific feature of most, if not all, topic models is that the size of "topic" (or latent) space must be defined by the user. Indeed, the question of what is the "best" or optimal size of the topic space comes up often, and no clear criteria exist (Glaser, Glanzel, & Scharnhorst, 2017). However, we propose this as a feature rather than a bug. Increasing the size of the latent space increases the granularity in which ideas are represented by the model. Considering all journal publications as the corpus, a topic space of size 5 would presumably decompose documents into the highest-level disciplines one may think of (*ex.* biomedical science, the physical sciences, social sciences, humanities, *etc.*) A latent space dimension of one or two hundred may decompose only well-established fields (*ex.* medicine, molecular biology, physics, economics, sociology, history, ...). But running the dimension up into the thousands, or even tens of thousands, allows one to identify very specific groups of research and researchers. For example, those working on a specific form of cancer within a specific model organism. Thus, the question of what is the correct number of topics should never be asked, but rather, one should ask, what is the proper number of topics to tackle your specific question.

Despite the many interesting questions that could potentially be attacked by pushing topic models to high dimensional topic spaces, they are rarely employed with latent dimension greater than, perhaps, a few dozen. This does not arise, however, from a lack of vision, but rather as one increases the number of latent dimensions the model, eventually, becomes unstable. To be very specific, at some point the exact same algorithm, with the exact same parameters, on exactly the same data, but with a different random seed will produce a quantitatively and qualitatively different set of document-topic and topic-term vectors (Belford, Namee, & Greene, 2017). In the topic modelling literature, a variety of information-theoretic measures have been proposed for estimating the extent to which topic-term vectors vary from run to run. However, it is indeed that case that changes in the topic-term vectors may *not* preclude stability when considering only document-document similarities. That is, even if the topics themselves are inconsistent from one training to the next, the measure of pairwise similarity may not change.

Analysis

To be suitable for application in scientometrics it is our view that it must be demonstrated that topics models possess three properties:

- Statistical robustness. That is, running the same model on the same data with the same parameters should produce the same, or at least highly similar results.
- Descriptive power increases with the size of the latent dimension. That is, changing the number of topics should alter the results both qualitatively and quantitatively.
- Reflect reality. That is, the results produced by topic modelling, be they document-document similarities or clustering or otherwise, must be consistent with patterns and relations known to exist within and across research domains.

Below we propose and execute specific statistical tests concerning the first two, while for the third we provide preliminary evidence and highlight paths for further work. However before getting into the analysis, we will define the specific data and context in which we are working.

Data and Methods

In the analysis below a document is the career output of a researcher and the terms are Medical Subject Headings (MeSH). The output of each researcher is taken from the Author-ity disambiguation of PubMed carried out by Torvik and Smallheiser (2009). For each researcher, we extract from their publications all assigned MeSH. To be explicit, the document-term vector resulting from this procedure is one in which each vector entry corresponds to the number of times a given MeSH was assigned to the given researcher's publications across the entirety of his or her career. Disambiguation is thus crucial for reliability of the models. We deal with careers starting 1974 or later, noting that our data terminates in 2009 as that is when the disambiguation ends. Our full corpus comprises about 147000 researchers, however in many analyses below we focus on a subset of 13900 researchers in the Neurosciences. This choice was based purely on a desire to reduce the scale of the analysis to one in which all pairwise similarity scores can be calculated in a manageable amount of time.

As mentioned previously we explored the performance of three topic modelling approaches: NMF, LDA and doc2vec. We performed tests with similar latent space sizes (around 400), multiple iterations and negative sampling when available.

Statistical Robustness

We propose to evaluate the statistical robustness of a topic model via the extent to which it produces consistent estimates of pairwise document-document similarities. Being more specific, as a given model is retrained using the same parameters and data, one can track the mean and standard deviation of the cosine similarity of each pair of documents. If perfectly robust, a model would produce exactly the same similarity each time. An imperfect, but still useful, model will produce slight variations in the each pairwise similarity, but over many retrainings, converge to a specific similarity value for each pair.

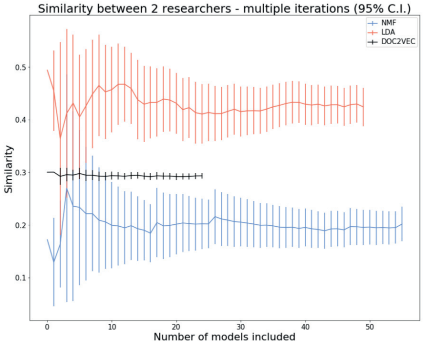


Figure 1 Cosine Similarity between 2 researchers across different models averaged under many retrainings.

Figure 1 shows the behaviour of a specific researcher-researcher similarity score produced by NMF, LDA and doc2vec under many retrainings. First note that each of the three topic models produces a different similarity score, despite having the same number of topics (NN). Second, and most importantly, note that the similarity score for NMF and LDA have a far larger range of results than doc2vec, and display weaker convergence. It is indeed the case that this figure is representative of the behaviour of the three models across all researcher-researcher pairs as well as a wide range of latent space sizes.[iii] Thus NMF and LDA are not statistically robust, while doc2vec warrants further analysis.

Descriptive Power

To get a handle on the descriptive power of doc2vec (or any topic modelling approach) we propose a straightforward procedure based on principal component analysis (PCA). In this approach PCA is first carried out on the document-topic vectors (researcher-topic vectors in this instance). The principal components are then ordered by explained variance and their cumulative explained variance is plotted in Figure 2.

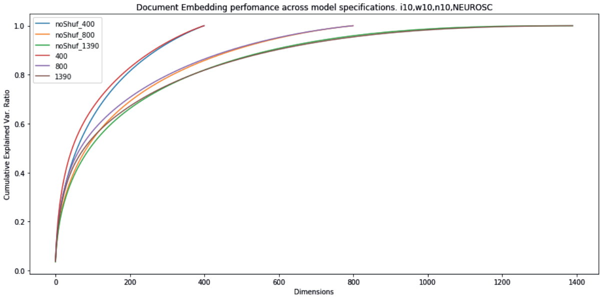


Figure 2 Explained Variance (%) of PCA-transformed embeddings

The PCA explained variance plot allows us to understand the extent to which each dimension allows differentiation among documents vis-à-vis the latent space. For example, a perfectly straight line running from the lower left to the upper right would indicate that each dimension contributes equally to explaining the variation among researchers, and hence, allow for the differentiation between researchers. On the other hand, a curve that quickly reaches 1.0, perhaps after only N dimensions, indicates that only those first N dimensions are actually contributing to explaining the variance.[iv] Or in other words, all dimensions beyond the first N are useless. Thus, the explanatory power of a given topic model can be measured by the area over curve (AOC) in such an explained variance plot. In Figure 2 we see that the doc2vec models do diverge from perfect, but still, a good 20% of the variance does reside in the last ~900 topics of even the most extreme models considered.

Concordance with reality

Properly reflecting reality is, of course, the most important criteria for approach for generating an abstract representation of data. It is often also the most difficult however. Here we provide two small examples as evidence that, at the very least, the results do not directly oppose expectations.

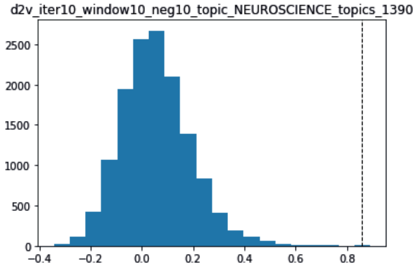


Figure 3 Distribution of cosine similarities between a specific researcher and all other researchers. Dotted line highlights similarity with this researcher's early-career PI.

In Figure 3 the distribution of cosine similarity between a given researcher and each other researcher in the corpus is shown. On the far right, the dotted line highlights the researcher's similarity with a PI from her or his early career. As one would expect, this researcher's highest similarity is, indeed, with his or her early career mentor.

Going forward we are pursuing 3 main avenues of analysis for evaluating the extent to which document similarities produced by doc2vec reflect reality. The first avenue involves using external information to identify pairs of documents that should be highly similar and valid on the doc2vec based similarity measures. In the case of researchers, as shown above, early career mentors represent a set that should be highly similar. In the case of individual publications, those arising from the same grant represent a group of highly similar documents (Boyack Kevin W. AND Newman, 2011). The second avenue involves validation by individuals with significant domain knowledge in each of a variety of fields. Third, using similar methodology clustered at the journal-year level, we observe how the same journal in consecutive years show high similarity.

Conclusion

Topic models are powerful statistical techniques with great potential to contribute to scientometrics, especially as textual data becomes more available going forward. However, they also suffer specific flaws that must be carefully weighted against the benefits. In particular, establishing statistical robustness is challenging, evaluating their descriptive power is key, and ultimately verifying they reflect reality is clearly necessary.

In this manuscript we have proposed a simple approach for estimating the statistical robustness of topic models that is based on pairwise similarity scores between documents. Applying that method, we found that Non-negative Matrix Factorisation and Latent Dirichlet Allocation do not appear to be especially robust for large latent spaces (dimension far greater than 10). doc2vec, a neural network-based approach does, on the other hand, appear to produce relatively stable estimates of pairwise similarity.

We further proposed a principal component analysis-based approach for assessing the descriptive power of topic models. Applying that method to researcher-topic vectors obtained from doc2vec we find that, while doc2vec does not produce perfect results descriptive power does persist into the highest dimensions of the latent space.

In terms of the extent to which doc2vec results reflect reality, many questions remain. We provided two small pieces of evidence that what doc2vec produces is not completely out of bounds. However, careful quantitative validation is still required.

Bibliography

Belford, M., Namee, B., & Greene, D. (2017). Stability of Topic Modeling via Matrix Factorization. *CoRR, abs/1702.07186*.

Blei, D., Ng, A., & Jordan, M. (2003, 3). Latent Dirichlet Allocation. *J. Mach. Learn. Res., 3*, 993-1022.

Borner, K., Chen, C., & Boyack, K. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179-255.

Boyack Kevin W. AND Newman, D. (2011, 2). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLOS ONE, 6*(3), 1-11.

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science, 122*(3159), 108-111.

Glaser, J., Glanzel, W., & Scharnhorst, A. (2017, 5). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics, 111*(2), 979.

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *CoRR, abs/1405.4053*.

Lee, D., & Seung, H. (1999, 2). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*, 788 EP -.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR, abs/1301.3781*.

Torvik, V., & Smalheiser, N. (2009, 7). Author Name Disambiguation in MEDLINE. *ACM Trans. Knowl. Discov. Data, 3*(3), 11:1–11:29.

van der Maaten, L., & Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. *The Journal of Machine Learning Research*.

Velden, T., Boyack, K., Glaser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017, 5). Comparison of topic extraction approaches and their results. *Scientometrics, 111*(2), 1169-1221.

[i] In addition to issues discussed below, they also fail to exploit citation data. A gap we are working to fill with further work.

[ii] Thus, this is not, strictly speaking, a bag-of-words approach. However, we have reproduced each analysis within this manuscript shuffling the order of the terms and all results are similar. Although this *is* curious and begs further consideration at an appropriate moment in the future.

[iii] although at smaller latent space sizes, around 10, NMF and LDA will also converge perfectly well

[iv] This is, indeed, a result that can be obtained for NMF, LDA, and especially doc2vec for various specific parameters

Professional Standards in Bibliometric Research Evaluation? Results from a Content Analysis of Evaluation Studies in Europe.

Arlette Jappe[1]

[1] *jappe@uni-wuppertal.de*
IZWT Interdisciplinary Centre for Science and Technology Studies,
University of Wuppertal Gaußstraße 20, D-42119 Wuppertal (Germany)

Abstract

The paper examines current performance assessment practice in Europe: first, it identifies those bibliometric methods that are used most often, and second, it identifies those actors who have been able to define current practice. The paper draws on Abbott's theory of professions and argues that indicator-based research assessment constitutes a potential jurisdiction for both individual experts and expert organizations. The paper presents a search methodology that yielded 82 professional evaluation studies from 14 EU countries, covering the period 2005 to 2014. Using structured content analysis, our findings are as follows: (1) Bibliometric research assessment is most frequently used in the Netherlands, the Nordic countries, and Italy. (2) The Web of Science (WoS) is the dominant database for the assessment of public research in Europe. (3) Expert organizations invest in improvement of WoS citation data and are able to set technical standards with respect to data quality. (4) Field averages are the most frequent frame of reference for citation impact. (5) The WoS classification of science fields (SCs) functions as a de-facto reference standard for research performance assessment. In light of these findings, the paper discusses the importance of data access and transparency for the development of professional bibliometric assessment.

Introduction

There is a growing demand for practicable methods in research evaluation on the part of research organizations and research funding agencies, including metrics based on publication and citation data (bibliometric indicators). Among the scientific communities affected by performance assessment, bibliometric indicators have remained controversial (Adler, Ewing, & Taylor, 2009; Bonaccorsi, 2018; Cagan, 2013; de Rijcke, Wouters, Rushforth, Franssen, & Hammarfeldt, 2016). So far, there are no widely agreed upon standards for bibliometric research assessment. Against this background, the present paper investigates two research questions: First, which methods of bibliometric performance assessment are prevailing in research evaluation practice? Second, if certain de facto standards of bibliometric research assessment can be observed, which social actors have been able to define them?

While there have been several reviews on scientific developments in the area of evaluative citation analyses in recent years (EC, 2010; Mingers & Leydesdorff, 2015; Todeschini & Baccini, 2016; Waltman, 2016; Wilsdon et al., 2015), there is little overview so far on which methods are actually used in bibliometric assessment practice. Methodologically, the present study resembles a meta-evaluation or evaluation synthesis. The expression 'meta-evaluation' is commonly used in the literature to denote systematic reviews of evaluation studies with respect to their methodological quality and results (Cooksy & Caracelli, 2005; Stufflebeam, 2001). The present study also analyses the methods used in existing evaluation studies from a meta-perspective. But rather than assuming predefined methodological standards of good practice and evaluating published studies accordingly, the purpose is to find out what methodological standards existed in the field of assessment practice during a certain period and who shaped them. These are referred to as professional de facto standards.

This paper is part of a project seeking to understand the development of bibliometric assessment methods from the perspective of Abbott's sociological theory of professions (Abbott, 1988, 1991). This theory was chosen to investigate how particular methodological choices become socially established as professionally legitimate means of treating certain types of evaluation problems. More specifically, this framework is used to address the question of professional control in bibliometric assessment. In Abbott's terminology, the increasing demand for practicable and efficient assessment of academic performance constitutes a problem amenable to expert service. Thus, research assessment constitutes a potential jurisdiction for professional experts who are capable at defining the nature of assessment problems as well as offering solutions that effectively adress client´s needs. A recent paper investigated empirically whether the academic research area of evaluative citation analysis (ECA) has been able to define scientific standards for the practice of bibliometric research evaluation in the period 1972-2016 (Jappe, Pithan, & Heinze, 2018). Based on the theory of intellectual fields as reputational organizations (Whitley, 2000) and on an organizational network analysis, this study concluded that the research area of ECA has been characterized by low levels of reputational control, as evident from high shares of outsider contributions and new actors entering the field over the whole period. It was argued that the observed lack of reputational control within the academic research area is at least consistent with observed difficulties in establishing scientific authority for bibliometric assessment practice.

Continuing this line of inquiry, the current paper investigates whether professional evaluation practice is characterized by at least de facto standards of bibliometric methods. For this purpose, we conducted a content analysis of professional evaluation studies published in European countries in the period 2005-2014. We focus on the measurement of citation impact as practiced in real evaluation studies. Other topics of bibliometric assessment such as emerging research topics, research profiles, or international collaboration are not part of this study. Our initial assumption was that leading organizations within the expert field take the role of defining de facto standards, first because they have a high market share of assessment services and second because they serve as a legitimate role model that is then imitated by other bibliometric experts. Our study sample includes a total of n=82 individual studies evaluating either research organisations (RO) or research funding instruments (FI) from 14 European countries plus EU framework programs. An important selection criterion was that each study had been conducted for purposes of decision-making in research policy or research management. The paper first presents theoretical considerations concerning the application of professional sociology to the field of bibliometric research evaluation, then describes data and methods of our analysis, including search strategies, selection criteria and the content analysis of evaluation studies with respect to their methodological design and metrics, followed by a presentation of the empirical findings and a discussion of the results in the light of the theoretical framework. The present version of the paper includes two selected data tables. More detailed results will be shown in the conference presentation.

Theoretical Considerations

Abbott's theory of professions is a sociological framework for analyzing how professional expertise is socially constructed and institutionalized in modern societies (Abbott, 1988, 1991). The analysis starts with a societal problem amenable to expert service and with groups of professional actors who claim relevant expertise for treatment of this problem. The theory distinguishes cognitive claims for expert knowledge from social claims for jurisdiction for problem diagnosis and treatment that professionals need to establish in various social arenas, including the legal system, the public, and the workplace. The concept of a professional jurisdiction goes beyond a merely economic notion of a market for expert services by inquiring into the development of expert control concerning appropriate problem definitions and treatments from a socio-historical perspective. The framework is also suitable for cross-national comparisons since it does not make specific theoretical assumptions concerning the role of the nation state for the eventual settlement of professional jurisdictions in different countries. The conceptual framework was chosen for the present project because it is suitable for the investigation of emerging professions that do not yet possess recognized domains of expertise, which might eventually be protected by state licences, but are still engaged in an ongoing competition with other professional actors for the appropriation of relatively new jurisdictions or tasks.

Applying this theoretical framework, we assume that a demand for professional services in the realm of quantitative research assessment arises mainly on the part of research organizations and research funding agencies as the two most important groups of potential clients. These organizations have a demand for reliable information concerning the performance of their scientists, research groups or funded projects for purposes of decision making (Miller, 2013; H. F. Moed, 2005), and for accountability and legitimacy (Power, 1997; Strathern, 1996). Thus we assume that demand is mostly located at a meso-level of organizations in the public research system. Although private firms also use bibliometrics, information on research performance assessment in the private sector is not systematically accessible and therefore not part of the present study. Several European countries are also experimenting with the introduction of bibliometric methods for performance based institutional funding on a national scale (Ancaiani et al., 2015; Hicks, 2012; Sivertsen, 2016, 2017). The country with the most extensive use of bibliometric performance measurement during our observation period was Italy. The Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR) was created in 2006 with the mandate to evaluate all public research, an exercise called “Valutazione della Qualità della Ricerca” (VQR) (Ancaiani et al., 2015). Bibliometric reports for the first round of the VQR, covering the period 2004-2010, were included in this meta-evaluation.[1] National evaluations with a disciplinary scope by the Nordic Institute for Studies in Innovation, Research and Education NIFU were also included. In the UK, the introduction of bibliometric methods for the Research Excellence Framework (REF) has been intensely debated since 2006 (Arnold et al., 2018). The methodological framework for the upcoming REF exercise in 2021 prescribes that disciplinary sub-panels may use citation data which will be centrally provided by the REF “where appropriate and available (…) to inform their assessment of output quality” ((REF, 2019): 36; 50).

According to Abbott, the work of professionals can be generally described as the application of abstract knowledge to complex individual cases. Abstract knowledge is an important source of legitimacy for claims to jurisdiction because it ties professional work to the general values of logical consistency, rationality, effectiveness, and progress. This scientific legitimacy includes a definition of the nature of problems, rational means of diagnosing them, and the delivery of effective treatment. In addition, abstract knowledge enables the instruction and training of students entering the profession and is oriented towards generating new mechanisms of diagnosis, inference, and treatment. Abstract knowledge is typically produced by an academic sector closely related to the profession. A recent study investigated the academic research area of ECA as the academic sector which is closely aligned with bibliometric evaluation practice (Jappe et al., 2018). Abstract knowledge is also stored in specialized artefacts which Abbott refers to as expert commodities. In the case of evaluative bibliometrics, citation databases, such as Web of Science (WoS) or Scopus, are the most important artefacts for professional work.

The present study observes professional practice by means of examining a set of professional evaluation studies. These studies evaluate the performance of either RO or FI in Europe based on publication and citation data and have been published either in the format of study reports (grey literature) or as journal articles. Bibliometric experts and expert organizations are encompassed as authors of these studies, while RO and FI are encompassed as evaluation objects, but also in many cases also contracting entities of the respective studies. This definition of professional practice excludes all usage of bibliometric indicators that is less explicitly codified, as for example in cases when RO use internal metrics to evaluate staff performance, or in cases when funding agencies use the journal impact factor or the h-index to make selection decisions among program applicants without publishing them. The study is confined to Europe, i.e. evaluation objects must be located in a European country. In this way, the analysis of widespread assessment practices also contributes to the investigation of commonalities in an European research area.

Data and Methods

The current section describes the selection criteria for inclusion of studies in our analysis, the search strategies to identify such studies, and the coding scheme and procedures used to extract methodological information from each individual study. The sampling strategy did not aim at statistic representativeness but at the identification of as many incidents of bibliometric evaluation as possible from diverse sources. The selection criteria were as follows:

1. Each evaluation must include a publication and citation analysis. The sample includes both studies that rely exclusively on bibliometric data as well as multi-dimensional evaluations that combine bibliometric data with other information such as peer

evaluations, financial and personnel data, case studies etc. (Martin, 1996; H. Moed & Halevi, 2015). In either case, only the bibliometric analyses are object of the content analysis.

2. The objects of evaluations are either research organizations, ROs (typically universities and/ or their departments or faculties or extra-university public research institutes) or funding instruments, FI (typically aimed at supporting research projects or individual researchers at public RO, sometimes involving private firms, and sometimes supporting more long-term investments such as excellence schemes).
3. Evaluation objects must be located in Europe.
4. The evaluation was conducted with the stated purpose to inform decision making on behalf of the respective RO or FI. Purely academic studies on the basis of bibliometric data were excluded from the study sample.
5. The evaluation has been published between 2005 and 2014, including grey literature (project reports) as well as journal articles.

Multiple search strategies were combined in order to identify as many studies as possible:

1. Prior research identified expert organizations and individual experts with a central position in the academic research area of ECA (Jappe et al., 2018). A request for non-confidential evaluation studies was sent to experts in 35 research organizations in 13 European countries. 18 organizations responded (51 %) of which 11 (31 %) contributed evaluation reports or shared information on published studies. In total, 16 studies (20 %) of the eventual sample have been identified in this way.
2. Evaluation studies were extracted from a set of WoS publications identified as "follow-up research on citation impact indicators" (Jappe et al., 2018). This set includes all publications in WoS that cite any of 169 specified citation impact indicators, a total of 2757 publications from 2005-2014. Keywords and journal titles were analysed in order to identify relevant studies from this sample. From an initial set of 315 potentially relevant publications, 15 evaluation studies have been retained (18 % of study sample).
3. The "Science and Innovation Policy Evaluation Repository" (SIPER) was searched for bibliometric studies. This publicly accessible database contains meta-data and in part also original documents of evaluation studies. A search for citation analysis retrieved 24 potentially relevant documents, three of which were eventually retained (4 % of study sample).[2]
4. Prior research investigated the history of CWTS as an expert organization in the field of evaluative bibliometrics (Petersohn & Heinze, 2018). In the course of fieldwork, the authors obtained 24 evaluation reports by CWTS (29 % of study sample), as well as three more evaluations by other bibliometric authors (4 %).
5. The Italian public agency ANVUR has the legal mandate to evaluate the quality of activities by all RO receiving public money, as well as FI aimed at research and innovation (ANVUR, 2013). The first round of VQR, covering the period 2004-2010, was included in the content analysis, while the reports of the second and third round were published after 2014. Among 14 disciplinary areas in the Italian system, nine applied bibliometric assessment. Since each disciplinary committee has the mandate to decide upon the appropriate evaluation criteria within its field(s) of research, the reports for the different sectors were treated as individual bibliometric exercises for the purpose of this content analysis. In this way, the VQR contributed nine individual studies to the study sample (11 % of study sample).
6. The worldwide web was searched for evaluation reports by funding agencies. Some countries and agencies follow high standards of transparency concerning the evaluation of

public research, including e.g. the Swedish Council for Science (Vetenskapsradet VR), the Swedish Environmental Protection Agency (SEPA), the Danish Council for Strategic Research, or the British Wellcome Trust, among others. In total, web searches identified 12 relevant and publicly available evaluation reports (15 % of study sample).

The final sample includes 82 distinct bibliometric studies of which 58 (71 %) evaluate RO and 24 (29 %) evaluate FI. Three expert organizations stick out in particular. CWTS is the organization with the largest share of studies (n=24), followed by ANVUR VQR (n=9) and NIFU (n=7). Since the studies from the same organization use identical citation impact metrics (CWTS, NIFU), or share at least important characteristics (VQR), the respective subsets are analyzed separately on some dimensions from the remaining 42 studies, the latter subset is referred to as studies "by other bibliometric experts". All evaluation studies are documented in the Annex to the paper as currently under preparation for journal publication.

Methodologically, our study is based on a structured content analysis. The bibliometric design of each individual study was analyzed with a scheme of 37 coding questions in ten topical areas. The topics include (1) bibliographic information on the individual study, (2) the professional setting, (3) the object of evaluation, (4) the citation databases, (5) quality enhancement of bibliometric raw data, (6) sampling strategy and data collection, (7) research areas under study, (8) definition of citations, (9) citation impact indicators, and (10) statistical methods used. The definition of variables follows the methodological literature on citation analysis, especially (H. F. Moed, 2005; Todeschini & Baccini, 2016; Waltman, 2016). Most items are on nominal level of measurement, i.e. non-ordered qualitative characteristics. Five items are formulated as open questions, so that raters could write down more detailed information. This coding scheme was developed in an iterative procedure beginning with a partial sample. In order to test interrater reliability, the initial coding scheme was applied by two raters to an initial sample of 20 different studies. Where differences in coding became apparent, they were discussed among the two raters and the items improved in order to reduce their ambiguity. The remaining studies were each coded once by the author.

Results

1. Bibliometric research assessment is most frequently used in the Netherlands, the Nordic countries, and Italy.

Although we found instances of bibliometric evaluation in many European countries, more regular use of bibliometric assessments has been concentrated in a few countries during the observation period. Overall, the sample includes studies from 14 countries plus Framework Programs of the European Union, approximately 60 % of the studies come frome the Netherlands, Italy, Sweden, Norway, Finland, and Denmark. The Netherlands and the Scandinavian countries are medium-size research systems with a strong performance (particularly Denmark and the Netherlands) in international comparison. Among the larger public research systems in Europe, Italy is the only one with national-scale bibliometric assessment. The UK Research Excellence Framework uses bibliometric data only to inform peer review (Arnold et al., 2018; REF, 2012, 2019), and Germany has no national framework for research evaluation. Our search strategy did not yield any bibliometric assessments for France.

Coverage of individual nations does not only represent the diffusion of bibliometric methods but may also result from different publication policies. For example, evaluations of Max-Planck-Institutes in Germany are usually kept confidential. By contrast, in Sweden, publication is mandatory under national transparency rules, so that many evaluation reports are available on the internet. However, there is no reason to assume that bibliometric techniques differ systematically between confidential sources and published reports, except for the reported level of aggregation. For example, reports by CWTS and the Italian VQR do not contain information on individual researchers or research groups.

2. The Web of Science (WoS) is the dominant database for the assessment of public research in Europe.

During the observation period, the citation indices contained in the WoS provided the basis for the bibliometric evaluation of public research in Europe. 90 % of all studies in the sample rely on WoS, while only 15 % use Scopus or a combination of WoS and Scopus. In some cases, designated databases such as PubMed or MathSCInet are employed, but these alternative citation databases exist in only a few disciplines. In other cases, citation data are complemented by national databases that are more comprehensive in terms of research products, but do not contain original citation data. Two examples are the Norwegian Current Research Information System (Cristin), which includes a larger array of document types such as books and book chapters (Sivertsen, 2016), and the Italian VQR that includes all types of research outputs, e.g. software, patents, maps or artworks (ANVUR, 2013).

3. Expert organizations invest in cleaning and improvement of WoS citation data and are able to set technical standards with respect to data quality.

Citation raw data as provided by WoS or Scopus require considerable processing before they are adequate for the assessment of authors and research organizations. The main problems are the ambiguity of author names and institutional addresses as well as the unambiguous assignment of authors to research institutions. In particular, variants of institutional names require detailed knowledge of national research systems for correct disambiguation. These and other technical problems also lead to certain proportions of false citation linkages in the raw data. Expert organizations such as CWTS and NIFU, but also the Italian CNR-IASI, the German Max Planck Society and the German Competence Center Bibliometrics currently deal with this situation by buying raw data from database providers (Clarivate Analytics, formerly Thomson Reuters, sometimes complemented by Scopus Elsevier) and then construct in-house databases with improved data quality. The studies by the other bibliometric experts also frequently mention the effort required for disambiguation of author names and institutional addresses. Also, one of the main practical arguments for the H-index is its alleged robustness with regard to incomplete publication and citation data.

4. Field averages are the most frequent frame of reference for citation impact in professional evaluation studies.

In order to analyse how the frame of reference was conceived in bibliometric studies, we analysed their evaluation objects and choice of metrics in more detail. Concerning evaluation objects, ROs were differentiated according to scale (number of institutes or universities) and scope (mono- vs. multi-disciplinary), while FI were distinguished according to funding units (research projects, scientists, ROs, portfolios). Concerning the choice of metrics, we distinguished impact metrics based on “international field averages” from impact metrics based on “national rankings” and “other”. International field average means that observed citation rates are assessed with reference to citation rates for the same research field and the same period (sometimes also the same document type) in the entire database. This is also refered to in the literature as a comparison of observed to expected citation rates, where the database field average stands for average expected citation rate (H. F. Moed, de Bruin, & van Leeuwen, 1995; Waltman, 2016). This type of metric is used in 65 % of the study sample, including all 31 studies by CWTS and NIFU. National rankings represent a different approach where the relative national position is the frame of reference for research performance. This type of metric is used only in the evaluation of Italian university departments, or 12 % of the studies. Other frames of reference include mainly journal impact based measures and h-type indices, but also some quasi-experimental group comparisons (funded vs. non-funded scientists) in FI studies (Table not included).

5. The WoS classification of science fields (SCs) functions as a de-facto reference standard for research performance assessment.

In principle, field normalization is applicable to different types of citation metrics (Waltman, 2016), including journal impact, arithmetic mean, highly cited percentiles, as well as h-type indices, or indirect citation metrics. Solely source normalized impact metrics are construed as a methodological alternative in order to avoid problems of field normalization (Waltman & van Eck, 2013). Yet only some of these possible combinations are actually found in our sample (Table 1). Field normalized arithmetic means are the single most frequent type of metric (65 %), often combined with top-percentiles, as practiced by CWTS, among others. The h-index and other h-type indices (g-index) occurred in 18% of studies, mostly by other bibliometric experts; a field-normalized h-index appears only once. Source normalized indicators were not used by a single study, and only one study applied an indirect citation indicator. The category of "other indicators" includes different metrics. For example, in the case of VQR this refers to the construction of composite indicators for university rankings. In the case of CWTS, this category includes the normalization of citation impact with reference to the journal set in which a group has published. In general, it can be concluded from the overview in Table 1 that few of the methodological improvements recently proposed in the academic debate on impact metrics (Jappe et al., 2018; Todeschini & Baccini, 2016) have made their way into research assessment practice during the period observed.

Journal impact continued to be used quite frequently (46 %), even though the substitution of the impact of the publishing journal for the actual number of citations of an article has repeatedly been critized for lack of validity (Table 1). Journal impact is used in order to substitute missing data, either because publications are so recent that actual citations are not yet available (e.g. VQR), or because articles are published in journals that are not covered by the database (e.g. NIFU, VQR). Another reason is comparatively easy access of journal impact metrics via Journal Citation Reports, which is relevant especially for bibliometricians without fully licensed access to citation databases. Sometimes journal impact is used in a pragmatic way, just distinguishing two levels of journal quality (NIFU).

In total, field normalization, here including the field normalized arithmetic mean and other field normalized percentiles, was used by 84 % of all studies (Table 2). Of those 69 studies, 84 % rely on the WoS classification of science fields (WoS subject categories), plus an additional 3 % that use the related Essential Science Indicators classification by Clarivate Analytics (formerly Thomson Reuters). The Scopus science classification is mentioned mainly in the Italian VQR reports. 13 % of field normalizations are based on self-defined journal sets, or keywords in combination with self-defined journal sets, and some studies used more than one field classification. Alternative field taxonomies proposed in the academic literature have attained little influence so far. Few studies use discipline-specific databases and subfield classifications. It can be concluded that the WoS classification of science fields has attained the status of a de facto reference standard, at least for the time period under study.

Table 1: Types of impact metrics used in evaluation studies

	Type of metric	*VQR*	*CWTS*	*NIFU*	*Other bibliometric experts*	*Studies total*	*% studies*
1.	Journal impact	9	0	6	23	38	46
2.	Field normalized arithmetic mean	0	24	7	22	53	65
3.	Other field related percentiles (e.g. top highly cited)	9	21	0	13	43	52
4.	H-index and h-type indices	0	1	0	14	15	18
5.	Indirect citation impact	1	0	0	0	1	1
6.	Source normalization	0	0	0	0	0	0
7.	Other metrics	9	23	5	8	45	55
8.	More than one type of metric	9	24	6	27	66	80
	Studies total	9	24	7	42	82	100

Source: Meta-evaluation study set

Table 2: Classification of science fields used for field normalization

	Taxonomy	*VQR*	*CWTS*	*NIFU*	*Other bibliometric experts*	*Studies total*	*% studies*
1.	WoS SC classification	6	24	7	21	58	71
2.	Scopus classification	5	0	0	1	6	7
3.	Essential Science Indicators	0	0	0	2	2	2
4.	Alternative academic classification*	0	0	0	4	4	5
5.	Self-defined journal sets	4	0	0	3	7	9
6.	Keywords combined with journal sets	0	0	0	3	3	4
7.	Other	1	0	0	2	3	4
8.	More than one classification	6	0	0	7	13	16
9.	Studies with field normalization	9	24	7	29	69	84
	Studies total	9	24	7	42	82	100

Source: Meta-evaluation study set

* This category includes field classifications proposed in bibliometric literature.

Discussion

The findings of our analysis support our initial assumption that leading expert organizations have an important role in defining de facto standards. The prominent position of individual expert organizations in the field is well documented by the study set. The two most prominent organizations according to this study are CWTS and NIFU, both of which have regularly conducted bibliometric assessments for many years and produced important shares of the data set. These expert organizations are able to define technical standards with regard to enhanced quality of publication and citation data. Following the example of CWTS, not only NIFU, but also other expert organizations such as the Italian CNR-IASI, the German Max-Planck Society, the German Competence Center Bibliometrics, and other institutes invest heavily in inhouse-databases in order to clean WoS raw data. The same level of data quality cannot be attained by bibliometric experts without equivalent databases, at least not for larger publication quantities.

Our assumption that prominent expert organizations are imitated and that their approach gains legitimacy within a field as an instance of recognized expert practice is most likely true with respect to the broad dissemination of the field-normalized arithmetic mean. In this case, the role model is CWTS which, although not having invented the idea of observed versus expected mean values, was first in Europe in starting to use such indicators systematically on the basis of an inhouse database in the early 1990s (H. F. Moed et al., 1995). This finding is qualified by the fact that the Italien VQR deviates from the professional standards as modeled by CWTS and more generally from the methodological debates in the academic field of ECA.

Many alternatives and refinements of field-normalized mean indicators have been proposed in the academic literature, including the h-index as the most cited methodological alternative and source normalized indicators as a way to avoid problems of field normalization altogether (Jappe et al., 2018) This debate is consistent with the Abbott's proposition that the academic sector would invent new methods and that experts would make competing cognitive claims. The subset of studies by "other bibliometric experts" displays more heterogeneity with regard to citation impact metrics and field classifications than the total sample. This greater methodological diversity suggests that the expert field as a whole did not produce a clear challenge to the predominant approach in the emerging jurisdiction. Thus despite existing criticisms, neither of these alternatives was able to overtake the prominence of the field-normalized arithmetic mean as a recognized professional practice. But neither did the expert field collectively adopt one standard methodology for bibliometric evaluation.

Beneath a surface of methodological diversity and academic openness, this analysis unequivocally documents the predominance of the commercial database WoS with respect to the definition of methodological standards for performance assessment of public research in Europe. During the observation period, the provider of WoS assumes the most important role in defining de facto standards for bibliometric assessment. All expert organizations, including CWTS and NIFU, base their citation analyses on data licensed by Clarivate Analytics, and this is also true for almost all other bibliometric experts. But not only does Clarivate Analytics (via its licencing policy) regulate to which extent different user groups have access to citation data, but WoS science categories function as de facto reference standards for bibliometric assessment. In addition, there is the effective dissemination of selected impact indicators via the Journal Citation Reports and Incites. It appears that all efforts on the part of academic bibliometricians to develop alternative categorizations of scientific fields (Shu et al., 2019) or more complex impact indicators have had little impact on professional practice so far because they cannot be distributed alongside with citation data. The few examples for the use of alternative or supplementary sources such as the specialized citation database Medline/Pubmed or the research documentation system Cristin underline that bibliometric evaluation practice depends first and foremost on the data sources that are accessible for comparative analyses of research performance. While some studies in the literature investigated within-field homogeneity and across-field heterogeneity of citations in WoS subject categories (Albarrán, Crespo, Ortuno, & Ruiz-Castillo, 2011; Leydesdorff & Bornmann, 2016; Wang & Waltman, 2016), the adequacy of this classification as the main frame of reference for the evaluation of scientific impact has not been established.

References

Abbott, A. (1988). *The system of professions: An essay on the division of expert labor*. Chicago: University of Chicago Press.

Abbott, A. (1991). The Future of Professions: Occupation and Expertise in the Age of Organisation. *Research in the Sociology of Organisations, 8*, 17-42.

Adler, R., Ewing, J., & Taylor, P. (2009). Citation Statistics: A Report from the International Mathematical Union (IMU) in Cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). *Statistical Science, 24*(1), 1-14.

Albarrán, P., Crespo, J. A., Ortuno, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics, 88*, 385-397.

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., & al., e. (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. *Research Evaluation*, 242-255.

ANVUR. (2013). *Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010). Rapporto finale ANVUR Parte Prima: Statistiche e risultati di compendio.* Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca ANVUR.

Arnold, E., Simmonds, P., Farla, K., Kolarz, P., Mahieu, B., & Nielsen, K. (2018). *Review of the Research Excellence Framework. Evidence Report*. Technopolis Group.

Bonaccorsi, A. (Ed.) (2018). *The evaluation of research in social sciences and humanities. Lessons from the Italian experience*: Springer International Publishing.

Cagan, R. (2013). The San Francisco Declaration on Research Assessment. *Disease Models & Mechanisms, 6*, Editorial. doi:doi:10.1242/dmm.012955

Cooksy, L. J., & Caracelli, V. J. (2005). Quality, Context, and Use. Issues in Achieving the Goals of Metaevaluation. *American Journal of Evaluation, 26*(1), 31-42.

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfeldt, B. (2016). Evaluation practices and effects of indicator use. A literature review. *Research Evaluation, 25*(2), 161-169.

EC. (2010). *Assessing Europe´s University-based Research. Expert Group on Assessment of University-based Research* (EUR 24187 EN). Brussels: European Commission, DG Research.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*(2), 251-261.

Jappe, A., Pithan, D., & Heinze, T. (2018). Does bibliometric research confer legitimacy to research assessment practice? A sociological study of reputational control, 1972-2016. *PLOS One, 13*(6), e0199031.

Leydesdorff, L., & Bornmann, L. (2016). The Operationalization of "Fields" as WoS Subject Categories (WCs) in Evaluative Bibliometrics: The Cases of "Library and Information Science" and "Science & Technology Studies". *Journal of the Association for Information Science and Technology, 67*(3), 707-714.

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics, 36*(3), 343-362.

Miller, P. P., M. (2013). Accounting, organizing and economizing: connecting accounting research and organization theory. *The Academy of Management Annals, 7*(1), 557e605.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research, 246*, 1-19.

Moed, H., & Halevi, G. (2015). Multidimensional Assessment of Scholarly Research Impact. *Journal of the Association for Information Science and Technology, 66*(10), 1988-2002.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

Moed, H. F., de Bruin, R. E., & van Leeuwen, T. N. (1995). New Bibliometric Tools for the Assessment of National Research Performance - Database Description, Overview of Indicators and First Applications. *Scientometrics, 33*(3), 381-422.

Petersohn, S., & Heinze, T. (2018). Professionalization of bibliometric research assessment. Insights from the history of the Leiden Centre for Science and Technology Studies (CWTS). *Science and Public Policy*(45), 565-578.

Power, M. (1997). *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

REF (2012/01). *Panel criteria and working methods*. Research Excellence Framework (REF) 2014.

REF (2019/02). *Panel criteria and working methods*. Research Excellence Framework (REF) 2021.

Shu, F., Julien, C.-A., Zhang, L., Qiua, J., Zhang, J., & Lariviere, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics, 13*, 202-225.

Sivertsen, G. (2016). Publication-Based Funding: The Norwegian Model. In M. Ochsner, S. Hug, & H. Daniel (Eds.), Research Assessment in the Humanities: Towards Criteria and Procedures (pp. 79-90). Zürich: Springer Open.

Sivertsen, G. (2017). Unique but still best practice? The Research Excellence Framework from an International Perspective. *Palgrave Communications, 3*, 17078. doi:10.1057/palcomms.2017.78

Strathern, M. (1996). *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy*. London: Routledge.

Stufflebeam, D. L. (2001). The Metaevaluation Imperative. *American Journal of Evaluation, 22*(2), 183-209.

Todeschini, R., & Baccini, A. (2016). *Handbook of bibliometric indicators: quantitative tools for studying and evaluating research*: Wiley VCH, Weinheim, Germany.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*, 365-391.

Waltman, L., & van Eck, N. J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics, 96*, 699-716.

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics, 2016*(10).

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences. 2nd Edition*. Oxford: Oxford University Press.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE.

[1] <http://www.anvur.it/attivita/vqr/> accessed 28 Jan 2019.

[2] http://si-per.eu/Home/About; last accessed 28 January 2019. Most SIPER entries were excluded because the respective studies contained citation data for the national level but not for RO or FI. In other cases evaluations were mentioned in SIPER but the respective reports were not accessible.

ISSI 2019 Enhancing knowledge of Research Organizations: An analysis of their current classification, collaboration schemes and research impact

Sonia Mena, Tobias Nosten, Juan Pablo Bascur and Clara Calero-Medina[1]

[1] *s.d.mena.jara@cwts.leidenuniv.nl; t.j.a.nosten@umail.leidenuniv.nl; j.p.bascur.cifuentes@cwts.leidenuniv.nl; clara@cwts.leidenuniv.nl,*
Centre for Science and Technology Studies, Faculty of Social and Behavioral Sciences, Leiden University, Leiden, The Netherlands

Abstract

As the figure and role of Research Organizations is increasingly changing in the last decades, the way we categorize this kind of institution is becoming more challenging. This article aims to differentiate profiles of European Research Organizations considering their performance in terms of bibliometric indicators. The indicators are associated with scientific impact, international collaboration and industry links. The results of a k-means cluster analysis shows three distinctive groups of institutions that can be described by their levels of impact, international collaboration and industry collaboration. As expected, we found a correlation between organizations collaborating internationally and having a high impact research output. The analysis opens up the possibility of identifying well defined groups of Research Organizations based on bibliometric indicators.

Introduction

The terms ‘Public Research Organizations’ (PROs) or Public Research Institutions (PRIs), are broadly used concepts to define this institution type. However, they are slowly and inevitably becoming difficult to apply when addressing them, especially when referring to their different degrees of publicness. According to Bozeman & Bretschneider (1994) few, if any, complex organizations are purely public or purely private. Instead, some mix of public and private authority influences the behavior of most organizations. The term Research Organizations as such, once represented by a more homogenous group, have experienced a diversification of role, mission, source of funding and the collaboration schemes at national and international level. This transition has been well documented in recent decades, whether in institutional reports or scientific articles (OECD, 2011; Cruz-Castro, Jonkers & Sanz-Menéndez (2015); Bodem *et al.* (2006); Peñasco, Sanz-Menéndez, Cruz-Castro & Martínez (2017). In this article, we address them by using the term Research Organizations (RO) as a more inclusive concept that encompasses increasingly privately funded and autonomous institutions referred to as “semi-public”.

In 2015 the OECD published the last edition from The Frascati Manual, with its first version dating back to 1963. This manual is considered the main guideline for evaluating R&D activities, being the bibliometric analysis one of the indicators that provide users and producers of R&D statistics a context for setting measurements within the framework of the overall S&T system. These indicators can be applied at different levels, assessing research performance of individuals/research teams, institutions and countries, to identify national and international networks, and to map the development of new (multidisciplinary) fields of science and technology (OECD, 2002). In terms of institutional sectors, the Frascati manual encloses the following units: *Business enterprise (BE), Government (GOV), Higher Education (HE) and Private non-profit (PNP).*

The roles and structures of research organizations are constantly transforming, and as such, novel paths and models have risen seeking to close the gap between this traditional classification and new exploratory approaches to group and distinguish them. In 2010, Arnold *et al.,* identify three types of PRO: government laboratories, scientific research institutes and research and technology organizations. Afterwards categories proposed by OECD (2011) involving MOCs (Mission-Oriented Centres), PRCs (Public Research Centres and Councils), IRIs (Independent Research Institutes) and RTOs (Research Technology Organizations) are seen as ideal types of PROs. While some organizations fit into these categories in a straightforward manner, the classification becomes complex for other organizations. This is mainly due to the nuances or ownership and control by government bodies. As stated in the Frascati Manual "*It should be noted that in some cases the definition of control is challenging, because the power to decide on the allocation and amount of funding can be a major means of control. Therefore, it can be appropriate to use the major source of funding as an additional criterion to decide whether the institution is government controlled or not*". It is worth pointing out that finding information related to the source of funding for organizations implies an arduous effort and most of the time this information is not reachable using conventional methods.

As seen in figure 1, ROs spread and interact throughout different sectors, emphasizing the multisector profile of the Research Institutions. The links depicted in figure 1 are based on the definition and identification of every category extracted from the Frascati Manual (2015) and OECD (2011).

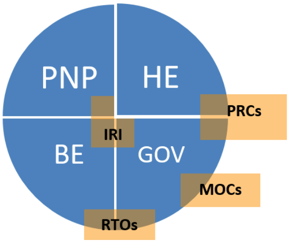


Figure 1. Relationship between unit sectors from FRASCATI (2015) and Public Research Organizations (OECD, 2011) (Source: Author)

In 2017, Peñasco *et al.* developed a novel attempt of providing new categories that would explain the main characteristics of Research Organizations. This classification was based on qualitative aspects such as their mission, legal status and ownership among other criteria. A cluster analysis determined a total of four different groups as follows: Hybrids (HYB), Research Councils (REC), Technology Oriented Centres (TOCs) and Government Laboratories (GOL).

One of the criteria that may not be fully captured in the previous classifications, or at least not directly, is the process of internationalisation/Europeanisation of institutions. This, among other reasons, is an adaptation strategy of a changing research policy scenario (Cruz-Castro, 2015). According to a study developed by OECD and published in 2011, a surveyed list of Public Research Institutions described the methods of international linking as varied, ranging from informal exchange and researcher interaction on projects, to collaborative centres. The same study observed the tendency of having more diversified international linkages for entities with multiple research areas and intensive academic orientations. There are also instances of nationally funded organizations that have been able to evolve into a type of global actor and change some of their functioning rules (for example, the German Max Plank Society (MPG) or the French National Center for Scientific Research (CNRS). As a general rule, those PROs fundamentally financed through project-based funding and have small proportion of their funding portfolio coming directly from the state, face greater incentives to internationalise via fundraising or fund applications. Zacharewicz, Sanz Menendez & Jonkers in 2017 claimed Research Technology Organizations (RTOs) have progressively extended the scope of their activities outside their country of origin, motivated by producing excellent and high impact R&I while solving societal challenges and boosting industry's competitiveness.

The research in progress presented here aims to explore underlying structures or groups of ROs, relying on bibliometric indicators as the source of analysis. This, in order to achieve a deeper understanding of ROs in terms of publicness, internationalization and impact from the research performance perspective. For the purpose of obtaining quantitative measurements of the three aforementioned topics we selected one bibliometric indicator per matter. In this way, the **PP(Industry)** indicator**,** was chosen to quantify links with the private sector and therefore a way of indirectly measuring the degree of publicness shared by ROs. The **PP(Int_collab)** reflects international collaboration among institutions and **PPtop5%** to approximate measurement of scientific impact based on citation counts. Further efforts aim to determine to which extent these different profiles of ROs can relate to other current classification of Research Organizations.

Method

This study focuses on publications in journals processed for the Web of Science's (WoS) database produced by Clarivate Analytics. The indexes used are the Science Citation Index (SCI), Social Science Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI). The Conference Proceedings Citation Index-Science (CPCI-S) and Emerging Sources Citation Index (ESCI) databases within the WoS are not included in this study.

In our analysis we focus exclusively on European research institutions, in an attempt to maintain some degree of organizational similarity in the style of governance. This is under the assumption that within the EU, research organizations adhere to a similar structure compared to research organizations in other countries**.** Additionally, we filtered the list of European Research Organizations, isolating the 'parent' institutions in these webs, defined as the most central nodes in the structure of umbrella organizations; those that lie at the top of their organizational hierarchies. We then attributed all the publications of the 'child' organizations to their respective parent organizations. The advantage of this parent/child system is twofold: a simplification of the previously chaotic network benefits the analysis. Additionally, this resolves an underlying source of error inherent to this type of data. Namely, publications are occasionally incorrectly attributed to their parent institutes. The limitation of this dichotomy is that we lose information about the nuances of the child institutes, as all of their publications are attributed to the parents. As a result of applying all the mentioned filters, a total of 643 institutions were selected.

Afterwards, we calculated three bibliometric indicators, only taking publications and reviews into account, between the years 2009 – 2016. The analysis implements bibliometric indicators related to scientific impact and two types collaboration; with international actors and the association with the industry sector. The chosen indicators are all size-independent: **PPtop5%** (citation impact indicator): The proportion of a research organization's publications that, compared with other publications in the same field and in the same year, belong to the top 5% most frequently cited. **PP(Int_collab)** (Collaboration Indicator): Proportion of international collaborative publications. This means the proportion of a research organization's publication that has been co-authored by two or more countries. **PP(Industry)** (Collaboration Indicator):

The proportion of a research organization's publication that have been co-authored with one or more industrial organizations. All private sector, for profit business enterprises, covering all manufacturing and services sector, are regarded as industrial organizations.

We used K-Means clustering to cluster our data, set the number of clusters to 3 and used the Z-score of the parameters. Z-score assumes a normal distribution, so we visually inspected the distribution of the parameters. Two of our parameters had a non-normal distribution: **PP(Industry)** and **PPtop5%**. We have these parameters normally distributed by transforming **PP(Industry)** into logarithm of **PP(Industry)** with base 2 (**log_Industry**) and **PPtop5%** into square root of **PPtop5%** (**sqrt_top5)**. Therefore, we made the K-Means with the Z-score of **PP(Int_collab)**, **log_Industry** and **sqrt_top5**. We made the decision tree in order to facilitate the further interpretation of our clusters, by fitting a decision tree to the data points with deep 2 and the Gini Impurity as split criteria.

Results and preliminary conclusions

Through the K-means clusters, we observe the International collaboration variables as being the best predictor to create the clusters (Figure 2). This is supported with the information provided by the decision tree (Figure 3) which shows this variable at the top the tree to discriminate the first two nodes leading to the different clusters, where the decision rule set the range of values as ≤ 0.592.

In the Figure 2 one can observe there is a correlation between both variables, where an increase of collaboration with ROs from abroad is related to an increment in the scientific impact. This varies for specific institutions in the sample. These results are in line with many bibliometric studies carried out by CWTS (AKA, 2015; NWO-WOTRO, 2017). Conversely, it seems there is no direct correlation between the variables, International collaboration and Industry ties. Literature suggest as one of the motivations for internationalization having access to new markets. This allows a diversification of resources that may expand the economic revenue for the Organizations and avoid to rely solely on national funding sources (Zacharewicz *et al.*, 2017).

From the clustering results (Figure 2) we can notice that despite the differentiation of the sample in tree clusters, the tree groups of organizations share characteristics, this is clearer for the impact indicator (PPtop5%). In the other hand, it is possible to identify a small group of organizations performing as outliers, specifically for clusters 2 and 3. In general terms, these clusters have the same response for the Industry indicator showing two well defined groups, with the vast majority of the organizations scoring low values (94% values between 0 - 0,2) and a minuscule portion where almost all the publications are shared with the private sector. When analysing the organizations belonging to this small group, it is possible to identify Research Technological Organizations (RTO), considered as performing in the semi-public sector and with tights links to firms. It is worth to note that, the high percentage of ROs scoring low values, may also indicate that, when assessing research collaboration through the **PP(Industry),** this metric may not reflect the engagement of ROs with the private sector and perhaps other forms of collaboration among institutions, rather than publishing, are being developed, such as informal exchange or patents creation.

Organizations performing with high scores for International collaboration and scientific impact were represented for different ROs profiles, including MOCs, IRIs and RTOs. Regarding OECD (2011) IRIs are in many cases highly innovative in organisational terms and some of them have outstanding performances.

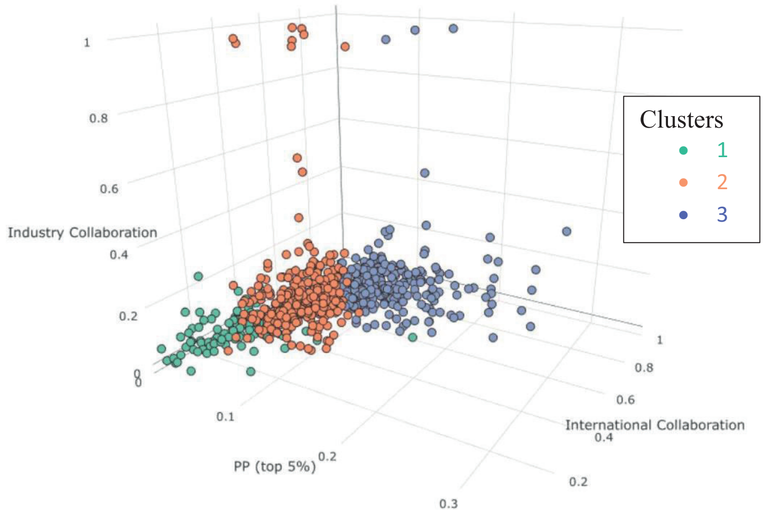


Figure 2. Scatter plot representing K-means output for a 3 dimensional space. The dots represent Research Organizations and the colours are clusters.

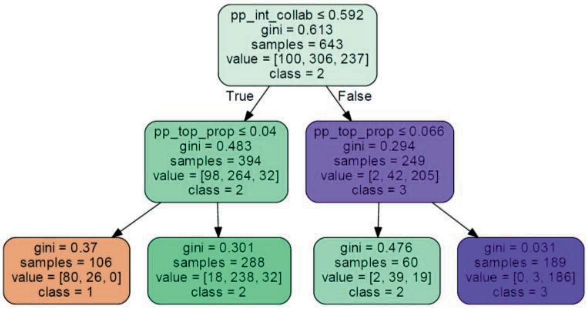


Figure 3. Decision tree output from K-means cluster model.
The top 3 nodes are decision nodes. The bottom 4 nodes are end nodes. The first row in a decision node is the logic test of the node. If true, go left, if false, go right. The second row is the Gini impurity of the node. The third row is the number of samples in each node. The fourth row is the number of samples from each class in the row (class 1, 2 and 3, respectively). The fifth row is the class predicted by the node. The colours represent the predicted class in the node, red is class 1, green is class 2 and purple is class 3. The intensity of the colours represents the relative frequency of that class in the node.

While our analysis does show that there are observable differences between each cluster of ROs, there were some limiting factors to our analysis. As previously mentioned, measuring industry collaboration is not straightforward due to the nature of these collaborations. We include collaborations that are based on accessible bibliometric information, however the true scope of industry collaboration may well exceed what is presently reported.

This article is a work-in-progress, and presents a preliminary exploration of the structures which underlie Research Organizations. More in-depth research is needed before we can make strong claims about the different types of ROs.

Acknowledgments

This work was supported by RISIS – Research Infrastructure for Research and Innovation Policy Studies, an EU FP7 Research Program Project, grant agreement no: 313082 and KNOWMAK - Knowledge in the making in the European society funded by European Union's Horizon 2020 under grant agreement 726992.

References

Academy of Finland (AKA) (2015) Bibliometric impact analysis of the Academy of Finland’s Centre of Excellence Programmes. *CWTS B.V. Bibliometric report, Centre for Science and Technology Studies, Leiden University*. Retrieved May 20, 2019 from:https://www.aka.fi/globalassets /31huippuyksikot/cwts_bibliometric_impact_coe_programmes_2015.pdf

Arnold, E., Barker, K., & Slipersæter, S. (2010). Research Institutes in the ERA. *Technopolis group*, Retrieved December 18, 2018 from: http://ec.europa.eu/research/era/docs/en/research-institutes-in-the-era.pdf

Boden, R., D. Cox, and M. Nedeva (2006), ‘The appliance of science? – New public management and strategic change’, *Technology Analysis and Strategic Management* , 18 (2), 125–241

Bozeman, B., & Bretschneider, S. (1994). The ‘publicness puzzle’ in organization theory: A test of alternative explanations of differences between public and private organizations. *Journal of Public Administration: Research and Theory*, 4(2), 197–223.

Cruz-Castro, Laura, Koen Jonkers, and Luis Sanz-Menéndez (2015) The internationalization of Research Institutes. In *Towards European Science Dynamics and Policy of an Evolving European Research Space*, eds. Linda Wedlin and Maria Nedeva, 175–198. Cheltenham: Edward Elgar.

NWO-WOTRO (2017) Assessing bibliometric performance of NWO-WOTRO funded research. *CWTS B.V.bibliometric report. Centre for Science and Technology Studies, Leiden University.* From: https://www.nwo.nl/en/documents/wotro/nwo-wotro-cwts-bibliometric-study-2017

OECD (2002), Frascati Manual 2002: Proposed Standard Practice for Surveys on Research and Experimental Development, *The Measurement of Scientific and Technological Activities, OECD Publishing*, Paris, https://doi.org/10.1787/9789264199040-en

OECD (2015), Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, *The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing*, Paris, https://doi.org/10.1787/9789264239012-en.

OECD (2011), Public Research Institutions: Mapping Sector Trends, *OECD Publishing*. http://dx.doi.org/10.1787/9789264119505-en

Peñasco, C, Sanz-Menéndez, L., Cruz-Castro, L., Martínez, (2017) The classification of Research Organisations: Taxonomical explorations with a set of “dedicated” Research organisations. *Annual Conference of EU-SPRI Forum.* 6-8 June 2017- Paris

Zacharewicz, T., Sanz Menendez, L., Jonkers, K. (2017), The Internationalisation of Research and Technology Organisations, *EUR 28442 EN*, doi: http://dx.doi.org/10.2760/944413

Tracking content updates in Scopus (2011-2018): a quantitative analysis of journals per subject category and subject categories per journal

Frederique Bordignon[1]

[1] *frederique.bordignon@enpc.fr*
Ecole des Ponts ParisTech, Direction de la Documentation, Champs-sur-Marne, France

Abstract

The aim of this study is to track Scopus content updates since 2011 and more particularly the distribution of journals into subject areas. An unprecedented corpus of data related to sources indexed in Scopus has been created and analyzed. Data shows important fluctuations regarding the number of journals per category and the number of categories assigned to journals. Those variations are very irregular, invisible to the average user and unpredictable over time. They question the reliability of studies based on Scopus data if no precautions are taken. The suggestion is made that category changes should not systematically be applied to all previously indexed publications of a journal, but only to those that will be indexed in Scopus after the new assignment is made.

Introduction

As far as scholarly literature is concerned, two levels of aggregation can be used to delimit scientific areas: the article level and the journal level. Both journals and articles can be classified into fixed sets of subject areas but the delineation of journals at the disciplinary level plays a major role in scientometrics, mainly for analyses based upon the extraction of scientific outputs from databases. That is why those information and reference sources need to be organized through an appropriate and consistent classification scheme (Gómez-Núñez, Batagelj, Vargas-Quesada, Moya-Anegón, & Chinchilla-Rodríguez, 2014). It serves as the basis of profiling authors, research groups, institutions or countries and helps in the making of comparisons and rankings. It is also useful in the calculation of standards for relative citation indicators. And beyond those research evaluation perspectives, journal classifications can also be used in describing the structure of scholarly publication and designing maps of science.

The two following tenets were formulated a long time ago: comparisons should be made in terms of "like with like" (Martin & Irvine, 1983) and over time in terms of fixed journal sets (Narin, 1976). But bibliographic databases do not take these principles into account and the data that is made available is used unsuspectingly by analysts in organizations.

The aim of this study is to track Scopus coverage updates since 2011 and more particularly the distribution of journals into subject areas. Scopus classification of documents is based upon the All Science Journal Classification (ASJC) whose structure does not evolve over time whereas the content of the different categories fluctuates substantially.

Background

Classification for evaluation purposes

(Archambault et al., 2011) stated that no international standard classification scheme exists that supports bibliometric research, and no single classification scheme has been widely adopted by the bibliometric community. Even research funders don't have a standardized classification

system to assess the impact of the funds distributed across different scientific fields (Katz & Hicks, 1995).
Among many initiatives to elaborate efficient classification systems for evaluation purposes, we can mention the following:

- the Steunpunt Onderwijs & Onderzoek Indicatoren (SOOI) implemented for the evaluation unit in Leuven (Glänzel & Schubert, 2003),
- the CHI Research classification (from Computer Horizons Inc) designed for the US National Science Foundation (NSF) (Carpenter & Narin, 1973) and also used by the Canadian Observatoire des sciences et des technologies (OST)
- the Australian Research Council Evaluation of Research Excellence (ERA) classification, abandoned since 2012.

However, it seems that the two most commonly used systems are those on which the Web of Science and Scopus databases are built.

The number and more particularly the diversity of classification schemes complicate comparative analyses (Gómez, Bordons, Fernández, & Méndez, 1996) because of the dual problem of matching categories and delineating journals comparable sets.

Mono vs multi-disciplinary classification systems

Some systems provide a way to classify publications with a great level of details in a restricted research area: for instance the widely used JEL (Journal of Economic Literature) classification system in economics, the Chemical abstracts service in chemistry or the MeSH (Medical Subject Headings) hierarchical system in medicine.
On the opposite, others are appended to multidisciplinary databases that index articles from journals and offer the possibility to retrieve them according to the field(s) the journals are assigned to. Journal level classification systems are of course very convenient but they are known as well to be sometimes too fuzzy, at least not so accurate as article level classifications. Indeed it is well known that most journals contain articles dealing with a relatively broad range of themes, in spite of their "main subject". Thus, a subject delimitation based on journal classification will probably contains some articles weakly related with the target subject, while some pertinent articles will be missing (Bensman, 2001; Gómez et al., 1996). And (Pudovkin & Garfield, 2002) said about the Web of Science classification system that "journals are assigned to categories by subjective, heuristic methods. In many fields these categories are sufficient but in many areas of research these ‘classifications’ are crude and do not permit the user to quickly learn which journals are most closely related."

Multiaffectation classification systems

Another limit of journal level classifications is due to the fact that many journals are assigned to multiple categories to better represent the scientific themes their articles deal with. A mutually exclusive classification is of course more convenient, in particular because it prevents a journal to being counted more than once. What is more, those classifications are generally not well documented (Archambault et al., 2011), and therefore there is no indication about why one or more categories were chosen for a particular journal. (Wang & Waltman, 2016) found that a significant share of the journals in both databases, but especially in Scopus, seem to have assignments to too many categories and then suggested to adopt a stricter policy supported by the use of citation analysis when assigning journals to categories.

Multiaffectation is supposed to reflect interdisciplinarity but in the end, multidisciplinary journals (eg: *Nature*, *PNAS*, and *Science*) are the most poorly managed and that is what leads (Wang & Waltman, 2016) to reconsider journal classification systems at a more fundamental level and warn an increasing share of publications cannot be properly classified at the journal level because of the increasing popularity of large multidisciplinary journals (eg: *PLoS ONE*). Scopus allows the assignment of a journal to several ASJC categories.

Stability over time

Exploring the limits of existing classification schemes and trying to improve them has given rise to many studies in the field of bibliometrics and scientometrics. But among them, the problem of the changes over time has more rarely been tackled and assessed. At least, the problem of fitting new journals into existing schemes has been dealt with: (Leydesdorff, 2002) with the idea of comparing structural changes in a database with reorganizations of relations among previously included journals concludes that "if one does not systematically account for redelineation in the groupings over time but uses "fixed journal sets" instead, one risks making a prediction of performance with reference to an outdated unit.". Despite this conclusion, Scopus (more particularly the possibility to request for Scopus data according to preset corpus of journals in different subject areas) is still the easiest way to retrieve data to produce reports in many organizations.

The question is whether these analyses are reliable when they cover different periods of times and subject areas if the different sets of journals are not stable and if the changes are not clearly reported. Indeed, queries in Scopus do not take into account any journal assignation update according to the publication year the query is based upon.

In this study, we investigate 2 kinds of potential changes in Scopus: (1) number of journals per categories (2) number of categories per journal, both impacting the delineation of categories and therefore the data retrieved from Scopus.

Methods

The All Science Journal Classification scheme

Scopus journal classification system is called the All Science Journal Classification (ASJC). There seems to be no official description about the way it is constructed. It has always been freely available online either from a dedicated page on the Elsevier website and an Excel file available for download, or from the former JournalMetrics website. It can also been downloaded from Scopus database (*Browse sources* page).

It is commonly described as consisting of two levels, but there is actually a third level above all, differently called Top-Levels, Supergroups or Subject areas (depending on time periods and downloadable files). This uppest level is not used at all in Scopus but can be used to filter out journals in the Excel file. The lowest level has 307 subfields, and the intermediate level includes 26 fields called *Subject areas* in Scopus. There is another field and another subfield for the *Multidisciplinary* category. All the subfields are assigned a 4-digit code.

We do not know how the assignment of journals to fields and subfields is decided but we can infer it is done by the Scopus Content Selection and Advisory Board, "an international group of scientists, researchers and librarians who represent the major scientific disciplines" (Elsevier website, 2019[1]).

Table 1. ASJC journal classification system

Supergroups	***Fields***	***No. of Subfields***
-	Multidisciplinary	1
Health Sciences	Medicine	48
Health Sciences	Nursing	23
Health Sciences	Veterinary	4
Health Sciences	Dentistry	6
Health Sciences	Health Professions	16
Life Sciences	Agricultural and Biological Sciences	11
Life Sciences	Biochemistry, Genetics and Molecular Biology	15
Life Sciences	Immunology and Microbiology	6
Life Sciences	Neuroscience	9
Life Sciences	Pharmacology, Toxicology and Pharmaceutics	5
Physical Sciences	Chemical Engineering	8
Physical Sciences	Chemistry	7
Physical Sciences	Computer Science	12
Physical Sciences	Earth and Planetary Sciences	13
Physical Sciences	Energy	5
Physical Sciences	Engineering	16
Physical Sciences	Environmental Science	12
Physical Sciences	Materials Science	8
Physical Sciences	Mathematics	14
Physical Sciences	Physics and Astronomy	10
Social Sciences	Arts and Humanities	13
Social Sciences	Business, Management and Accounting	10
Social Sciences	Decision Sciences	4
Social Sciences	Economics, Econometrics and Finance	3
Social Sciences	Psychology	7
Social Sciences	Social Sciences	22
4 supergroups	**26 fields**	**307 subfields**

The *Multidisciplinary* field and its unique subfield is dedicated to journals with a very broad multidisciplinary scope like *Nature*, *Science* or *Scientific reports.*

In all fields, there are 2 quite similar subfields:

- one whose label starts with the "*General*" mention and code ends with 00,
- the other whose label ends with the "*(miscellaneous)*" mention and code ends with 01.

It is impossible to say what led the Scopus experts panel to choose between the *General* subfield or the *Miscellaneous* corresponding subfield. There are many examples of journals assigned to both (for example, *Biology Letters*, assigned to the "*General Agricultural and Biological Sciences*" subfield and the "*Agricultural and Biological Sciences (miscellaneous)*" subfield.

Like other classification schemes, the ASJC has been criticized, most frequently because of confusing subfield labels (*Linguistics & Language* and *Language & Linguistics*, (Wang & Waltman, 2016)) or strong imbalanced distribution of journals and therefore documents among

[1] https://www.elsevier.com/solutions/scopus/how-scopus-works/content#content-policy-and-selection, available online on February 2019

the fields (Jacsó, 2013). There have been attempts to improve it (Gómez-Núñez et al., 2014; Jacsó, 2013) but still few considerations about the impact of coverage changes over time.

The nomenclature structure itself is stable since 2011, no new field or subfield has been created over the period we are interested in. Codes remained the same and names of fields and subfields as well, excepted for the *General* field which changed name in 2016 and was renamed *Multidisciplinary* and all the subfields ending with the "*(all)*" mention (eg: *Agricultural and Biological Sciences (all))* that changed name and have been started with "*General*" since 2017 (eg: *General Agricultural and Biological Sciences*).

Data

We retrieved all the title list files Elsevier has released twice a year since 2011 to investigate what content is included in Scopus. Journals, trade journals, conferences and book series are listed but we only focus on journals in this study. Most of those files are still available online thanks to the Wayback machine website. They are the best way to retrieve the metadata needed to our study. We only kept one file a year (the one published at the end of each year) and compiled the 8 files into a single dataset. The aggregated data (Bordignon, 2019) used for this study is available for reuse and further investigation (SNIP values and Open Access status are included in the dataset even if they are not analyzed in our study).

Results

Inclusion and withdrawal of journals at the category level

Table 2. Number of journals included in Scopus and annual growth

	2011	2012	2013	2014	2015	2016	2017	2018
No. of journals	28 335	29 561	31 154	32 332	33 058	33 810	34 772	36 189
Annual growth	-	+4,3%	+5,4%	+3,8%	+2,2%	+2,3%	+2,8%	+4,1%

As far as Scopus content is concerned, the most significant change since 2011 is the increasing number of journals indexed in the database (+28% between 2011 and 2018, with most important increases in 2012 (+4,3%) and 2013 (+5,4%)). Very few journals are merely dropped (min=27;max=318). And even inactive journals are sometimes added to the index (inactive either because they changed name, merged with another journal, splitted or simply ceased to publish anything).

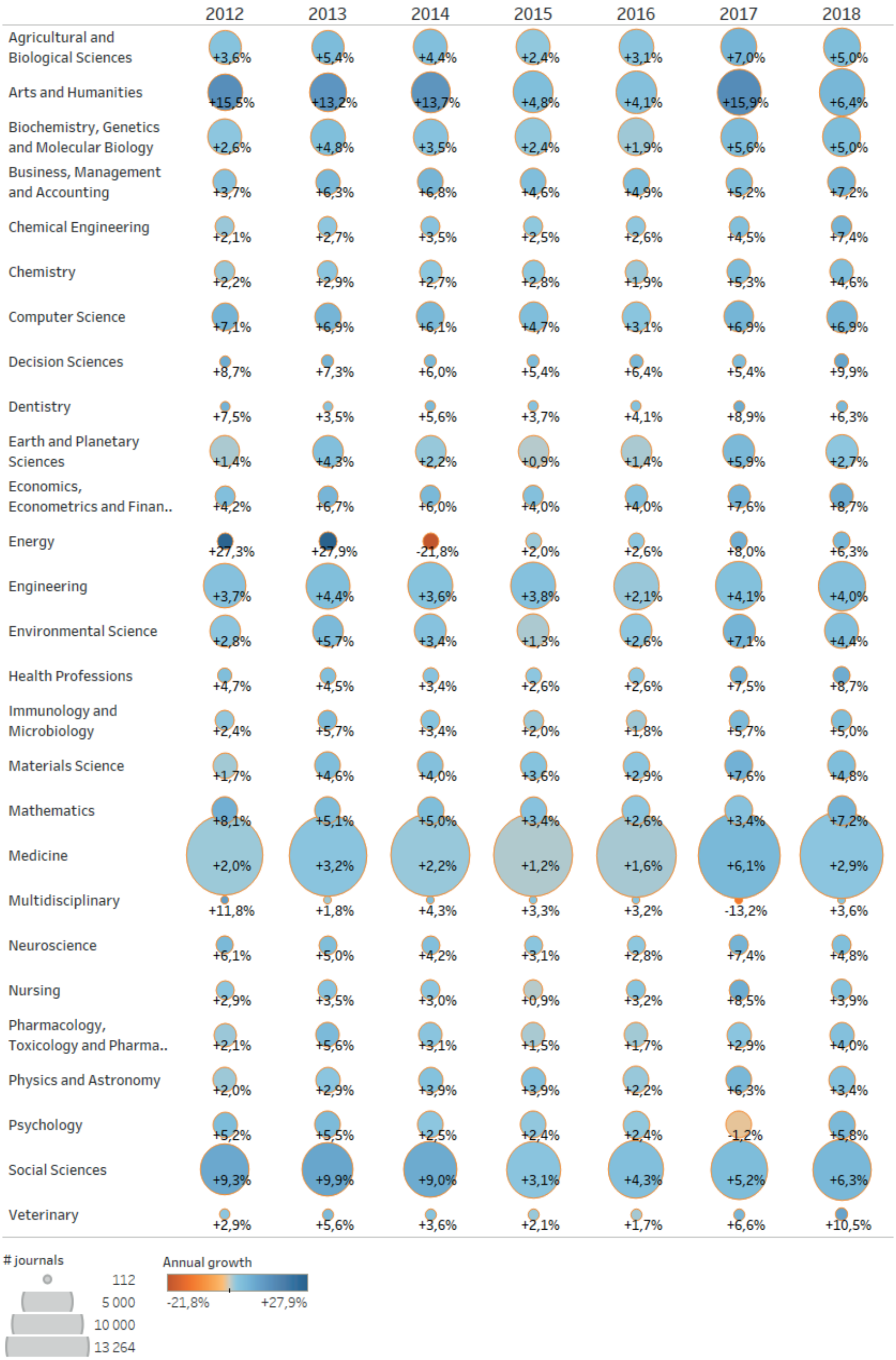


Figure 1. Number of journals per field and annual growth

The evolution of the number of journals is contrasted from one field to another. Here are the highlights Figure 1 reveals:

- in 2012, 2013 and 2014, among the largest fields (2000+ journals), the fields *Arts & humanities* and *Social sciences* had the highest increase (from 9% to 15.5% annual growth). Another large inclusion of sources also occurred in 2017 in *Arts & humanities* (+15.9%, ie: 506 journals added)
- in general, there has been no significant increase in any field in 2015 and 2016, with a maximum of +5.4% in 2015 and a maximum of +6.4% in 2016 (both concerning the *Decision sciences* field).
- in 2017, 500+ journals were added to the *Arts & humanities* field (+15,9%)
- few fields are undergoing decreases:
    - *Psychology* in 2017, -1.2% but this only represents 15 journals, 73 were finally added the following year
    - *Multidisciplinary* in 2017 also with -13.2%, but this amounts to only 17 dropped journals
    - the *Energy* field must be considered as a particular case: indeed it recorded both the highest increase and the highest decrease over the entire period, with the addition of 126 journals in 2013 and the same amount of sources withdrawn the year after. Out of the 126 journals added in 2013, 91 were removed from Scopus in 2014 (all belonging to the *General Energy* subfield).

These updates are unpredictable and have inevitably an impact on comparative studies that are conducted on those fields at different periods of time.

Apart from these fluctuations (mainly additions) in the number of journals per field, it is important to know whether these additions are newly included journals or whether they "come from" other fields, in other words whether journals would change field/subfield, be assigned to more fields/subfields or withdrawn from any field/subfield.

Table 3. Annual percentage of journals whose assignment to fields and subfields has been updated

2012		2013		2014		2015		2016		2017		2018	
subfield	field	subfield	field	subfield	field	subfield	field	subfield	field	subfield	field	subfield	field
0,1%	0,1%	0,2%	0,2%	1,0%	1,0%	0,5%	0,5%	0,3%	0,3%	**10,6%**	**6,9%**	1,5%	1,0%

Table 3 clearly shows that field or subfield shifts are very unusual. This means that a journal is only very rarely reassigned to more or less fields/subfields. The changes mentioned above are therefore only due to additions or withdrawals of journals from the Scopus index. However, a significant change can be observed between 2016 and 2017; Table 4 focuses on this time period and shows that, apart from the *Multidisciplinary* field (and only 33 sources involved), the field for which most shifts are detected is *Psychology* with a significant share (21.6%) being reassigned to more or less subfields.

Table 4. Number and percentage of journals per field with subfield shifts between 2016 and 2017

Fields	No. of journals with subfield shifts	% of all journals in the field
Multidisciplinary	33	25,60%
Psychology	274	21,6%
Immunology and Microbiology	132	18,0%
Biochemistry, Genetics and Molecular Biology	439	17,7%
Health Professions	85	17,7%
Pharmacology, Toxicology and Pharmaceutics	172	16,3%
Neuroscience	99	15,9%
Veterinary	33	13,60%
Medicine	1630	13,4%
Nursing	88	12,5%
Environmental Science	239	12,1%
Social Sciences	652	11,2%
Engineering	376	9,5%
Computer Science	136	8,7%
Earth and Planetary Sciences	149	8,1%
Agricultural and Biological Sciences	186	8,0%
Arts and Humanities	201	6,3%

The Scopus search interface does not allow to query or filter on subfields. But skilled analysts who use the source list file can do this sorting after having exported the bibliographic data; their comparative analyses are likely to be biased because of too important changes between 2016 and 2017.
In addition, some world university rankings by subject are based on citation indicators collected across several subfields. The results of these rankings are necessarily skewed by these significant updates in Scopus.

Number of categories per journal

Our 2018 data shows that the maximum number of fields assigned to a journal is 9 (*The Bulletin of mathematical biophysics*) whereas the highest number of subfields is 13, assigned to *Journal of Geophysical Research*. This example reveals several technical problems actually: first of all, this journal is organized in 7 disciplinary sections (eg: *JGR: Atmospheres*, *JGR: Biogeosciences* etc.). Those sections are not integrated into the Elsevier title list, this might be the reason why so many subfields are assigned to this source. On the other hand, when querying Scopus sources index about *Journal of Geophysical Research,* there are 2 answers: one for the *stem* journal without any mention of sections, the other for a single specific section (*Solid Earth*). And finally, when searching for any documents with *Journal of Geophysical Research* as source title, relevant results indicate the complete correct titles of all the discipline sections. Even if further investigation is needed to measure the extent of the issue, it seems that the 3 sources of information about Scopus content are not consistent.

As far as our dataset is concerned, unsurprisingly, it shows an increasing average number of fields (+1,49% since 2011) and subfields (+3,11% since 2011) assigned to journals, which seems to attest the increasing interdisciplinarity of science (Morillo, Bordons, & Gómez, 2003). Our calculation of the average number of fields assigned to journals is consistent with (Wang & Waltman, 2016) results (2.1 in Scopus) as shown in Table 5.

But consistently with what we stated earlier, this increase is not due to updates at the journal level but almost exclusively due to the addition of journals to the database index. Since 2013, those newly included sources have always been assigned to more fields and subfields on average than those previously indexed.

Table 5. Average number of fields and subfields per journal for added or previously included ones

		2011	2012	2013	2014	2015	2016	2017	2018
Avg number of fields	Added	1,48	1,51	1,56	1,57	1,61	1,66	1,67	1,66
	Previously included	1,53	1,53	1,53	1,53	1,53	1,54	1,54	1,54
Avg number of subfields	Added	1,91	2	2,02	2,08	2,26	2,37	2,45	2,38
	Previously included	1,99	1,98	1,98	1,99	2	2,01	2,01	2,03

This is a global tendency and further studies will be able to reveal differences among fields. It should also been worth checking whether newly created journals (and not newly included ones) are being assigned more fields and subfields as well.
It seems unlikely that interdisciplinarity will only arise on newly added journals. This means that the journals already in the database should be re-examined by the Elsevier experts panel. And of course this reinforces the idea that indexing at the article level better reflects reality.

Discussions and perspectives

Whether analysts work directly in Scopus or use the data Elsevier makes available in downloadable files, they cannot perform the time-consuming analysis work that would assess if updates to Scopus coverage are not too substantial and if comparing reports produced from one year to the next is still possible.

Moreover, the large volume changes we have highlighted in some categories certainly do not reflect the scientific reality of the field but rather Elsevier's objectives to increase its coverage. And yet, the consequences can be significant: for example, on SNIP values due to an unstable scope of journals and therefore a very unstable citations rate, or on international thematic university rankings whose evaluation criteria are based partly on the collection of citations and outputs according to subject areas.

Without giving up the extension of the coverage and the necessary updating of the database, Elsevier should inform the user of Scopus content updates in order to prevent potential impacts on the resulting analyses. This is obviously something very complex to set up in the interface, but one cannot assume that all users regularly consult the title list file. One possibility is to reflect category changes of a journal only on newly added publications (recently published or not) and not on all publications already present in the database. It will mitigate the bias for university rankings or the calculation of indicators.

As for the increase in the average number of fields and subfields per journal, since it is limited to additions, it cannot be said that it can be used to support work on interdisciplinarity.
On this particular point, it should be reiterated that a journal can be added to the list of indexed sources even if it has a long-established history, or even if it is inactive. Therefore, there is a

limit to our analysis since we would have to examine whether the increase in the average number of fields/subfields per journal is true for all journals added to the index or more particularly for those created and included at the same period of time.

Conclusion

We know that classifications at the article level are more relevant, but since bibliographic databases offer the possibility of queries, analyses and data exports based on the classification of journals, it is important to know to what extent this could impact their reliability for bibliometric analyses.

We revealed the existence of very important updates in the Scopus database which can have a significant impact, depending on the scope of the analyses carried out. We also showed that these fluctuations were very irregular, invisible to the average user and unpredictable. That is why we suggested that category changes should not systematically be applied to all previously indexed publications of a journal, but only to those that will be indexed in Scopus after the new assignment is made.

References

Archambault, É., Beauchesne, O. H., Caruso, J., Archambault, É. ;, Beauchesne, O. H. ;, & Archambault, C. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. *Proceedings of the 13th international conference of the international society for scientometrics and informetrics. E.C.M. Noyons, P. Ngulube, J. Leta (Eds.)* (pp. 66–77). Retrieved December 23, 2018, from www.sciencemetrix.com

Bensman, S. J. (2001). Bradford's Law and Fuzzy Sets: Statistical Implications for Library Analyses. *IFLA Journal*, *27*(4), 238–246.

Bordignon, F. (2019). *Scopus sources title list: aggregated data*. *Mendeley Data*.

Carpenter, M. P., & Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science*, *24*(6), 425–436.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Gómez, I., Bordons, M., Fernández, M. T., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics*, *35*(2), 223–235.

Gómez-Núñez, A. J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., & Chinchilla-Rodríguez, Z. (2014). Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, *8*(2), 369–383.

Jacsó, P. (2013). The need for end-user customization of the journal-sets of the subject categories in the SCImago Journal Ranking database for more appropriate league lists. A case study for the Library & Information Science field. *El Profesional de la Informacion*, *22*(5), 459–473.

Katz, J. S., & Hicks, D. (1995). The Classification of Interdisciplinary Journals: A New Approach (Version 2.0). *Proceeding of The Fifth Biennial Conference of The International Society for Scientometrics and Informatics, Rosary College, River Forest, Il, USA, June 7-10, 1995*. IEEE.

Leydesdorff, L. (2002). Dynamic and evolutionary updates of classificatory schemes in scientific journal structures. *Journal of the American Society for Information Science and Technology*, *53*(12), 987–994.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinary in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, *54*(13), 1237–1249.

Narin, F. (1976). *Evaluative Bibliometrics: the use of publication and citation analysis in the evaluation* *of scientific activity*. *Computer Horizons, Inc Project No. 704R - Contract report to the National Science Foundation*.

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113–1119. John Wiley & Sons, Ltd.

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364.

3D printing as a research domain: mapping the main areas of knowledge

Andréia Cristina Galina[1] and Jacqueline Leta[2]

[1]*andreia.galina@bioqmed.ufrj.br*
Federal University of Rio de Janeiro, Av. Brigadeiro Trompowisky s/ nº, Prédio do CCS, Bloco B – sala 39,
CEP 21941-590, Rio de Janeiro (Brazil)

[2]*jleta@ bioqmed.ufrj.br*
Federal University of Rio de Janeiro, Av. Brigadeiro Trompowisky s/ nº, Prédio do CCS, Bloco B – sala 39,
CEP 21941-590, Rio de Janeiro (Brazil)

Abstract

The 3D printing is an emerging technology whose principle is to deposit material in thin layers until the piece takes on the projected shape. 3D printing has the potential to change society as well as some scientific fields. Hence, considering this last aspect, the present paper aims to map the most prevalent research areas in 3D printing domain during the 1983-2017 period analyzed. A total of 119,800 documents on 3D printing were collected from Web of Science and Scopus. Results show that the number of documents on 3D printing increased 13.65%, when considering Scopus and Web of Science together. The areas with the highest ratio of 3D printing documents are engineering, computing and materials science. The analysis of keywords co-occurrence and field trends of magazines and conferences revealed also a close linkage between 3D printing and health sciences, where the concepts of bio-ink, bioprint and bioplotter are introduced. These results reinforce the role of 3D printing as a research domain that is been structured upon different areas.

Introduction

Emerging technologies have the potential to create a new sector or to transform an existing sector; they may be also considered as discontinuous technologies derived from radical innovations or by the convergence of previously separated research chains (Day, Schoemaker & Gunther, 2009). An example of this type of technology is the 3D printers, whose first patents were granted in the middle of 1980's.

The 3D printing technology is based in the additive manufacturing process, since it works under the principle of material deposition in thin transverse layers until the designed shape is acquired. It differs from the conventional additive manufacture that is grounded in the subtractive process and the material removal from the block to be formed, which requires a variety of specific tools and equipment. Comparing the additive and the subtractive manufacturing, in the second we have a longer time production as well as a higher energy and material consumption and a higher amount of waste and pollutants generated.

According to Berman (2012), the usage of 3D printing is divided into three phases: in the first, its greater use aims at the construction of prototypes or models; in the second phase, its use aims at the creation of products for marketing test, with variations of size, styles, and colors. This phase is also known as direct digital manufacturing or rapid tooling. In the third phase, the main use is made by final consumers, who may have their individual printers and, instead of buying a trade product, they can purchase the product file and print it by their own. The progress to the third phase depends on the ease of software execution for developing each part, as well as the cheapening of inputs and equipment.

It is noteworthy highlight the economic impact of 3D printing, estimated to be between US $ 230 billion and US $ 550 billion per year by 2020, which would have a higher influence on the final consumer, i.e. finished products that are cost-effective compared to the acquisition in retail (Manyika et al, 2013).

Since its development, 3D printing has been considered a technology with a high potential to change the world around us. Rifkin (2012) and many authors believe that it is responsible for the 4th industrial revolution. For other authors (Prince, 2014, Garrett, 2014, Gleber, Uiterkamp & Visser, 2014), 3D printing is about to bring a radical change in contemporary culture due to the advance in manufacturing products applied to industry, art, medicine and domestic environment. Also, they believe it has the power to provoke changes in the world economy by changing business models, production sites and supply chains as well as promoting changes in work structures.

The impact of 3D printing is already perceived in our society through a range of products, including parts to build a house or components to assemble cars and airplanes. Through this technology it is also possible to produce clothes and accessories, household utensils, food and medicines. More recently, we witnessed products coming from 3D printing that may represent a paradigmatic chance in society: the impression of skin, bone, vascular grafts, tracheal splints, cardiac tissue, cartilaginous structures, models of molecules for research (Murphy, Atala, 2014). Such products and others not yet available have the potential to change the form and time of medical treatments. In fact, the application of 3D printing in solving issues in regenerative medicine will represent its most disruptive use (the one that causes interruption of the regular course of a process), being the impression of organs its main challenge.

Taking into account the current (and the future) uses of 3D printing in generating products that impact different areas and social sectors, we decided to investigate whether this technology figures out a domain, in the sense of Hjorland & Albrechtsen (1995). Hence, considering 3D printing as a scientific domain, in which their actors share thoughts, discourses and communication forms, among other aspects, we are interested, in this paper, to map the most prevalent research areas in this domain. Although the technology dates back to the 1980s, to our knowledge, there is no similar study within the scientometric or bibliometric literature. We understand, therefore, that this is a pioneering study, in which the mapping of the most prolific areas on 3D printing along the last decades will allow us to better understanding how research is been developed in this domain.

Materials and Methods

The search strategy included three main steps. The first search strategy contained a couple of words found in articles and news about 3D printing plus some synonyms listed in MESH terms and their variations. Then, based in the results of the first strategy, a co-occurrence analysis of keywords was performed, which allowed to identify some other words and uses of 3D printing techniques applied to specific areas that started to have a proper term as, for example, bioprinting. Finally, words and their variations regarding the seven additive manufacturing families listed in ISO / ASTM 52900: 2015, or formerly ASTMF2792 (Standard, A.S.T.M., 2012) were added to the search. The final search strategy presented 49 word or expressions for 3D printing, as following: "Material Extrusion" OR "Fused Deposition Model*" OR "Fused Filament Fabricat*" OR "Directed Energy Deposit*" OR "Laser Engineer*" OR "Net Shap*" OR "Digital Light Process*" OR "Continuous Liquid Interface Production" OR "Continuous Digital-Light Process*" OR "Selective Laser Sintering" OR "3d print*" OR "threedimensional print*" OR "three dimensional print*" OR "3 dimensional print*" OR "Fast Prototyp*" OR "Solid Free Form" OR "Solid Freeform" OR "Rapid Prototyp*" OR "Additive Manufacturing" OR "VAT

Photopolymerisation" OR "Stereolithography" OR "ExOne" OR "Powder Bed Fusion" OR "Direct Metal Laser Sinter*" OR "Selective Laser melt*" OR "Electron Beam Melt*" OR "Material Jet*" OR "Multi-Jet Fus*" OR "Binder Jet*" OR "Ultrasonic Additive Manufactur*" OR "Voxeljet" OR "Drop On Demand" OR "Nano Particle Jet*" OR "Polyjet"

OR "Sheet Laminat*" OR "Laminat* Object Manufactur*" OR "Selective Deposit* Laminat*" OR "Electron Beam Additive Manufactur*" OR "Laser Metal Deposit*" OR

"Direct Metal Deposit*" OR "prototyping" OR "bioplot*" OR "bioprint*" OR "bio-print*". The detail of this process is the focus of a scientific article in development.

The data on the scientific production in 3D printing were retrieved from the two leading scientific bases: Web of Science (WoS) and Scopus. An update search was carry on January 21st and 22nd of 2019.

A total of 88,571 documents were retrieved from Scopus and 50,074 documents from WoS. The documents of both databases were joined up using the R version 3.5.2 programming language (R Core Team, 2018) and the bibliometrix package, version 2.0.2 (Aria, Cuccurullo, 2017). To remove duplicities, we considered three criteria: title, journal name and year of publication. As can be seen in the Venn-Euler diagram (figure 1), the intersection between the Scopus and WoS documents was 18,845 documents; the duplications represent 13.89% of the sum of the documents of both databases. Once the duplicity was withdrawn, the union between Scopus and WoS summed of 119,800 documents.

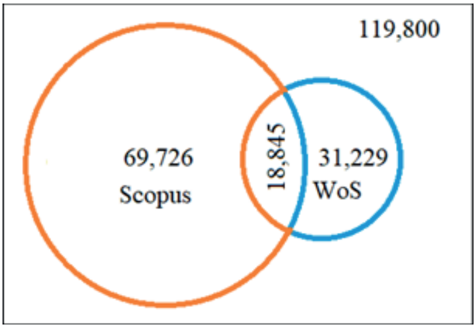


Figure 1. Venn-Euler diagram with the totals of documents on 3D printing retrieved from the databases Scopus and WoS.

The period analysis 1983 - 2017 was divided into seven quinquennia to visualize the evolution of the data studied, namely: 1983 - 1987, 1988 - 1992, 1993 - 1997, 1998 - 2002, 2003 - 2007, 2008 - 2012 and 2013 - 2017.

In order to evaluate if the growth on the topic 3D printing was significant, we compared it with the growth of the total number of documents indexed in each database throughout the studied period. Data on total number of documents in each database was obtained on April 07th of 2018.

We used VOSviewer software version 1.6.9 for elaborating graphs and the analysis of networks (van Eck, Waltman, 2009). For the co-occurrence network, we used all keywords listed in the 119,800 documents, but a cleanup was performed to remove nonsignificant terms (like 'na'). Cuts were performed by increasing the minimum number of occurrence of the word so that the network had a maximum of 1000 nodes, allowing visual analysis to be possible.

Due to the large number of data and analyzes as well as the limitation of space of the present paper, we chose to perform a cut in some analyzes selecting only the odd quinquennia.

For the calculation of the average growth rate in the period (35 years), we used the following equation:

$$\text{Growth Rate}= [((\text{valor final} / \text{start value})^{\wedge}1/\text{period}) - 1]*100 \qquad (1)$$

Results and Discussion

The results are presented in two groups. In the first, we evaluated the growth of scientific production on 3D printing, the countries responsible for these publications and the distribution towards the type of production. We understand that this first block of analyzes is necessary to better contextualize the following section, which is the focus of our study. In the second group of results, we present analyzes on keywords co-occurrence and trends of areas regarding journals and annals that allow us to identify the most prolific areas in 3D printing publications.

Growth, affiliation, and typology of scientific publications on 3D printer

The number of documents per period can be visualized in figure 2, where the total number of documents in both databases is represented by the bars, while the number of documents on 3D printing found in the two databases without duplicates (n = 119,800) is presented on the black line. For comparison, the figure also includes the annual totals of 3D documents at each database (Scopus - blue line; WoS - orange line).

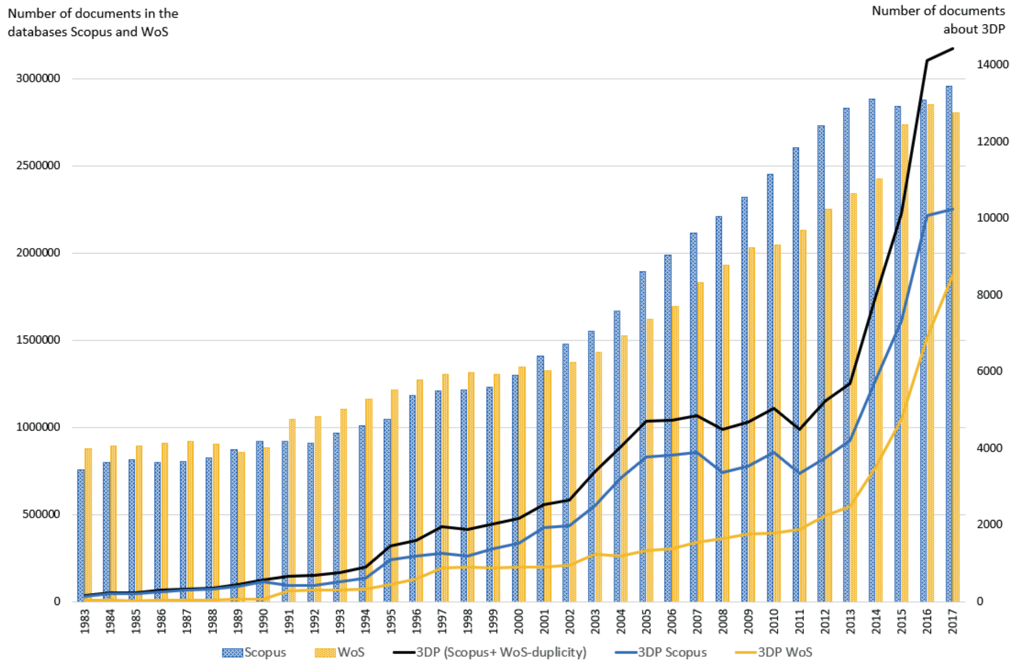


Figure 2. Number of total documents in Scopus and Web of Science (blue and yellow bars), number of 3D printing documents in each database (blue and yellow lines) and the sum of 3D printing in both databases without duplicity (black line), 1983-2017.

It can be noted a continuous and little-accentuated growth of total documents registered in the each database from 1983 to 2017 (blue and yellow bars). A different profile is observed for 3D printing documents (black line), which starts with a slow growth until 1993, but from this point onwards, there are three marked growth movements: one starts in 1994, another in 2002 and a last growth, where can be seen an explosion, that occurs by 2011.

The growth waves observed for the scientific production on this technology may have been influenced by events and initiatives that promoted and spread the concept of 3D printing, including: (a) the launch of the RepRap community in 2008, which began to teach, through free videos available on the internet, how to assemble a low-cost printer, which is the first open source 3D printer; (b) the foundation of the Thingiverse site, the first dedicated to file sharing for 3D printing using open source hardware (Michael et al, 2013); and (c) a patent breach in 2009 of the fused deposition modeling (FDM) technique, which allowed, in 2012, the launch of the first desktop printer. Afterwards, an explosion occurred in the manufacturing and supply of printers in the market using the FDM technique.

In order to evaluate the growth of the documents in 3D printing, we calculated the average growth rate in the period, as shown in table 1. The average growth rates of total documents are 3.98% and 3.39% in Scopus and in WoS, respectively. Documents on 3D printing displays an average growth of 13.07% and 18.00% in Scopus and in WoS, respectively. The sum of Scopus and WoS documents on 3D printing, withdrawn duplicity, presents an average growth rate of 13.65%, that is, more than three times higher than the growth observed for each database, indicating that 3D printing is emerging as a research domain.

Table 1. The average growth rate of the number of documents in the Scopus, WoS and of the sum of both databases on 3D printing documents in the period 1983-2017.

Source	Average growth rate in the period
Scopus	3.98%
3 DP Scopus	13.07%
Web of Science	3.39%
3 DP Web of Science	18.00%
3 DP (Scopus + WOS) - duplicity	13.65%

Once the previously results showed the number of 3D printing documents increased notably during the studied period, the document by type and by country of the authorship was analyzed in order to get evidences of how this domain are structured.

Figure 3 shows the total number of documents with single (blue) or multiple authorship (orange) by one of the 15 countries that contributed the most in four quinquennia. One first remark is that majority of documents are single authored, that is, the level of collaboration in 3D printing domain is very low.

As for the countries, it is clear the role of the United States that stands out in the number of publications in all quinquennia. In the last period, as evidenced, the country shows a lower performance. China, which did not appear in the period 1983-1987, appeared in the period 2003-2017 at second in the ranking position and remains in this position in the subsequent period with more than double documents as the previous period. Other Asiatic countries figured among the top-15 countries: Japan (in all periods), Korea and Singapore (since period 1993-1997) and Twain (since 2003-2007).

Two Latin American countries are included the ranking of top 15 countries with the highest number of documents on 3D printing are: Argentina, with a coauthored publication in the first period, and Brazil, which appeared in the 13th position (with 269 single authored documents) in the period 2003-2007 and in the 14th position (with 595 single authored documents) in the period 2013-2017.

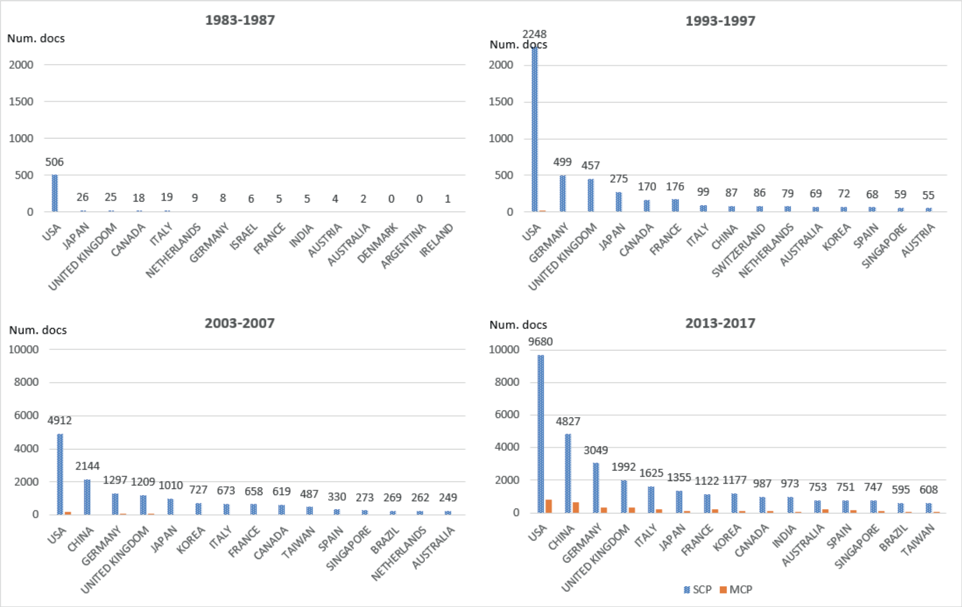


Figure 3. Number of documents on 3D printing retrieved from WoS and Scopus with single-authored (blue) or multiple-authored (orange), according to the author's country of affiliation, in different quinquennia.

In order to get a description of the communication forms of 3D printing domain, we The number documents by sources and by different types are shown in table 2. The distribution of documents by type is shown as a percentage of the total number of documents per quinquennium, whereas the classification in the ranking (right side) considers the total in the period.

It can be noted a remarkable increase in the number of sources in the period, with an average growth rate of 10.18%. We may consider the emergence of magazines on the 3D printing as well as magazines devoted to diffuse other types technological applications.

In table 2, we can also verify that the scientific article is the most frequent typology among documents on 3D printing. This typology presents the highest number in four of the six quinquennia, representing 43.21% of the total. The typology with the second highest frequency is conference paper, which represents 34.33% of the total and, for two consecutive periods, 2003-2007 and 2008-2012, it has the largest number of documents. It is well known that conference papers is the most common means of communication of some areas, such as engineering, since it brings more speed in the diffusion for the new knowledge or for the new technology, thus guaranteeing, in a more agile way, the priority of the discovery or invention.

Table2. Number of documents on 3D printing retrieved from WoS and Scopus by sources and by typology in different quinquennia, 1983 – 2017.

Period	*1983* *1987*		*1988* *1992*		*1993* *1997*		*1998* *2002*		*2003* *2007*		*2008* *2012*		*2013* *2017*		*1983* *2017*	
Sources (Journals, Books, etc.)	432		957		2268		3480		6210		7979		12838		27179	
Document types	**% of 1244**		**% of 2677**		**% of 6650**		**% of 11229**		**% of 21703**		**% of 23956**		**% of 52341**		**% Total**	
Article	53,30	1	52,86	1	45,46	1	46,89	1	33,82	2	36,34	2	48,44	1	43,209	1
Conference Paper	43,57	2	36,50	2	32,33	2	29,17	2	44,61	1	41,22	1	27,95	2	34,333	2
Inproceedings			7,28	3	18,51	3	20,76	3	15,86	3	18,83	3	17,85	3	17,576	3
Conference Review	1,93	3	2,09	4	2,24	4	1,35	4	1,20	5	1,20	4	1,11	6	1,260	4
Review	0,32	5	0,67	5	0,56	5	0,97	5	2,58	4	0,56	6	1,12	5	1,207	5
Note	0,08	7	0,15	7	0,06	9	0,28	6	0,54	7	0,09	9	0,70	7	0,459	6
Short Survey			0,04	8	0,08	8	0,19	7	0,56	6	0,31	7	0,29	8	0,315	7
Book	0,64	4	0,04	8	0,29	7			0,08	10	0,21	8	0,21	10	0,174	8
Editorial							0,12	9	0,11	9	0,05	10	0,26	9	0,153	9
Erratum							0,01	12	0,04	12	0,05	10	0,09	11	0,058	10
Letter			0,04	8	0,02	10	0,01	12	0,01	13	0,01	12	0,08	12	0,044	11
Report	0,16	6	0,19	6	0,38	6	0,15	8							0,041	12
Incollection					0,02	10	0,04	11	0,06	11	0,02	11	0,04	14	0,037	13
Article in Press													0,05	13	0,023	14
Book Chapter			0,15	7	0,06	9	0,05	10	0,51	8	1,11	5	1,79	4	0,011	15
Business Article									0,01	13					0,003	16
Abstract Report													0,00	15	0,001	17

Primary fields of scientific publications on 3D printer

In this section, we present two main analyzes, keyword co-occurrence and trends of areas regarding journals and annals, in order to find out the most prolific areas in 3D printing publications which will help revealing aspects of the domain identity.

From Table 2, it can be observed that 77.5% of documents are classified under the typology articles or conferences. In order to qualify the scope of these documents, we present in table 3 the 15 top journals and 15 top annals of conferences with the highest number per period. For this study, due to the space limitation, we only present the data referring to four quinquennia.

The percentage that these 15 journals and conferences represent in relation to the total in the period is given by the sum of the percentages in the period removed the fields without filling. This value is presented in the last line of each quinquennium, whose result for journals varies from 11.68% to 25.34%, while for conferences the value varies from 25.48% to 37.74%.

It is observed that journals from engineering remains in all quinquennia, whereas journals from physics in the first quinquennia only, journals from computer science in the first two and journals from materials science in the second and third. In the last period, for the first time, journals from health area appear among the top 15 journals, including Tissue Engineering, Biofabrication and Lab on the chip.

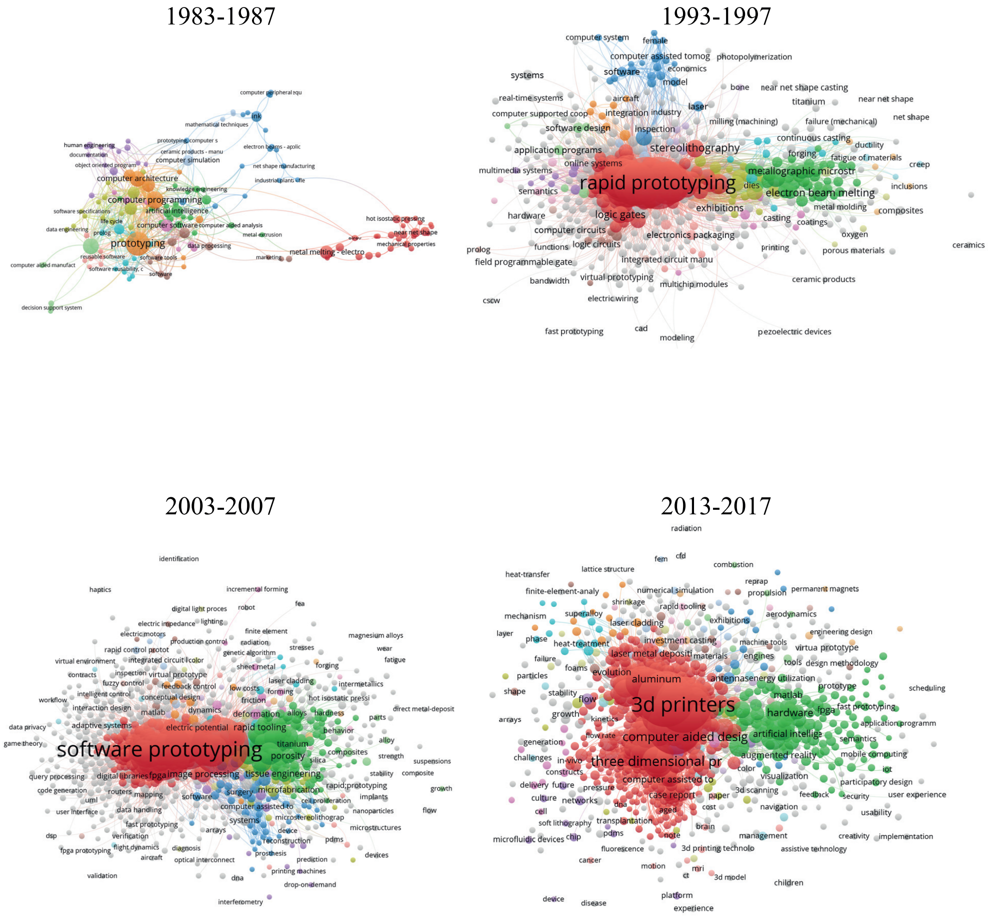
For conference proceedings, annals from computer science and engineering appear in all quinquennia, while annals from materials science in the first, third and fourth and from physics only in the second. In this type of publications, we do not observe the presence of annals from other areas, probably because of the nature of this type of communication, which, as it is well known, has much prestige in areas with a more technological approach and also in some fields of exact sciences and mathematics.

Table 3. Top 15 journals and Conference proceedings with the largest number of articles and conference papers retrieved from WoS and Scopus on 3D printing per quinquennia.

Articles			*Conference*		
1983 - 1987	*% of 663*		*1983 - 1987*	*% of 542*	
J. Of Metals	3,47%	1	Lecture Notes In Computer Science (+Lecture Notes In Artif. Intellig. And Lecture Notes In Bioinf.)	3,32%	1
Acm Sigplan Notices	2,56%	2	Proc.Of The Hawaii International Conference On System Science	2,77%	2
Metal Powder Report	2,41%	3	Sae Technical Papers	2,58%	3
Ibm Technical Disclosure Bulletin	2,26%	4	Proc.- Ieee Computer Society'S International Computer Software & Applications Conference	2,40%	4
Ieee Transactions On Software Eng.	1,96%	5	Progress In Powder Metallurgy	2,03%	5
Information And Software Technology	1,96%	6	Digest Of Technical Papers - Sid International Symposium (Society For Information Display)	1,85%	6
Proc. Of Spie - The Intern. Society For Optical Eng.	1,66%	7	Iee Colloquium (Digest)	1,48%	7
Datamation	1,51%	8	Proc.- International Conference On Software Engineering	1,48%	7
Industrial Heating	1,36%	9	Conference Record - Electro	1,29%	8
Ibm J. Of Res. And Development	1,21%	10	Ieee Proc.Of The National Aerospace And Electronics Conference	1,29%	8
Russian Metallurgy. Metally	1,21%	10	Nasa Conference Publication	1,11%	9
Proc. Of The Sid	1,06%	11	Simulation Series	1,11%	9
Communications Of The Acm	0,90%	12	Afips Conference Proceedings	0,92%	10
Information And Management	0,90%	12	Materials Research Society Symposia Proceedings	0,92%	10
JOM	0,90%	12	Modern Developments In Powder Metallurgy	0,92%	10
Unfilled Filds	1,21%		Unfilled Filds	43,54%	
Sum	25,34%		Sum	25,46%	
1993 - 1997	*% of 3023*		*1993 - 1997*	*% of 2150*	
J. Of Mat. Proces. Technology	1,89%	1	Proc.Of Spie - The International Society For Optical Engineering	8,37%	1
Rapid Prototyping J.	1,69%	2	Lecture Notes In Computer Science (Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics)	7,30%	2
Communications Of The Acm	0,76%	3	Conference On Human Factors In Computing Systems - Proceedings	4,00%	3
Machine Design	0,69%	4	Sae Technical Papers	3,72%	4
Manufacturing Eng.	0,69%	5	Proc.Of The International Workshop On Rapid System Prototyping	2,98%	5
Computer	0,66%	6	Iee Colloquium (Digest)	2,00%	6
Stahl Und Eisen	0,66%	6	Advances In Powder Metallurgy And Particulate Materials	1,95%	7
Assembly Automation	0,63%	7	Proc.Of The Ieee International Conference On Systems, Man And Cybernetics	1,53%	8
Information And Software Technology	0,63%	7	Asee Annual Conference Proceedings	1,16%	9
Ceramic Eng. And Science Proc.	0,60%	8	Icassp, Ieee International Conference On Acoustics, Speech And Signal Processing - Proceedings	1,07%	10
Cirp Annals - Manufacturing Technology	0,56%	9	Lecture Notes In Computer Science (Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics)	1,07%	10
Computers In Industry	0,56%	9	Annual Technical Conference - Antec, Conference Proceedings	0,93%	11
J. Of Mat. Proces. Tech.	0,56%	9	Tms Annual Meeting	0,93%	11
Mat. Science And Eng. A	0,56%	9	Proc.- Ieee International Conference On Robotics And Automation	0,88%	12
Computer-Aided Eng.	0,53%	10	Proc.- International Conference On Software Engineering	0,84%	13
Unfilled Filds	0,20%		Unfilled Filds	0,84%	
Sum	11,68%		Sum	38,74%	
2003 - 2007	*% of 7339*		*2003 - 2007*	*% of 9681*	
Rapid Prototyping J.	2,33%	1	Lecture Notes In Computer Science (+ Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics)	7,61%	1
J. Of Mat. Proces. Technology	2,18%	2	Proc.Of Spie - The International Society For Optical Engineering	5,29%	2
Intern. J. Of Advanced Manufacturing Technology	1,58%	3	Virtual Modelling And Rapid Manufacturing - Advanced Research In Virtual And Rapid Prototyping	1,93%	3
Lecture Notes In Computer Science (+ Subseries In Artificial Intelligence And In Bioinformatics)	1,50%	4	Sae Technical Papers	1,81%	4
Jisuanji Jicheng Zhizao Xitong/Computer Integrated Manufact. Systems	1,02%	5	Proc.Of The 3Rd International Conf. On Advanced Research In Virtual And Rapid Prototyping: Virtual And Rapid Man. Adv. Res.Virtual And Rapid Prototyping	1,29%	5
Jixie Gongcheng Xuebao/Chinese J. Of Mechanical Eng.	0,82%	6	Materials Science Forum	1,28%	6
Proc. Of The Institution Of Mechanical Engineers Part B-J. Of Eng. Manufacture	0,79%	7	Acm International Conference Proceeding Series	1,14%	7
Xitong Fangzhen Xuebao / J. Of System Simulation	0,75%	8	Lecture Notes In Computer Science	0,97%	8
Mat. Science And Eng. A	0,68%	9	Proc.Of The Asme Design Engineering Technical Conference	0,97%	8
Mat. Science And Eng. A-Struct. Mat. Prop. Microstructure And Proc.	0,67%	10	Proc.Of The International Workshop On Rapid System Prototyping	0,93%	9
Zhongguo Jixie Gongcheng/China Mechanical Eng.	0,67%	10	Conference On Human Factors In Computing Systems - Proceedings	0,81%	10
Nuclear Inst. And Methods In Physics Res., Section A: Accelerators, Spectrom., Detectors And Associated Equipment	0,63%	11	Asee Annual Conference And Exposition, Conference Proceedings	0,66%	11
Intern. J. Of Machine Tools \& Manufacture	0,57%	12	Key Engineering Materials	0,64%	12
Intern. J. Of Machine Tools And Manufacture	0,57%	12	European Space Agency, (Special Publication) Esa Sp	0,56%	13
Advanced Mat. And Processes	0,52%	13	Proc.- Ieee International Conference On Robotics And Automation	0,54%	14
Unfilled Filds	0,00%		Unfilled Filds	0,00%	
Sum	15,27%		Sum	26,42%	
2013 - 2017	*% of 25353*		*2013 - 2017*	*% of 14629*	
Hongwai Yu Jiguang Gongcheng/Infrared And Laser Eng.	4,10%	1	Proc.Of Spie - The International Society For Optical Engineering	4,29%	1
Rapid Prototyping J.	1,43%	2	Lecture Notes In Computer Science (Including Subseries Lecture Notes In Artificial Intellig. And Lecture Notes In Bioinf.)	3,24%	2
Internat. J. Of Advanced Manufacturing Technology	1,37%	3	Applied Mechanics And Materials	1,85%	3
Mat. \& Design	1,07%	4	Acm International Conference Proceeding Series	1,68%	4
Tissue Eng. Part A	0,97%	5	Procedia Cirp	1,65%	5
Mat. And Design	0,96%	6	Proc.Of The Asme Design Engineering Technical Conference	1,59%	6
Scientific Reports	0,88%	7	Asme International Mechanical Engineering Congress And Exposition, Proc.(Imece)	1,52%	7
Additive Manufacturing	0,88%	8	Proc.Of The International Conference On Progress In Additive Manufacturing	1,44%	8
Biofabrication	0,81%	9	Advanced Materials Research	1,41%	9
Abstracts Of Papers Of The American Chemical Society	0,74%	10	Asee Annual Conference And Exposition, Conference Proceedings	1,41%	9
Mat. Science And Eng. A-Struc. Mat. Prop. Microstruc. And Proces.	0,67%	11	Aip Conference Proceedings	1,36%	10
J. Of Mat. Proces. Technology	0,62%	12	Conference On Human Factors In Computing Systems - Proceedings	1,34%	11
Mat. Science And Eng. A	0,56%	13	Iop Conference Series: Materials Science And Engineering	1,20%	12
Lab On A Chip	0,54%	14	Procedia Engineering	1,05%	13
Plos One	0,54%	14	Key Engineering Materials	0,98%	14
Unfilled Filds	0,00%		Unfilled Filds	0,00%	
Sum	16,14%		Sum	26,01%	

In order to better characterize the articles on 3D printing in terms of their priority areas, we performed a co-occurrence analysis based in all keywords presented in the 119,800 documents.

In Figure 4, it is possible to observe that the density of word networks increases from the first period to the last, as a result of the increase in the number of keywords with at least three occurrences (from 86 to 999) and the increase of connected items (from 151 to 989). It is also observed that the number of clusters grew very intensely up to the third quinquennium (from 12 to 225), but was reduced to 99 in the last period.



1983-1987

1993-1997

2003-2007

2013-2017

Figure 4. Keyword network of documents retrieved from WoS and Scopus on 3D printing per quinquennia.

The period 1983-1987 has 12 clusters; the largest cluster (red) displays 28 connected items that are clearly related to mechanical engineering. This cluster is shifted from the center of the network and is weakly connected to it. In the orange cluster, the "prototyping" node is the most co-occurring one within the network, a concept that gains prominence over time. If we isolate this node, we see a large number of connections between keywords related to computation, a central characteristic of the 3D printing technique, since it is necessary to use software for generation images and planning layers of decomposition of the material.

From period 1993-1997, the largest cluster (in red) assumes the central position, having the smallest distance between all elements of the network. In this quinquennium, there are 176 clusters, and the bulkiest cluster (red) has 109 items. Unlike the previous map, here the item of greater co-occurrence is "rapid prototyping". There appears to be an explosion of prototyping studies, printing techniques and also about computational part for designing the layers, which are now in the same cluster. It seems that the blue and red clusters of the previous period were combined and formed the green cluster, the second largest in this period. The composition of this cluster, points to issues related to engineering, physical and material science. The blue cluster refers to imaging techniques, and the element that links it to the central cluster is image processing.

The map of period 2003-2007 has 225 clusters, with a substantial number of small clusters. The largest cluster (red) maintains the characteristics of the previous period, being related to prototyping and computation. The cluster in green, the second largest in this period, has words related to 3D printing techniques such as stereolithography and lithography, biomaterials and tissue engineering, indicating an initial approximation with the area of health science.

Finally, the period 2013-2017 displays 99 clusters with 989 elements connected, the largest cluster (red) has 474 elements, which are mostly synonyms of the expression 3D printing. In this period, it appears, for the first time, some words as bioprint, bioplotter and bioink in the largest cluster. These new terms indicate a true approximation of 3D printing domain by the area of health science.

Conclusions

The 3D printing is an emerging technology with great potential to impact economy as well as to transform the routine of various social sectors. There is a high expectation that it will introduce many changes in society, including in the productive chain. In health, for example, the impression of organs may cause a revolution by impacting the productive chain of drugs, since the traditional treatment can be substituted by the implantation of a new organ printed in the 3D printer. It may also impact several scientific areas, once the printing of organs to treat diseases, research project related to these diseases may no longer be necessary or priority.

As we did not find studies in the field of Scientometric or Bibliometric regarding 3D printing, we decided to carry on this present study on the mapping of the most prevalent research areas at this domain. The design of this study may be considered extensive, since the volume of analyzed data included documents retrieved from a strategy search that used 44 terms. This strategy allows the development of future studies, with specific cuts to some of the topics identified, for example, aerospace identified in the conferences from 2003 to 2007, or health sciences evidenced in the period 2013-2017 and in the co-occurrence of keywords.

As main results, we noted a remarkable growth of 3D printing publications, which are mainly published in types of documents, articles and annals, which are totally compatible with the demand of this subject towards a fast flow of communication. As for the countries that have contributed more with this research domain, as expected, the United States appears in a prominent position, a result not only of its large tradition in research but also in patenting most of the 3D printing technologies. One positive and not expected finding was the contribution of Brazil, a developing and peripheral country in science with little tradition in technological development.

Regarding the areas, the data show that engineering, computer science and material science carry out most of the research on this technology, while health sciences has emerged in the last analyzed periods. Indeed, it was expected a closer relation between 3D printing and the areas of engineering, computer science and material science, since these are areas in which the technology is grounded. In other words: 3D printing is a equipment (engineering) that requires sophisticated software (computer science) and different inputs (materials science).

Due to the nature of the analyses presented in this study, we identified only the most prominent areas, but some other areas (with lower number of documents) may play significant roles for the establishment of 3D printing domain and must be understood through specific cuts. As a future perspective, we intend to look closer at some engineering specialties, in order to highlight other areas that may be masked by the large volume of material analyzed in this work. Another possibility is to investigate to the most prolific authors as well as their main works in order to identify where they are positioned in the history of the 3D printing domain development.

We believe that the enlargement of the analyses will make it possible to get a more comprehensive framework on the main areas, institutions and actors of 3D printing domain.

Acknowledgment

The authors are grateful to Pedro Accioly, founder of Maker Factory, for listing the significant events that helped increasing the use of 3D printing and to Gonzaga Sillas for aiding the R code that unified the two databases. Andreia Cristina Galina receives a PhD fellowship from Capes.

References

Aria, M. & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959-975.

Berman, B. (2012). 3-D printing: The new industrial revolution. *Business horizons*, *55*(2), 155-162.

Day, G. S., Schoemaker, P. J., & Gunther, R. E. (2009). *Gestão de Tecnologias Emergentes: A visão de Wharton School*. Bookman Editora.

Garrett, B. (2014). 3D printing: new economic paradigms and strategic shifts. *Global Policy*, *5*(1), 70-75.

Gebler, M., Uiterkamp, A. J. S., & Visser, C. (2014). A global sustainability perspective on 3D printing technologies. *Energy Policy*, *74*, 158-167.

Hjørland, B. & Albrechtsen, H (1995). Toward a new horizon in information science: Domain‐analysis. *Journal of the American society for information science* 46 (6), 400-425.

Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., & Marrs, A. (2013). *Disruptive technologies: Advances that will transform life, business, and the global economy* (Vol. 180). San Francisco, CA: McKinsey Global Institute.

Michael, S., Sorg, H., Peck, C. T., Koch, L., Deiwick, A., Chichkov, B., ... & Reimers, K. (2013). Tissue engineered skin substitutes created by laser-assisted bioprinting form skin-like structures in the dorsal skin fold chamber in mice. *PloS one*, *8*(3), e57741.

Murphy, S. V., & Atala, A. (2014). 3D bioprinting of tissues and organs. *Nature biotechnology*, *32*(8), 773.

Prince, J. D. (2014). 3D printing: an industrial revolution. *Journal of electronic resources in medical libraries*, *11*(1), 39-45.

Standard, A.S.T.M.(2012). ISO/ASTM 52900: 2015 Additive manufacturing-General principles-terminology. *ASTM F2792-10e1*.

Rifkin, J. (2012). The third industrial revolution: How the internet, green electricity, and 3-d printing are ushering in a sustainable era of distributed capitalism. *World Financial Review*, *1*(1), 4052-4057.

Team, R. C. (2018). R: a language and environment for statistical computing. The R project for statistical computing. Vienna, Austria.

van Eck, N., & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

Mapping the translational process of Her-2 studies with the pioneer's publications

Yuxian Liu[1], Ewelina Biskup[2] Yueqian Wang[3], Fengfeng Cai[4] and Xiaoyan Zhang[5]

[1]*yxliu@tongji.edu.cn*
Tongji University, Tongji University Library, Siping Road 1239, 200092 Shanghai (China)
School of History and Archives, Yunnan University, 650091 Kunming (China)

[2] *ewelinabiskup@yahoo.de*
University Hospital of Basel, Department of Internal Medicine, Petersgraben 4, 4051 Basel (Switzerland)
Shanghai University of Medicine and Health Sciences, Zhouzhu Hwy 279, 201318 Shanghai (China)

[3]*1532757@tongji.edu.cn*
Tongji University, School of Life science and Technology, Siping Road 1239, 200092 Shanghai (China)

[4]*caifengfeng@tongji.edu.cn*
Tongji University, Yangpu Hospital, Department of Breast Surgery, Tengyue Road 450, 200090 Shanghai (China)

[5]*xyzhang@tongji.edu.cn*
Tongji University, School of Life Sciences and Technology, Shanghai East Hospital, Research Center for Translational Medicine, Siping Road 1239, 200092 Shanghai (China)

Abstract

Translational research (TR) aims to "translate" findings in fundamental research into the clinic, with the ultimate goal of better health outcomes. The term however is ambiguous and qualitative control tools have only been developed recently. They include identification and description of TR stages, as well as evolving from one stage to another. A proper evaluation of the TR in Her2 domain has not been performed thus far. Given the significance of the Her2 research and its implications into guidelines and pharmaceutical policies, we conducted a concise clustering of the pioneering publications of Dennis Slamon. Moreover we analyzed the keyword co-occurrence of the network extracted by the publications. By tracking the linkage strength within and between clusters, their dynamics and connected terms, we show how different TR stages occur and how they relate to each other. Our comprehensive mapping of the translational process of Her-2 research has implications for policy definition and implementation, as well as future TR evaluation framework

Introduction

Molas-Gallart et al. proposed a new framework for the evaluation of translational research (TR), suggesting to analyze how one TR stage is translated to another. In a medical context, TR aims to bring findings of fundamental research into medical and nursing practice, as well as needs from the clinic to the laboratory.

The application of this framework faces severe challenges, such as how to mark different phases of the translational process and how to chart the pathways along these phases. Small (2000) captioned clusters gathered together by citation links in a specific research area, then charted pathways between these clusters so that the discovery process could be logically followed. His approach was retrograde from existing discoveries back to the core, identifying paths that connect the clusters earlier. Cambrosio et al. (2006) mapped the emergence and development of translational cancer research. They found that journals or concepts on the same research levels such as basic research or clinical practice are more likely to cluster together via inter-citation networks and semantic network maps. These maps displayed links between concepts that co-occurred within the title and the abstract of a given article. Can the clusters of publications on a specific topic reflect different TR phases?

Her-2 is a gene located at the long arm of human chromosome 17. It was identified in the mid-1970s by transfection studies with DNA from chemically induced rat neuroglioblastomas by Shih, Padhy, Murray and Weinberg. Soon, the gene was found to be pro-oncogenic. Amplification or over-expression of the Her-2 gene occurs in approximately 15-30% of breast cancers, which thus were labeled "Her-2 positive" and was the most aggressive type before Herceptin was developed, approved by the FDA in 1998 and been used in the clinical practice ever since. The invention of this effective substance was based on what was called TR and represents one of the most prominent examples of bench to bed translations with an immense prognostic improvement for targeted patients.

In this investigation study, we use the concept of terms' co-occurrence in the publications on Her-2, an acronym for human epidermal (growth factor) receptor 2. Our goal was to using scientometrics tool to map the translational process of Her2 research. Some scientists compare the TR to crossing a valley of death, underlining the significant gap between basic science and clinical reality, as well as burden and barriers along the translation process, such as policies. Her2 research is of an uttermost significance, being implemented constantly into the clinic: patients' prognostic screening, therapy choice, pharmaceutical developments etc. we clustered the publications of the domain's pioneer, Dennis Slamon, and created co-occurrence keyword networks with cross-link comparison to reenact the translational processes, from the lab, through the valley of death, to the successful implementation with survival benefits.

Methods

With bibliometric data, we focused on and term co-occurrence network to see the map and the clusters of Her-2 research. We further do link analysis on the network. With all these methods, we will map the translational process of Her-2 research.

Mapping and clustering in a term co-occurrence network

We conducted mapping and clustering of terms in co-occurrence network of Her2-research. Terms in each article represented the concepts the article described. Two terms were considered as linked if they both occurred on the same article. The terms and their co-occurrence links constituted a term co-occurrence network, used to provide a graphic visualization of potential relationships between concepts represented within the collection of articles on a specific topic. Mapping and clustering techniques are frequently used to study such networks. Mapping is the creation of maps. We created a graphic, symbolic representation of significant features, using the distance between two items to reflect the strength of their relation. A smaller distance indicated a stronger relation. Clustering assigned a set of elements into clusters of similar elements, with a high density of within-group links and a lower density of between-group links. The stronger the link strength between two terms, the more likely was the categorization into the same cluster. Mapping and clustering techniques were based on similar principles to enhance the analysis transparency and to avoid unnecessary technical complexity and inconsistencies between the results produced.

Link analysis

We used link analysis - a data-analysis technique used to evaluate relationships between terms and clusters. How links occur within and between clusters and which terms are connected by these links, are the key factors to understand the relationship of the clusters and properties of a cluster.

Result

Mapping and Clustering Slamon's publications

Dennis Slamon and his team found a correlation between relapse and survival with amplification of the Her-2 oncogene, which became a significant prognostic factor. Since Her-2 is an oncogene that controls the growth of cancer cells themselves, Slamon tried to find a protein, which binds to the Her-2 and prevents from relaying a signal for cancer cells stimulation.

The Web of Science contains 410 publications authored or co-authored by Slamon. We use VOSviewer to map and cluster Slamon's publications (Figure 1).

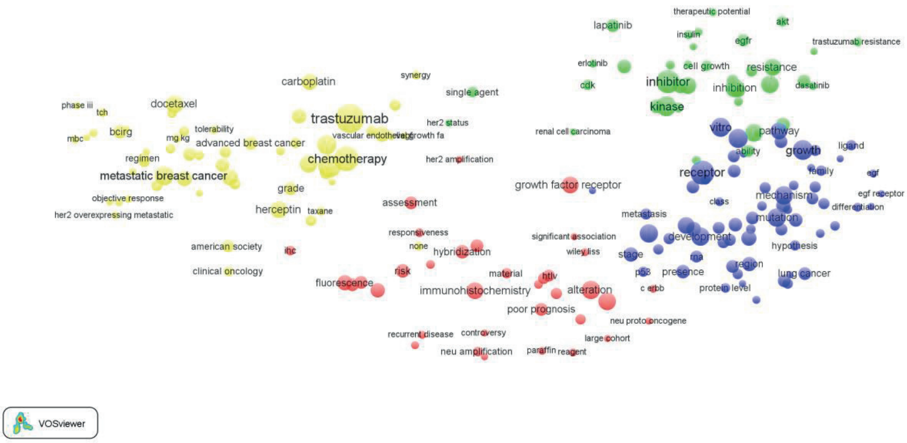


Fig 1. The clusters of Slamon's publications drawn by VOSviewer

The dots in the map represent terms drawn from the Slamon's publications. Different colours represents different clusters. So the landscape drawn by Slamon's publications has four clusters: The clusters of the green and the blue are on the right side of the landscapes, whose topics refer to the basic biological properties of the Her-2 oncogene, thus to fundamental research; The cluster of the yellow and red are on the left side, whose topics concern clinical practice. We name the clusters according to the terms in each clusters as showed in table 1.

Table 1. The summary of the four clusters's characteristics.

Colors	*clusters' name*	*Subclusters' name*
Green	Biological inhibiting mechanism	Her-2’ s biological property Development of cancer
Blue	Biological correlation mechanism	
Yellow	Clinical & pharmacology	
Red	Detection & diagnosis	

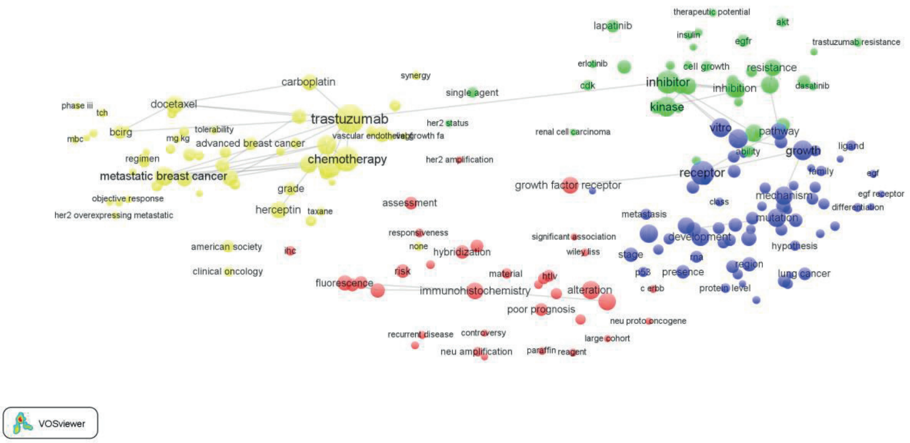
In the knowledge landscape drawn by Slamon's publications, the blue cluster emphasizes the biological mechanism of revealing the relation between biological properties of the Her-2 oncogene and the development of breast cancer. Hence the cluster has two parts: the basic biological property of oncogenes Her-2, which is at large near the green cluster; and the relation of these biological properties with the development of cancer, which is at large near the red cluster. However, the two parts are not completely separated. The biological mechanism constitutes the foundation of monoclonal antibodies therapy, in which Slamon mainly engaged. So we name this cluster biological correlation mechanism.

The links analysis for the landscape drawn by Slamon's publications

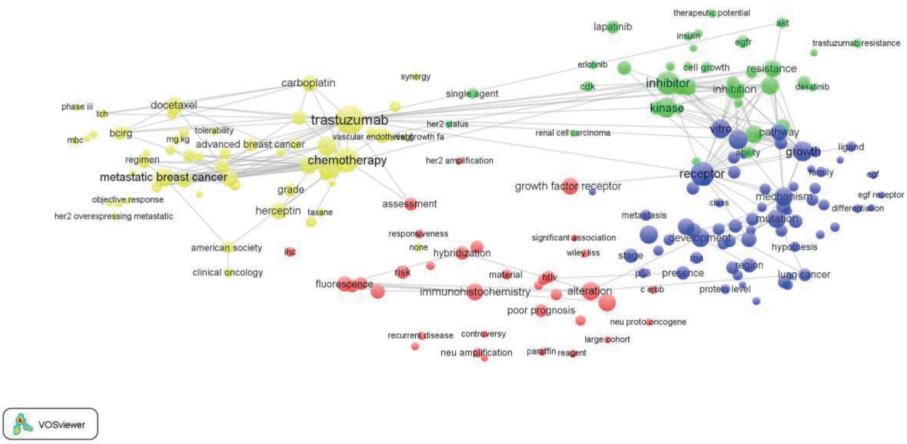
Link strength is the key factor to understand the formation of clusters in a map and the relation between clusters. We ranked the links by their linkage strength from strongest to weakest and studied them continuously increasing the groups (going further in the ranks).

A translation process does not occur till the discovery of a relation is found to be reliable. The reliable results can be verified and repeated. Repetition will make the linkage strength stronger and stronger.

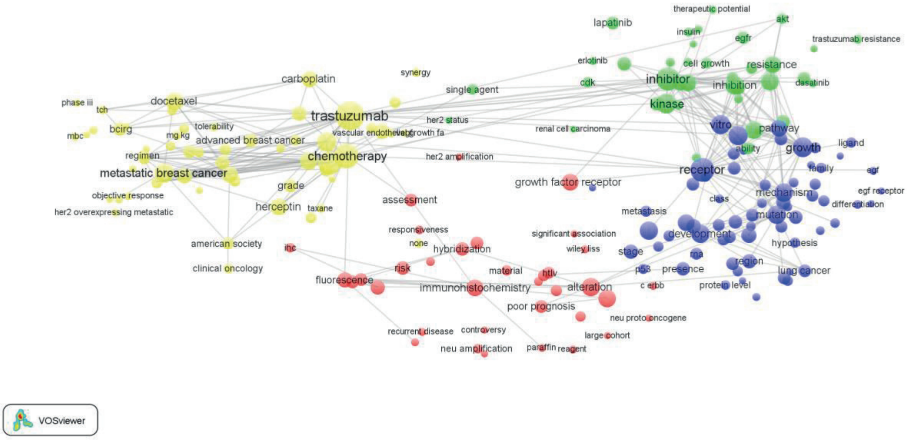
Figure 2 shows how the first 50, 200, 300 strongest links appear in the network.



(a)



(b)



(c)

Figure 2. Slamon's publication keyword co-occurrence network showing the 50(a), 200(b), 300(c) strongest links

The links within clusters

We noticed that indeed the strong links are more likely to occur within clusters: In Figure 2a, 23 out of the 50 strongest links occur within the clinical and pharmacology cluster (yellow). The words linked together by the strongest links were *trastuzumab, chemotherapy, carboplatin, cyclophosphamide, docetaxel, bcirg, dose, safety, efficacy* and *metastatic breast cancer (MBC)*, indicating that Slamon performed many trials to study the efficacy of the therapeutic regimes of trastuzumab with different chemotherapy antitumor agents.

Within the biological inhibiting mechanism cluster (green), two centers are formed by the first 50 strongest links. One is the term *inhibitor (inhibition)*, and the other *activation*. Inhibitor links to words such as *proliferation, apoptosis, panel, kinase* (which is an enzyme that catalyzes the transfer of phosphate groups from high-energy, phosphate-donating molecules to specific substrates, based on which the Her-2 proliferation can be inhibited). The term *activation* links to *resistance* and *kinase*. It further links to the word *pathway* in the blue cluster. *Inhibition* and *activation* describe reverse processes. Interestingly, though the two centers do not link directly (two centers are linked via the word *kinase*), each of them links to a word that is opposite to itself, hence to words that are similar to the other. This indicates Slamon's effort to find the apoptosis mechanism of Her-2's proliferation by activating or resisting some biological substances via kinase.

Only a few of these 50 strongest links appear in the biological correlation mechanism cluster (blue). Two parts of this cluster do not link together by the strongest links: in the part of basic biological property of oncogenes Her-2, the term *growth* are linked with the terms of *receptor*, *model* and *mechanism*. There are three links between the terms *receptor* and *pathway*, the terms *model* and *vitro*, the terms *loss* and *mutation* respectively in this part. In the part of the relation of these biological properties with the development of cancer, only one link connect the terms, *nsclc* and *non small cell lung cancer* (nsclc). These strongest links in this cluster do not make a lot of sense: *growth* and *receptor*, are the words in the phrase of *growth factor receptor*, *nsclc* is the acronym for *non small cell lung cancer*. Two parts in the biological correlation mechanism cluster (blue) remain separated even if the number of links increases to 100 (Figure 2b). However, we notice that from the link from the term *mutation* in the middle of the cluster stretches down to the term *loss*, which is tangled together with the part of the relation of these biological properties with the development of cancer. We know that loss or mutation of tumour suppressor gene is one of the reasons that lead to tumorigenesis. Two parts in the biological correlation mechanism were not connected when the number of links increase to 100. However, when it increase to 110, two important links from the term *mutation* to the phase *non small cell lung cancer*, and the term *mechanism* to the term *development* appear. As the number of links increase, five radiate centers, the terms of *receptor, growth, mechanism, mutation and development*, are formed.

In the detection and diagnosis cluster (red), only the terms describing the methods of detecting Her-2 status such as the *immunohistochemistry* and *FISH (fluorescence in situ hybridization)* are connected by some of the first 50 strongest links. The term *growth factor receptor* and *assessment* are in the middle of the map, but, the term *assessment* does not link to other words that describe Her-2 status' detection skills until in Figure 2(c), with the first 300 strongest links, the links from *assessment* to *immunohistochemistry* and from *immunohistochemistry* to *paraffin* appear within this cluster. G*rowth factor receptor* has no connection with the other terms in this cluster till the number of links increases to 600 via a link to *immunohistochemistry*. Weaker links provide cohesion to this cluster.

The links between clusters

Only six out of the fifty strongest links appeared between clusters (Figure 2a). Five of them stretch out from the blue cluster. Four of them appear between the biological inhibiting mechanism (green) and the biological correlation mechanism cluster (blue). These two clusters have higher link strength than others. This also explains why these two clusters are mapped closer to each other. Another one of the strongest links between clusters appears in the biological correlation mechanism cluster (blue) and the detection and diagnosis cluster (red) via a line between *receptor* and *growth factor receptor*. This link represent the bridge between the part of Her-2's biological property and biological inhibiting mechanism. Another important link between these two clusters is that from the term *development* to the term *alteration*, which appears when the number of links increase to 150, representing the bridge between the part of development of cancer and the biological inhibiting mechanism.

One of the six strongest links between clusters appears in the biological inhibiting mechanism cluster (green) and the clinical and pharmacology cluster (yellow) via a link between *trastuzumab* and *inhibitor*, which points to therapeutic principle used in clinical practice: trastuzumab is used as an inhibitor of the growth of the Her-2 oncogene.

These six links connect four clusters. The clinical and pharmacology cluster (yellow) is not connected to the detection and diagnosis cluster (red) (Figure 2a) despite their proximity on the map. The clinical & pharmacology cluster (yellow) links to the detection and diagnosis cluster (red) via links from *chemotherapy* and *trastuzumab* to assessment, when the number of strongest links increases to 200 (Figure 2b). In Figure 2c with the first 300 strongest links, an important link, from *trastuzumab* to *fluorescence*, is added between the clinical and pharmacology cluster (yellow) and the detection and diagnosis cluster (red). This link goes further to *neu gene amplification* in the detection and diagnosis cluster. The new links imply two methods to detect the status of Her-2: one being immunohistochemistry, which is based on the protein expression of Her-2 oncogene, and one being the gold standard, FISH (Fluorescence in Situ Hybridization), based on the Her-2 gene. This suggest that the detection of Her-2 status is crucial to decide on whether the drug trastuzumab should be applied or not.

The clusters at the end of diagonal lines, the clinical and pharmacology cluster (yellow) versus the biological correlation mechanism cluster (blue) and the biological inhibiting mechanism cluster (green) versus the detection and diagnosis cluster (red), are not linked to each other by the first 50 strongest links in Figure 2a. The clinical and pharmacology cluster (yellow), is linked to the biological correlation mechanism cluster (blue), via the term *trastuzumab* to the term *vitro* (when the number of links is smaller than 100), *model* and *receptor* (when the number of links is larger than 150 but smaller than 200). The red cluster, via the term *growth factor receptor*, is connected to the term *inhibitor* in the biological inhibiting mechanism cluster (green) by one of the 250 strongest links.

Discussion

In order to analyze how one TR stage is translated to another during bringing findings of fundamental research into medical and nursing practice, or, vice versa, we need to know how to mark different phases of the translational process and how to chart the pathways along these phases.

Clusters and the phases of Translation research

Her2 gene and its implication in the breast cancer treatment is one of the most representative examples of a knowledge transfer from the bench to the clinic and back. Since the Her-2 oncogene was identified, scientists have begun to study its biological properties. Using the knowledge of medicine-based molecular biology, they tried to discover the biological mechanism of linking the Her-2 protein with cell-growth of breast tissue. The mechanism is called biological inhibiting mechanism. Through this mechanism, scientist find the her-2 gene can stimulate human's own immune system to attack cancer cells or by giving immune system components, such as man-made immune system proteins (American Cancer Society, 2016) to cure the cancer. Since then, scientist began to study Her-2's biological process for immunotherapy. Based on these two mechanisms, Herceptin was created to break this link. When this succeeded in animal experiments and clinical trials, the drug was approved by the FDA and entered into the clinical practice. Hence, her-2 research come to a stage of clinical & pharmacology. For deciding on whether the drug can be used in patients depends on patients' Her-2 status, different methods were invented to detect Her-2 status. Based on detection, different therapeutic regimes were used to treat breast cancer in clinic. Detection and diagnosis are link together in the clinic practice.

However, the blue cluster in Slamon's landscape emphasizes the biological mechanism of revealing the relation between biological properties of the Her-2 oncogene and the development of breast cancer. The biological mechanism constitutes the foundation of monoclonal antibodies therapy, in which Slamon mainly engaged. The drug invented by Slamon, namely trastuzumab, is a monoclonal antibody against the Her-2 protein. It attaches to and blocks antigens on cancer cells so that they stop growing.

Slamon was the first scientist to find a correlation between the Her-2 oncogene and development of breast cancer. He made a thorough investigation of the biological mechanism of revealing the relation between biological properties of the Her-2 oncogene and the development of breast cancer.

So, we may say that the clusters of Her-2 publications exactly represent the phases through which the research topic went. The methodology of clusters and links between them allows identifying the relations between the topics. The maps then visualize the pathways connecting the clusters.

links and the bridge between the valley of death in the process of translation

As expected, the strongest links occured within the clusters. 23 out of the 50 strongest links were found within the clinical and pharmacology cluster (yellow). The other 27 were in the other three clusters. Term *inhibitor (inhibition)* and the term *activation* are the opposite centers formed by the first 50 strongest links within the biological inhibiting mechanism cluster (green). Only a few of these fifty strongest links appear in the biological correlation mechanism cluster (blue). Two parts of this cluster did not bind together till the strongest links increased to 200. In the detection and diagnosis cluster (red), only the terms describing the methods of detecting Her-2 status such as the *immunohistochemistry* and *FISH (fluorescence in situ hybridization)* were connected by some of the first 50 strongest links. The term growth factor receptor and assessment are in the middle of the map. The term *assessment* did not link to the other words that describe Her-2 status' detection skills until the number of strongest links increase to 300. *Growth factor receptor* had no connection to other terms in this cluster till the number of links increases to 600 via a link to immunohistochemistry.

Only 6 out of the 50 strongest links connect the clusters. Four of them appear between the biological inhibiting mechanism (green) and the biological correlation mechanism cluster (blue). Most of the links ranked by strength ranked between the 51st -100th strongest appear within or between these two clusters, which are also mapped closely. All these findings indicate these two clusters are closely related.

One of the six strongest links were between clusters appears in the biological inhibiting mechanism cluster (green) and the clinical and pharmacology cluster (yellow) via a link between *trastuzumab* and *inhibitor.* Another one of the strongest links between clusters appears in the biological correlation mechanism cluster (blue) and the detection and diagnosis cluster (red) via a line between *receptor* and *growth factor receptor*. This underlines the translational process that lead to the development of this medicine from the basic science profile.

We have also detected weaknesses of the TR processes, such as in the link between the clinical/pharmacology cluster (yellow) and the detection/diagnosis cluster (red). They are connected via links of *chemotherapy*, *trastuzumab* and *assessment* only when we increase the number of strongest links to 200. With the first 300 strongest links, an important link, from *trastuzumab* to *fluorescence*, is added between the clusters, which goes further to *neu gene amplification* in the red cluster. In general, the links between the cluster of the Her-2 status detection (red) and other clusters are the weakest. This proves that the importance of Her-2-detection came up only later in the TR process. The biological inhibiting mechanism cluster mainly deals with how over-expressed Her-2 in genes can be inhibited; the biological correlation mechanism cluster reveals the relation between Her-2 oncogene and the development of breast cancer. Correlation between the Her-2 oncogene and development of cancer inspired scientists to find ways to inhibit Her-2 oncogene as a therapeutic principle. The weakness of the links imply two methods to detect the status of Her-2: one being immunohistochemistry, which is based on the protein expression of Her-2 oncogene, and one being the gold standard, FISH (Fluorescence In Situ Hybridization), based on the Her-2 gene. This suggests that the detection of Her-2 status is crucial to decide on whether the drug trastuzumab should be applied or not. The separation of the links reflects the fact that Her-2 status test was not strictly conducted in the clinical practice at the beginning. The gold standard for Her-2 status detection – FISH - was approved by the FDA in 2002, four years after the drug has been used in clinical practice. In 2007, the Her-2 status test in breast cancer was recommended by ASCO-CAP. In 2013, ASCO-CAP convened an Update Committee that included co-authors of the 2007 guideline to conduct a systematic literature review and update recommendations for optimal HER-2 testing(Wolff, et al.2007,2013). The links between the cluster of the Her-2 status detection (red) and other clusters are also weak. This also indicates that Her-2 status detection was not considered to be important in the beginning. However, nowadays it has become a standard in breast cancer diagnostic and therapy choice.

The clinical and pharmacology cluster (yellow), the biological correlation mechanism cluster (blue), the biological inhibiting mechanism cluster (green) and the detection/diagnosis cluster (red) are not linked to each other by the first 50 strongest links.

The weaker links between the cluster of the Her-2 status detection and diagnosis and between two parts within the clusters of biological correlation mechanism are very important to understand the translation process of Her-2 research. The relations not only influence the usage of the drug trastuzumab, but also indicate the trends of therapeutic development: Herceptin was approved by the FDA only in 1998 for treating metastatic breast cancer; since then, a series of clinical trials evaluating the potential use of Herceptin for the adjuvant treatment of early-stage HER2-positive breast cancer in patients with early-stage HER2-positive, node-positive breast cancer; in 2008, Herceptin was approved as a single agent for the adjuvant approach. Finally in 2010, it was approved to treat patients with HER2-overexpression in metastatic gastric or gastroesophageal junction adenocarcinoma.

Conclusion

Her-2 research is typical translational research. This has an uttermost importance since Her2 research is finding its way to various clinical practices and policies. We have established the research clusters, linkage strengths between them and discovered weaknesses of some connections, which reflect the chronological transition of scientific processes and policies. Further investigation of the relation between the linkage strength and translation process is warranted. Our work is the first to target Her2 research in terms of systemic analysis, research design and concepts. For the first time, as to our knowledge, we establish structured models and processes describing Her 2 translational research, describing the results on the background of the translation into practice. It is important for both practitioners and researchers to have a control tool and be able to follow the translation of evidence-based Her2 guidelines into routine clinical-, community-, and policy-based practice.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (No.81573023 and 71173154). We also thank Tongji University's Funding for attending the international conferences.

References

American Cancer Association (2007). The personal approach. *Triumph Magazine*, (fall-winter), 13-16.

American Cancer Society. (2016). *Cancer immunotherapy*. Retrieved September 4, 2016 from: http://www.cancer.org/acs/groups/cid/documents/webcontent/003013-pdf.pdf

Bazell, R. (2011). *Her-2: The making of herceptin, a revolutionary treatment for breast cancer*. Random House.

Cambrosio, A., Keating, P., Mercier, S., Lewison, G. & Mogoutov, A. (2016). Mapping the emergence and development of translational cancer research. *European Journal of Cancer,* 42, 3140-3148.

Herceptin® (Trastuzumab) Development Timeline. Retrieved September 4, 2016 from: https://www.gene.com/patients/medicines/herceptin

Molas-Gallart, J., D'Este, P., Llopis, O., Rafols, I. (2015). Towards an alternative framework for the evaluation of translational research initiatives. *Research Evaluation,* 25,235-243.

Shih, C., Padhy, L., Murray, M. & Weinberg R. A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature*, 290, 261.

Slamon, D. J., Clark G. M., Wong, S. G., Levin, W. J., Ullrich, A., McGuire, W. L.(1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science,* 235, 177-182.

Small, H. (2000). Charting pathways through science: exploring Garfield's vision of a unified index to science. In *the web of knowledge: A Festschrift in honor of Eugene Garfield*, (pp. 449-473).

Wolff, A. C., Hammond, M. E. H., Hicks, D. G., Dowsett. M., McShane, L. M., Allison, K. H., Allred, D. C., Bartlett, J. M., Bilous, M. & Fitzgibbons, P. (2013). Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *Archives of Pathology and Laboratory Medicine,* 138, 241-256.

Wolff, A. C., Hammond, M. E. H., Schwartz, J. N., Hagerty, K. L., Allred, D. C., Cote, R. J., Dowsett, M., Fitzgibbons, P. L., Hanna, W. M. & Langer, A.(2007). American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Archives of pathology & laboratory medicine*, 131,18-43.

Decreasing the noise of scientific citations in patents to measure knowledge flow

Fangfang Wei[1*], Guijie Zhang[2*], Lin Zhang[3], Yikai Liang[4] and Jianben Wu[5]

[1] *weifftju@163.com*
Business School, University of Jinan, Jinan 250002 (China)
* Corresponding author

[2] *zgjzxmtx@163.com*
School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014 (China)
*Corresponding author

[3] *18622791972@163.com*
School of Management, Harbin Institute of Technology, Harbin 150001 (China)

[4] *yikailiang@qq.com*
School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014 (China)

[5] *hiteriter@gmail.com*
School of Data Science, City University of Hong Kong, Kowloon, Hong Kong 999077 (China)

Abstract

This paper proposes a method of reducing the noise of scientific citations in patents in order to show the real reference relationships and thus ensure the accuracy of the results of the follow-up study. First of all, based on the analysis of the sources of patent citations, and citation purposes as well as roles, this paper establishes that some of the scientific citations have lower similarity with patent documents. This we name citation noise. Then drawing upon the vector space model, this study builds a noise reduction model of scientific patent citations, and calculates the degree of similarity between the patents and their citations in order to filter the citations with low similarity. In the process of text similarity calculations, we consider semantics based on latent semantic indexing to increase the accuracy. The results contribute to accurate analysis of the knowledge flow between scientific research and technical innovation and provide data support for further research.

Introduction

Scientific knowledge is the driving force behind technological development and economic growth (He and Deng, 2007; Klitkou and Gulbrandsen, 2010; Callaert and Grouwels et al., 2011). Understanding the nature of the relationship between scientific research and technological innovation contributes to the guidance of science policy making, the improvement of technological performance and the realization of national prosperity (Guan and He, 2007). Consequently, substantial research has been undertaken to investigate the interaction between science and technology (Li and Chambers et al., 2014; Sung and Wang et al., 2015; Chen, 2017).

In the process of exploring the technological value of scientific research, the common quantitative methods include citation analysis (Hegde and Sampat, 2009; Li and Chambers et al., 2014; Chen, 2017), author-inventor name matches (Breschi and Catalini, 2010; Wang and Guan, 2011), performance of academic inventors (Meyer, 2006; Lissoni, 2010) and author-inventor co-publications (Klitkou and Gulbrandsen, 2010). Among the various citation analysis techniques, science linkage, which refers to the average number of scientific non-patent references (SNPRs) of patent documents, provides a useful perspective and consequently has gained popularity in scientometric literature (Narin and Hamilton et al., 1995; Narin and Hamilton et al., 1997; Huang and Yang et al., 2015; Sung and Wang et al., 2015; Fukuzawa and Ida, 2016). The SNPRs represent the knowledge flows from science to technology or more specifically, indicate the contribution of scientific knowledge to technology advances (Shearer and Lundeberg et al., 1997; Van Looy and Magerman et al., 2007; Criscuolo and Verspagen, 2008; Sung and Wang et al., 2015; Chen, 2017; Ding and Hung et al., 2017). Given the extensive and consistent availability of patent databases, science linkage provides a systematic view and empirical evidence to illustrate the interactions between science and technology (Callaert and Grouwels et al., 2011; Li and Chambers et al., 2014).

Although SNPRs data serve as an effective instrument in studying knowledge flow between science and technology (Hu and Chen et al., 2007; Lo, 2010; Magerman and van Looy et al., 2010; Ding and Hung et al., 2017), there is a problem that should not be ignored: the SNPRs data contain noise and little has been done to filter it (He and Deng, 2007; Li and Chambers et al., 2014). According to US patent laws, the applicant should include the prior art when filing a patent application (Hicks and Breitzman et al., 2000; Tijssen, 2001; Criscuoloa and Verspagen, 2008). However, prior art cited in the patent application document might cause its disapproval if it overthrew the novelty of the application (Jaffe and Trajtenberg et al., 1993). In consideration of the legal consequences, inventors might omit relevant information strategically to gain an economic interest (Alcácer and Gittelman et al., 2009; Lampe, 2012; Li and Chambers et al., 2014), thus some of the prior art cannot indicate science linkage objectively. Accordingly, some scholars suggested that a future study should take into account the intrinsic limitation of the patent citations and further develop interpretation of the data (e.g., He and Deng, 2007; Jaffe et al., 2000).

US patent law requires both the applicant and the patent examiner to provide the related prior art during the patent application and examination period (Chen, 2017). These references serve different purposes, respectively. The applicant references are for the purpose of demonstrating prior art for the invention generally, while examiner references are for restricting the patent claims (Azagra-Caro and Mattsson et al., 2011). Based on the analysis of the citation motivations of quoters, Li et al. (2014) contend that the noise of SNPRs concentrates on the non-self-citation by inventor/applicant, and thus this part of references cannot indicate science linkage objectively. In contrast, the self-citation by inventor/applicant and examiner citations can disclose prior art and describe technological background more accurately and comprehensively, and consequently are effective to measure science linkage (Li and Chambers et al., 2014; Chen, 2017). These authors shed light on the noise of SNPR data, which is helpful for us better to understand the knowledge diffusion from science literature to patent technology. In fact, patent citation behavior is extremely complex and it is unbelievable that all non-self-citation by inventor/applicant are strategic citations and completely useless for us to study the science linkage (Alcácer and Gittelman et al., 2009; Azagra-Caro and Mattsson et al., 2011; Li and Chambers et al., 2014).

In the era of data explosion, the development of science and technology is becoming more data-driven and high-quality databases are particularly essential to achieve research breakthroughs (Yu and Ding et al., 2015). As the SNPR data accompanied by an amount of noise and overload of irrelevant information which will bring negative influences to the results, simply applying SNPR data to indicate knowledge diffusion is illogical and unreasonable (Li and Meng, 2010; Wang and Yu et al., 2012). Under these considerations, we employ the vector space model (VSM) (Salton and Wong et al., 1975), which is an important, mature and popular text similarity method of filtering the noise of SNPR, reducing the negative influences and achieving a better analysis of patent citations (Ahlgren and Colliander, 2009; Magerman and van Looy et al., 2010). Specifically, in order to increase the accuracy of text similarity calculations, we consider content analysis based on latent semantic indexing in the patent text mining process. Through integrating citation analysis with semantics analysis, this study will reduce the noise of scientific citations in patents effectively and reveal the real referential relationships.

The rest of this paper is structured as follows. The Section 'Literature review' discusses the topics related to the science linkage discussed in previous literature to develop a theoretical foundation for the follow-up study. The Section 'Data Collection' describes the details of the data gathering and preprocessing. The Section 'Research methods' introduces the vector space model. The Section 'Results and application' applies the model in a practical case study and discusses the results. Conclusion and future work directions are illustrated in the 'Conclusion and discussions' section.

Data collection

With reference to the prior studies (Li and Chambers et al., 2014; Huang and Yang et al., 2015; Sung and Wang et al., 2015; Chen, 2017; Ding and Hung et al., 2017), all the patents information used in this paper comes from the USPTO (http://patft.uspto.gov/netahtml/PTO/index.html), which provides reliable information on patent documents and is generally accepted by scholars. We adopted the following search terms which are used in previous researches (Zhang and Feng et al., 2017a; Zhang and Yu et al., 2017b). We constructed the origin database by retrieving and downloading the full text documents of patents. There are many types of documents referenced in patents, but not all of them are considered as scientific output (Ding and Hung et al., 2017). In order to process the follow-up study smoothly, following the approach of previous authors such as Lo (2010), Li and Meng (2010), Sung et al. (2015), and Ding et al (2017), we conducted a thorough selection process by removing the extra information, such as notice of allowance for U.S. application, international search report and written opinion, office action in corresponding application. As a result, only scientific papers were retained in the database.

Based on the patent citation information provided by USPTO (Alcácer and Gittelman, 2009; Yasukawa and Kano, 2015), we distinguished between the citations cited by application or examiner. Employing Web of Science and Google Scholar, we conducted a research of the details of scientific literature, such as authors, authors' organization, title, abstract and keywords. Drawing upon that information, we then further divided the citations cited by application into self-citation by inventor/applicant (literature written by inventors) and non-self-citation by inventor/applicant (literature written by other scholars). A total of patents in these two fields were identified, which are 423 and 376, respectively. The details of the data set are as shown as Table 1.

Table 1 Basic information of patent data in these fields

Research field	Number of patents	Number of citations by examiner	Number of self-citations by applicant	Number of non-self-citations by applicant
Pharmaceuticals	423	321	358	8235
Biosensor	376	144	208	4026

Research Methods

In the vector space model, the text information is transformed into the feature vector.

$$\mathrm{V}(\mathrm{d_j}) = (\mathrm{t_1}, \mathrm{w_{1j;}}\, \mathrm{t_2}, \mathrm{w_{2j;}}\ \ldots\ldots ; \mathrm{t_n}, \mathrm{w_{nj}}) \quad (1)$$

Where $t_i(i = 1, 2, \ldots\ldots, n)$ are the feature items corresponding to text d_j , and $w_{ij}(i = 1, 2, \ldots\ldots, n)$ are the weight of $t_i(i = 1, 2, \ldots\ldots, n)$ in d_j.

Matrix X is used to describe the relationship between text and feature items:

$$\mathrm{X} = \begin{bmatrix} \mathrm{w}_{11} & \mathrm{w}_{12} \cdots\cdots & \mathrm{w}_{1n} \\ \mathrm{w}_{21} & \mathrm{w}_{22} \cdots\cdots & \mathrm{w}_{2n} \\ \cdots\cdots & \cdots\cdots\cdots\cdots & \cdots\cdots \\ \mathrm{w}_{m1} & \mathrm{w}_{m2} \cdots\cdots & \mathrm{w}_{mn} \end{bmatrix} = (\mathrm{X}_1, \mathrm{X}_2, \ldots\ldots, \mathrm{X}_n) = \begin{bmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \cdots\cdots \\ \mathrm{Y}_m \end{bmatrix} \tag{2}$$

Where both column vector X_j and row vector Y_i represent the corresponding weights of the feature items.

w_{ij} is used to define TF − IDF factor by the following formula:

$$\mathrm{w}_{ij} = \frac{\log\left(\frac{\mathrm{N}}{\mathrm{df}_{ij}}\right) * \log(\mathrm{tf}_{ij}+1.0)}{\sqrt{\sum_{j=1}^{n}[\log\left(\frac{\mathrm{N}}{\mathrm{df}_{ij}}\right) * \log(\mathrm{tf}_{ij}+1.0)]^2}} \tag{3}$$

Where N is the total number of texts; df_{ij} is the number of texts which has feature item t in N; tf_{ij} is the frequency of a certain feature item in text d_j. Since many words appear less frequently, some elements in the matrix are zero. In order to reduce the amount of computation and storage space, and improve the efficiency of the computation, this study uses Latent Semantic Indexing (LSI) to reduce the dimension of the matrix.

In this process, singular value decomposition (SVD) is used to decompose the term-document matrix *X* into three matrices: a term concept matrix *U*, a singular value matrix *C* and document-concept matrix *V*: $X=UCV^T$ (Kontostathis and Pottenger, 2006). The top *K* dimensions of matrix *X* and *V* serve as the best approximation to the original matrixes. Each document is represented by a vector that shows the least square distances to the top *K* dimensions. There are three major advantages of LSI: synonymy, polysemy, and term dependence. Synonymy means that the same concept can usually be described by different terms. In LSI, the concept in question and other documents related to it are all described by a weighted combination of indexing variables. Polysemy refers to words that have a different meaning in a diverse context. An SVD of the term similarity matrix is available to combine with cluster analysis in order accurately to determine the sense of a particular word. LSI factors are orthogonal by definition, and words are positioned in the reduced space in a way that represents the correlations of their use across documents, which determines the characters of term dependence.

There are many formulas for text similarity, which include the inner product, cosine, Pearson and Manhattan distance formula. This paper employs the cosine formula which is as follows:

$$\mathrm{SIM}(\mathrm{d_i}, \mathrm{d_j}) = \frac{\sum_{k=1}^{n} \mathrm{w}_{ik}\mathrm{w}_{jk}}{\sqrt{\sum_{k=1}^{n} \mathrm{w}_{ik}^2}\sqrt{\sum_{k=1}^{n} \mathrm{w}_{jk}^2}} \tag{4}$$

Where $0 \le \mathrm{SIM}(d_i, d_j) \le 1$.

Results and Analysis

In this part, we apply the text similarity computing models described above to filter the noise in the non-self-citations by application to the fields of pharmaceuticals and biosensors. After that, in order better to describe the actual citation relationship, we calculate the science linkage using the following formula:

$$\text{NumSL\_NF=NumCE+NumSCA+NumNSLA\_NF} \tag{5}$$

In Eq. (5), $\mathrm{Num}_{\mathrm{SL\_NF}}$ is the science linkage after noise filtering; $\mathrm{Num}_{\mathrm{CE}}$ is the number of citations by examiner; $\mathrm{Num}_{\mathrm{SCA}}$ is the number of self-citations by applicant; $\mathrm{Num}_{\mathrm{NSLA\_NF}}$ is the number of non-self-citations by applicant after noise filtering.

We then set several thresholds for comparison. For example, if the threshold is 0.5, we will remove the citations having similarities with the citing patent which are less than or equal to

0.5. Overall speaking, the test similarities between patents and their scientific literature citations are relatively low, so we choose 0 and 0.05 as the threshold values.

To observe the distribution of the cited times of journals more directly, we provide the cited details of journals in these two fields. For these two fields Pharmaceuticals, and Biosensors, the total number of cited journals is 1215, and 780, respectively. The number of journals cited more than 100 times is 8, and 4, respectively. The number of journals which are cited between 50 and 100 times is 13 and 10, respectively. The number of journals of which the number of citations is between 10 and 50 are 123, and 70, respectively. The number of journals cited less than 10 times is 1071, and 696, respectively.

In order to observe the effect of noise reduction more intuitively, we provide the details of the top 15 journals listed in the field of pharmaceuticals and biosensors with the cited quantity under three conditions: without threshold, with the threshold of 0 and of 0.05, respectively. The details are shown in Table 2~Table 4.

Table 2~table 4 show that when processing with different thresholds, the journals' sequence and cited quantity changes with the total number of journals. The maximum value of similarity in the fields of pharmaceuticals and biosensors is 0.57 and 0.61, respectively. This result means that the test similarities are generally relatively low. According to the results in table 2~table 4, for the inventors in the field of pharmaceuticals, attention should be focused on the following journals: Journal of Medicinal Chemistry, PNAS, Journal of Biological Chemistry, Nature, Journal of Organic Chemistry, Science, Tetrahedron Letters, Journal of the American Chemical Society, and Tetrahedron. For the inventors in the field of biosensors, attention should be focused on the following journals: Analytical Chemistry, Sensors and Actuators B: Chemical, Biosensors & Bioelectronics, Journal of the American Ceramic Society, PNAS, Nature, and Science. As the top journals in the world, Science and PNAS appear in both the lists, which means that these journals record a large number of outstanding achievements in basic research.

Table 2 The top 10 journals list in the field of Pharmaceuticals and Biosensors (no threshold)

	The most cited journals and the cited quantity in Pharmaceuticals	The most cited journal and the cited quantity in Biosensors
1	Journal of Medicinal Chemistry (260)	Analytical Chemistry (356)
2	PNAS (210)	Sensors and Actuators B: Chemical (218)
3	Journal of Biological Chemistry (206)	Biosensors & Bioelectronics (172)
4	Nature (138)	Journal of the American Ceramic Society (171)
5	Journal of Organic Chemistry (137)	PNAS (86)
6	Science (130)	Nature (85)
7	Tetrahedron Letters (111)	Science (83)
8	Journal of the American Chemical Society (109)	Journal of Biological Chemistry (74)
9	Tetrahedron (96)	Journal of Electroanalytical Chemistry (70)
10	Journal of Virology (79)	Langmuir (59)

Table 3 The top 10 journals list in the field of Pharmaceuticals and Biosensors (threshold 0)

	The most cited journals and the cited quantity in Pharmaceuticals	The most cited journal and the cited quantity in Biosensors
1	Journal of Medicinal Chemistry (225)	Analytical Chemistry (340)
2	Journal of Biological Chemistry (189)	Sensors and Actuators B: Chemical (193)
3	PNAS (178)	Journal of the American Ceramic Society (169)
4	Journal of Organic Chemistry (121)	Biosensors & Bioelectronics (162)
5	Science (119)	Nature (84)
6	Nature (116)	Science (80)
7	Journal of the American Chemical Society (103)	Journal of Biological Chemistry (69)
8	Tetrahedron (93)	Journal of Electroanalytical Chemistry (65)
9	Tetrahedron Letters (87)	Langmuir (59)
10	Journal of Virology (76)	PNAS (52)

Table 4 The top 10 journals list in the field of Pharmaceuticals and Biosensors (threshold 0.05)

	The most cited journals and the cited quantity in Pharmaceuticals	The most cited journal and the cited quantity in Biosensors
1	Journal of Medicinal Chemistry (218)	Analytical Chemistry (237)
2	PNAS (167)	Sensors and Actuators B: Chemical (136)
3	Journal of Biological Chemistry (163)	Biosensors & Bioelectronics (155)
4	Nature (108)	Journal of the American Ceramic Society (113)
5	Journal of Organic Chemistry (111)	Nature (58)
6	Science (93)	Journal of Biological Chemistry (45)
7	Journal of the American Chemical Society (99)	Science (44)
8	Tetrahedron (91)	PNAS (41)
9	Journal of Virology (76)	Langmuir (33)
10	Tetrahedron Letters (75)	Journal of Electroanalytical Chemistry (32)

Discussion

Based on the patent data of four representative fields from the USPTO, this paper focuses on the method of reducing the noise in patents' scientific literature citations. In terms of theoretical contributions, firstly, drawing upon analysis of the sources of patent citations, the citations' purposes as well as their roles, we found the main source of citation noise, which laid the foundations of the preceding analysis. Secondly, according to the characteristics of patent citation noise, this paper proposed the text similarity computing method based on the vector space model and latent semantic indexing to filter the patents' citation data. The results help to reflect the real citation relationship, measure the knowledge diffusion from scientific literature to technical innovation, and ensure the accuracy of the follow-up study.

With respect to practical contributions, firstly, filtering the noise of patents' scientific literature citations contributes to evaluating accurately the contribution of scientific journals. Based on different thresholds, we observed the change of the cited scientific journals' sequence and quantity, which helps to choose the appropriate scientific journals during technical innovation. Secondly, drawing upon the concept of journal impact factor (JIF), a technological impact factor (TIF) is proposed as a new indicator to evaluate the importance of a specific journal from the aspect of scientific research's contribution to practical innovation, and thus describe knowledge flow between journal papers and patents (Huang and Huang et al., 2014). However, this indicator ignores the noise in patents' citations, and the method proposed in this paper is very useful for improving the indicator.

The data used in this paper comes from the United States Patent and Trademark Office (USPTO). Different patent offices have diverse patent review mechanisms. The follow-up research can collect the patent citation data from other patent databases, such as the European Patent Office (EPO) and the Japan Patent Office (JPO) and make a comparison. The results can be applied to strengthen the discussion of the noise in patents' citations and illustrate the relationship between scientific literature and technological innovation.

Acknowledgments

This study was supported by National Natural Science Foundation of China (Nos. 71804091 and 71701078), Humanities and Social science Foundation of the Ministry of Education of China (No. 18YJC630237), Natural Science Foundation of Shandong Province (No. ZR2019QG007), the University Scientific Research Plan Projects of Shandong Province (Nos. J18KA342 and J18RA049) and Social Science Foundation of the University of Jinan (Nos. 18YY02 and 17YB06).

References

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49-63.

Alcácer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in US patents: An overview and analysis. *Research Policy*, *38*(2), 415-427.

Azagra-Caro, J. M., Mattsson, P., & Perruchas, F. (2011). Smoothing the lies: The distinctive effects of patent characteristics on examiner and applicant citations. *Journal of the American Society for Information Science and Technology*, *62*(9), 1727-1740.

Breschi, S., & Catalini, C. (2010). Tracing the links between science and technology: An exploratory analysis of scientists' and inventors' networks. *Research Policy*, *39*(1), 14-26.

Callaert, J., Grouwels, J., & Van Looy, B. (2011). Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references. *Scientometrics*, *91*(2), 383-398.

Chen, L. (2017). Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations. *Journal of Informetrics,* 11(1), 63-79.

Criscuolo, P. & B. Verspagen (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy,* 37(10), 1892-1908.

Ding, C. G., Hung, W. C., Lee, M. C., & Wang, H. J. (2017). Exploring paper characteristics that facilitate the knowledge flow from science to technology. *Journal of Informetrics*, *11*(1), 244-256.

Fukuzawa, N., & Ida, T. (2016). Science linkages between scientific articles and patents for leading scientists in the life and medical sciences field: The case of Japan. *Scientometrics*, *106*(2), 629-644.

Guan, J. & He Y. (2007). Patent-bibliometric analysis on the Chinese science - technology linkages. *Scientometrics, 7*2(3), 403-425.

He, Z. & Deng M. (2007). The evidence of systematic noise in non-patent references: A study of New Zealand companies' patents. *Scientometrics,* 72(1), 149-166.

Hegde, D. & Sampat B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters,* 105(3), 287-289.

Hicks, D., Breitzman Sr, A., Hamilton, K., & Narin, F. (2000). Research excellence and patented innovation. *Science and Public Policy*, *27*(5), 310-320.

Hu, D., Chen, H., Huang, Z., & Roco, M. C. (2007). Longitudinal study on patent citations to academic research articles in nanotechnology (1976–2004). *Journal of Nanoparticle Research*, *9*(4), 529-542.

Huang, M. H., Yang, H. W., & Chen, D. Z. (2015). Increasing science and technology linkage in fuel cells: A cross citation analysis of papers and patents. *Journal of informetrics*, *9*(2), 237-249.

Huang, M. H., Huang, W. T., & Chen, D. Z. (2014). Technological impact factor: An indicator to measure the impact of academic publications on practical innovation. *Journal of Informetrics*, *8*(1), 241-251.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, *108*(3), 577-598.

Klitkou, A., & Gulbrandsen, M. (2009). The relationship between academic patenting and scientific publishing in Norway. *Scientometrics*, *82*(1), 93-108.

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*, *42*(1), 56-73.

Lampe, R. (2012). Strategic citation. *The Review of Economics and Statistics* 94(1), 320-333.

Li, R. and Meng L. (2010). On the framing of patent citations and academic paper citations in reflecting knowledge linkage: A discussion of the discrepancy of their divergent value-orientations. *Chinese Journal of Library and Information Science,* 3, 37-45.

Li, R., Chambers, T., Ding, Y., Zhang, G., & Meng, L. (2014). Patent citation analysis: Calculating science linkage based on citing motivation. *Journal of the Association for Information Science and Technology*, *65*(5), 1007-1017.

Lissoni, F. (2010). Academic inventors as brokers. *Research Policy,* 39(7), 843-857.

Lo, S. S. (2010). Scientific linkage of science research and technology development: a case of genetic engineering research. *Scientometrics,* 82(1), 109-120.

Magerman, T., Van Looy, B., & Song, X. (2009). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289-306.

Meyer, M. (2006). Are patenting scientists the better scholars? *Research Policy,* 35(10), 1646-1662.

Narin, F., Hamilton, K. S., & Olivastro, D. (1995). Linkage between agency-supported research and patented industrial technology. *Research Evaluation*, *5*(3), 183-187.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research policy*, *26*(3), 317-330.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Shearer, B. A., Lundeberg, M. A., & Coballes-Vega, C. (1997). Making the connection between research and reality: Strategies teachers use to read and evaluate journal articles. *Journal of Educational Psychology*, *89*(4), 592.

Sung, H. Y., Wang, C. C., Huang, M. H., & Chen, D. Z. (2015). Measuring science-based science linkage and non-science-based linkage of patents through non-patent references. *Journal of Informetrics*, *9*(3), 488-498.

Tijssen, R. J. W. (2001). Global and domestic utilization of industrial relevant science: Patent citation analysis of science-technology interactions and knowledge flows. *Research Policy,* 30, 35-54.

Van Looy, B., Magerman, T., & Debackere, K. (2007). Developing technology in the vicinity of science: An examination of the relationship between science intensity (of patents) and technological productivity within the field of biotechnology. *Scientometrics*, *70*(2), 441-458.

Wang, G. and Guan J. (2011). Measuring science-technology interactions using patent citations and author-inventor links: an exploration analysis from Chinese nanotechnology. *Journal of Nanoparticle Research,* 13(12), 6245-6262.

Wang, M., Yu, G., Xu, J., He, H., Yu, D., & An, S. (2012). Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics*, *6*(4), 586-599.

Yasukawa, S. & Kano S. (2015). Comparison of examiners' forward citations in the United States and Japan with pairs of equivalent patent applications. *Scientometrics,* 102(2), 1189-1205.

Yu, Q., Ding, Y., Song, M., Song, S., Liu, J., & Zhang, B. (2015). Tracing database usage: Detecting main paths in database link networks. *Journal of Informetrics*, *9*(1), 1-15.

Zhang, G., Feng, Y., Yu, G., Liu, L., & Hao, Y. (2017a). Analyzing the time delay between scientific research and technology patents based on the citation distribution model. *Scientometrics*, *111*(3), 1287-1306.

Zhang, G., Yu, G., Feng, Y., Liu, L., & Yang, Z. (2017b). Improving the publication delay model to characterize the patent granting process. *Scientometrics*, *111*(2), 621-637.

Models of parenting and its effect on academic productivity: Preliminary results from an international survey

Gemma E.Derrick[1], Adam Jaeger[5], Pei-Ying Chen[2], Cassidy R.Sugimoto[2], Thed van Leeuwen[4], Vincent Lariviere[3].

[1] g.derrick@lancaster.ac.uk
Centre for Higher Education Research & Evaluation, Lancaster University, LA1 4YD (United Kingdom)

[2] sugimoto@indiana.edu & peiychen@iu.edu
School of Informatics, Computing, and Engineering, Indiana University Bloomington (United States)

vincent.lariviere@umontreal.ca
School of Library and Information Science, University of Montreal, Montreal, H3C 3J7, Quebec (Canada)

leeuwen@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

jaeger@math.wichita.edu
Department of Mathematics, Statistics and Physics, Wichita State University, Wichita, KS, 67260 (United States)

Abstract

This preliminary paper investigates the cost of parenting engagement on academic productivity and impact. Instead of investigating the relationship between gender and academia, this study focuses on time invested in parenting as the lead factor underpinning productivity differences for both men and women. Survey responses from 17,519 first and last authors publishing between 2007 and 2017 yielded four distinct parenting types: Lead parents; Satellite parents; Sole parents; and Dual parents. In addition a free text box in the survey allowed for the analysis of 5976 qualitative responses about participant's experiences balancing parenting with their partners, and academic careers. Results show a significant difference across all types of parenting relative to gender for the number of papers produced, as well as for the proportion of papers published in top journals. In addition, for men and women who take on dual parenting roles (a hypothetical 50/50 split), the productivity cost is higher for women. Conversely, there is a significant cost for men and women who take on the role of Lead parent. Further qualitative investigation highlights the incidence of an 'invisible burden'in self-identified dual parenting families, wherein there is a significant amount of unacknowledged labor that is undertaken by females. This invisible labor may contribute to the difference in productivity between men and women in dual-parenting relationships.

Introduction

The gender gap in academe has been the focus of several analyses, which span from describing the various forms taken by this gap (Larivière et al., 2013; Nittrouer et al., 2015) to the mechanism that drive it (Leslie et al, 2015; Moss-Racusin,et al., 2012). Among those, differences in academic rank, organisational approaches, and the extent of specialisation (Leahey, 2006), the main explanation remains an assumption that women are faced with the majority of childcare (Beddoes & Pawley, 2014). As a consequence this restricts their ability to engage fully in the demands of the academy, directly influencing their academic productivity both in the short and long term.

No one denies that raising children demands a considerable amount of time and effort that diminishes the time and energy that can be devoted to scholarship and on academic earnings.

Barriers for women when pursuing academic careers in science include systematic barriers such as child rearing and the inability of research systems to allow the flexibility necessary to juggle research with home-responsibilities (Feeney et al, 2014; van Anders, 2004). However, this centralisation of women as the primary caregiver, and hence the majority of the time burden in previous research further blurs an understanding of the productivity cost of parenting. This is especially when variables other than childcare perpetuate the gender gap in academia.

The concept of "balanced" parenting is also a relatively new social change (Bright Horizons, 2017) that questions the relevance of past studies of how commitments to the academy and parenting are fulfilled. A more modern perspective on parenting also acknowledges that both parents, irrespective of gender or marital status, involved in a degree of parenting. In addition it also recognises parenting strategies that incorporate non-parental figures (e.g. grandparents/extended family; formal childcare provision etc) that take an active role in maintaining the work-life balance in academia. These more modern parenting models allow families to strategise children, childcare and full time careers (for both parents potentially) around academic demands. What is missing in an understanding of how parenting influences academic labor is the productivity and performance costs of modern parenting strategies.

Using a world-wide survey of academic parents (n=17,519 respondents), this research in progress uses the term "parent" as gender neutral, acknowledging that modern parenting is a joint, or multiple-partner endeavour. As such, this research aims to avoid ascribing loss of productivity on a single individual "parent" alone, in order to further investigate the parenting cost on academic productivity.

Methods

Survey

Web of Science was used to sample all first and last authors who had published at least one article in the period 2007-2017. All authors were then invited to complete an online survey. The first survey question used skip logic to eliminate all potential respondents who were not parents. In total, 17,519 individuals who met this initial filtering requirement responded. Data cleaning was done to exclude unfinished responses and erroneous responses (e.g., doctoral degrees obtained before birth) and to account for missing data resulted in a final sample of 10,444.

Survey questions included: demographic information on children and partners; contribution to childcare; the balance of parental labor with other caregivers; and their perception of the relationship between childcare and academic careers. The underlying hypothesis guiding survey construction was that it was not the parental status, but rather time allocated to parenting that would lead to decreased productivity.

Quantitative Methods

The analytic set included the 10,444 respondents with complete surveys. ANOVA was used to test the null hypothesis that the mean productivity (i.e., number of papers) and impact (i.e., proportion of published papers that are considered highly cited relative to field and year (PPTop)), is the same for parenting type relative to gender (gender/parenting type categorisation). Permutation tests were used as a post hoc test comparison to further test the relationship between the gender/parenting type categorisation. Here, two test statistics were used; one that measures the square distance between each observation and the group mean; and the other measuring the difference between the group median.

Qualitative data

A free text section was included at the end of the survey that encouraged participants to *"Please feel free to add any additional comments you have regarding childcare and scientific labor, drawing upon your own experiences".* In total, 5976 participants completed this section. To analyse this, a random sample of 500 was selected and coded thematically using a grounded theory-informed approach. Themed categories were developed (n=59) and then collapsed into 8 overarching thematic codes capable of facilitating the manual coding of large numbers of responses.

Results

A breakdown of how respondents described their involvement in parenting is shown below in Table 1.

Table 1. Proportion of respondents (Male and Female) in different parenting styles

Parenting style	*Male*	*Female*
I am the primary caregiver to my child(ren)	4.2	31.8
My partner is the primary caregiver to my child(ren)	33.2	4.0
The majority of childcare is performed by non-parental caregiver(s)/other	5.8	12.5
I share equal parenting roles with my partner	55.1	46.6
I share equal parenting roles with non-parental caregiver(s)/other	1.7	5.1
TOTAL	100	100

Here, 31.8% of female respondents indicated that they were the 'primary caregiver' to their children, compared with 4.2% of male respondents. In contrast, 33.2% of men indicated that their 'partner is the primary caregiver' to their children. A relatively equal proportion of male (55.1%) and Female (46.6%) respondents indicated that they 'share equal roles with my partner'. These results hid whether respondents were able to share parenting duties as a married and/or partnered relationship, and those were not but still "shared" parenting. Therefore the results of Table 1 were cross-referenced with marital stat gender/parenting type categorisation us to create the following parenting classifications.

Table 2. Classification of parenting types and proportion of respondents in each category

Parenting type	*Definition*	*Male*	*Female*
Sole parent	Are *'primary caregiver'* to their children AND are single/widowed/divorced or separated	1.1	6.5
Lead parent	Are within married/partnered relationships AND *'the primary caregiver'*	2.9	24.1
Satellite parents	Any relationship arrangement AND have a *'partner who is a primary caregiver'; or 'the majority of childcare is performed by a non-parental caregiver(s)/other'*	38.8	17.4
Dual parents	Any relationship arrangement AND have *'share equal parenting roles with a partner'*; AND *'share equal parenting roles with non-parental caregivers(s)/other'*	57.2	51.9

Although both genders report engaging in dual-parenting arrangements and/or as a satellite parent; 24.1% of women are acting as the Lead-parent whereas less than 3% (2.9%) of men act the same.

Figure 1. Number of papers by parenting type and gender (outliers removed)

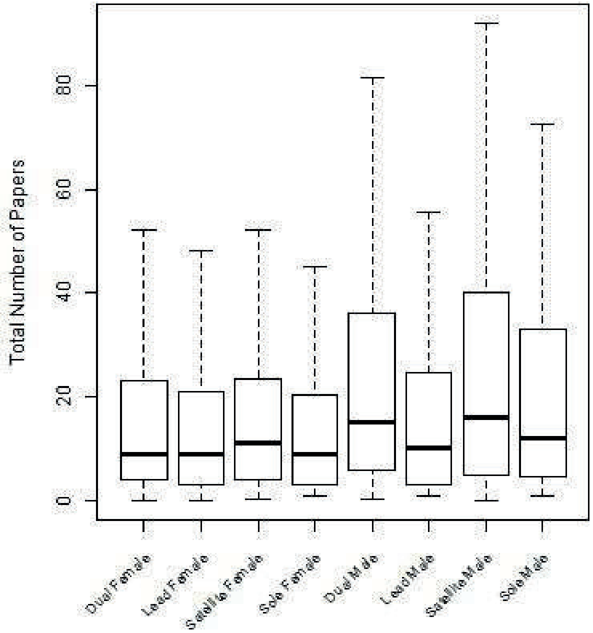
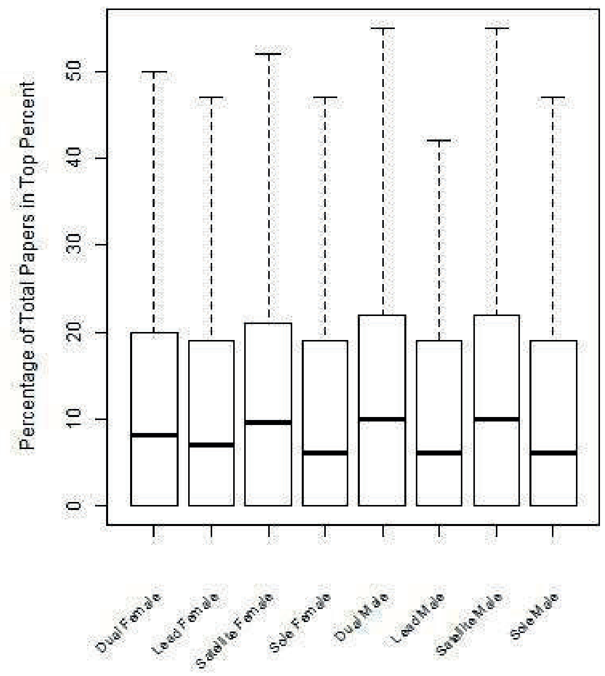


Figure 1 shows the average number of papers, with the outliers removed, for each parenting type relative to gender. A one-way ANOVA test was performed across all gender/parenting type categorisation showing a significant difference (F(leadparent) F=25.31 p<0.001), and a further two way test, showed interactional effects indicating that the effects of taking on the Lead parenting role on academic productivity are different for men and women F(gender:leadparent) F=3.34, p=0.01. Further, using a permutation test for the number of papers, the probability that the expected mean number of papers and median number of papers is not different for at least one gender/lead parenting level is almost 0.

In addition, Figure 2 shows the proportion highly cited papers relative to field an year (PPTop) for each gender/parenting type categorisation. The single factor analysis shows how the expected percentage is different for at least on gender/parenting type categorisation (F=3.40, p=0.001). The two-factor interaction ANOVA indicates that while there is a difference in the expected percentage between males and females (F(gender:leadparent) F=0.15, p=0.92) , there is no difference for the lead parent role regardless of gender. Permutation tests verified this finding by showing a probability that the means and medians were the same for each gender/lead parent categorisation of close to 0 (=10^{-4}), p=0.0016.

Figure 2. Proportion of parent types by gender in PPTop



Qualitative results

A free text question allowed respondents to comment further on the survey or experiences balancing parenting with an academic career. In these responses, participants reflected on the flexibility of an academic career as being useful for allowing the time necessary to engage in parenting activities. However when the spouse was not an academic this flexibility was taken for granted as outwardly it seemed that they were *"not busy"*, resulting in parenting tasks being unconsciously conducted by the academic parent.

> *"It is hard to balance academic work and home life - as in many cases your partner does not understand that reading and working on your computer is your job. Thus, you find that you have various tasks (family, children, house, errands) thrown to you by your spouse who works a "regular" job because you are "not busy".*

The assumed-flexibility of an academic career also served to ingrain practices that burdened academic women with the majority of responsibility for parenting;

> *"I didn't want to miss out on anything and had the more flexible career, so I did most of the parenting roles. However, this eventually just became the habit of "how we did things" and my husband had time for hobbies while every moment of my time was taken up by work and kids."*

In addition, the invisible burden of parenting on academic women, and the flexibility of academic work that allowed parents to appear *"not busy"* by working at home, infiltrated the reasoning of dual-parents around who would assume the majority of the parenting responsbilities; “*Inevitably, we both feel that if a sacrifice must be made, it is my schedule”*. This decision was made irrespective of salary considerations.

In many cases, the adoption of invisible parenting labor was not a result of a conscious decision about how to divide roles between parents, but still incorporated a large temporal and emotional burden;

> *The mental labor of researching and remembering EVERYTHING related to kids activities and school falls to me - including selecting locations, remembering deadlines for sign-ups, getting proper equipment: summer camps, swimming lessons, dance, after school care, parties at school (bringing snacks/valentines etc.), field trips. It is constant, exhausting, and under-appreciated.*

Men who were part of dual-parenting arrangements acknowledged the existence of this invisible parenting labor burden on women; “*Although I try to be active in child care and share responsibilities equally, my wife still takes care of more child care tasks than I do”*. A further analysis of men in satellite-parenting arrangements also reinforced the benefits they accrue in academic productivity when their partner takes on the lead parenting role. Men in self-declared dual-parenting relationships also acknowledged the invisible burden on their partners; *I like to think we shared, but the wife apparently did more.*

Finally, there are benefits for women who adopt a lead-, and dual-parenting arrangement, provided that their partner takes a lead or dual-parenting role as well; “*The system was not perfectly equal in all regards, but he made every effort to make it as fair to both of us as possible. That is a big reason why I have had a successful career in science.*

Discussion

The results showed that there is a connection between the amount of parental responsibility assumed by an individual and research productivity as measure by the number of papers, and the proportion of papers considered highly cited for the field and year (PPTop). The model also show that there is a significant interaction with gender, suggesting that the link between parenting arrangements and productivity differs is different for men than it is for women. This study demonstrates how the level of parental responsibility is a powerful variable to explain academic productivity differences between men and women. Further research is currently underway to investigate these effects in more detail, which also includes a deeper understanding of the nature of the invisible parenting labor burden, and its interactions with the parenting typologies and academic productivity and impact.

References

Beddoes, K., Pawley, A.L. (2014) Different people have different priorities: work-family balance, gender, and the discourse of choice. Studies in Higher Education, 39(9), 1573-1585.

Bright Horizons. (2017) Modern Family Index 2017. Available at https://solutionsatwork.brighthorizons.com/~/media/BH/SAW/PDFs/GeneralAndWellbeing/MFI_2017_Report_v4.ashx

Feeney, M.K., Bernal, M., Bowman, L. (2014) Enabling work? Family-friendly policies and academic productivity for men and women scientists. Science and Public Policy, 41(6), 750-764.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. Nature, 504(7479), 211.

Leahey, E. (2006). Gender differences in productivity: Research specialization as a missing link. Gender & Society, 20(6), 754-780.

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. Science, 347(6219), 262-265.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of Sciences, 109(41), 16474-16479.

Nittrouer, C. L., Hebl, M. R., Ashburn-Nardo, L., Trump-Steele, R. C., Lane, D. M., & Valian, V. (2018). Gender disparities in colloquium speakers at top universities. Proceedings of the National Academy of Sciences, 115(1), 104-108.

Van Anders, S,M., (2004) Why the academic pipeline leaks: Fewer men that women perceive barriers to becoming professors. Sex roles, 51(9-10): 511-521.

Identifying Research Fronts in a Fine-grained Way: A Case Study in the Field of Artificial Intelligence

Bentao Zou[1], Yuefen Wang[2] and Jiajun Cao[3]

[1] *zoubentao@njust.edu.cn*
Nanjing University of Science & Technology, School of Economics & Management, Nanjing (China)

[2] *yuefen163@163.com*
Nanjing University of Science & Technology, School of Economics & Management, Nanjing (China)
Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Nanjing (China)

[3] *jscjj95@126.com*
Nanjing University of Science & Technology, School of Economics & Management, Nanjing (China)

Abstract

To satisfy the needs of the development of science and technology for diversified fronts exploration, this paper proposed a fine-grained way to identify different types of research fronts from the author-level aspect. We selected high yield authors and applied burst detection algorithm to detect burst terms from representative prolific authors, then extracted additional attributes from their papers. Based on the attributes, we created some indicators to define different types of research fronts. A case study in the field of Artificial Intelligence (AI) had been done and we got 13 emerging research fronts, 19 growing research fronts and 38 hot research fronts. Emerging research fronts depict specific research directions or definite applications, growing research fronts reflect the transition from the emerging research fronts to the hot research fronts, while hot research fronts mostly represent basic method or theories in the field. Our results show enlightenments to related researchers and give suggestions to policy makers, administrators or funders. Further works will focus on testing much more high yield authors to explore the impact on the results with regard to the number of authors.

Introduction

The study of research front has become a hot topic in recent years. From 2013 to 2017, Institutes of Science and Development, Chinese Academy of Sciences (CAS) and Clarivate Analytics joint released a series of reports of *Research Fronts* based on the Clarivate Analytics database Essential Science Indicators (ESI). These reports have gained widespread attention from around the world, the *Physics World* ever cited fronts in physics identified by *Research Fronts 2016* (Physics World, 2016).

Understanding the latest developments or tracking emerging specialty areas provides a distinct advantage for administrators, policy makers, and others who need to monitor, support, and advance the conduct of research in the face of finite resources. In the other hand, developments of science and technology have seen rapid increases in publications in years. In this condition, there is a commensurate increase in the need for scientific and technical intelligence to discover research fronts. However, new development in science and technology pushes new needs of diverse research fronts detection. Our study selects papers of high-yield authors as input to detect burst terms and classifies detected terms into different types of research fronts based on certain attributes, which puts emphases on identifying research fronts in a fine-grained way.

Related Works

The Definition of Research Fronts

The concept of research fronts was first introduced by Price (1965), he pointed out that there was a tendency for the most-cited papers to be also the most recent, and clusters of recent published papers with high citations indicated the nature of the scientific research front.

Persson (1994) believed that articles that were similar in terms of citing the same literature formed a research front. Morris, Yen and Wu et al. (2003) defined research fronts as clusters of documents that tend to cite a fixed, time invariant set of base documents. Besides these two views about the definition of research fronts, some researchers defined it from words or topics level. Braam, Moed and Van Raan (1991) regarded it as a coherent set of subject-related research problems and concepts upon which attention is focussed by a number of scientific researchers. More recently, Chen (2006) defined a research front as an emergent and transient grouping of concepts and underlying research issues. However, there is still no general consensus about the definition of research fronts. Considering that citation analysis methods are inadequate in depicting new emerging research fronts (namely the time lag), we identify research fronts from burst terms in this study.

Methods on Identifying Research Fronts

As Figure 1 shows, we classify existing approaches into qualitative methods and quantitative ones. Qualitative methods mainly include Delphi study, analysis of technology policy and comparative analysis etc. (Ran, Su & Zhao, 2017). They were widely used in the early stage of detecting research fronts, and were mostly applied to guide the work of making government technology strategies. However, those expert-based approaches, which utilize the explicit knowledge of domain experts, are often time-consuming and subjective in this information-excess era. Quantitative methods contain: (1) citation-based methods, including co-citation analysis (Small, 1973; Griffith, Small & Stonehill et al., 1974; Persson, 1994; Shibata, Kajikawa & Takeda et al., 2009), bibliometric coupling (Kessle, 1963; Persson, 1994; Huang & Chang, 2014), direct citation (Kajikawa, Fujimoto & Takeda et al., 2009; Shibata, Kajikawa & Takeda et al., 2009) and burst reference (Kleinberg, 2003; Fang, 2015; Hou, Yang & Chen, 2018); (2) word-based methods, including word frequency (Mane, 2004), burst term (Chen, 2006; Chen, Dubin & Kim, 2014; Song, Zhang & Dong, 2016) and co-occurrence (Callon, Michel & Turner et al., 1983; Rip & Courtial, 1984). However, as yet no consensus has emerged as to which of them has better detection performance (Huang & Chang, 2015). In our study, we use burst term detection and statistical attributes for identifying research fronts, which ensures both objectivity and timeliness of identified research fronts to a certain extent.

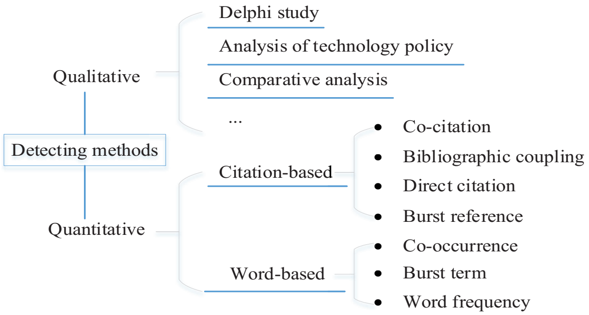


Figure 1. Taxonomy of methods on identifying research fronts

Types of research fronts

Looking back on existing researches, only a few researchers detected different types of research fronts. Upham and Small (2009) measured the growth rates of research fronts and categorized them as growing, shrinking, stable, emerging, or existing fronts. CAS and Clarivate Analytics (2017) joint released the *Research Fronts 2017*, and they selected hot research fronts and emerging research fronts from 9690 research fronts in 21 ESI fields. Total citations and the average year of fronts' core papers were regarded as two key factors in identifying hot research fronts, while research fronts whose core papers dated to the second half of 2015 or more recently (on average) and total citations were considered to define emerging research fronts. Based on previous work, we specify the definitions of several types of research fronts and make them more in line with the research needs in this era.

Data and Methodology

Research design

The research schema is depicted in Fig. 2:

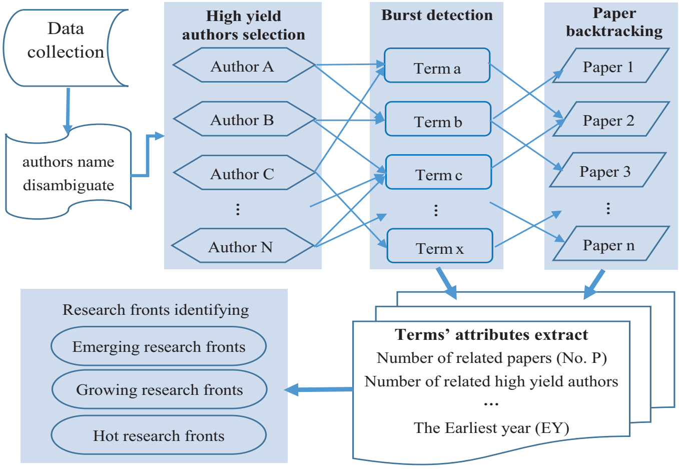


Figure 2. Research schema of this study

A main principle of identifying research fronts from a fine-grained way is that we do not use all publications in a field. Instead, we regard publications belonging to high yield authors as the research object. Follow the schema shows in Fig. 2, (1) we first combine authors' co-authors and institutes to disambiguate names, (2) then set the threshold and select prolific authors; (3) For each high yield author, burst algorithm was applied to the author's publications to detect burst terms; (4) for each burst term, backtrack related papers, match each burst term to related papers when the term appears in the title, the abstract, the keywords, or the keywords plus of a paper; (5) extract the publication year, the number of cited times and other attributes of a paper; (6) based on those attributes and further indicators, we define different types of research fronts and make detailed illustrations.

Note that burst term itself only provides limited information about the research front, so we focus on the author-level perspective and assign informative attributes to burst terms. We highlight it as an innovation of our study, since most previous studies using burst terms for detecting research fronts merely describe and explain burst terms but without combining other useful attributes. Table 1 shows descriptions of the attributes.

Table 1. Descriptions of terms' attributes

Attributes (Abbr.)	*Descriptions*
Number of related papers (No. P)	No. of related papers to a term
Number of related high yield authors (No. A)	No. of related high yield authors to a term

The earliest year (EY)	The earliest year that a term appeared (in our dataset)
The latest year (LY)	The latest year that a term appeared (in our dataset)
Average year (AY)	The average year of a term
Number of the earliest related high yield authors (No. EA)	No. of the earliest related high yield authors to a term
Number of the latest related high yield authors (No. LA)	No. of the latest related high yield authors to a term
Total citations of a term's related papers (TC)	No. of the total cited times of a term's related papers

After finishing those processes, we can form a matrix (Table 2), via which we define three types of research fronts in this study: (1) Hot research fronts, are identified by the PAC indicator we proposed in this paper:

$$PAC = \frac{No.P}{\max(No.P)} + \frac{No.A}{\max(No.A)} + \frac{TC}{\max(TC)} \qquad \text{(a)}$$

To detect hot research fronts, equation (a) not only considers the citations of a burst term, but also combines the number of related papers and related high yield authors. A larger PAC indicator always means a more likely the term be a hot research front. (2) Growing research fronts, are identified by the G indicator depicted as equation (b):

$$G = \frac{No.LA - No.EA}{LY - EY} \qquad \text{(b)}$$

The G indicator is the average number of increased researchers per year. The larger G value of a term, the higher growth rate a term has, and more likely be a growing research front. (3) Emerging research fronts, are only determined by time factor. We set a threshold of the earliest year of a term's appearance to define this type of research front. Above all, for each burst term, we have these three indicators to determine which kind of front it belongs to.

Table 2. An example of burst terms matrix

Terms	*No. P*	*No. A*	*EY*	*LY*	*AY*	*No. EA*	*No. LA*	*TC*
Term A	9299	877	1996	2017	2009	42	252	123547
Term B	3956	552	1996	2017	2009	6	95	74318
…	…	…	…	…	…	…	…	…

Data collection and processing

First, we used WosDownload.exe to collect bibliographic records from 1996 to 2017 with the search strategy as "WC = Computer Science, Artificial Intelligence", WoS core collection and BIOSIS Citation Index (BCI) were selected. All searches were done within May, 2018 and we got 726597 records. Second, after disambiguating authors' names, we got 1085 authors who published more than 50 papers (full-count) from 1996 to 2017 and defined them as high yield authors in this study. Considered that bibliographic records of every single author's publications would be an input for each author's burst detection, it would be a huge task if we did the same thing for all 1085 authors, so we took 70 representative authors from 14 countries (or regions) as our research sample. Those countries (or regions) were Top 14 with the greatest number of high yield authors and each country's Top 5 authors with the most publications were selected. Since research fronts depict new trends or new movements in a field, the burst detection selected those authors' papers published within the latest 10 years (2008-2017). To find the earliest year of a burst term, we match terms with papers of 1085 high yield authors with papers' publication year ranging from 1996 to 2017. Table 3 shows authors we selected.

Table 3. Selected high yield authors

Ranks	*Country /Region*	*No. of prolific authors*	*Top 5 authors (No. of publications)*
1	China	144	Jiao, Licheng(272)、Li, Xuelong(247)、Cao, Jinde(189)、Zhang, David(180)、Zhang, Lei(174)
2	America	75	Abraham, Ajith(235)、Shen, Dinggang(181)、Narayanan, Shrikanth S.(175)、Zhang, Huaguang(163)、Chellappa, Rama(155)
3	Japan	46	Watada, Junzo(126)、Fukuda, Toshio (104)、Ishibuchi, Hisao (104)、Ishiguro, Hiroshi (104)、Cichocki, Andrzej (99)
4	Spain	40	Herrera, Francisco (246)、Grana, Manuel (157)、Bustince, Humberto (144)、Herrera-Viedma, Enrique (110)、Bajo, Javier (108)
5	German	32	Navab, Nassir (185)、Schuller, Bjoern (119)、Cremers, Daniel (118)、Ney, Hermann (118)、Knoll, Alois (100)
6	England	27	Yao, Xin(179)、Hancock, Edwin R.(177)、Pantic, Maja(135)、Zisserman, Andrew(109)、Jin, Yaochu(106)
7	Australia	23	Tao, Dacheng(366)、Nahavandi, Saeid(134)、Lu, Jie(131)、Zhang, Guangquan(105)、Lim, Chee Peng(104)
8	Italy	15	Caldwell, Darwin G.(123)、Murino, Vittorio(102)、Sebe, Nicu(93)、Roli, Fabio(81)、Loia, Vincenzo(79)
9	Canada	15	Pedrycz, Witold(345)、Shi, Peng(165)、Wu, Q. M. Jonathan(97)、Bouguila, Nizar(96)、Sabourin, Robert(92)
10	Singapore	14	Yan, Shuicheng(211)、Li, Haizhou(129)、Er, Meng Joo(97)、Tan, Chew Lim(95)、Suresh, Sundaram(92)
11	South Korea	13	Kweon, In So(136)、Cho, Sung-Bae(106)、Lee, Minho(88)、Jo, KangHyun(77)、Lee, Kyoung Mu(77)
12	India	12	Das, Swagatam(154)、Pal, Umapada(132)、Deb, Kalyanmoy(125)、Jawahar, C. V.(119)、Panigrahi, Bijaya Ketan(108)
13	Taiwan	12	Hong, Tzung-Pei(218)、Pan, Jeng-Shyang(137)、Chen, Shyi-Ming(119)、Chang, Chin-Chen(113)、Tseng, Vincent S.(112)
14	France	10	Prade, Henri(127)、Schmid, Cordelia(89)、Dubois, Didier(85)、Paragios, Nikos(79)、Ogier, Jean-Marc(75)

Results

Results of burst detection

For the 70 representative authors we got 212 burst terms after manually filtering terms having no specific meanings, such as *the-art method (the state-of-the-art method), novel method* and *good performance* etc. Limited by space, Table 4 only lists burst terms of 14 authors with the most publications in the corresponding country (or region), for full version please email to us.

Table 4. Excerpt from representative high yield authors' burst detection results

Authors	*Terms*
Tao, Dacheng	linear discriminant analysis; learning algorithm; dimensionality reduction; discriminative information; human visual system; local geometry; image classification; sparse representation; objective function
Jiao, Licheng	evolutionary algorithm; benchmark problems; spectral clustering; neural network; real-world data set; memetic algorithm;
Watada, Junzo	DNA computing; fuzzy random variable; neural network
Herrera, Francisco	fuzzy rule-based classification system; feature selection; membership function; nonparametric statistical tests; prototype selection; nearest neighbor classifier; noisy data; big data; fuzzy set; multi-class problems
Schuller, Bjoern	recurrent neural net; automatic speech recognition; long short-term memory; neural network; recurrent neural network
Yao, Xin	negative correlation learning; benchmark problems; many-objective optimization; evolutionary algorithm
Murino, Vittorio	generative model
Pedrycz, Witold	Neural network; genetic algorithm; feature selection; differential evolution; information granularity; main objective; optimal allocation; particle swarm optimization; justifiable granularity; time series
Yan, Shuicheng	learning algorithm; face recognition; visual classification; objective function; training data; learning framework; action recognition
Cho, Sung-Bae	activity recognition; mobile environment; bayesian network
Das, Swagatam	invasive weed optimization; multimodal optimization; search space; artificial bee colony; particle swarm optimization; differential evolution
Hong, Tzung-Pei	multiple minimum support; fuzzy data mining; mining algorithm; membership function; tree structure; real-world application; data mining; utility mining; high utility itemset; objective function; execution efficiency; pre-large concept; static database; original database; transaction deletion; synthetic dataset; search space; candidate generation; high-utility itemset
De Baets, Bernard	edge detection
Prade, Henri	fuzzy set; formal concept analysis

Finally, we collected those 70 authors' burst detection results as Table 5 shows. It includes Top 10 burst terms with the greatest number of related papers. Further calculations to define different types of research fronts are based on this table, and further discussion about this terms-attributes matrix will be introduced in the next part.

Table 5. Excerpt of the Terms-Attributes matrix

Terms	*No. P*	*No. A*	*EY*	*LY*	*AY*	*No. EA*	*No. LA*	*TC*
Neural network	9299	877	1996	2017	2009	42	252	123547
Genetic algorithm	3956	552	1996	2017	2009	6	95	74318
Support vector machine	3353	697	1998	2017	2010	1	105	63102
Face recognition	3005	467	1996	2017	2010	1	99	70280
Optimization problem	2984	665	1996	2017	2011	2	125	62349
Learning algorithm	2939	752	1996	2017	2009	5	127	48517
Computer vision	2726	579	1996	2017	2011	8	110	59396
Data mining	2457	471	1996	2017	2009	2	63	36280
Evolutionary algorithm	2430	318	1997	2017	2011	2	81	57629
Feature extraction	2401	628	1996	2017	2010	5	80	39395

Three types of research fronts

The most cutting edges are (or partly are) emerging research fronts. We find 13 emerging research fronts in AI. From Table 6, we can conclude that most of emerging research fronts are specific research directions oriented or definite application situations oriented. For example, *bat algorithm* is used to solve the global optimal solution, *high-utility itemset* is a kind of research directions of pattern mining, and *multi-component robotics system* is a sub-research direction of robotic system. In general, these emerging research fronts are characterized by less attention (the number of the earliest authors is less than 5, and the number of the latest authors is less than 7) and low citations (only three of them have total citations more than 800).

Table 6. Emerging research fronts

Terms	*No. P*	*No. A*	*EY*	*LY*	*AY*	*No. EA*	*No. LA*	*TC*
High-utility itemset	33	7	2014	2017	2016	5	6	89
Bat algorithm	26	12	2014	2017	2016	3	6	55
Transaction deletion	12	8	2014	2017	2016	2	4	29
Target individuals	9	2	2014	2017	2015	1	1	36
Actual sampling pattern	5	2	2013	2015	2014	2	2	314
Hesitant fuzzy linguistic term set	35	8	2011	2017	2015	2	5	838
Nonlocal self-similarity	23	21	2011	2017	2014	2	7	1267
Visual emotion challenge	20	11	2011	2016	2013	3	5	164
Deep learning	515	294	2010	2017	2016	1	106	3540
Malware detection	26	10	2010	2017	2015	1	4	143
Justifiable granularity	26	6	2010	2017	2014	3	1	218
Adversarial settings	8	3	2010	2016	2014	1	2	87
Multi-component Robotic system	7	1	2010	2015	2012	1	1	39

Before developing into hot research fronts, some emerging research fronts may turn into growing research fronts, whose growing rate (the G indicator) is no less than 3. G = 3 is a threshold could be acceptable in our test, which leads to a result of 19 growing research fronts. Table 7 depicts attributes of Top 10 research fronts with the most G value, and compared with emerging research fronts, growing research fronts gain much more attention from high yield authors.

Table 7. Excerpt from growing research fronts

Terms	*No. P*	*No. A*	*EY*	*LY*	*AY*	*No. EA*	*No. LA*	*TC*	*G*
Deep learning	515	294	2010	2017	2016	1	106	3540	15
Neural network	9299	877	1996	2017	2009	42	252	123547	10
Convolutional neural network	687	292	2001	2017	2016	1	109	4379	6.75
Optimization problem	2984	665	1996	2017	2011	2	125	62349	5.86
Learning algorithm	2939	752	1996	2017	2009	5	127	48517	5.81
Support vector machine	3353	697	1998	2017	2010	1	105	63102	5.47
Computer vision	2726	579	1996	2017	2011	8	110	59396	4.86
Big data	295	202	2004	2017	2015	1	62	1827	4.69
Face recognition	3005	467	1996	2017	2010	1	99	70280	4.67
Image classification	990	373	1997	2017	2012	1	91	17159	4.5

Note: the rest are: particle swarm optimization, genetic algorithm, sparse representation, evolutionary algorithm, feature selection, feature extraction, object function, dimensionality reduction and computational cost.

The threshold we select is PAC = 0.5, we define a term as hot research front when its PAC indicator is larger than 0.5, with which we get 38 hot research fronts in the end. In table 8, these research hotspots include some fundamental terms and method theories in AI, such as *neural network*, *learning algorithm*, *support vector machine*, *sparse representation* and so on.

Table 8. Excerpt from hot research fronts

Terms	*No. P*	*No. A*	*EY*	*LY*	*AY*	*No. EA*	*No. LA*	*TC*	*PAC*
Neural network	9299	877	1996	2017	2009	42	252	123547	3
Genetic algorithm	3956	552	1996	2017	2009	6	95	74318	1.66
Support vector machine	3353	697	1998	2017	2010	1	105	63102	1.66
Optimization problem	2984	665	1996	2017	2011	2	125	62349	1.58
Learning algorithm	2939	752	1996	2017	2009	5	127	48517	1.57
Computer vision	2726	579	1996	2017	2011	8	110	59396	1.43
Face recognition	3005	467	1996	2017	2010	1	99	70280	1.42
Feature extraction	2401	628	1996	2017	2010	5	80	39395	1.3
Feature selection	1774	576	1996	2017	2011	1	84	30604	1.1
Image segmentation	1730	497	1996	2017	2010	7	61	42367	1.1

Note: Table 8 lists Top 10 fronts with the most PAC indicator, for full version please email to us.

In general, these terms, which represent the essential theories and methods of the field of artificial intelligence, are the basis for any author to turn to this field for research. Therefore, from the perspective of a single author, these burst terms are the embodiment of his research interest or change of research direction, while from the perspective of the whole field, these research fronts represent the research hotspots in the field.

Conclusions

It is of highly valuable to research detections of fine-grained research fronts. In this paper, we identify several types of research fronts in a fine-grained way from the author-level aspect. Different from previous studies, we extract additional attributes of burst terms from their related papers, e.g., the number of related papers of a term, the number of related high yield authors, and the earliest year a term appeared. Based on these attributes, the earliest year of a term's appearance is used to define emerging research fronts, the G indicator for defining growing research fronts and the PAC indicator for hot research fronts. The results of our case study in AI have enlightenments to related researches: we get 13 emerging research fronts, illustrating specific research directions or definite applications; 19 growing research fronts, reflecting the transition from the emerging research fronts to the hot research fronts and attracting much attention at present and is likely to evolve into hot research fronts in the future; 38 hot research fronts, representing fundamental methods and theories in the field. The informative results could guide decision-makings for policy makers, administrators and other stakeholders. We only applied burst terms from the 70 representative high yield authors to explore research fronts, which brought the limitation that a larger dataset may lead to different results. Further works will put emphases on the result of more prolific authors, in the meanwhile, considering AI experts' advices in the process of defining research fronts from burst terms.

Acknowledgments

This study is supported by major bidding project of National Social Science Fund of China(16ZDA224) and Postgraduate Research & Practice Innovation Program of Jiangsu Province(SJCX18_0135).

References

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3), 359-377.

Chen, C., Dubin, R., & Kim, M. C. (2014). Orphan drugs and rare diseases: A scientometric review (2000–2014). *Expert Opinion on Orphan Drugs*, 2(7), 709-724.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, 22(2), 191-235.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?. *Journal of the American Society for Information Science and Technology*, 61(12), 2389-2404.

Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991). Mapping of science by combined co‐citation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science*, 42(4), 252-266.

Fang, Y. (2015). Visualizing the structure and the evolving of digital medicine: a scientometrics review. *Scientometrics*, 105(1), 5-21.

Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literatures II: Toward a macro-and microstructure for science. *Science studies*, 4(4), 339-365.

Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: a document co-citation analysis (2009–2016). *Scientometrics*, 115(2), 869-892.

Huang, M. H., & Chang, C. P. (2014). Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics*, 98(3), 1721-1744.

Huang, M. H., & Chang, C. P. (2015). A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation. *Scientometrics*, 102(3), 2041-2057.

Institutes of Science and Development, Chinese Academy of Sciences, The National Science Library, Chinese Academy of Sciences & Clarivate Analytics. (2017). *Research Fronts 2017*. Retrieved January 10, 2019, from http://clarivate.com.cn/research_fronts_2017/2017_research_front_en.pdf

Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics*, 65(2), 245-263.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4), 373-397.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation,* 14(1), 10-25.

Kajikawa, Y., Fujimoto, S., Takeda, Y., Sakata, I., & Matsushima, K. (2009, July). Detection of emerging research fronts in solar cell research. *In 12th International Conference on Scientometrics and Informetrics* (ISSI2009).

Mane, K. K., & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5287-5290.

Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American society for information science and technology*, 54(5), 413-422.

Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 510-515.

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986–1990. Journal of the American society for information science, 45(1), 31-38.

Physics World. (2016). *China forges ahead in global research*. Retrieved January 10, 2019, from https://physicsworld.com/a/china-forges-ahead-in-global-research/

Ran, W., Su, C., & Zhao, X. (2017). Comparison and application of research frontier identification method. *Chinese Journal of Medical Library and Information Science*, 26(11), 14-22.

Rip, A., & Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6), 381-400.

Song, J., Zhang, H., & Dong, W. (2016). A review of emerging trends in global PPP research: analysis and visualization. *Scientometrics*, 107(3), 1111-1147.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for information Science and Technology*, 60(3), 571-580.

Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2009). Detecting emerging research fronts in regenerative medicine by citation network analysis of scientific publications. *Portland International Conference on Management of Engineering & Technology*. IEEE.

Upham, S., & Small, H. (2009). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, 83(1), 15-38.

Man-woman collaboration practices and scientific visibility: how gender affect scientific impact in economics and management

Abdelghani Maddi[1], Vincent Larivière[2] and Yves Gingras[3]

[1] *abdelghani.maddi@hceres.fr*
Observatoire des Sciences et Techniques, Hcéres, Rue Albert Einstein, Paris, 75013 (France); CEPN, UMR-CNRS 723, Université Paris 13

[2] *vincent.lariviere@umontreal.ca*
Chaire de recherche du Canada sur les transformations de la communication savante, Université de Montréal

[3] *gingras.yves@uqam.ca*
Chaire de recherche du Canada en histoire et sociologie des sciences, UQÀM

Abstract

The question of the place of women in the various sciences is now widely discussed and studied. The field of economics and management is no exception and also requires a reflexive analysis of its practices. This study contributes to a better understanding of the place of women in these disciplines by characterizing the difference in levels of scientific collaboration between men and women (as measured by joint publications) in economics and management. First, the results show for the first time on an empirical basis that the practices of collaboration between men and women are quite different in management sciences compared to discipline of economics. Second, a regression analysis shows that there is a negative and statistically significant relationship between the Normalized Citation Score and the proportion of women per article in economics, which is not the case in management sciences. Results also show that international collaboration and the choice of journals significantly affect normalized citation scores.

Introduction

The question of the place of women in scientific research is widely discussed in all disciplines. The fields of economics and management are no exception and also require a reflexive analysis of its practices. Similarly, reflections on the productivity and "quality" of scientific research have become ubiquitous since the 1980s. Nowadays, no scholar can escape the evaluation of its activities (Pansu, 2013, Gingras, 2016). At the individual level, the "quality" of a scientific publication, an abstract property that some consider unmeasurable, can in fact be approached by measuring its "visibility", i.e. the number of citations it receives in publications by other members of the scientific community (Cronin, 1984). In the bibliometric literature, there are numerous analysis of scientific production both in terms of the choice of indicators and the analysis of the determinants of productivity and visibility of researchers, particularly the question of the difference between men and women (Cole and Zuckerman 1984, Xie et al., 1989, Leahey, 2006, Castilla and Bernard, 2010, Baccin et al., 2014, Mairesse and Pezonni, 2015, Nielsen, 2016, Nielsen, 2018). Despite the multiplicity of bibliometric studies devoted to the evaluation of research in economics and management, the analysis of the determinants of the visibility of articles has rarely been analyzed in these disciplines. Even less the study of the links between the impact factor of the journal, the social characteristics of the authors and the number of citations received.

Judge et al. (2007) analysis of the citation determinants of articles published in the top 21 management journals shows that the main factor in the visibility of an article is the journal in which it is published. Harzing (2016) shows that it is rather the topic studied by the article, as well as the profile of the author, that influences the visibility of publications in management. Starbuck (2005) and Singh et al. (2007) conclude that the evaluation of research articles based solely on the impact of journals provides erroneous results as to the "quality" of publications, given intra-review variability.

With regard to the link between scientific performances and gender, several recent studies have shown that gender gaps still persist in favor of men both in terms of productivity and scientific visibility (measured by the number of citations). Nielsen (2016) analyzes gender disparities in production, impact and scientific collaboration. It examines a sample of 3,293 Danish researchers (7,820 publications) of which 65% are men and 31% are women (4% indefinite). It shows the persistence of the gender gap in these indicators. Nielsen (2016) concludes that his findings raise deep concerns about the management of research organizations, characterized by an asymmetrical gender structure. This would call into question the validity of meritocratic explanations of discrepancies. For example, the age of the beginning of a scientific career directly affects the level of production of a researcher, as well as family commitments. According to Mairesse and Pezonni (2015), the gaps in production between men and women disappear if one control for differences in access to jobs and different working conditions between men and women.

Based on an econometric study, Nielsen (2017) analyzes the differences in academic impact in management sciences by gender. In a sample of nearly 27,000 publications and more than 6,500 authors, he concludes that women have a slightly greater impact than men, while remaining cautious about the representativeness of the sample and the possibility of generalization. Similarly, women have a larger share in the decile of the most cited publications in this area. However, in a more recent publication, Nielsen (2018) considers that the mere use of quantitative indicators can be very dangerous for the recruitment and promotion of researchers. Although these indicators may appear to be objective and reinforcing the "story of meritocracy", they are often biased ex-ante by the gender barriers. According to Nielsen (2018), these indicators must absolutely be accompanied by a qualitative assessment by peers.

Based on a large sample, the present study aims to (1) investigate the practices of collaboration between men and women in economics and management and (2) its effect on scientific visibility of the publications using an econometric model (Tobit regression). Our data, extracted from the Web of Science (WoS), cover global production as indexed in 300 journals in economics and 330 journals in management, with respectively 79,078 and 90,022 articles published between 2008 and 2015. A Tobit regression model was used to measure the relations between the different variables analyzed and the normalized score of citations.

Collaboration practices in economics and management

At the global level, scientific collaboration, measured by the number of authors per article, is relatively stronger in management than in economics. The proportion of articles co-published by at least two authors is 81% in management (almost half with at least 3 authors) against 66% in economics (see Figure 1). This is a global average and the results vary somewhat by country. This is a first interesting difference between these two disciplines.

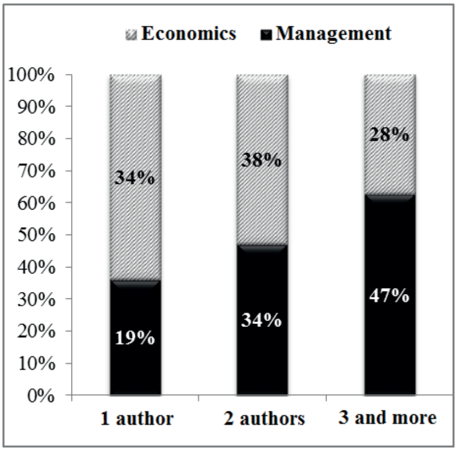


Figure 1: Number of authors per publication in economics and management

Figure 2 shows that the proportion of man-woman collaboration is much higher in management than in economics, 49% versus 27%. It also shows that the proportion of publication by women alone (without men collaboration) is three times higher in economics than in management (7% versus 21%). In economics, the majority of publications (52%) are signed by men and only 21% of women have published alone or with other women. In management, nearly half of the papers (49%) are the result of male-female collaborations. This is a second important observation on the disciplinary differences in male-female collaboration in these two disciplines.

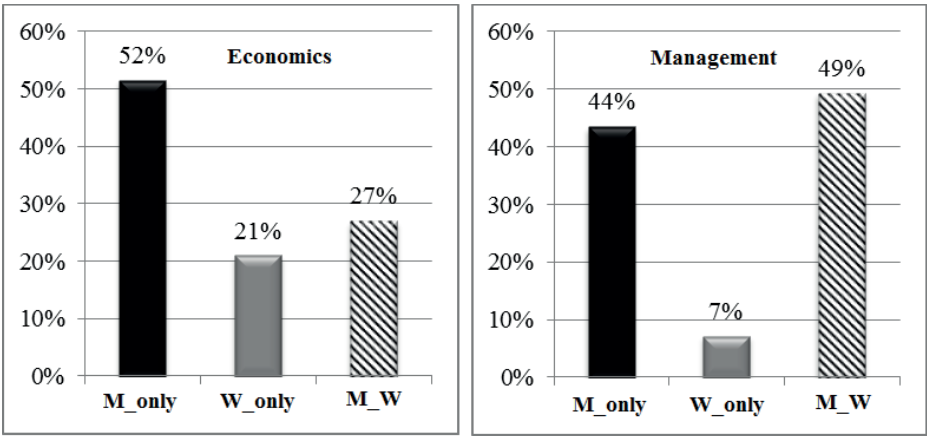


Figure 2: Collaboration between men and women in economics and management

Although the proportion of articles written in collaboration is greater in management, the distribution between national and international co-publications by gender is similar in both disciplines. This is true for co-publications that include only women, only men, or both (Figure 3). It should be noted, however, that women have a lower proportion of international publication than men.

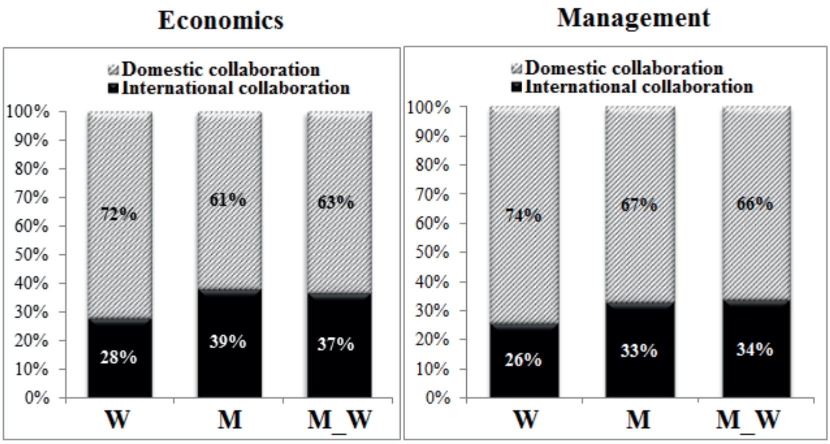


Figure 3: Proportion of articles co-authored according to the type of collaboration and gender

Figure 4 shows the gender distribution according to the CNRS classification of the journal in which articles are published. It can be seen that the proportion of articles published by women in the highest ranked CNRS journals (category 1) is slightly lower than that of men in the two disciplines, especially in management. Women in this discipline also publish more than men in journals classified in category 4, which are much less important in economics. We also observe that collaboration with men allows women to publish in higher-ranking journals.

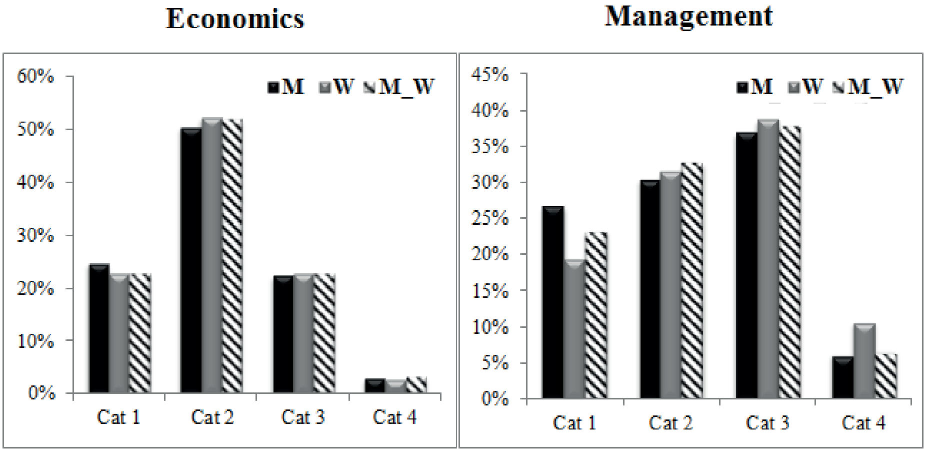


Figure 4: Proportion of articles by sex and CNRS categories of journals

Regression analysis

The database used for the regression analysis includes several information's about authors (names, gender and number), articles (publication year, title, Normalized Citation Score) and journals in which they are published (title, country of publisher, 2 years Journal Impact Factor and CNRS journal classification).

Dependent variable and model choice

The dependent variable is the logarithm of Normalized Citations Score (labelled Log (NCS)) received by each publication during the period 2008-2015. To retain the zeros, we have added 1 to the NCS before making the logarithmic transformation. Log (NCS) is a continuous variable with a lower boundary at zero and an upper boundary at infinity. Thus, a left censored Tobit regression model is used (see, McDonald and Moffitt, 1980) to account for the disproportionate number of observations with zero values, because a significant proportion of the observations in our sample are zeros. Tobit regressions avoid inconsistent estimates from Ordinary Least Square (OLS) regression.

Independent variable

In this study, we seek to analyze whether the gender of authors have an incidence on the number of citations received by scientific publications. To represent gender in scientific publications, we used the proxy of the proportion of women per publication labelled $Women\_Prop$. The authors' gender is assigned based on the methodology presented in Larivière et al. (2013), which uses the author's first name to assign a gender to them. For each of the articles in the two domains, we calculated the proportion of authors belonging to the feminine gender, using as denominator the sum of the authors to whom we assigned a gender. For example, an article with 5 authors, including two women, two men, and one unknown, was assigned a proportion of female authors of 0.5, leaving unknown cases out of the calculation. For an article co-signed by men only, the proportion is 0, and for an article whose all authors are women the proportion is 1. The values between 0 and 1 represent articles co-authored by both men and women. The higher the number of women per publication, the more the proportion is closer to 1.

For both disciplines, the proportion of women is lower than that of men. It is 32% in economics and 26% in management. These distributions are similar to the average of the global distribution of women researchers which is around 30% (UNESCO, 2018).

Control variables

A number of control variables are included in the model. The choice of control variables comes from the literature that shows that they are potentially associated with the number of citations received by publications. First, we control for the number of authors (Nbr_Authors) and the number of countries by publication, a proxy of international collaboration (Internat_collab). Second, we have controlled for the geographic origin of journals by building two dummy variables. The country of publisher was used as proxy of country of journals. US_Journal takes the value 1 if journal is American. $EU\_Journal$ takes the value 1 if journal is European. The non-American and non-European journals are the reference variables. Third, we control the impact of journals in which articles are published. In addition to the 2 years Impact Factor of the journal, we have constructed dichotomous variables to control for the journal classification of CNRS (2015). The CNRS classifies journals according to their degree of selectivity and importance in economics and management, thus providing a measure of their "quality". Four dummy variables were created. From $CNRS\_rank\_1$ that refers to the most selective journals in both disciplines, to $CNRS\_rank\_4$ that represent the least selective journals.

The regression model is written as follows:

$$\mathrm{Log}(NCS)_i = \beta_0 + \beta_1 Women\_Prop_i + \beta_2 Nbr\_Authors\ _i + \beta_3 Internat\_collab_i + \beta_4 US\_Journal_i + \beta_5 EU\_Journal_i + \beta_6 IF\_2_i + \beta_7 CNRS\_rank\_1_i + \beta_8 CNRS\_rank\_2_i + \beta_9 CNRS\_rank\_3_i + \beta_{10} CNRS\_rank\_4_i + \varepsilon_i$$

Using exactly the same variables, two distinct regressions were used for both disciplines. The aim is to analyze the differences between economics and management regarding the impact of gender on the citation score. The Table 1 resumes variables of model.

Table 1: Dependent, explicative and control variables of model

Dependent variable	
$\mathrm{Log}(NCS)_i$	Log transformed of NCS (Normalized Citations Score) by publication *i*
Explicative variable	
$Women\_Prop_i$	Proportion of women by publication. For example, for an article cosigned by 3 authors: 2 women and 1 man, the value will be 0.66 (66% of women). The value is between 0 and 1 (1 if all authors are women).
Control variables	
$Nbr\_Authors\ _i$	Number of authors by publication
$Internat\_collab_i$	International collaboration measured by the number of countries by publication
$US\_Journal_i$	It is dummy variable indicating the fact that the publisher of journal is American. It equal to 1 if it is.
$EU\_Journal_i$	It is dummy variable indicating the fact that the publisher of journal is European. It equal to 1 if it is.
IF_2	2 years Journal Impact Factor
$CNRS\_rank\_1_i$	It is a dummy variable representing categories 1, 1e, 1eg of the CNRS categorization of journals in Economics and Management. This category includes the most selective journals. It equal to 1 if it is, 0 otherwise
$CNRS\_rank\_2\ to\ 4$	Like $CNRS\_rank\_1$, variables $CNRS\_rank\_2\ to\ 4$ represents the journals of rank 2 to 4 of the CNRS classification. the degree of selectivity of journals decreases as the category increases

Before estimating the coefficients, we have verified the existence of multicollinearity. Multicollinearity is a problem that occurs when more than one of the model's predictor variables measures the same phenomenon. We are talking about multicollinearity when one of the explicative variables of model is a linear combination of one or more of the other variables introduced in the model. The absence of perfect multicollinearity is one of the conditions required to estimate a linear model. Two collinear variables are characterized in particular by a strong correlation. However, a strong correlation is not necessarily synonymous with collinearity. Both variables must, in addition, measure the same phenomenon. For example, the two variables $US\_Journal$ and $EU\_Journal$ are very negatively correlated (see Figure 5). This is normal since in our database a journal cannot be both American and European. On the contrary, the impact factor and the CNRS classification of journals measure more or less the same thing; the impact of journal. The difference between the two is that the impact factor is an objective measure based on the citations received by the journal, and that the CNRS classification incorporates a subjective dimension related to peer appreciation. The correlation test indicates that there is no strong correlation between the variables of model; all the correlation coefficients are less than 0.6 (see Figure 5: the larger and darker the bubble size, the higher the correlation).

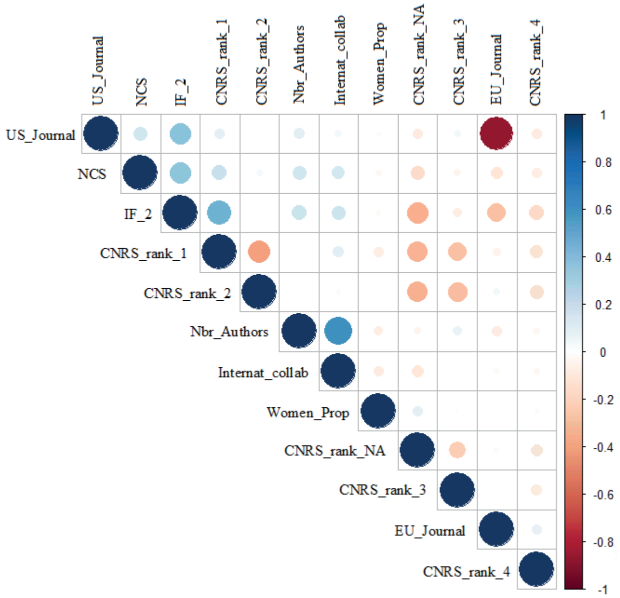


Figure 5: Correlation test of model variables

Results of regression analysis

In order to observe the interaction between the variables of the model, we chose to proceed by iteration. We can distinguish three types of explicative variables: sociological (proportion of women by publication, authors number and international collaboration), geographical (the fact that the journal is American or European) and bibliometric (impact of journals measured by Impact Factor and CNRS journal classification). This makes it possible to define three regression models; denoted respectively M1, M2 and M3. Tables 2 and 3 show, respectively, the regression results for economics and for management.

Table 2: Tobit maximum likelihood estimation, results for economics

Variables	(M1)		(M1) + (M2)		(M2) + (M3)	
	Coefficient	Pr (>|z|)	Coefficient	Pr (>|z|)	Coefficient	Pr (>|z|)
$Women\_Prop_i$	-0.065***	**6.71e^{-13}**	-0.057***	**1.8e^{-10}**	-0.034***	**3.41e^{-05}**
$Nbr\_Authors_{\ i}$	0.076***	**< 2e^{-16}**	0.065***	**< 2e^{-16}**	0.058***	**< 2e^{-16}**
$Internat\_collab_i$	0.043***	**< 2e^{-16}**	0.044***	**< 2e^{-16}**	0.011***	**0.00215**
$US\_Journal_i$	-	-	0.546***	**< 2e^{-16}**	0.176***	**1.65e^{-13}**
$EU\_Journal_i$	-	-	0.276***	**< 2e^{-16}**	0.107***	**< 2e^{-16}**
IF_2	-	-	-	-	0.290***	**< 2e^{-16}**
$CNRS\_rank\_1$	-	-	-	-	0.304***	**< 2e^{-16}**
$CNRS\_rank\_2$	-	-	-	-	0.290***	**< 2e^{-16}**
$CNRS\_rank\_3$	-	-	-	-	0.147***	**< 2e^{-16}**
$CNRS\_rank\_4$	-	-	-	-	0.010	0.55468
Wald-statistic	1425	**< 2.22e^{-16}**	3147	**< 2.22e^{-16}**	1.081e+04	**< 2.22e^{-16}**
Log-likelihood	-4.901e^{+04}		-4.818e+04		-4.471e^{+04}	

**** significant at 1% / ** significant at 5% / * significant at 10%.*

Tableau 3: Tobit maximum likelihood estimation, results for management

Variables	(M1)		(M1) + (M2)		(M2) + (M3)	
	Coefficient	Pr (>|z|)	Coefficient	Pr (>|z|)	Coefficient	Pr (>|z|)
$Women\_Prop_i$	-0.026***	**0.00114**	-0.026***	**2.47e-10**	0.001	0.823
$Nbr\_Authors_{\ i}$	0.040***	**< 2e^{-16}**	0.037***	**< 2e^{-16}**	0.026***	**< 2e^{-16}**
$Internat\_collab_i$	0.059***	**< 2e^{-16}**	0.062***	**< 2e^{-16}**	0.028***	**5.10e^{-15}**
$US\_Journal_i$	-	-	0.185***	**< 2e^{-16}**	0.086***	**< 2e^{-16}**
$EU\_Journal_i$	-	-	0.085***	**0.000529**	0.040***	**1.16e^{-05}**
IF_2	-	-	-	-	0.259***	**< 2e^{-16}**
$CNRS\_rank\_1$	-	-	-	-	0.201***	**< 2e^{-16}**
$CNRS\_rank\_2$	-	-	-	-	0.161***	**< 2e^{-16}**
$CNRS\_rank\_3$	-	-	-	-	0.100***	**< 2e^{-16}**
$CNRS\_rank\_4$	-	-	-	-	0.059***	**2.12e^{-08}**
Wald-statistic	1008	**< 2.22e^{-16}**	1592	**< 2.22e^{-16}**	1.081e+04	**< 2.22e^{-16}**
Log-likelihood	-9.039e+04		-9.01e+04		-8.314e+04	

**** significant at 1% / ** significant at 5% / * significant at 10%.*

Tables 2 and 3 show four important results that we can summarize as follows:

The impact of gender on citation scores

In economics, there is a negative and statistically significant relationship between the Normalized Citation Score and the proportion of women per article. In other words, the number of citations decreases as the proportion of women increases. The value of the coefficient (-0.034) of the variable $Women\_Prop$ means that when the proportion of women increases by one unit (1%), the NCS decreases by 3.4%. Thus, for example, for an article with three authors, 2 men and 1 woman ($Women\_Prop$ = 0.33), if the number of women increases by one unit ($Women\_Prop$ = 0.50), the NCS decreases by 5.78% (17% * 3.4%). It should be noted that the value of the coefficient decreases as one includes in the regression new groups of variables. We also note that the coefficients are statistically significant in the three models.

For management, the finding is different. Table 3 shows that if we do not take into account the impact of journal (M1 and M2) there is a negative and significant relationship between the Normalized Citation Score and the proportion of women per article. As soon as the variables Impact Factor and CNRS journal classification are integrated (M3), the coefficient of the variable $Women\_Prop$ becomes statistically insignificant. Therefore, all things equal otherwise, there is no evidence of a significant relationship, either positive or negative, between the gender and citation impact in management. This result may be due to the very strong collaboration between men and women in management: more than half of the publications in this discipline are the corollary of the collaboration between men and women (see Figure 2).

The importance of collaboration

Regression results show that, for both economics and management, citations are positively and significantly shaped by: the number of authors and the number countries involved in a publication. This result is true for the three models M1, M2 and M3.

However, some differences between the two disciplines are worth noting. In economics, the number of authors per publication is a stronger factor than the number of countries. The Normalized Citations Score increases by 5.8% when the number of authors per article increases by one, while the number of citations increases by 1.1% when the number of countries involved in the publication increases by one. In management both number of authors and number of countries, have similar coefficients (M3). Normalized Citations Score increase by nearly 3% when the number of authors or countries involved in the publication increases by one.

The country of journal and citation level

For both disciplines, the country of journal has a significant impact on citations. Thus, the fact that the journal in which the article is published is American or European increases the number of citations, which is not the case for journals published in any other country (that is outside US and EU). It is important to note also that citations increase faster if the journal is American than if it is European. In economics, if the country of the journal's publisher is American, the Normalized Citations Score is 17.6% higher than if it is neither American nor European. The percentage is 10.7 if the journal is European. In management, the percentages for American and European journals are much lower (respectively 8.6% and 4%).

The importance of the impact of the journal

The academic impact of the journal is the variable that most influences the number of citations received by articles. The more the journal in which the article is published has a high impact factor (or well classified by the CNRS), the higher the number of citations. Thus, in economics, citations increase by 29% if the impact factor increases by one. The rate is comparable (26%) in management. Likewise, citations increase as the journal is in the first classes of CNRS categories. However, for economics, the fact that the journal is classified in Category 4 of CNRS (class of least selective journals), has no positive or negative effect on citations. This means that there is no obvious gain from publishing in these journals (compared to journals not classified by the CNRS). This shows that this category 4 is very subjective and does not reflect a consensus within the community on the “quality” of these journals.

Discussion and conclusion

In this paper we have investigated the relationship between gender and citations received by academic papers in economics and management. The most striking results relate to the fact that the author's gender does effect the citations received. We observe that as the proportion of women per article increases, the citations tend to decrease, especially in economics. These results are consistent with previous works that has shown that, across all disciplines, women have less international collaboration than men and that the level of citations is higher for articles written in international collaboration (Larivière et al, 2011; -Salinas et al, 2011). This result is also valid in the natural sciences and engineering as well as in the health sciences (Beaudry and Larivière 2016). In a recent article, Mairesse and Pezonni (2015) showed that, in the case of physics in France, the difference in productivity of female physicists vanishes when other variables are taken into account, particularly inequalities in the chances of promotion of women to positions of full professor, and family commitments. One may wonder if the academic status (lecturer/assistant professor versus full professor) also influences the level of visibility. However, data is lacking to measure such an effect in our sample of nearly 170,000 articles covering two disciplines worldwide. Also, the choice of research objects may be different according to the gender. To take this effect into account and to neutralize it, it would be necessary to normalize the number of citations received by the subfield to which it belongs, which would require the topic of each article to be determined.

Otherwise, the lower impact of articles with a high proportion of women as co-authors may also be due to the fact that women are dealing with topics that have less prestige in the discipline. In field of management, Nielsen et al. (2019) have shown that women are well-represented in social- and human-centered areas of management, while men comprise the vast majority in areas addressing more technical and operational aspects.

Our data also highlight for the first time that the practice of collaboration between genders is quite different in economics and in management. While men publish among themselves in a similar way in both disciplines (about half of the articles are written between men only), we observe that in economics there are much less men-women collaborations (27%) than in management (49%) and therefore more collaboration between women only (21%) compared to only 7% in management. Explanations of such practices would require an in-depth, interview-based qualitative study, but highlighting such differences in collaborative practices is in itself an important result.

Our results also indicate that the visibility of research articles in economics and management is closely linked to the visibility of the journal in which they are published. This was to be expected because we know that there is a Matthew effect related to the journal impact factor (Larivière and Gingras, 2010). A more important result in the current context of bibliometric evaluation of research is the weight of American journals in the visibility of research articles both in economics and management. Indeed, if the journal is American, the citations to the articles will nearly double compared to a journal is European. It is likely that the important role of US journals (as the country of publication of the journal) in determining publication visibility as measured by citations is related to the fact that the WoS database (just like that of SCOPUS by elsewhere) has a strong Anglo-Saxon bias (Gingras and Khelfaoui, 2018). It remains true, however, that researchers' evaluations are, in fact, based on these databases. Our results are therefore all the more important as they may in turn influence the future publication practices of scholars in order to improve their "score" of citations.

References

Beaudry C., Lariviere V. (2016), “Which gender gap? Factors affecting researchers’ scientific impact in science and medicine”. Research Policy, 45(9), p.1790-1817.

Bornmann L., Daniel H.D. (2008) “What do citation counts measure? A review of studies on citing behavior”, *Journal of Documentation*, 64 (1) (2008), pp. 45-80.

Gingras Y. (2016). *Bibliometrics and Research Evaluation: Uses and Abuses*, MIT Press, Cambridge Massachusetts, London.

Gingras Y., Khelfaoui, M. (2018). “Assessing the effect of the United States’ “citation advantage” on other countries’ scientific impact as measured in the Web of Science (WoS) database”, *Scientometrics*, 114(2), pp. 517-532

Harzing A.W. (2006). “What, Who, or Where? Rejoinder to “Identifying Research Topic Development in Business and Management Education Research Using Legitimation Code Theory”, Journal of Management Education, 40 (6). pp. 726-731. ISSN 1052-5629.

Judge T.A., Cable D.M., Colbert A. E., and Rynes, S.L. (2007). “What causes a management article to be cited—article, author, or journal?”. *Academy of Management Journal*, 50(3), p.491-506.

Laband D.N., Piette M. J. (1994). “The Relative Impacts of Economics Journals: 1970-1990”, *Journal of Economic Literature*, 32 (2), p. 640-66.

Larivière V., Gingras Y. (2010). “The impact factor's Matthew effect: a natural experiment in bibliometrics”. *Journal of the American Society for Information Science and Technology*, vol. 61, no 2, p.424-427.

Larivière V., Ni C., Gingras Y., Cronin B., Sugimoto C.R. (2013). “Global gender disparities in science”. Nature, 504, p.211-213.

Larivière, V., Vigola-Gagné E., Villeneuve C., Gélinas P., Gingras Y. (2011). “Sex differences in research funding, productivity and impact: an analysis of Québec university professors”, *Scientometrics*, vol. 87 no 3, p. 483-498.

Leahey E. (2006). Gender differences in productivity – Research specialization as a missing link, Gender & Society, 20 (6) ,p.754–780.Merton R-K. (1968). “The Matthew effect in science: the reward and communication systems of science are considered”. *Science*, 159, p.56-63.

Mairesse, J. and Pezzoni M. 2015. “Does Gender Affect Scientific Productivity? A Critical Review of the Empirical Evidence and a Panel Data Econometric Analysis for French Physicists.” Revue économique 66(1): 392–96.

McDonald, J. F., and Moffitt, R.A. (1980). The uses of tobit analysis. The Review of Economics and Statistics, 62(2), 318–321.

Mingers J., Xu F. (2010) “The drivers of citations in management science journals”, *European Journal of Operational Research*, 205 (2) (2010), pp. 422-430.

Nielsen, M.W. (2016). Gender inequality and research performance: moving beyond individual-meritocratic explanations of academic advancement. Studies in Higher Education, 41(11), 2044–2060.

Nielsen M.W. (2017). Gender and citation impact in management research, *Journal of Informetrics*, Volume 11, Issue 4, November 2017, Pages 1213-1228

Nielsen, M.W. (2018). Scientific performance assessments through a gender lens: A case study on evaluation and selection practices in academia. Science and Technology Studies [preprint]. https://sciencetechnologystudies.journal.fi/forthcoming/article/60610/24863.

Nielsen, M.W. and Börjeson L. (2019). Gender diversity in the management field: Does it matter for research outcomes?, Research Policy (in press). https://www.sciencedirect.com/science/article/abs/pii/S0048733319300691

Rossiter-Margaret W. (1993). “The ~~Matthew~~ Matilda Effect in Science”, *Social Studies of Science*, Londres, Sage Publ, p. 325-341.

Singh G., Haddad K. M. and Chow C.W. (2007). “Are articles in “top” management journals necessarily of higher quality?”. *Journal of Management Inquiry*, 16(4), p.319-331.

UNESCO (2018). Women in science, Fact Sheet No. 51 June 2018 FS/2018/SCI/51.

Varying resonance chambers: A comparison of citation-based valuations of duplicated publications in Web of Science and Scopus

Stephan Stahlschmidt[1] and Dimity Stephen[2]

[1] *stahlschmidt@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a
10117 Berlin (Germany)

[2] *stephen@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a
10117 Berlin (Germany)

Abstract

Web of Science (WoS) and Scopus serve similar functions to the academic community as bibliometric databases, however their content and coverage is known to vary due to their owners' different business philosophies. In this paper, we investigated the impact of the differences in content and coverage between WoS and Scopus on citation-based bibliometric indicators. We calculated the excellence rates for ten countries and the sectors of the German science system to examine the macro-level effects of the differences. We then examined the impact at the micro-level by comparing the normalised citation counts between databases for the publications indexed in both WoS and Scopus by OECD discipline, using German publications from 2009 as an example. We found that WoS "favoured" base research as German sectors in this domain received higher citation-based impact values than in Scopus, while sectors that focused on applied research performed better in Scopus. WoS and Scopus apparently constitute two different resonance chambers and consequently the same content, given its orientation towards applied or base research, will receive a diverging citation-based impact value. Researchers need to consider these differences in focus when selecting a database for bibliometric analyses.

Background and purpose

Web of Science (WoS) and Scopus are both subscription-based databases of multidisciplinary bibliographic information that were developed to provide similar services to the scientific community, that is: to enable retrieval of relevant primarily scientific publications, to identify key journals, papers, authors, or institutions, and to facilitate bibliometric analyses. However, Clarivate Analytics and Elsevier, as owners of WoS and Scopus respectively, fundamentally differ in their business models, which influences the results of analyses using each database. Clarivate Analytics maintains that WoS will sufficiently capture the majority of important research simply by indexing the key journals in each discipline (Testa, 2018). As such, only approximately 10% of the 3,500 journals assessed for indexing in WoS each year are accepted into the 3 top-tier indices – Science Citation Index Expanded (SCIE), Social Science Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI; Testa, 2018). Conversely, Elsevier intended that Scopus should contain the largest number of records possible and so indexes 30%-70% of the approximately 25-250 journals assessed in any given month (Elsevier, 2017). As a result of their differing business models, WoS had indexed over 20,900 journals as of February 2019 (Clarivate Analytics, 2019) while Scopus had indexed more than 21,950 journals in the latest figures from August 2017 (Elsevier, 2017), with approximately 600 journals added each year since then (Elsevier, n.d.). This also translates to Scopus adding substantially more documents per year, although a large subset of publications are indexed in both databases, as can be seen in Figure 1.

These different business philosophies between Clarivate Analytics and Elsevier have implications for the content of the databases and the outcomes of analyses using them. While differences resulting from the variations in coverage between the databases have been studied extensively (e.g. Mongeon and Paul-Hus, 2016; Aman, 2016), the consequences for bibliometric indicators have mainly been studied in the context of assigning journals to disciplines (Donner, 2016; Wang and Waltman, 2016). Only recently in a case study on information retrieval, Bar-Ilan (2017) linked the differences in coverage to differences in citation counts. This study highlighted that a greater coverage not only increases the included publications of any entities to be analysed, but also increases the overall coverage and therefore alters the environment for any citation-based evaluation of scientific impact of the entities. Further, the changes in the overall coverage and in the coverage of an entity's publications might not be equivalent, but might result in either beneficial or adverse effects for the analysed entities.

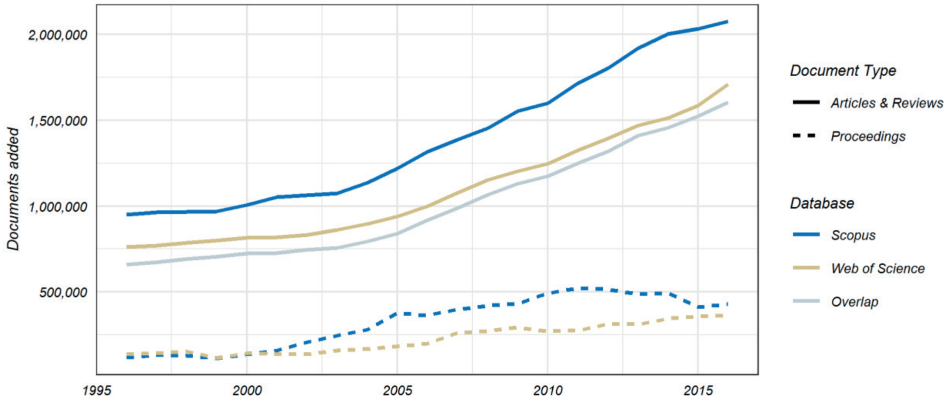


Figure 1. The number of articles and reviews and proceedings indexed in Scopus and WoS, and the overlap – the number published in both databases – annually between 1996 and 2016.[1]

That said, the bibliometric indicators currently commonly applied to gauge this citation-based scientific impact are mostly size-independent and implement this relative approach by relating a publication to a specific environment of similar publications. Due to differences in coverage, WoS and Scopus apply different environments to appraise the same publication, which is one driver of potential macro-level differences between the databases. The impact of different environments on the valuation can be measured by analysing the so-called "duplicated" publications covered jointly in both databases and comparing the database-specific valuation of these publications.

This is because the core differences between both databases consist of the respectively exclusive content. That is to say, any differences in the valuation of the same content result from differences in the respective environment, i.e. the exclusive content. Hence a comparison of the diverging valuation of the same content does not inform on the content itself, but on the exclusive content causing any differences and therefore the databases themselves. Therefore by understanding the deviations at the micro-level from general patterns observed between the databases at the macro-level, we not only explain these abnormalities, but en passant learn about any general difference in the coverage-based "character" of the databases which can facilitate a purposeful, instead of opportunistic, choice of a particular database to satisfy a particular research or evaluation end.

[1] The number of documents added to each database annually was calculated from the databases maintained by the German Competence Centre of Bibliometrics (http://www.bibliometrie.info). The method of determining the overlap in publications indexed in both databases is as described in the Methods section, however without restricting to German authors.

As such, here we examine two research questions. First, what are the macro-level differences in citation-based impact values between the WoS and Scopus databases, and secondly can the cause of the macro-level differences be revealed by comparing individual "duplicate" publications.

Method

The data included in the following analyses are document types 'article' and 'review' (jointly referred to as publications hereafter) published in journals for the relevant publication years. Data were extracted from the Scopus and WoS (SCIE, SSCI and A&HCI indices) databases maintained by the German Competence Centre of Bibliometrics. Fractional counting was used to assign publications to countries or sectors based on each author's reported institution. A citation window of 3 years was used for citation-based indicators. Self-citations have not been excluded.

We first calculated excellence rates to gauge the impact of the differences between the databases at the macro-level of countries and sectors of the German science system. Excellence rates, also known as Highly Cited publications or PP(Top10), are the share of publications belonging to the 10% most highly cited publications in each discipline. We calculated annual excellence rates from 2007 to 2015 based on citations from WoS and Scopus for ten countries with high levels of academic output, and the six German research sectors: Universities of Applied Sciences, or "Fachhochschulen" (FH); Helmholtz Association of German Research Centres with its focus on research infrastructure (HGF), universities (UNI); Frauenhofer Society in applied science (FhG); Max-Planck Society in basic science (MPG); and Leibniz Association that connects basic and applied science (WGL). We used the method described by Waltman and Schreiber (2013) to proportionally assign publications on the 90th percentile threshold to achieve exactly the top 10% most frequently cited publications.

To gain insight into the cause of any macro-level differences observed, we analysed how the same content is evaluated by both databases. As a case study, we identified duplicate publications, i.e. publications indexed in both WoS and Scopus, published by authors in Germany in 2009 and calculated the database-specific valuation of these publications. Publications indexed jointly in WoS and Scopus were identified by comparing hash values on a subset of the available metadata strings. Fractional counts of publications were then aggregated to the OECD Fields of Science and Technology disciplines using the database providers' official mappings to convert the respective database-specific classification to the OECD classification. Items in Scopus' 'multidisciplinary' category have been excluded as this category is not mapped to any OECD category. Also, a small proportion of items unclassified in WoS were also excluded. Due to differences in the mappings of the database providers' classifications to the OECD classification, some duplicated publications might be assigned to different disciplines and are consequently included in the databases' discipline-specific shares of exclusive publications.

We then normalised every duplicated German publication from 2009 in the two environments defined by the exclusive publications. In detail, we computed for every duplicated article i affiliated to a German address in 2009 the ratio

$$\frac{obtained\ citations_i^{(s)}}{expected\ citations_i^{(s)}}, \qquad (1)$$

where s denotes the source of citation and expected citations counts, i.e. WoS or Scopus. Expected citations were computed as the mean number of citations received by articles from

2009 in the three-year post-publication period 2009-2011 that were assigned to the same OECD discipline.

By varying s we obtain a separate citation-based valuation of a duplicated German publication i. Hereby the variation in obtained citations and expected citations could differ, because the exclusive share of publications by each source s might have a varying effect on the general citation level in a discipline expressed in the expected citation counts and the particular impact of a publication i in that context, which might be stronger or weaker than the change in expected citations. Indeed differences in the change of obtained and expected citations facilitate changes in the valuation of publication i, as a uniform increase (or decrease) in obtained and expected citations would not change the ratio and the resulting valuation of the respective publication would stay constant.

Finally, given equation (1), we individually contrasted every duplicated 2009 German article with its corresponding expected citation count, and calculated the percentage of duplicated German articles with difference Δ in normalised citations $norm.\ cit.$ based on

$$\Delta\ norm.\ cit. = \frac{obtained\ citations_i^{(Scopus)}}{expected\ citations_i^{(Scopus)}} - \frac{obtained\ citations_i^{(WOS)}}{expected\ citations_i^{(WoS)}}.$$

Results

The excellence rates for the selected countries are depicted in Figure 2, and show that differences are evident between the databases at this macro-level. For example, the Netherlands and Italy clearly benefited from the use of Scopus, while for China higher impact values were observed in WoS.

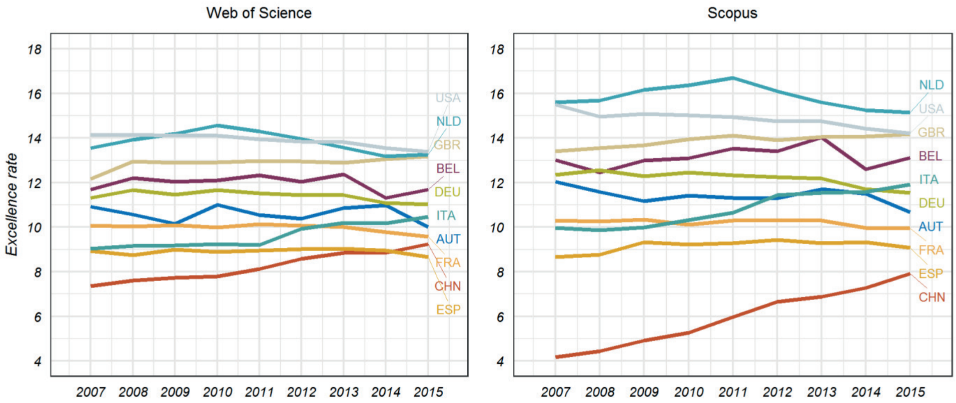


Figure 2. The excellence rates of selected countries between 2007 and 2015 calculated based on publications indexed in WoS (left panel) and Scopus (right panel).

Figure 3 likewise illustrates the excellence rates of the German research sectors. Similar effects to Figure 2 can be seen: while the MPG maintained a relatively high and stable citation-based impact in WoS, in Scopus it declined since 2009. At the same time we might identify a level shift in the impact of the FhG, as its impact was constantly and relative to all other sectors higher in Scopus than in WoS. It is also apparent that the largest effect of the database choice was a uniform level effect in which nearly all entities analysed in Figures 2 and 3 found their citation-based scientific impact was higher in Scopus than WoS. However, as seen, some entities benefit (or lose) relative to other entities from a particular choice of a database.

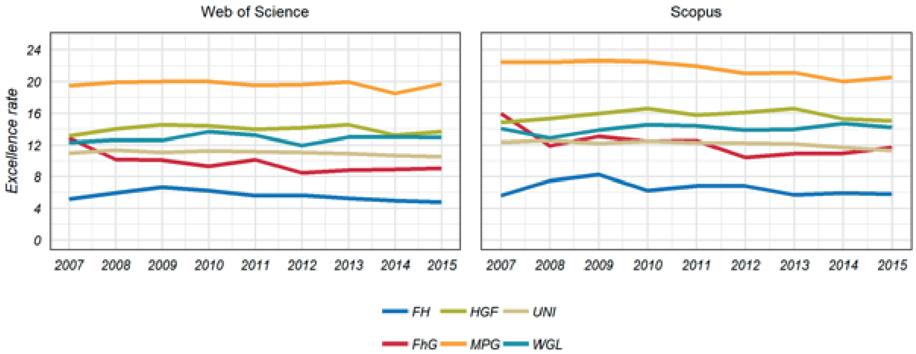


Figure 3. The excellence rates of sectors of the German science system between 2007 and 2015 calculated based on publications indexed in WoS (left panel) and Scopus (right panel).

We sought insight into the cause of these macro-level differences by examining the number of duplicated German publications and contrasting their citation-based values between the databases to observe the particularities induced by the databases' respective exclusive publications. Figure 4 presents the number of overlapping publications and the exclusive publications indexed in only one of the two databases by OECD disciplines. A strong focus on natural sciences and medical and health sciences can be observed. The share of more application-oriented engineering and technology sciences was lower, still, they maintained a substantial share of indexed publications. On the contrary, relatively few publications were indexed for agricultural sciences, social sciences or humanities.

Comparing the distribution of duplicated publications with the exclusive publications, we observed few differences. In terms of publication numbers also, the natural sciences and medical and health sciences had by far the highest share of exclusive publications, the engineering and technology sciences obtained a lower share, and agricultural sciences, social sciences and humanities the lowest share. Furthermore the share of additionally indexed publications in Scopus exceeded in nearly every discipline the exclusive share of the WoS by a wide margin. While, for example, around 2.25 million articles and reviews were assigned to the discipline *Clinical medicine* in both WoS and Scopus for the years 2007 to 2016, Scopus listed another 3.25 million records in this category, while WoS added less than 0.5 million additional records to this particular category. In addition Scopus indexed another 2.25 million records in the category *Other medical sciences*, which does not exist in the WoS mapping. Consequently, Scopus held a much larger number of articles and reviews in the OECD field *medical and health sciences*, a higher-level agglomeration of all medical and health-related OECD disciplines, than WoS and this observation also holds for all other OECD fields.

In general the additional share of publications indexed in Scopus mimicked the distribution of the overlap, adding proportionally more publications by discipline. The numbers of exclusive publications of the WoS were smaller in magnitude and more evenly distributed, at least among the natural sciences, medical and health sciences and engineering and technology. Exclusive publications in agricultural sciences, social sciences and humanities were hardly detectable. The exceptions to these observations were mostly of the residual classes *Other* * and occasions in which mappings do not make uniform use of a certain OECD discipline. In general the corpus of publications indexed by WoS hardly diverged much from the corpus of duplicates, while Scopus presented a much larger corpus and deviated proportionally from the overlap.

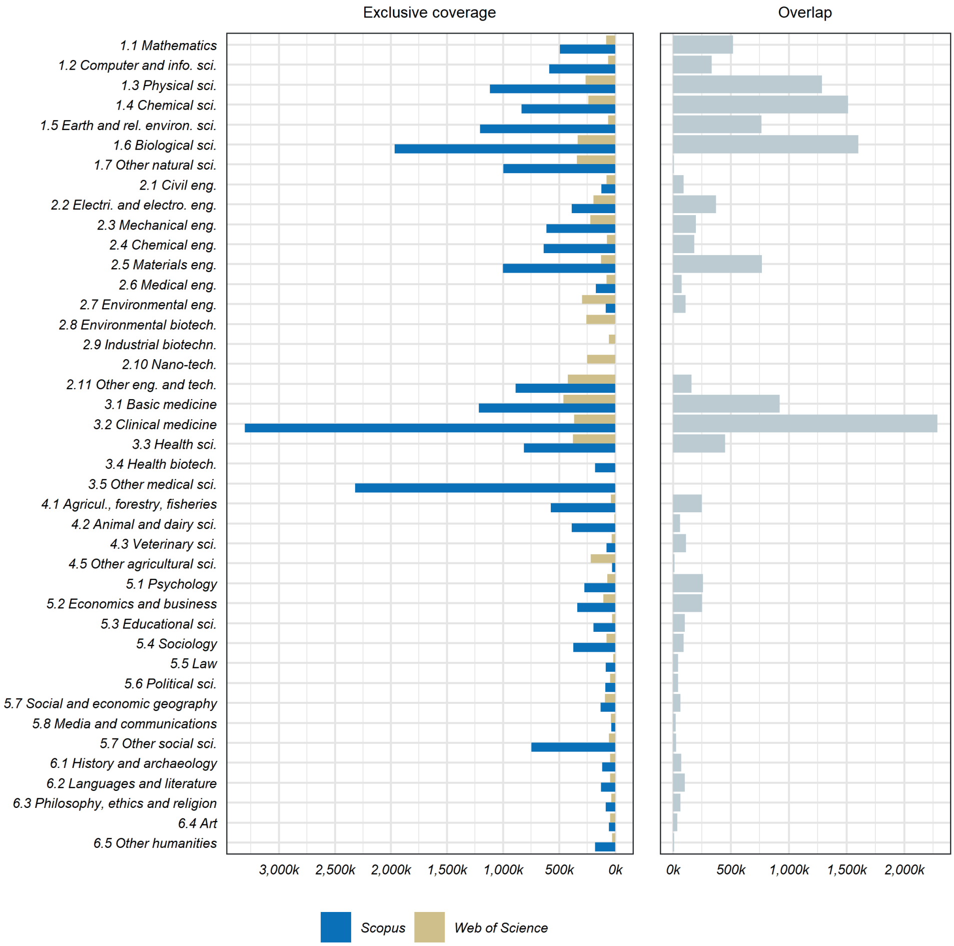


Figure 4. The number of articles and reviews exclusive to Scopus and WoS (left panel), and the number indexed in both databases (right panel) by OECD discipline for the years 2007-2016.

We then compared absolute citation counts and expected citation counts from the WoS and Scopus, as presented in Figure 5. In both graphs there is a strong positive correlation underlining findings, that in general both databases presented a fairly similar picture of national bibliometric evaluation. However some publications obtained more (or fewer) citations in Scopus than postulated by the otherwise linear relation. Also the expected citations of the mapped OECD disciplines followed a clearly positive correlation, although most disciplines possessed a higher expected citation count in either WoS or Scopus than postulated by the linear relationship. These deviations in the obtained or expected citation counts from the mean line caused a non-uniform change in the ratio of equation (1) and consequently a different valuation of German publications in the respective setting.

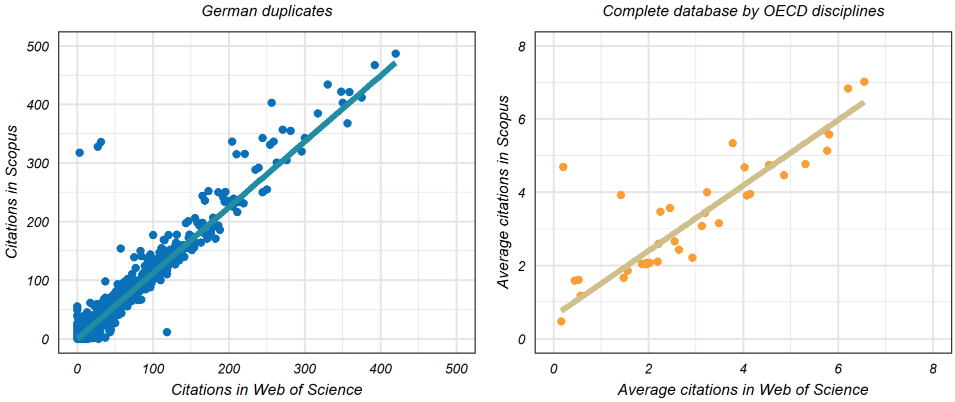


Figure 5. The number of citations for duplicated 2009 German articles in Scopus and WoS (left panel), and the average number of citations by discipline from each database (right panel).

We explored any accruing discipline-specific deviations from the linear trend in Figure 6. To ensure the sample size of publications for a discipline is sufficiently robust for interpretation, we present results for the 19 OECD disciplines with more than 400 duplicated articles affiliated with a German institution in 2009. The difference in raw citation counts is depicted by the blue bars, the mean difference in raw citations is the orange dotted line, and the dark yellow line shows any difference in the expected citation counts. More than 40% of such publications in the OECD discipline *Economics and business* obtained exactly the same number of citations in both databases. Twenty percent of German duplicates in *Economics and business* received one additional citation in Scopus. While in all disciplines the largest share of publications obtained exactly the same number of citations, we also consistently found a right-skewed distribution, in which the aforementioned German publications received in general more citations in Scopus than in WoS. The orange dotted lines depict the mean of these differences and consequently summarise the distributional effects in the blue bars in a single number. By comparing disciplines it might be noted that Scopus favoured German publications in raw citation counts in some disciplines more than in others. Duplicated German publications in *Clinical medicine*, *Economics and business* or *Computer and info. science* seemed to benefit especially from Scopus, while duplicated German publications in *Veterinary science* obtained few additional citations in Scopus.

Any difference in the raw citation count of a duplicated publication represents an altered standing of these duplicated publications in the different environments. General changes in the environments are expressed via differences in the expected citation counts, which are depicted via the dark yellow line in Figure 6. In general these changes in the expected citation counts seemed less pronounced and were sometimes even negative. In any case the additional, exclusive publications by Scopus altered the expected counts because they differ structurally, e.g. in citations over time or between disciplines, citations to non-indexed publications or different citation potentials due to longer or shorter reference lists, from the overlapping set of publications. Via these general changes in expected counts, as well via changes to the raw citation counts of duplicated German publications, the content exclusive to each database defines differences in the national bibliometric evaluation of any analysed entities.

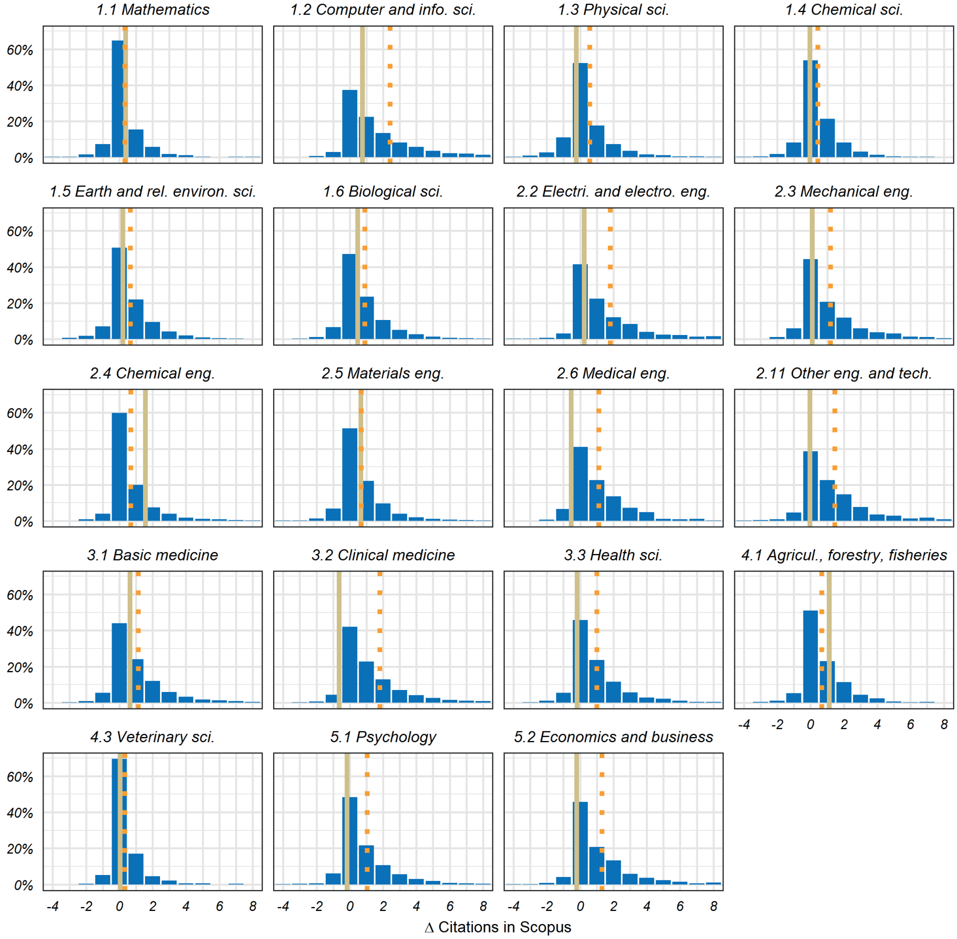


Figure 6. The difference in raw citation counts (blue bars), mean difference in raw citations (dotted line), and difference in expected citation counts (dark yellow line) between WoS and Scopus by OECD discipline.

The distribution of the percentage of duplicated 2009 German article with difference *Δ* in normalised citations $norm.\ cit.$ between databases, as denoted by the blue bars, is depicted in Figure 7. Accordingly more than 60% of German duplicated articles in *Veterinary science* found their normalised citation impact altered in the range of -0.2 to 0. The blue dotted line indicates the mean of this distribution of differences. As to be expected by Figure 6 the mean difference was positive for most disciplines. Duplicated German articles in the discipline *Agriculture, forestry, fisheries* observed on average the most severe negative effect on their citation-based impact stemming from the use of Scopus and its set of exclusive publications. This might also be seen in Figure 6, where the average in additional raw citations for publications in this discipline was outrun by the change in expected citation counts. The same observation also holds for *Mathematics* and *Chemical eng.*, while articles in most other disciplines observed a positive average effect in line with the higher shares in the German excellence rate reported for Scopus than WoS in Figure 2.

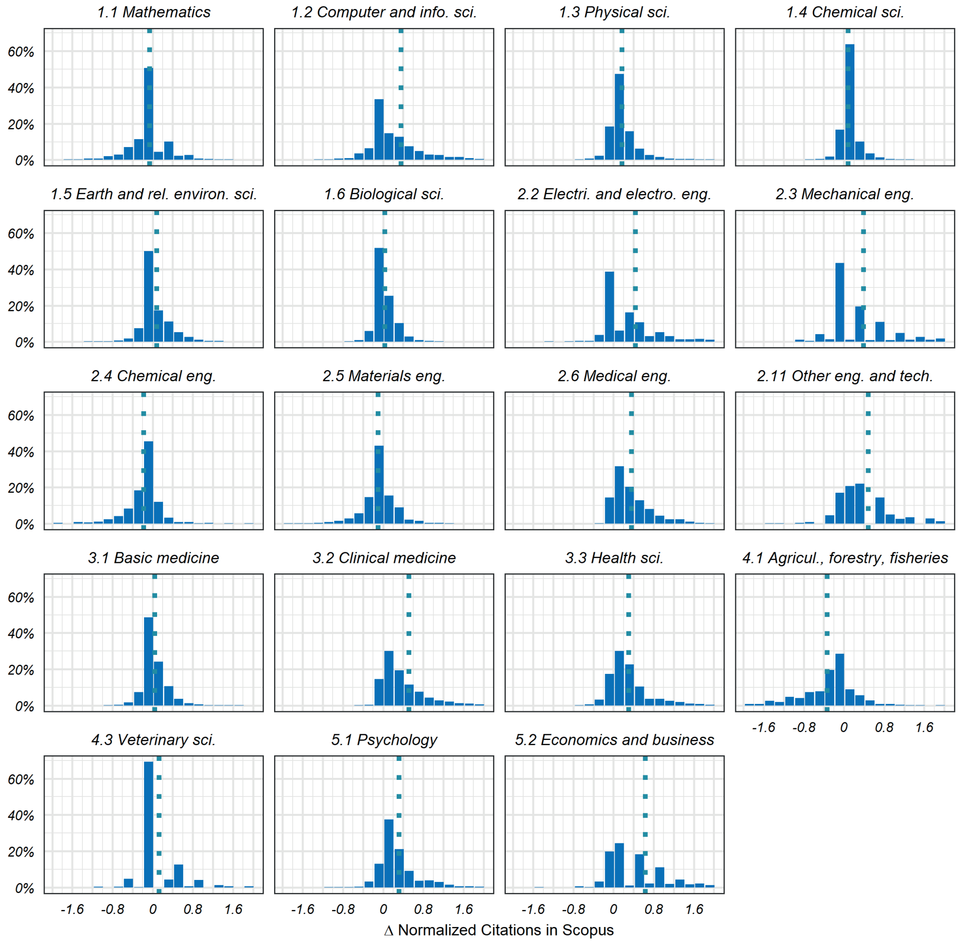


Figure 7. The distribution of duplicated German articles with difference Δ in normalised citations by OECD discipline.

The publication level rationale for the increase in the excellence rate for Germany can be observed in Figure 8. The top panel shows the distribution of $\Delta norm.cit._{(i)}$ for every duplicated 2009 German article. The 40%-60% quintile, depicted in the darkest blue shade, starts right above the zero line of no effect and almost reaches up to the 0.1 line. Given that a normalised citation count of 1 for a publication is commonly interpreted as exactly reaching the discipline-specific citation expectation an increase of 0.1 might be interpreted as 10% increase in this indicator. Consequently the 20% most strongly affected German duplicates improved on their normalised citation impact by over 25% in Scopus, while the 20%-40% quintile of affected German publications had reductions on their normalised citation impact of -5% to zero. Consequently the distribution is skewed to the right allowing duplicated German publications to obtain on average a higher normalised citation impact in Scopus than WoS.

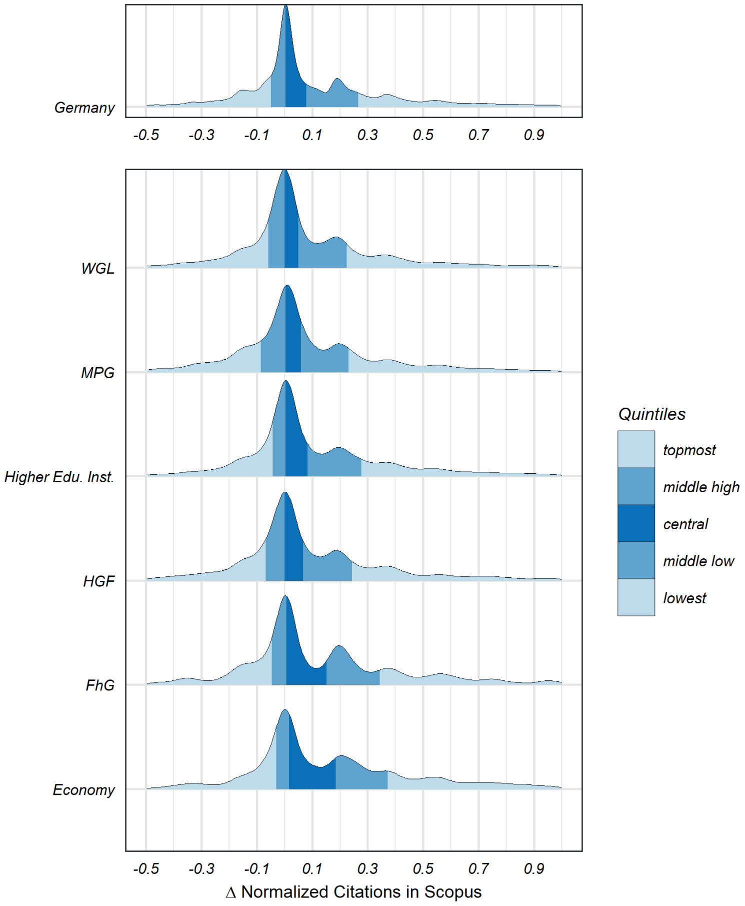


Figure 8. The distribution of $\Delta norm.cit._{(i)}$ for every duplicated 2009 German article (top panel), and by research sector (bottom panel) with quintiles denoted.

The lower panel of Figure 8 subdivides the duplicated 2009 German articles by the aforementioned sectors. In general all sectors showed similar distributions to the national one. The distribution of the Higher Education Institutions sector, which is dominated by German universities, but also includes universities of applied sciences or specialised schools/colleges, most closely resembled the national distribution, as these institutions are responsible for the largest share of German publications. Surprisingly the MPG obtained the weakest effect on increased normalised citations of all sectors, although its excellence rate clearly outperformed all other sectors. On the contrary the strongest effect was observed for the publications affiliated with the German business sector (denoted by "Economy") and the FhG, which according to their excellence rate, found themselves on the lower half of citation-based impact assessment.

Discussion

The purpose of this study was to first gauge the magnitude of macro-level differences in citation-based impact values between the WoS and Scopus databases due to differences in their content and coverage, and secondly determine the origin of these differences through a comparison of the valuations of publications indexed in both databases. With regard to the macro-level differences, in calculating excellence rates for countries and German research sectors, we found that nearly all of the entities examined experienced a uniform level effect with higher impact values in Scopus than WoS, although some entities gained or lost relatively to others dependant on the database. Further, a comparison of the citation-based valuations of duplicate publications between databases revealed the macro-level differences occur as the databases each “favour” a domain of research.

For example, the FhG and the German business sector might be distinguished by their missions, respectively intrinsic needs, from other German sectors by their greater orientation towards applied research, translation or technology. Their publications, as observed in Figure 3, lag behind most other sectors in their overall citation impact in WoS. But at the same time their duplicated publications indexed in WoS and Scopus observed the highest gains in Scopus among all other sectors. Consequently the resonance of these duplicated publications stemming from an applied research environment is higher in Scopus than in WoS. Hence the particular selection of exclusive publications by Scopus shows a stronger interest in applied research, translation and technologies and therefore ultimately the Scopus database itself might be differentiated by its stronger focus on applied research, translation and technologies.

On the other hand, the MPG excels in WoS citation-based scientific impact, as given its mission it focuses nearly exclusively on base research. However, its duplicated publications gain least by switching to Scopus. Further, its Scopus-based excellence rate in Figure 3 has constantly declined since 2009 despite the steadily increasing difference in indexed publications seen in Figure 1. The number of publications exclusively indexed by Scopus grows increasingly in the same timeframe, as the excellence rate of the MPG decreased, i.e. while Scopus increasingly differentiates from WoS and establishes its own “character”, the base-research-orientated MPG seems to increasingly lose from this differentiation while its excellence rate in WoS stays constant. Consequently WoS seems to focus more on relatively highly cited base research.

WoS and Scopus apparently constitute two different resonance chambers and consequently the same content, given its orientation towards applied or base research, will receive a diverging citation-based impact value. WoS focuses on relatively highly cited base research and will value any such research higher, while Scopus – with its stronger focus on national or regional journals which may seek to answer applied research questions in the local context context – seems to add an outer ring of applied research, translation or technologies and will appraise respective publications higher. It is up to the researcher then to decide which database best suits their project, given their research or evaluation objectives.

Acknowledgments

The authors like to thank the German *Commission of Experts for Research and Innovation (EFI)* for funding this study.

References

Aman, V. (2016). *Regional Coverage of Authorship in Wos and Scopus: Report Prepared for Thomson Reuters' Tender on ‘Web of Science vs. Scopus Comparison’*. Berlin: Institute for Research Information and Quality Assurance.

Bar-Ilan, J. (2017). Bibliometrics of ‘Information Retrieval’–A Tale of Three Databases. In P. Mayr, M. K. Chandrasekaran & K. Jaidka (Eds.), *Proceedings of the 2nd Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (Birndl)* (pp. 83–90). Tokyo: CEUR Workshop Proceedings.

Clarivate Analytics. (2019). Web of Science: Summary of Coverage. Retrieved April 3, 2019 from https://clarivate.libguides.com/webofscienceplatform/coverage

Donner, P. (2016). *WoS vs Scopus – Subject Area Coverage: Report Prepared for Thomson Reuters for the Tender 'Web of Science vs Scopus Comparison'.* Berlin: Institute for Research Information and Quality Assurance.

Elsevier. (2017). *Scopus: Content Coverage Guide.* Retrieved September 24, 2018 from: https://www.elsevier.com/__data/assets/pdf_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf.

Elsevier. (n.d.). How Scopus works: Content. Retrieved 3 April, 2019 from https://www.elsevier.com/solutions/scopus/how-scopus-works/content

Mongeon, P. & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics,* 106(1), 213–228.

Testa, J. (2018). *Journal Selection Process*. Retrieved September 24, 2018 from: https://clarivate.com/essays/journal-selection-process/.

Waltman, L. & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology,* 64(2), 372–379.

Wang, Q. & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, 10(2), 347–364.

Detecting Key Topics Shifts in Thermal Barrier Coatings (TBC) as Indicators of Technological Advancements for Aerospace Engines

K. A. Khor[1] and L. G. Yu[1]

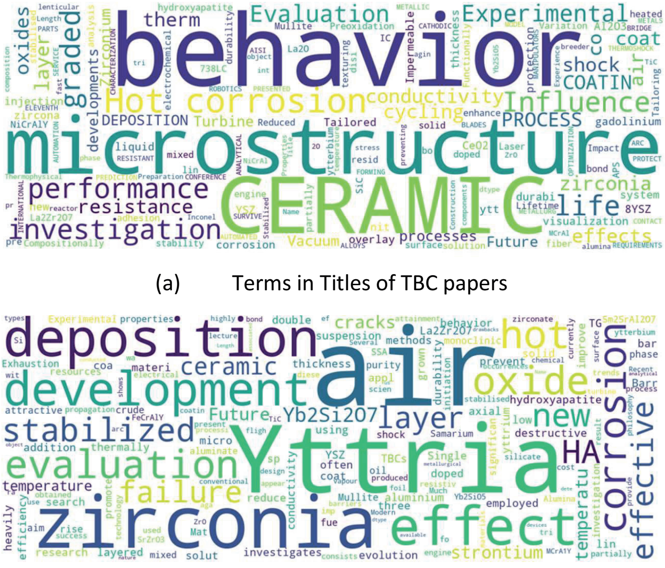
[1] *mkakhor@ntu.edu.sg; mlgyu@ntu.edu.sg*
Talent Recruitment and Career Support (TRACS) Office and Bibliometrics Analysis, Nanyang Technological University, #B4-01, Block N2.1, 76, Nanyang Drive, Singapore 637331 (Singapore)

Abstract

Thermal barrier coating (TBC) systems reduce the temperature of the metallic substrate of modern gas turbines, resulting in improved component durability and increased efficiency. This study examine the research publications and patents on thermal barrier coating in Scopus and USPTO patent database, respectively, during the period 1981-2018. Bibliometrics analysis reveal the developments in country and institution contribution to thermal barrier coating research as well as the trends in topics evolution and technology advancement. Results show that USA, China, Germany, Japan, India and United Kingdom are the top countries in thermal barrier coating research, and the Chinese Academy of Science, Beihang University, Forschungszentrum Juelich, NASA Glenn Research Center and Deutsches Zentrum fuer Luft- Und Raumfahrt are among the leading institutions that have high scholarly research output during 1981-2018. While USA dominated the TBC publication output during 1981 to 2010, China takes the lead after 2011 on TBC papers, and India took over Germany to become the third leading country on TBC research in 2018. The terms of the titles, keywords and abstracts of the publications, mapped visually using the VOSviewer reveal the progress of thermal spray technology. Results show that atmospheric plasma spraying (APS) and electron beam physical vapour deposition (EB-PVD) are well-established deposition techniques for TBC coating preparation, and plasma spray physical vapour deposition (PS-PVD) and low-pressure plasma spraying (LPPS) are two new emerging deposition techniques in TBC research. The journal papers cited by patents have low correlation to the highly cited papers in Scopus.

Introduction

Thermal barrier coatings (TBC) are refractory and thermally insulating ceramic coatings deposited on substrates to protect components from high temperatures (Shiembob and Hyland, 1979; Miller, 1987; Goswami, Ray & Sahay, 2004; Guo, Gong & Xu, 2014). TBCs are primarily a two layer system consisting of a ceramic top coat layer of yttria partially stabilized zirconia (YSZ) and an alumina forming bond coat layer, primarily of NiCoCrAlY or NiAlPt based compositions (Guo, Vassen & Stover, 2004; Padture, Gell & Jordan, 2002; Darolia, 2013; Busso, Wright & Evans, 2007). Atmospheric plasma spraying (APS) and electron beam physical vapour deposition (EB-PVD) are the two main processes in preparing TBCs (Padture, Gell & Jordan, 2002; Bose & DeMasiMarcin, 1997). TBCs are widely employed to lend thermal protection to metallic surfaces in aeroplane engines and gas turbine parts. Over the years, the TBCs have gradually developed from merely insulating layers to more complex designs. By considering factors such as heat flux, heat transfer coefficients, backside cooling, part geometry and location, coating thickness and its thermal conductivity, higher thermal efficiency is achieved (Golosnoy, Cipitria & Clyne, 2009). New materials (Clarke & Phillpot, 2005; Wang, Lu, Huang & Xie, 2018; Gurak, Flamant & Laversenne, 2018), structures and manufacture processes (He, Guo & Peng, 2013; Bakan & Vaßen, 2017; Fauchais, Montavon & Lima, 2011; Markocsan, Gupta & Joshi, 2017) on TBCs are explored for application in future aeroplane engines. Currently, turbine engine heat-components based on Ni-based superalloys have reached their upper limit of temperature capabilities. Therefore, alternative materials such as SiC(fibre)/SiC ceramic matrix composites (CMCs) are investigated as next generation turbine engine heat-component (Murthy, Nemeth & Brewer, 2008). Environmental barrier coating (EBC) systems deposited on top of CMC components protects the Si-based CMCs from water vapor attack at high temperatures (Suzuki, Sodeoka & Inoue, 2008; Zhang, Zhou & Liu, 2017; Appleby, Zhu & Morscher, 2015). Along with the technology progress is a rapid growth in global research paper output and citation impact on thermal barrier coatings publications. Figure 1 give a Wordle representation of terms used in abstracts and titles of thermal barrier coating publications. This Wordle representation clearly shows that ytrria stabilized zirconia (a ceramic) as the dominant thermal barrier coating material, while new materials like $La_2Zr_2O_7$ is emerging as high temperature coatings.



(a) Terms in Titles of TBC papers

(b) Terms in Abstracts of TBC Papers

Figure 1. Terms in titles and abstracts of thermal barrier coatings

Research papers on thermal barrier coatings over the past 38 years (1981-2018) from Scopus provide an understanding of the global research in this field. The publication data enable analysis of annual scholarly outputs and impact, mainstream journals, leading countries and institutions. Patents published by USPTO on thermal barrier coating related topics are downloaded, and analysed. Special focus placed on the citations of journal papers in patents and the co-relation with highly cited papers on thermal barrier coatings in Scopus.

Method and Data Set

Papers on thermal barrier coatings

Publications on TBC found in Scopus during 1981-2018 are examined using the query “((TITLE-ABS-KEY (“thermal barrier" OR "thermal barriers") AND TITLE-ABS-KEY (coating OR coated OR coatings OR coat)))”. The total number of articles found is 8966. The extracted publications are analysed for countries and researchers (Institutions) collaborations and FWCI (by uploading these papers to SciVal).

Patents on thermal barrier coatings

The USPTO Patent Full-Text and Image Database (PatFT) Quick Search engine is the data source. The searched term “thermal barrier coating” is in “All Fields” and “1976 to present [full-text]” specified as search period. Analysis performed on over 2,940 patents. The abstract and reference citation information for the patents with "thermal barrier coating" in the patent title (around 340 patents) are downloaded for term and reference analysis. The titles, publication year and abstracts of the patents are inserted into the Web of Science publication data template for term analysis.

Journal and Conference Publications on Thermal Barrier Coatings

Leading Countries and Institutions

Figure 2 shows the yearly publications on TBC for the top 10 countries contributing to thermal barrier coatings. USA took the leading position during 1981-2008, while the contribution of China have grown rapidly after 2001, and it became the leading country during the past 10 years. The study of thermal barrier coating in China started in late 1980s on plasma spray TBC. In the mid-1990s, a group of researchers in Beihang University (formerly known as BUAA) set up a lab to carry out research in EB-PVD thermal barrier coatings in order to catch up with the global research frontiers on TBC. As China sets aviation industry as one of its national strategic key industries, strong support from the government poured into aerospace and air transport related research and development areas with several of its national 5-year plans, the research on TBC in China has progress aggressively, as shown in Figure 2. Other top countries include Germany, Japan, United Kingdom and France. India is another rapidly advancing country on thermal barrier coating research.

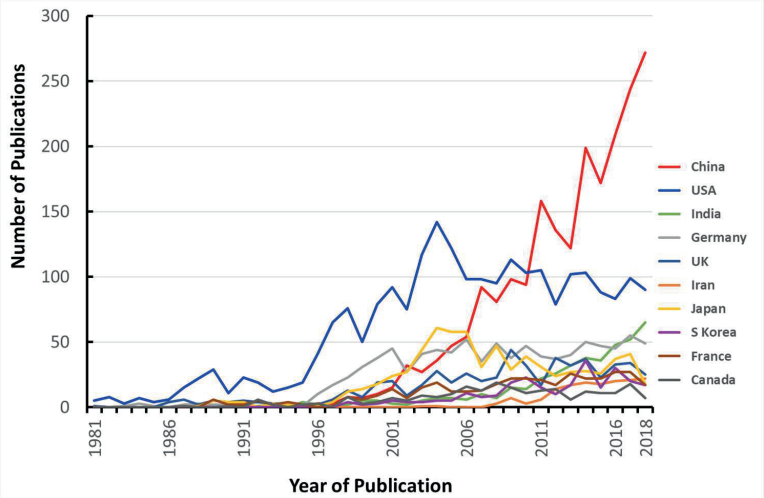


Figure 2. Yearly scholarly output for top 10 countries on thermal barrier coatings.

Among the institutions, German Aerospace Center (DLR), NASA Glenn Research Center are the two leading institutions on TBC research during the period 1996-2005; while the Chinese Academy of Sciences and Beihang University become the leaders in the past 10 years.

Figure 3 shows the 5-year average field-weighted citation impact (FWCI) for the publications on thermal barrier coatings in various subject areas. These subject area categories are high-level SciVal categories. Some papers may be assigned to more than one categories. Yet, this should not affect the results significantly, as the FWCI of each interdisciplinary paper has already count in the effect of different subject area categories. One can find that the FWCI of publications published in recent years generally have a lower FWCI than those published in the past. This reflects the fact that, the effect of thermal barrier coatings research is for the long term. From the first concept of thermal barrier coating in 1940s to the first application teste in the 1970s, it took over 30 years to put the idea into practical use. So past publications on thermal barrier coatings are still quite relevant to contemporary applications, and are highly cited. While publications in materials sciences have a higher 5-year average FWCI before the period 2004-2008, chemical engineering started with a low 5-year average FWCI during 2003-2007, and grew rapidly in the following periods. Compared to materials and structure design which takes a long time, chemical characterization of TBC coatings and combustion science take a much shorter cycle, so that publications in the last five or ten years are more likely get high attentions.

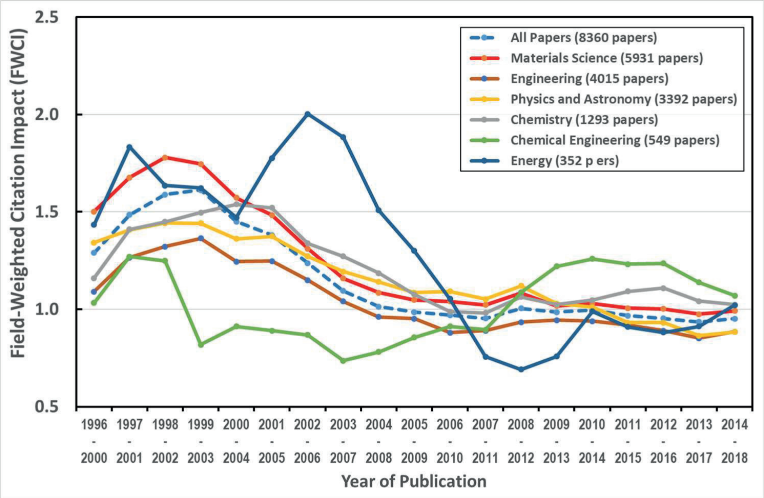


Figure 3. 5-Year Average FWCI for Top 6 Research Areas of Global TBC Research (Data from SciVal, Time Span: 1996 – 2018)

Figure 4(a) and (b) show the trends of country and researchers in TBC and their collaboration networks.

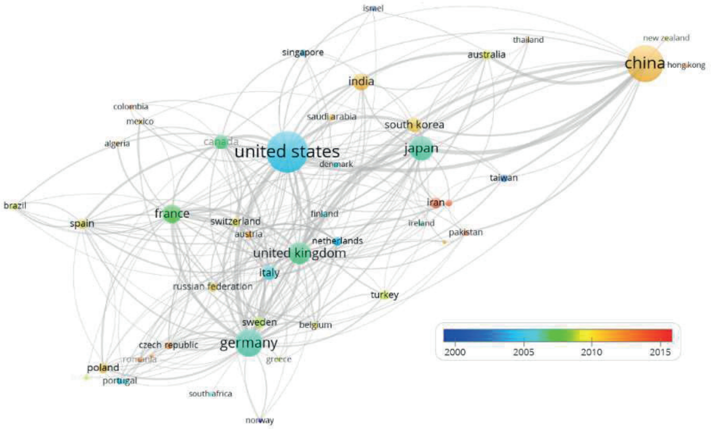


Figure 4(a). Mapping of Contributing Countries and their Collaborations in TBC Research (Total 192 countries, 48 countries with No. of publications > 5)

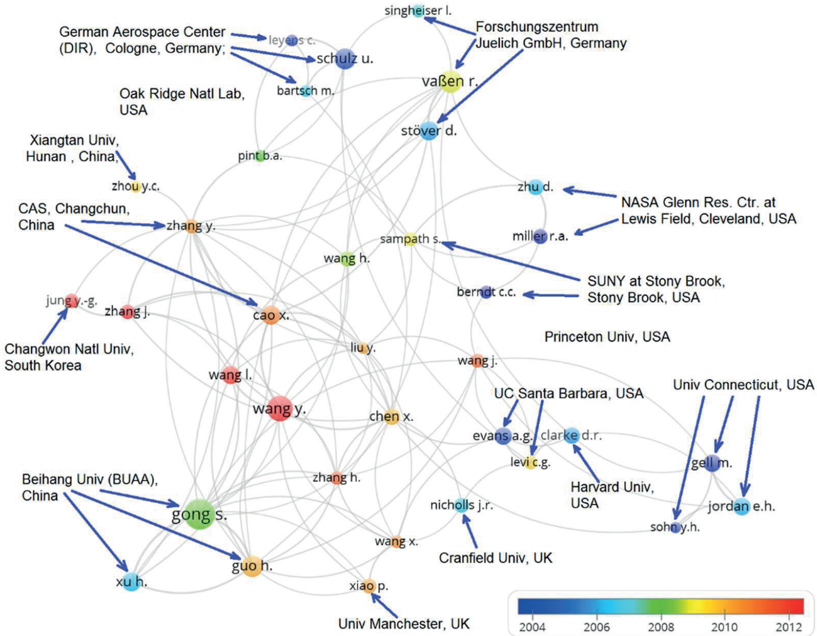
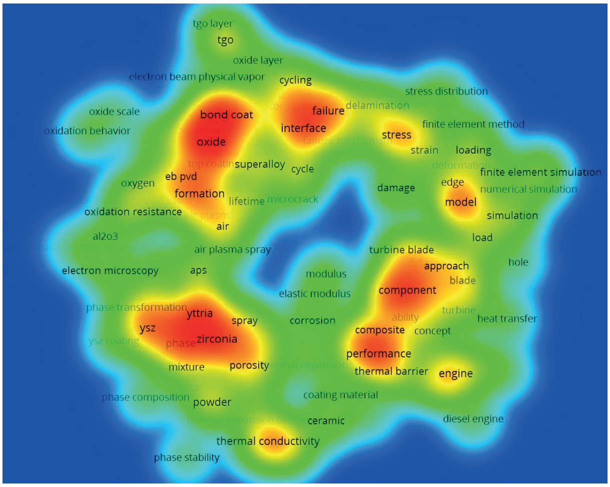


Figure 4(b). Mapping of Researchers and their Collaborations in Thermal Barrier Coating Research (Total 11681 Researchers, 35 researchers with No. of publications > 40)

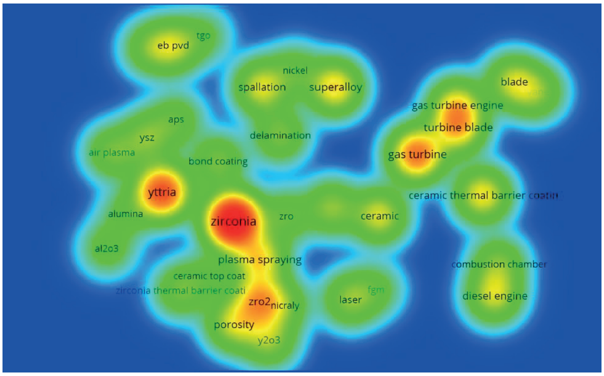
USA and China are the leading countries in different periods, and researchers from German Aerospace Center (DLR), NASA Glenn Research Center, Chinese Academy of Sciences and Beihang University contribute to significant portion of publications on TBC technology. The collaboration map also clearly shows the collaboration of TBCs research between USA, Germany and China or India over the years. Many of the Chinese researchers studied in USA or Germany on TBC technology. These researchers applied the knowledge they learned on their research upon return to China, which helps the rapid development of TBC research in the Chinese institutions.

Mapping of Technical Terms and Institutions

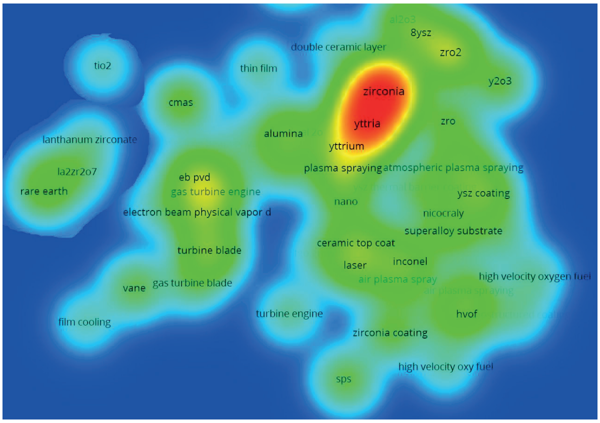
Figure 5(a) and (b) shows the research area trends by mapping the technical terms used in the titles and abstracts in the publications. Selection of terms limited to those that occurred 20 times or more in the year. Terms such as “EB PVD“, “Plasma Spraying”, “YSZ”, “Gas Turbine Engine” can be seen in the maps on the 1981-2000 period, while terms like “Lanthanum Zirconate”, “Rare Earth”, CMAS, “double ceramic layer”, SPS (suspension plasma spraying), HVOF are found in the map for the period 2011-2018, denoting the novel materials, fresh considerations and innovative processing technologies in thermal barrier coatings.



(a) 1981-2018, 85495 terms, Occurrences > 100, 314 terms, 150 selected



(b) 1981-2000, 18057 terms, Occurrences > 25, 207 terms, 35 selected



(a) 2011-2018, 43426 terms, Occurrences > 25, 518 terms, 57 selected

Figure 5. Terms used in thermal barrier coating publications published in Year (a) 1981-2016, (b) 1981-2000 and (c) 2011-2018.

Patents on Thermal Barrier Coatings

Analysis of Patents on Thermal Barrier Coatings

NASA initiates the concept of thermal barrier coating in 1940s, and its first application tested in the 1970s. Figure 6 shows the annual number of patents published by the USPTO on thermal barrier coating related research topics. Figure 6 shows the patents published by the USPTO increase steadily. There is a sharp increase during the period 2001-2003, which is in correspondence with the development of ultra-efficient engine technology by adopting low thermal conductivity TBC and new idea of environmental barrier coatings, i.e., coatings for uncooled silicon nitride (Si_3N_4) series of uncooled turbine blades.

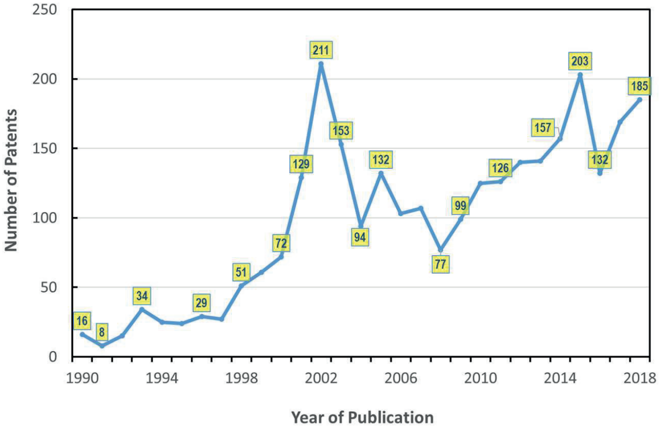
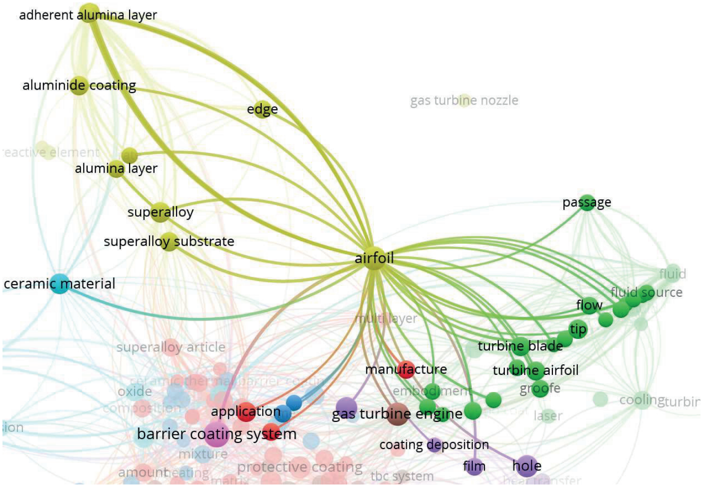
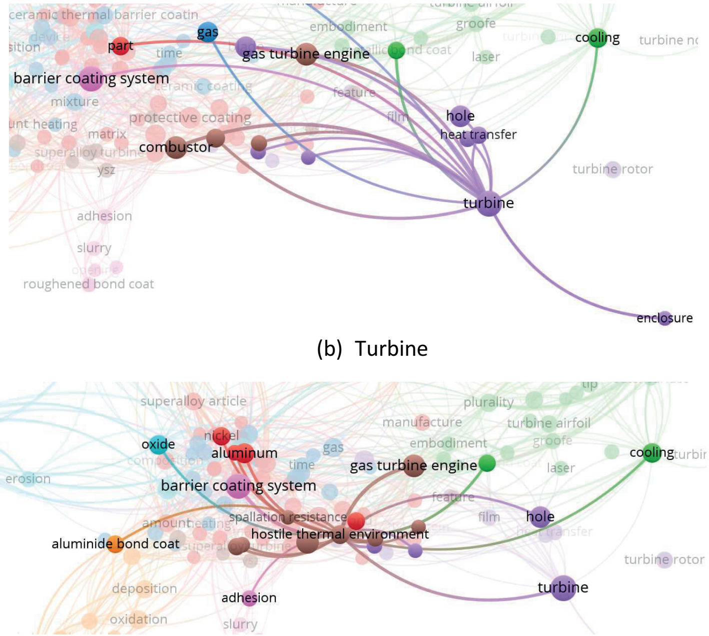


Figure 6. Yearly number of patents published by USPTO on thermal barrier coating related research areas, 1990-2018.

Figure 7 shows an analysis of the terms used in the titles and abstracts of patents published during 2001-2003 by the USPTO on TBC related research topics.



(a) Airfoil

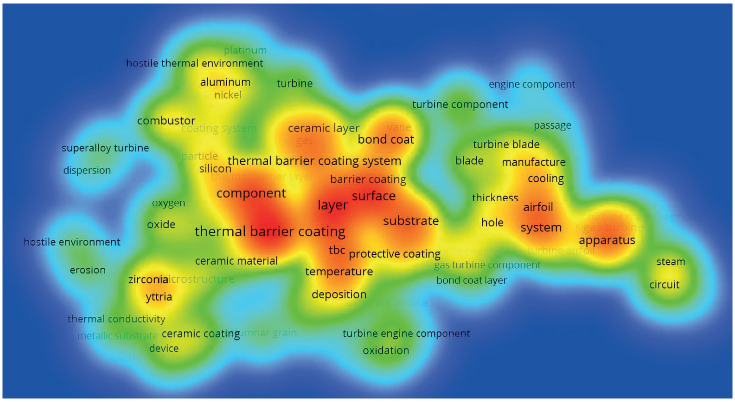


(b) Turbine

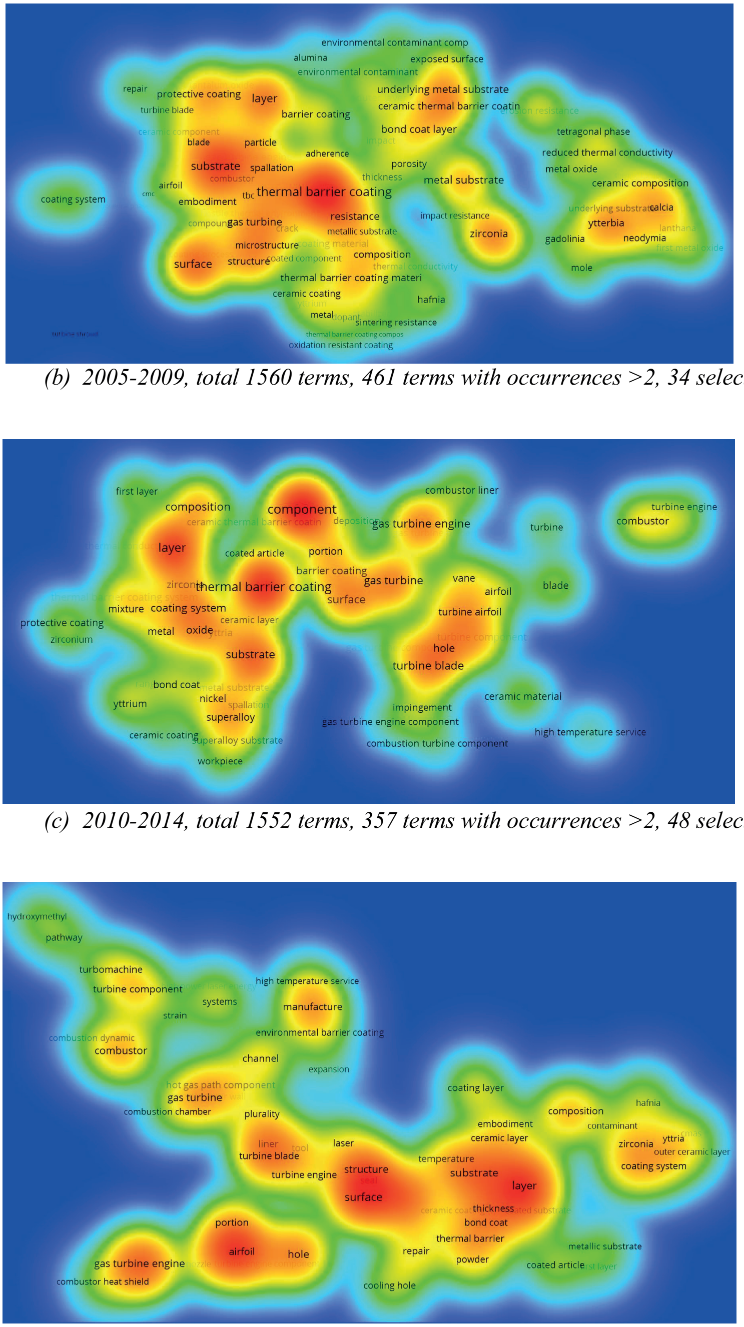
(c) Hostile thermal environment

Figure 7. Term maps for patents published by USPTO on TBC related research areas, 2001-2003.

In order to investigate the trends of patents on thermal barrier coating, the terms used in the titles and abstracts of patents published by USPTO during 2000-2018 are analysed. The term mapping are shown in Figure 8(a) to (d). Terms such as "thermal barrier coating", "Plasma Spraying", YSZ, "Gas Turbine Engine" can be seen during the 2000-2004 and 2005-2009 maps. Correspondingly, terms like "Lanthanum Zirconate", "Rare Earth", CMAS, "double ceramic layer", SPS (suspension plasma spraying), HVOF are found in the maps for the period 2011-2014, and subsequently, "self-healing", "environmental barrier coating" are found in the map for the period 2015-2018, denoting the new materials, new considerations and innovative processing technologies in thermal barrier coatings.



(a) 2000-2004, total 1739 terms, 562 terms with occurrences >2, 77 selected



(b) 2005-2009, total 1560 terms, 461 terms with occurrences >2, 34 selected

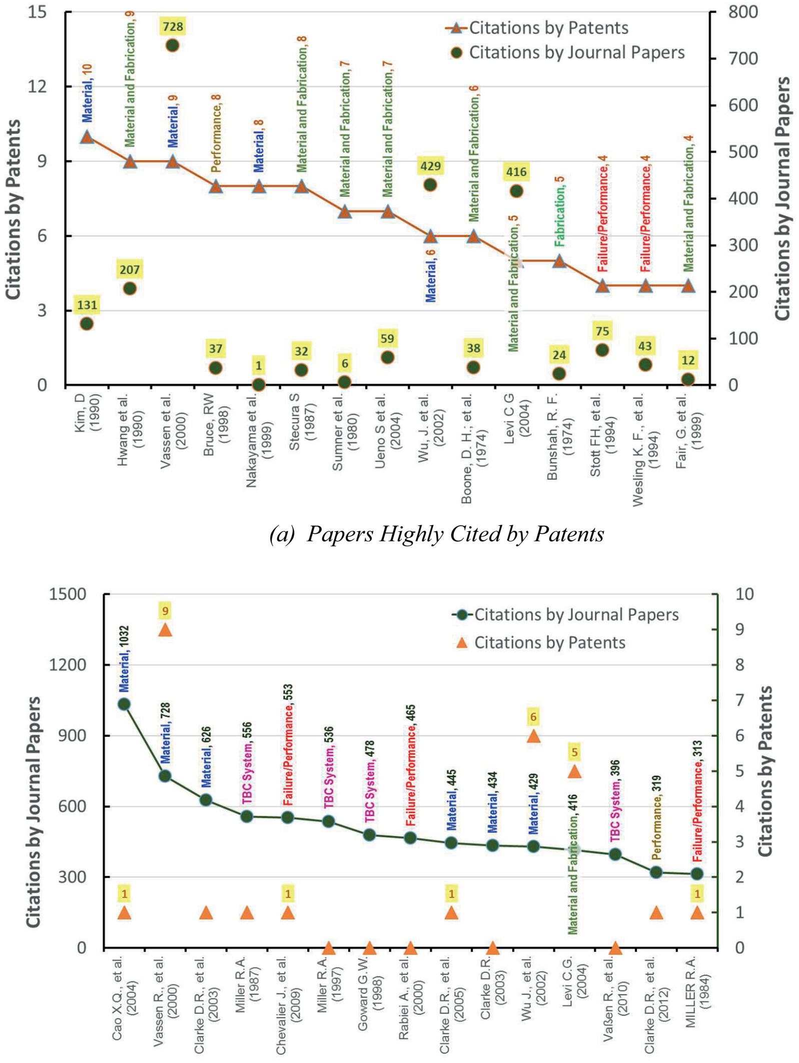
(c) 2010-2014, total 1552 terms, 357 terms with occurrences >2, 48 selected

(d) 2015-2018, total 1309 terms, 300 terms with occurrences > 2, 32 selected

Figure 8. Terms used in thermal barrier coating publications published in Year (a) 1981-2000, and (b) 2011-2018.

Papers on Thermal Barrier Coatings that are Highly Cited by Patents and Journal Papers

In order to ascertain the influence of journal papers on patents, the journal papers cited in patents of thermal barrier coating are studied. More than 400 patents with "thermal barrier coating" in their title are selected, and around 500 papers cited in these selected patents are identified. Figure 9(a) plots a list of the top 15 papers with high citations in patents. Among these 15 papers, only four of them have high (more than 200) journal citations. Conversely, among the top 15 highly cited TBC papers, only three papers received more than one citation by patents, as shown in Figure 9(b). This indicates that topics in patents have low correlation with academic research topics on TBCs.



(a) *Papers Highly Cited by Patents*

(b) *Papers Highly Cited by Journal Papers*

Figure 9. Top 15 Highly Cited Papers by Patents and by Journals.

Academic research interests are significantly broader than industry needs. Many novel ideas reported in TBC papers may not have the requisite potential or relevance to the specifications of the aerospace industry. In addition, most of the work reported in journals may not yet have results that appeal to an industrial application. There are distinctive difference on the topics between highly patent-cited papers and highly journal-cited papers. While highly patent-cited papers are more focused on materials' fabrication and performance, highly journal-cited papers are more focused on TBC system in general and its performance.

Summary

In the past 38 years, the number of research publications on thermal barrier coatings research rose steadily. The USA took a leading position during 1999-2008, and China took over the lead from 2009 onwards, indicated by the rapid rise of publications by Chinese researchers, and showing China's ambition in its aviation industry development. The FWCI of publications in recent years generally have lower values than those published in the past. While publications in materials sciences have a higher 5-year average FWCI before the period of 2003-2007, chemical engineering started with a low 5-year average FWCI during 2003-2007, and grew rapidly in subsequent years. The FWCI trend variances in the TBC research topics reflect the shifting effective phases of research-application cycles of the corresponding research focus. The evolution trends of research topics revealed by mapping of the technical terms in titles and abstracts of the publications shows that new materials like "Lanthanum Zirconate", "Rare Earth"; new designs of coating systems like "double ceramic layer", and new processing technics like SPS (suspension plasma spraying) and HVOF are employed in thermal barrier coatings. These will promote the rapid development of TBC and make it highly relevant in future aircraft engines. Analysis of patent citations of journal papers shows that research topics in patents have low correlation with academic research topics on TBCs. Nevertheless, academic research interests on TBC are broader and based on new ideas on TBC systems and its performance. Most of which have yet to address industry concerns, and far from the final stage of industrial applications.

References

Appleby, M.P., Zhu, Dongming & Morscher, G.N. (2015). Mechanical properties and real-time damage evaluations of environmental barrier coated SiC/SiC CMCs subjected to tensile loading under thermal gradients. *Surface & Coatings Technology,* 284, 318-326. DOI: 10.1016/j.surfcoat.2015.07.042.

Bakan, E. & Vaßen, R. (2017). Ceramic Top Coats of Plasma-Sprayed Thermal Barrier Coatings: Materials, Processes, and Properties. *Journal of Thermal Spray Technology*; 26(6):992-1010. DOI: 10.1007/s11666-017-0597-7.

Bose, S & DeMasiMarcin, J. (1997). Thermal barrier coating experience in gas turbine engines at Pratt & Whitney. *Journal of Thermal Spray Technology,* 6(1), 99-104.

Busso, E.P., Wright, L., Evans, H.E., McCartney, L.N., Saunders, S.R.J., Osgerby, S. & Nunn, J. (2007). A physics-based life prediction methodology for thermal barrier coating systems. *Acta Materialia*, 55(5), 1491-1503. DOI: 10.1016/j.actamat.2006.10.023.

Clarke, D.R. & Phillpot, S.R. (2005). Thermal barrier coating materials. *Materials Today*, 8(6):22-29. DOI: 10.1016/S1369-7021(05)70934-2.

Darolia R. (2013) Thermal barrier coatings technology: critical review, progress update, remaining challenges and prospects, *International Materials Reviews*, 58:6, 315-348, DOI: 10.1179/1743280413Y.0000000019

Fauchais, P., Montavon, G., Lima, R.S. & Marple, B.R. (2011). Engineering a new class of thermal spray nano-based microstructures from agglomerated nanostructured particles, suspensions and solutions: An invited review. *Journal of Physics D: Applied Physics*; 44(9),Article number: 093001. DOI: 10.1088/0022-3727/44/9/093001.

Golosnoy, I.O., Cipitria, A. & Clyne, T.W. (2009). Heat Transfer through Plasma-Sprayed Thermal Barrier Coatings in Gas Turbines: A Review of Recent Work. *Journal of Thermal Spray Technology,* 18(5-6), 809-821. DOI: 10.1007/s11666-009-9337-y

Goswami, B., Ray, A.K. & Sahay, S.K. (2004). Thermal barrier coating system for gas turbine application - A review. *High Temperature Materials and Processes*, 23(2), 73-92.

Guo, H.B., Vassen, R. & Stover, D. (2004). Atmospheric plasma sprayed thick thermal barrier coatings with high segmentation crack density. *Surface & Coatings Technology*; 186(3), 353-363.

Guo, H.B., Gong S.K. & Xu H.B. (2014). Research progress of New High/Ultra-high temperature thermal barrier coatings and processing technologies. *Acta Aeronautica et Astronautica Sinica*, 35(10): 2722-2732. (In Chinese)

Gurak, M., Flamant, Q., Laversenne, L. & Clarke, D.R. (2018). On the Yttrium Tantalate – Zirconia phase diagram. *Journal of the European Ceramic Society*, Article in Press.

He, J., Guo, H., Peng, H. & Gong, S. (2013). Microstructural, mechanical and oxidation features of NiCoCrAlY coating produced by plasma activated EB-PVD. *Applied Surface Science*; 274:144-150. DOI: 10.1016/j.apsusc.2013.02.136.

Markocsan, N., Gupta, M., Joshi, S., Nylen, P., Li, X.H. & Wigren, J. (2017). Liquid Feedstock Plasma Spraying: An Emerging Process for Advanced Thermal Barrier Coatings. *Journal of Thermal Spray Technology*, 26(6), 1104-1114. DOI: 10.1007/s11666-017-0555-4.

Miller, R.A. (1987). Current status of thermal barrier coatings - An overview. *Surface and Coatings Technology.* 30(1):1-11

Murthy, P.L.N., Nemeth, N.N., Brewer, D.N. & Mital, S. (2008). Probabilistic analysis of a SiC/SiC ceramic matrix composite turbine vane. *Composites Part B-Engineering*, 39(4), 694-703. DOI: 10.1016/j.compositesb.2007.05.006.

Padture, N.P., Gell, M. & Jordan, E.H. (2002). Materials science - Thermal barrier coatings for gas-turbine engine applications. *Science,* 296(5566), 280-284. DOI: 10.1126/science.1068609.

Shiembob, L.T. & Hyland J.F. (1979). Development of a Plasma Sprayed Ceramic Gas Path Seal for High Pressure Turbine Applications. *NASA CR-159669.*

Suzuki, M., Sodeoka, S. & Inoue, T. (2008). Zircon-based ceramics composite coating for environmental barrier coating. *Journal of Thermal Spray Technology,* 17(3), 404-409. DOI: 10.1007/s11666-008-9178-0.

Wang, C.-A., Lu, H., Huang, Z. & Xie, H. (2018). Enhanced anti-deliquescent property and ultralow thermal conductivity of magnetoplumbite type LnMeAl11O19 materials for thermal barrier coating. *J Am Ceram Soc.*;101:1095–1104. https://doi.org/10.1111/jace.15285

Zhang, X.F., Zhou, K.S., Liu, M., Deng, C.M., Deng, C.G., Niu, S.P. & Xu, S.M. (2017). Oxidation and thermal shock resistant properties of Al-modified environmental barrier coating on SiCf/SiC composites. *Ceramics International*, 43(16), 13075-13082. DOI: 10.1016/j.ceramint.2017.06.167.

Has the 2008 Global Financial Crisis a lasting impact on universities and public research institutes in the European Union?

Marc Luwel and Thed N. van Leeuwen

**{luwel, leeuwen}@cwts.leidenuniv.nl*

Centre for Science and Technology Studies (CWTS),
University Leiden, Netherlands

Introduction

2018 marked the 10th anniversary of the 2008 Global Financial Crisis (GFC) followed by a global recession, for most economists the worst disaster since the great depression of the 1930s.

The GFC created in the Eurozone, the 19 EU member states which have adopted the euro as their common currency, an asymmetric shock, affecting particularly the heavily indebted countries Cyprus, Ireland, Greece, Ireland, Spain and Portugal. The result was the Eurozone Crisis. In the aftermath of GFC these countries were unable to refinance their sovereign debt and/or bail out their debt overloaded banks.

Not only in these six Eurozone countries but all over the EU this combination led to a substantial increase in the government debt-to-GDP ratio. To reduce budget deficits, government spending was reduced in the Eurozone. This affected severely the public sector in the heavily indebted countries, but also in countries such as Belgium that was confronted with the near collapse of its three major banks and Italy that had to prop up its banking sector and large firms. With their small open economies the Scandinavian EU member states were also confronted with the impact of the GFC, but the Nordic model turns out to be less vulnerable and more resilient (Gylfason et al., 2010). Of these countries, only Finland is a member of the Eurozone.

In the Eurozone cuts in government spending had also an impact on public R&D expenditures and in particular on the public funding of the higher education sector (EUA, 2011) and public research institutes. A number of authors studied the performance of public R&D funding systems at country level (Lepori, Reale & Spinello, 2018 and references therein). However 10 years after the start of the GFC little work has been done on the impact of the reduction in public funding on the research performance of the higher education sector and the public research institutes.

In the present study, for a panel of eight Eurozone countries and the two Scandinavian EU member states with their own official currency we investigate trends in the publication output of these institutes and the potential existence of causality between funding and output. To tackle these questions models and techniques developed in econometrics are applied on bibliometric data.

The remainder of this paper is organized as follows. Section 2 describes the data that are used in the model. In section 3 the model and the methodology are presented. Section 4 provides the quantitative results and in section 5 their policy relevance and follow up work are discussed.

Data

In most developed countries nearly all basic research is done at universities and public research institutes and to a large extent funded by public authorities. Although these organizations have multifaceted missions, most of the results of their research activities becomes publicly available.

Although they do not cover the total research output papers published in scholarly journals are often used as a proxy for this output. In this paper the proxy is further restricted to the publications in journals covered by the Web of Science (WoS), a bibliographic database produced by Clarivate Analytics. The WoS which coverage has been extended over the last decades, now covers all major journals within the natural and life sciences, medicine and the basic disciplines in applied sciences as well as peer reviewed conference proceedings. However scholarly journals in the humanities and social sciences remain poorly represented with a bias towards those published in English (van Leeuwen, 2013).

For this analysis the data on scientific publications were extracted from the WoS database licensed to the Centre of Science and Technology Studies, Leiden University. This version of the database includes the Science Citation Index Expanded, the Social Science Citation Index and the Arts and Humanities Citation Index. For this study only the publication types 'articles', 'reviews' and 'letters' were taken into account. In the counting scheme the first two publication types received 1 as weight and the letters were weighted as 0.25.

To study the publication output at the level of countries an important methodological issue is the applied counting scheme. An increasing number of journal publications are co-authored by researchers working in different countries (Luukonen et al., 1993; van Leeuwen, 2009). In this paper publications are aggregated at country level applying two counting schemes. The 'whole' counting scheme gives equal weight to all the countries mentioned in the by-line of a publication, is used. An alternative 'fractional' counting scheme allocate to each country a fractional weight based on the number of countries in the by-line (Perianes-Rodriguez, Waltman & van Eck, 2016). For example a publication with two addresses from Belgium, three addresses from Italy and one address from Greece, is counted as one third for each of the three countries. More sophisticated counting schemes have been devised but they are outside the scope of this study.

OECD publishes data on its member states' research funding. These data are produced by these countries using the methodology to collect statistics on research and development described in the Frascati Manual (OECD, 2015).

One of the statistics collected by the OECD is the total intramural expenditure on R&D performed in the national territory during a specific reference period, the Gross domestic expenditure on R&D (GERD). As a proxy for expenditure on research of a country's higher education institutes the component of the GERD incurred by units belonging to the higher education sector, called Higher education expenditure on R&D (HERD), is used. It is the measure of intramural R&D expenditure within the higher education sector.

Similarly, a proxy for the expenditure on research carried out by public research institutes and other governmental organizations is the measure of expenditure on intramural R&D within the Government sector during a specific reference period. This component of the GERD is labelled Government Expenditure on R&D (GOVERD).

To correct for differences in inflation rate and purchasing power among countries the data on the HERD and the GOVERD are expressed in 2010 US $ constant prices and purchasing power parity (PPP).

In the OECD Main Science and Technology Indicators database these data are publicly available on an annual basis (see: https://stats.oecd.org/). For some countries and some years there are missing values. For example for Sweden only biannual data on HERD and GOVERD are available for the later part of last century. In this study the missing values are estimated using linear interpolation. At the moment the study was made the 2017 data on HERD and GOVERD were not yet published; they were estimated using an extrapolation based on the 2015 and 2016 values. The sum of the HERD and GOVERD, further research expenditure, was used as a proxy for the public expenditure on research in the combined higher education and the government sector (Aksnes et al., 2017).

Methodology

In this study time series analysis is applied on research expenditure and publication data of countries. In this section concepts and techniques used in this study are succinctly described with the appropriate references to relevant literature.

The panel countries

For this study, eight Eurozone countries were selected based on the degree of severity of the impact of the GFC. These countries can be classified in three groups:

- Group 1: Countries strongly affected by the GFC: Ireland (IE), Greece (GR), Spain (ES), Portugal (PT);
- Group 2: Countries with severe problems in their banking sector: Belgium (BE) and Italy (IT);
- Group 3: The countries, more mildly affected by the GFC: Netherlands (NL) and Finland (FI).

To complete the panel the two other Scandinavian EU member states were added: Denmark (DK) and Sweden (SK). These two countries were also only mildly affected by the GFC.

Statistical tools

The data are analyzed with the statistical package EViews (version 10) developed by IHS Markit Ltd mainly for time-series oriented econometric analysis. The EViews 10 User's Guide I and II (EViews 10, 2017a, 2017b) provides a detailed description of the models and tests applied in this study and the relevant references to the scientific literature.

Stationary and non-stationary time series

The data on expenditures and on publications are both univariate time series, a sequence of measurements of the same variable indexed in time order. In this study, the data are available per annum.

In time series analysis a distinction is made between stationary and non-stationary series (Brockwell & Davis, 2016). For a stationary time series, the mean, variance and covariance are not a function of time. In the presence of a deterministic trend, i.e. a consistent directional movement, the time series is called trend stationary.

For non-stationary time series two models are generally used:

- The stochastic model with a drift $x(t)= \acute{\alpha} + \rho{*}x(t-1) + \acute{\varepsilon}(t)$,
- The stochastic model with a drift and a deterministic trend $x(t) = \acute{\alpha} + \rho{*}x(t-1) + \beta{*}t+ \acute{\varepsilon}(t)$

where $\acute{\alpha}$, β and ρ are a constant and $\acute{\varepsilon}(t)$ is white noise.

The parameter ρ plays an important role. It can be easily shown (Elder & Kennedy, 2001) that for

- $\rho<1$ a shock in the model gradually disappears;
- $\rho=1$ a shock persists in the system and never dies away;
- $\rho>1$ a shock becomes more influential over time.

Time series with $\rho=1$ are said to have unit roots. Unit roots have profound implications for statistical testing, especially regressions as the conditions required to apply the Law of Large

Numbers and the Central Limit Theorem are violated and, as a result, commonly used test statistics such as t-statistics can no longer be used or lead to spurious results.

If a time series has unit roots, successive differences, d, can transform the series in a stationary one. This series is called to be integrated of order d, or I(d). In most cases the order of integration is 0 (non-integrated stationary series), 1 or 2.

To check for the presence of unit roots a series of tests were developed and the most frequently used are the Augmented Dickey Fuller (ADF test) and the Phillips-Perron (PP) test using as null hypothesis the presence of a unit root (Neusser, 2016). Davidson and MacKinnon (2004) report that the PP test performs worse in finite samples than the ADF test. In this study the ADF test is used. A detailed discussion of these tests and other unit root tests is beyond the scope of this paper and can be found in the literature (Neusser, 2016; Kwiatkowski et al., 1992).

Structural breaks in time series and the Chow test

In a time series an unexpected shift can occur. For example countries' GDP had a shift due to the effects of 1973 oil price crisis. Such a structural break or structural change can lead to errors and unreliability of the model in general (Ho & Iyke, 2017).

Limiting to linear models there are several forms:

- A single break in mean with a known breakpoint;
- A known number of breaks in mean with unknown break points;
- An unknown number of breaks in mean with unknown break points;
- Breaks in variance.

A structural change in time series can influence the results of tests for unit roots (Perron, 1989). Methods have been developed for unit root testing in the present of structural breaks. A distinction can be made between an exogenous, a priori fixed break date and endogenous break dates (Vogelsang and Perron, 1998). As remarked by Perron (2017) the interplay between structural change and unit roots remains an important research topic with a number of open questions.

Visual inspection may suggest a structural break in the time series. To test this assumption the Chow test (Chow, 1960) is often used. The Chow statistics tests whether the single regression line or the two separate regression lines fit the data best with as null hypothesis H0 no structural break in the time series. The Chow statistics is distributed F(k, n1+n2-2*k) where k is the number of estimated parameters and n1 and n2 the number of observations in the two groups. If the Chow statistics is larger than the F-critical value H0 is rejected. The assumption to carry out the Chow test is that the errors in the regressions are serially uncorrelated and homoskedastic.

To test for autocorrelation the Breusch–Godfrey serial correlation LM test (LM test) (Asteriou & Hall, 2016) is used with the null hypothesis H0 of no serial correlation. If the probability p is larger than the critical value H0 cannot be rejected. To remove serial correlation a modified regression model can be used with a one period lag of the dependent variable as an additional independent variable.

The presence of heteroscedasticity, i.e. the variance of the residual in a regression model is not constant, is tested with the Breusch-Pagan-Godfrey (BPG) test or the White test (Asteriou & Hall, 2016). The null hypothesis H0 is homoscedasticity. If the probability p is larger than the critical value H0 cannot be rejected.

Causality

When studying time series one often looks at correlations to test whether and how strongly two series are related. However correlation does not imply causation. Causation indicates that one event is the (partial) result of the other. Granger (1969) developed a methodology to test bi-directional causality between two variables.

The Granger Causality (GC) test uses a bivariate linear autoregressive (VAR) model of two variables x_t and y_t:

$$y_t = \mu_0 + \sum_{i=1}^{k} \alpha_{1i} y_{t-i} + \sum_{i=1}^{k} \beta_{1i} x_{t-i} + \varepsilon_{1t}$$

$$x_t = \varphi_0 + \sum_{i=1}^{k} \gamma_{1i} x_{t-i} + \sum_{i=1}^{k} \delta_{1i} y_{t-i} + \varepsilon_{2t}$$

where

- k is the maximum number of lagged observations included in the model,
- $\mu_0, \alpha_{1i}, \beta_{1i}, \varphi_0, \gamma_{1i}$ *and* δ_{1i} *are* parameters of the model, and
- ε_{1t} and $\varepsilon_{2t}$are residuals (prediction errors) for each time series.

If the variance of ε_{1t} is reduced by the inclusion of the x_t terms in the first equation, then it is said that x_t Granger causes y_t, i.e. if the coefficients β_{1i} jointly significantly different from zero.

This can be tested by performing an F-test of the null hypothesis that β_{1i}= 0 for ∀ i ∈ [1,…, k], assuming the series x_t and y_t are stationary. GC is bidirectional as can be seen from the equation for x_t.

To carry out the GC-test the appropriate model order, i.e. the number of lags k in the regression has first to be determined. There are several approaches and in this paper two of the most frequently used in literature are applied: the Akaike Information Criterion (AIC, (Akaike, 1974)) and the Schwartz Information Criterion (SIC, (Schwartz, 1978)). In case of a difference in the estimated lag length, the AIC is used. In studies with small samples (n< 60) AIC is superior to other information criteria (Mishra, 2014).

The GC-test has a number of limitations. The two variables must be of the same order of integration and to be stationary. To obtain the latter first or higher order of differences of the variables are used in the VAR resulting in the loss of information.

To avoid the problems inherent to the traditional testing of GC, Toda and Yamamoto (1995) developed a procedure by fitting a VAR model in the levels of the variables and adding to the optimal lag length k of the VAR model the highest order of integration dmax of the two variables. A VAR(k+dmax) model is estimated and the coefficients of the last lagged dmax vector are ignored (Zapata & Rambaldi, 1997):

$$y_t = \mu_0 + \left(\sum_{i=1}^{k} \alpha_{1i} y_{t-i} + \sum_{i=k+1}^{d_{\max}} \alpha_{2i} y_{t-i}\right) + \left(\sum_{i=1}^{k} \beta_{1i} x_{t-i} + \sum_{i=k+1}^{d_{\max}} \beta_{2i} x_{t-i}\right) + \varepsilon_{1t}$$

$$x_t = \varphi_0 + \left(\sum_{i=1}^{k} \gamma_{1i} x_{t-i} + \sum_{i=k+1}^{d_{\max}} \gamma_{2i} x_{t-i}\right) + \left(\sum_{i=1}^{k} \delta_{1i} y_{t-i} + \sum_{i=k+1}^{d_{\max}} \delta_{2i} y_{t-i}\right) + \varepsilon_{2t}$$

where

- k is the maximum number of lagged observations included in the model,
- dmax is the maximum order of integration of the two variables in the model,
- $\mu_0, \alpha_{1i}, \alpha_{2i,} \beta_{1i}, \beta_{2i}, \varphi_0, \gamma_{1i}, \gamma_{2i}, \delta_{1i}$ *and* δ_{2i} *are* parameters of the model, and
- ε_{1t} and $\varepsilon_{2t}$are residuals (prediction errors) for each time series.

The null hypothesis of non-causality of x_t to y_t can be expressed as $H0 : \beta_{1i}$ = 0, i ∈ [1,…, k].

To test for GC Toda and Yamamoto developed the Modified Wald (MWald) statistics. The MWald statistics follows a Chi-square distribution asymptotically and the degrees of freedom are equal to the number of time lags (k+dmax).

To carry out the T-Y procedure to test for GC between two time series the following steps are needed:

1. Test each of the time series to determine their order of integration.
2. Let the maximum order of integration of the two of time-series be dmax. For example if one series is I(1) and the other I(2), dmax=2.
3. Make a VAR model in the levels of the data, regardless of the orders of integration of the time series.
4. Determine the appropriate maximum lag length for the variables in the VAR, k, using the appropriate tests such as AIC and SIC.
5. Test that the VAR model is well-specified. For example, test that for the residuals the serial correlation, the normality and the homoskedasticity. As already explained to test for the presence of serial correlation the LM test is used and for homoskedasticity the White test. The Jarque-Bera test (Jarque & Bera, 1980) tests for normality by measuring the difference of the skewness and kurtosis of the series with those from the normal distribution with the null hypothesis of a normal distribution. If necessary, the value of k can be increased until the autocorrelation issue are resolved.
6. In the VAR model to the lag length k dmax additional lags of the two variables are added into the two equations.
7. Test for Granger causality where the dmax lags that are added, are not included when carrying out the Wald test for they were added to assure that the Wald statistics is asymptotically chi-square distributed.
8. Rejection of null hypothesis supports the presence of GC.

Empirical results

Structural break in the publication output

Figure 1 shows the evolution between 1999 and 2017 of the number of publications, using full and fractional counting (further full- and fractional-counted publications). The overall impression is a fairly linear increase during the first part of the 19 year period followed by a linear increase with a shallower slope or even a decrease. This evolution is more pronounced for the fractional-counted publications. For Denmark (full counting scheme), the publication output shows the opposite behaviour, with a stronger increase over the last part of the period.

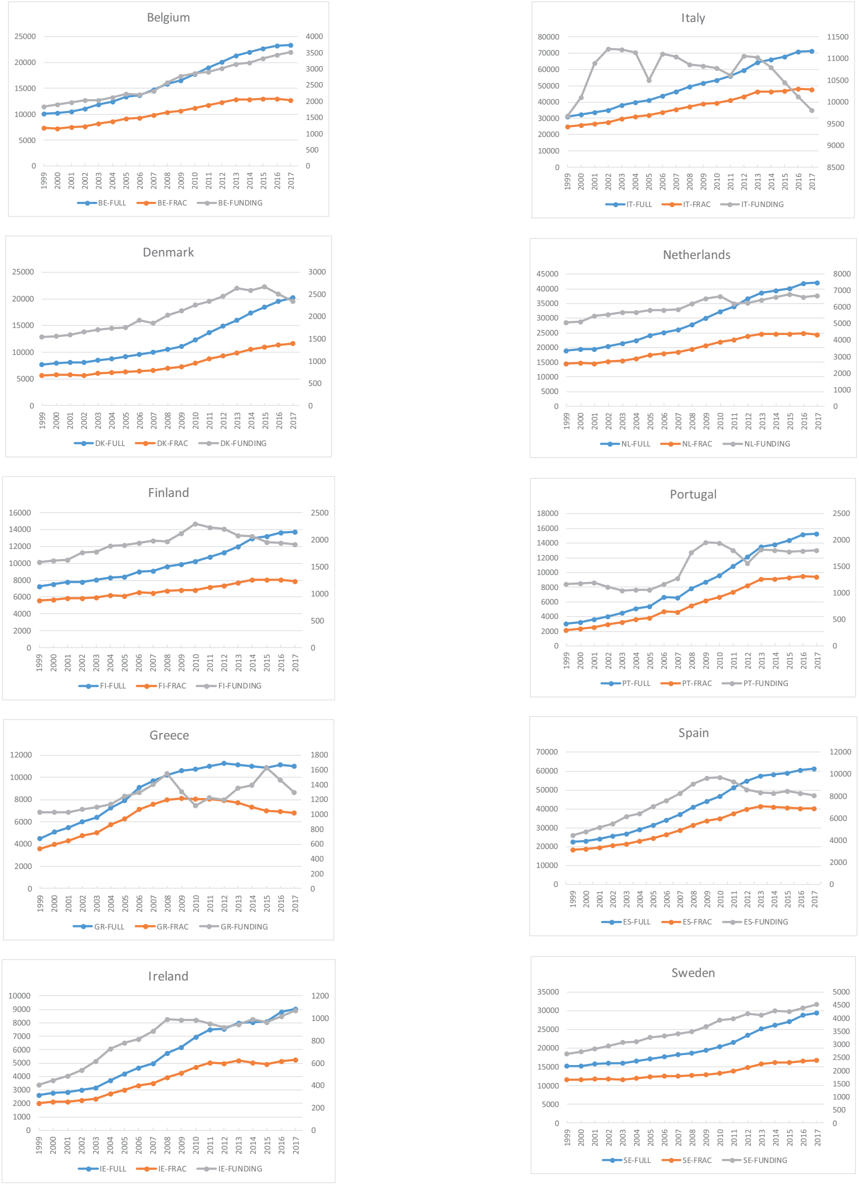


Figure 1- For each country in the panel the evolution of the number of publications, full counting scheme (XX-FULL) and fractional counting scheme (XX-FRAC) on the left Y-axis and the funding (XX-FUNDING), (HERD + GOVERD) at constant prices and PPP $ in $ million 2010 on the right Y-axis. The period is 1999-2017.

Table 1 – For evolution of the number of publications, full (Full) and fractional (Frac) counting scheme, the Chow statistics (Chow stat) is given for the breakpoint year (Breakpoint). At 5% significance level the critical value is F(3,13)= 3.41. The null hypothesis 'no break at the specified breakpoint' has to be rejected for values of the Chow statistic larger than 3.41. The residuals of the regression are tested for serial correlation (p LM-test) and homoscedasticity (p BPG test); for both tests the null hypothesis (no serial regression; homoscedasticity) has to be rejected for p-values smaller than the 5% significance level. The time period is 1999-2017.

CO	Pub count	Breakpoint	Chow stat	p LM-test	p BPG test
BE	Frac	2012	18.25	0.05	0.46
	Full	2012	6.19	0.08	0.18
DK	Frac	2010	13.08	0.10	0.24
	Full	2010	9.13	0.12	0.30
ES	Frac	2012	28.52	*0.01*	0.30
	Full	2011	28.68	*0.01*	0.09
FI	Frac	2013	10.54	0.13	0.07
	Full	2012	25.07	0.70	0.31
GR	Frac	2009	10.62	*0.04*	0.74
	Full	2011	101.29	0.15	0.78
IE	Frac	2011	12.03	*0.03*	0.20
	Full	2011	6.15	0.21	0.26
IT	Frac	2012	8.12	0.07	0.07
	Full	2013	7.41	0.61	0.39
NL	Frac	2012	9.41	*0.01*	0.11
	Full	2012	7.37	*0.01*	0.10
PT	Frac	2012	5.45	0.16	0.18
	Full	2011	2.90	0.52	0.60
SE	Frac	2012	6.45	0.07	0.47
	Full	2012	12.35	0.62	0.16

The Chow statistics is used to test whether there is a structural break in the countries' publication output. Chow test uses an exogenous break date. Table 1 confirms the observations from the visual inspection of the graphs. For Portugal the null hypothesis of no structural break cannot be rejected for the full counting scheme. A sensitivity test was done by calculating the Chow statistic for years around the selected break year resulting in no substantial changes in the Chow statistics.

For all countries in the sample a break is observed in 2010-2012 except Finland (fractional counting scheme, 2013), Greece (fractional counting scheme, 2009), and Italy (full counting scheme, 2013). In the interpretation of the Chow statistics some caution is needed for Spain and the Netherlands, and in the fractional counting scheme for Greece and Ireland as serial correlation in the residuals is observed.

The observed decrease in publications is not due to changes in the coverage of the WoS. In the period 1999-2017 the number of publications processed for the WoS increases by 1 million with some fluctuations over the years due to changes in the coverage of the database.

Visual inspection of the plots (Fig. 1) learns that in the same period the research expenditures decreased except for Belgium and Sweden. For these two countries the research expenditures increased over whole period 1999-2017 and no break is observed. For Ireland, Finland, Netherlands, Portugal, and Spain there is a break in the funding data around 2008-2009 and for Denmark around 2013. Italy's research expenditures data are rather flat with no pronounced trend till 2012 when the stagnation changed in a decrease. For Greece the research expenditures increase till 2008 to become in the following years rather erratic.

Todo-Yamamoto approach to Granger causality between research expenditure and publication output

To carry out a statistically meaning full GC-test the sample size must be at least 30 (Mackinnon, 1996) For most countries in the sample in the 80's of last century a break in the publication output can be observed with 1987-1988 as break date. To avoid the influence of this break on the results the time series were limited to a 30 year period: 1988-2017. In this paper only the results of GC-test between the research expenditures and the full-counted publications are presented.

To study the order of integration of the research expenditures and the full-counted publications two approaches are used:

- The standard unit root test with the ADF test with drift and trend options;
- The breakpoint unit root test with the ADF test with for the basic trend specification and for the breaking trend specification drift and trend options.

The reason for this approach is the presence of structural breaks which can lead to size distortion in standard ADF test (Ho & Iyke, 2017).

The tests were done with the EViews application for unit root testing. A detailed description of the tests can be found in the EViews 10 User's Guide II (EViews 10, 2017a)

All series are level stationary or first difference stationary except for Belgium's full-counted publication series that is stationary at second difference. As for each country in the sample at least one series is I(1), the parameter dmax is 1 for the TY approach for GC, except for Belgium with dmax=2.

Table 2 gives the optimal lag length k for the VAR model with the full-counted publications and the research expenditures. The LM-test, the Jarque-Bara test and the White test were done. No serial correlation or heteroscedasticity are found in the residuals which are normally distributed, except for Portugal. For this country the p-value of the Jarque-Bara test is smaller than the critical value resulting in the rejection of the null hypothesis of normality.

For each country with the optimal lag length k and the maximum order of integration of the 2 variables dmax a VAR (k+dmax) model was estimated with the additional lag(s) as exogenous variable. The bidirectional GC-test was performed with the null hypothesis that research expenditure does not Granger cause the full-counted publication output or alternatively that the full-counted publications output does not Granger cause research expenditures. Table 2 gives the MWald statistics and the p-values.

The results suggest that at the 5% significance level a unidirectional causality runs from funding to publications for Belgium, Spain, Greece and Ireland. Only for Ireland a bidirectional causality between the two variables can be observed. For Denmark, Finland, Italy, Netherlands, Portugal and Sweden the null hypothesis of no GC in both directions cannot be rejected.

At the 10% significance level for Denmark and Portugal the null hypothesis that research expenditure does not Granger cause full-counted publications can be rejected.

Table 2 – The reslts of the T-Y approach for GC for the countries (CO) in the panel: hypothesis (H0), lag length (k), maximum order of integration of the 2 series (dmax), modified Wald statistics (MWald), p-value and Decision using 5% significance level.

CO	H0	k	dmax	MWald	P-value	Decision (5% crit. value)
BE	EXP does not GC PUB	3	2	13.60	0.00	Reject
	PUB does not GC EXP	3	2	3.50	0.32	Not reject
DK	EXP does not GC PUB	2	1	5.30	0.07	Not reject
	PUB does not GC EXP	2	1	1.20	0.54	Not reject
ES	EXP does not GC PUB	2	1	11.00	0.00	Reject
	PUB does not GC EXP	2	1	0.70	0.72	Not reject
FI	EXP does not GC PUB	3	1	4.80	0.19	Not reject
	PUB does not GC EXP	3	1	3.70	0.29	Not reject
GR	EXP does not GC PUB	3	1	9.60	0.02	Reject
	PUB does not GC EXP	3	1	3.40	0.34	Not reject
IE	EXP does not GC PUB	3	1	26.30	0.00	Reject
	PUB does not GC EXP	3	1	15.80	0.00	Reject
IT	EXP does not GC PUB	1	1	0.10	0.81	Not reject
	PUB does not GC EXP	1	1	0.10	0.78	Not reject
NL	EXP does not GC PUB	1	1	4.00	0.14	Not reject
	PUB does not GC EXP	1	1	1.33	0.51	Not reject
PT	EXP does not GC PUB	3	1	6.80	0.08	Not reject
	PUB does not GC EXP	3	1	4.10	0.25	Not reject
SE	EXP does not GC PUB	2	1	2.60	0.27	Not reject
	PUB does not GC EXP	2	1	1.00	0.60	Not reject

Concluding remarks

For all countries in the panel except Sweden, the growth rate of the full-counted publication output was lower in the last 4-5 years of the period 1999-2017 compared to the years before.

As already mentioned in subsection ‘Structural break in the publication output’, these trends cannot be explained by changes in the coverage of the WoS. One of the possible causes may be the reduction in public research funding. However, there are exceptions. Although the public funding for higher education and public research institutes increased steadily between 1999 and 2017, for Belgium and to some extent Sweden the growth rate of the full-counted publications was lower in the last quarter of the period compared to the earlier years.

For the fractional-counted publication output the reduction in the growth rate was more pronounced compared to the full-counted number of publications and for some countries this growth rate was flat or even negative in the last quarter of the period. The difference in growth rates between the two counting schemes can be explained by an acceleration in the international collaboration. However, it is remarkable that this acceleration happened 3 to 5 years after funding reductions started to be implemented in 7 out of 10 countries in the panel. Moreover around the same year it is also observed in the two countries where research expenditures grew at a nearly steady rate over the entire period.

Part of the time lag between the changes in the funding and the publication output could be explained by the research cycle: obtaining a grant, carrying out the research work, submitting a manuscript and lastly its publication in a peer reviewed journal. Although there are differences between disciplines the publication process takes about a year (Bjork & Salomon, 2013) and in most countries the majority of research grants cover 2 to 4 years.

Further work is necessary to test this hypothesis taking into account the model's limitations. Governments are not the only funding source for research at higher education and public research institutes and their output is not limited to publications in journals covered by the WoS.

One approach to explore the causal relationship between funding and publication output is presented in the second part of this paper. The results can be summarized as follows: for the period under study at the 10% confidence level for 6 out of the 10 countries the Toda-Yamamoto approach for Granger causality shows a unidirectional causality between funding and publication output. At the same confidence level this relationship is even bidirectional for Ireland.

This results suggest that at a country level changes in public research funding affects with a certain delay the nation's publication output. But also more counter-intuitively that changes in publication output have an effect on public research funding. In some national performance-based research funding systems a university's publication output is linked to its further public funding (Hicks, 2012). In some countries to manage at arm's length public research institutes governments use key performance indicators linked to the allocation of funding. It remains an open question to what extent, if implemented, these incentive based schemes have at the national level an impact on the volume of public research funding.

The interpretation of the results of the analysis must however be done carefully. Besides the reduction of a complex world to a model with 2 proxy indicators and the problems related to the use of OECD funding data (Aksnes et al., 2017), the Granger causality test has its own methodological limitations such as its sensitivity on the selection of the optimal lag in the VAR model and the low number of available data points in the time series. Further work is necessary not at least to include multiple breakpoints to be able to use longer time series and more variables such as granted patents as a variable. Notwithstanding the necessary caution, for time series analysis powerful analytical tools have been developed and they are applied in many disciplines. To the best of our knowledge they have not yet been used to study bibliometric data.

References

Akaike, H. (1974), A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19: 716–723.

Aksnes, D.W., Sivertsen, G., van Leeuwen, T.N. & Wendt, K.K. (2017). Measuring the productivity of national R&D systems: Challenges in cross-national comparisons of R&D input and publication output indicators. *Science and Public Policy*, 44, 246-258

Asteriou, D. & Hall, S.G. (2016). *Applied Econometrics*. London: Palgrave (ISBN: 978-1-137-41546-2).

Bjork, B.C & Salomon D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, 7: 914-923.

Brockwell, P.J. & Davis R.A. (2016). *Introduction to Time Series and Forecasting.* Switzerland: Springer International Publishing AG (ISBN 978-3-319-29852-8).

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28, 591–605.

Davidson, R. & MacKinnon, J.G. (2004). *Econometric Theory and Models*. New-York: Oxford University Press (ISBN: 0-19-512372-7).

Elder J. & Kennedy P.E. (2001). Testing for unit roots: What should students be taught? *Journal of Economic Education*, 32, 137-146.

EUA (2011). Impact of the economic crisis on European universities. Retrieved January 2019 from: https://eua.eu/downloads/publications/impact%20of%20the%20economic%20crisis%20on%20european%20universities%20january%202011.pdf.

EViews 10. (2017a). User's Guide I. https://www3.nd.edu/~nmark/FinancialEconometrics/EViews%2010%20Users%20Guide%20I.pdf

EViews 10. (2017b). User's Guide II. https://www3.nd.edu/~nmark/FinancialEconometrics/EViews%2010%20Users%20Guide%20II.pdf

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37: 424–438.

Gylfason, T., Holmström, B., Korkman, S., Söderström, H.T., & Vihriälä, V. (2010). *Nordics in Global Crisis. Vulnerability and resilience.* Helsinki: The Research Institute of the Finnish Economy (ETLA)

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41, 251-261.

Ho, S.Y. & Iyke, B.N. (2017). On the causal links between the stock market and the economy of Hong Kong. *Contemporary Economics*, 11, 343-362.

Jarque, C.M. & Bera, A.K (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*. 6: 255–259.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P & Shin, Y.C. (1992). Testing the null hypothesis of stationarity against the alternative of a unit-root – How sure are we that economic time-series have a unit root. *Journal of Econometrics*, 54, 159-178.

Lepori, B., Reale, M. & Spinello, A.O. (2018). Conceptualising and measuring performance orientation of research funding systems. *Research Evaluation*, 27, 171-183.

Luukkonen, T., Tijssen, R.J.W., Persson & O. Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, 28, 15-36.

Mackinnon, J.G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, 11, 601-618.

Mishra, P.K. (2014). Gold Price and Capital Market Movement in India: The Toda-Yamamoto Approach. *Global Business Review*, 15: 37-45.

Neusser, K. (2016). *Time Series Econometrics*. Switzerland: Springer International Publishing (ISBN: 978-3-319-32861-4).

OECD (2015). *Frascati Manual 2015. Guidelines for collecting and reporting data on research and experimental development.* Paris: OECD Publishing.

Perianes-Rodriguez, A., Waltman, L. & van Eck, N.J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10, 1178-1195.

Perron, P. (1989). The Great Crash, the Oil Price Shock, and the Unit-root Hypothesis. *Econometrica*, 57, 1361-1401.

Perron, P. (2017). Unit Roots and Structural Breaks. *Econometrics*, 5, UNSP 22.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*. 6: 461–464.

Toda, H.Y. & Yamamoto, T. (1995) Statistical Interference in Vector Autoregressions with Possibly Integrated Processes. *Journal of Econometrics*. 66: 225-250.

van Leeuwen, T.N. (2009). Strength and weakness of national science systems: A bibliometric analysis through cooperation patterns. *Scientometrics* 79, 389-408.

van Leeuwen, T.N. (2013) Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship ? *Bibliometrie - Praxis und Forschung*, 2, 1-18

Vogelsang, T.J. & Perron P. (1998). Additional tests for a unit root allowing for a break in the trend function at an unknown time. *International Economic Review*, 39, 1073-1100.

Zapata, H.O. & Rambaldi, A.N. (1997). Monte Carlo evidence on cointegration and causation. *Oxford Bulletin of Economics and Statistics*. 9: 285-298.

Social media attention of the ESI highly cited papers: An Altmetrics-based overview

Jose A. Moral-Munoz[1], Alejandro Salazar[2], David Lucena-Antón[3], Pablo García-Sánchez[4] and Manuel J. Cobo[4]

[1] *joseantonio.moral@uca.es*
Department of Nursing and Physiotherapy, University of Cádiz, 11009 Cádiz (Spain)

[2] *alejandro.salazar@uca.es*
Department of Statistics and Operational Research, University of Cádiz, 11009 Cádiz (Spain)

[3] *davidmanuella@euosuna.org*
Department of Physiotherapy, University of Osuna, 41640 Seville (Spain)

[4] *pablo.garciasanchez@uca.es; manueljesus.cobo@uca.es*
Department of Computer Science and Engineering, University of Cádiz, 11202 Algeciras (Spain)

Abstract

The bibliographic database Web of Science (WoS) provides information about the top scientific actors through its Essential Science Indicators (ESI) product. One of the most important reports is the highly cited papers (HCP), that shows the Top 1% of papers cited in the 22 research field and publication year. New metrics based on online attention, the Altmetrics, have appeared. Using these metrics as a reference, our primary aim in this work is to offer an overview of the social media attention of the HCP. To do so, a match between the 148,767 WoS documents and their Altmetrics scores was performed using their DOIs in Altmetric.com (94,147 records). Then, a descriptive and correlation analysis were performed to describe the results. This analysis shows that Twitter is the leading social platform in which the HCP are disseminated and a source for News. Facebook is the third one, but probably it is not preferable for scientists. Our analysis also indicates that there are research fields with higher social attention and are those with a scientific output more transferable to society. The linkage between basic research and societal impact (patent and policy documents) seems to be low.

Introduction

It is well-known that the Web of Science (WoS) bibliographic database is for most researchers and institutions the primary source of information and the basis for academic career evaluation. The current provider, Clarivate Analytics, offers to the scientific community the InCites Essential Science Indicators (ESI). It is an analytic tool for identifying the top scientific actors through the WoS-indexed items, evaluating the impact of countries, institutes and scientists (Hu, Tian, Xu, Zhang, & Wang, 2018). Furthermore, a fundamental component of ESI is the Highly Cited Papers (HCP), which provides information about the most cited articles in each of the 22 research fields. In other words, it represents the Top 1% of papers cited in their field and publication year (Bauer, Leydesdorff, & Bornmann, 2016).

Nonetheless, although this tool provides relevant information about the influence of the research at different levels, it has its detractors, due to the procedure employed to identify the Top 1% HCP (Hu, Tian, Xu, Wang, et al., 2018; Hu, Tian, Xu, Zhang, et al., 2018; Miranda & Garcia-Carpintero, 2018). In that way, ESI does not consider the influence of the month of publication on the citation counts and establishes the citation window when it is published (without taking into account the online-first option). Moreover, they only use 22 research fields for more than 11,000 journals and is limited to the journal level. According to previous research (Ruiz-Castillo & Waltman, 2015), an article-level classification with 5119 research fields seems to be more suitable to identify the HCP. Finally, ESI takes into account two types of publications, reviews and articles, and it is known that reviews obtain three times more cites than articles (Glänzel, 2008). Despite this, it is interesting to analyse the characteristics of this set of documents, since they attract the attention of the scientific community (Bornmann, Bauer, & Schlagberger, 2018; Docampo & Cram, 2019; Kolditz, Dörhöfer, LaMoreaux, & Kolditz, 2018; Li, 2018; Zhang, Wan, Wang, Zhang, & Wu, 2018).

Although the classic bibliographic databases, such as WoS, are still an important source of information, several changes in the scientific output dissemination are happening. The agreement about the necessity of new forms of evaluating the research impact is clear (Mohammadi, Thelwall, Haustein, & Larivière, 2015). With social media, new actors have emerged (e.g. practitioners, undergraduate students, or lecturers with teaching or professional purposes), and even there are non-authors professionals, which read and perform critical analysis of research articles, and now also share them. Furthermore, new types of academic outputs have been appeared, such as datasets, posters, blogs, or online teaching. Thus, it is now accepted that the research output is not only disseminated within the scientific community as traditionally has been (Bornmann & Williams, 2013). In this sense, the appearance of social media is now highly present in academia, promoting new ways for measuring the impact or attention of the scientific production (Costas, Zahedi, & Wouters, 2015). These types of metrics are known as Altmetrics (Priem et al., 2011), and they are focused on the analysis of the article impact based on the social web. Currently, they do not substitute the traditional metrics (generally, citations), but they could be considered as complementary.

Although a book chapter about the differences between the Top 100 articles according to WoS and Altmetrics was published in 2016 (Banshal, Basu, Singh, & Muhuri, 2018), to our knowledge, no documents are providing an analysis of the whole set of HCP. Therefore, our primary aim is to show an overview of the social media attention of the HCP based on the Altmetrics. Three different subgoals were established to obtain the proposed overview: 1) to show the descriptive data of the social attention in all the platforms measured by Altmetric.com, 2) to analyse the existing correlation among the main social, policy, patent and news platforms, 3) to show the social, policy, patent and news social media attention for each research field, in number of mentions and percentage. To do this, this paper is organized as follows: i) Methods section, in which the procedures employed to obtain and analyse the information are described, ii) Results section, in which the results are detailed, showing the different aspects about the social media attention of the HCP, and finally, iii) Discussion and conclusions section, where the main implications of the study are stated.

Methods

In order to analyse the social media attention of the HCP, several tasks to obtain the data and process it were performed.

Dataset

The set of documents to carry out the analysis was obtained from the WoS list of HCP. In that way, it is important to take into account that historically it has been stated that WoS indexes the most important research output related to the different scientific disciplines since they are considered as a primary criterion in tenure, promotion and other professional decisions (Hodge & Lacasse, 2011; Seipel, 2003). Nevertheless, it has also been stated that this database does not cover some specialities adequately, such as social sciences and humanities (Mongeon & Paul-Hus, 2016).

To perform the proposed study, the list of the DOIs of the documents published by the HCP was obtained from WoS in December 2018. As shown in Table 1, a total of 148,767 documents were retrieved (1,338 documents did not have DOI). Then, they were matched with the data available in Altmetric.com. It is a commercial tool that monitor, analyses and records the online activity around research outputs from a set of online sources, such as blogs, Twitter, Facebook, Google+, news, media, and other sources (Adie & Roe, 2013; Costas et al., 2015). A total of 94,147 records were finally analysed, corresponding to the period 2008-2018. It is important to highlight that only 64% of the documents from WoS matched with Altmetric.com data.

Altmetrics analysis

In order to show the social media attention of the HCP and the correlation between the main social, policy, patent and news mentions, different analyses were carried out. First, the descriptive analysis of all the platforms measured by Almetric.com was performed for the whole set of documents; furthermore, statistical analysis based on correlations was employed to show the relationships among the main social, policy, patent and news mentions. Then, the number of mentions of each research field and the percentage of these mentions and WoS citations were analysed.

In that way, the correlation analysis applied to the whole period could be problematic. First of all, the Altmetric.com started to record the Altmetrics data in 2011 systematically, so we cannot consider the previous years (Robinson-Garcia, Arroyo-Machado, & Torres-Salinas, 2019). Furthermore, the Almetrics data corresponds to December 2018, so the very recent documents tend to have a fewer number of citations and mentions (Thelwall & Nevill, 2018). In order to solve this situation, the period 2012-2015 was selected to obtain a mature dataset (4-7 years old).

The statistical analysis was performed using the software IBM SPSS 25 (Armonk, New York, USA). First, the Kolmogorov-Smirnov test for all the variable mentions was $p<0.001$. Thus, they do not follow a normal distribution. Then, the Spearman correlation test was employed to analyse the correlation between the different variables.

Results

In this section, the characteristics of the 117,674 documents analysed are shown. First, in Table 2 the descriptive analysis of the mentions measured by the Altmetric.com is shown. These results are ranked by the number of mentions, being Twitter clearly the platform with the highest number of mentions (74.74%), followed by News (9.67%), Facebook (4.77%) and Patents (3.53%). In that way, the social platforms with the lowest number of mentions were Syllabi and LinkedIn.

Hereafter, the correlation study results of the main platforms are shown in Table 3. In this sense, the Policy mentions, although they are not in the Top 5 platforms concerning the number of mentions, were selected due to the relevance of this kind of documents. Overall, the mentions on Twitter and Facebook platforms are the most correlated, followed by Twitter and News. Conversely, the Policy and Patent mentions are less correlated. Nevertheless, several aspects will be discussed in the Discussion section.

Table 1. The number of documents with and without DOI and Altmetrics indexed in Web of Science database.

Research Field	*Items*	*Items with DOI*	*Items without DOI*	*With Altmetrics*	*%*	*Without Altmetrics*	*%*
AGRICULTURAL SCIENCES	4,112	4,069	43	2,513	62%	1,556	38%
BIOLOGY & BIOCHEMISTRY	7,431	7,407	24	5,977	81%	1,430	19%
CHEMISTRY	17,234	17,221	13	12,459	72%	4,762	28%
CLINICAL MEDICINE	27,253	26,941	312	15,692	58%	11,249	42%
COMPUTER SCIENCE	3,581	3,412	169	961	28%	2,451	72%
ECONOMICS & BUSINESS	2,720	2,671	49	2,208	83%	463	17%
ENGINEERING	12,712	12,689	23	3,696	29%	8,993	71%
ENVIRONMENT/ECOLOGY	4,887	4,871	16	3,798	78%	1,073	22%
GEOSCIENCES	4,434	4,421	13	2,581	58%	1,840	42%
IMMUNOLOGY	2,588	2,578	10	2,310	90%	268	10%
MATERIALS SCIENCE	8,173	8,158	15	4,516	55%	3,642	45%
MATHEMATICS	4,291	4,090	201	980	24%	3,110	76%
MICROBIOLOGY	2,092	2,089	3	1,670	80%	419	20%
MOLECULAR BIOLOGY & GENETICS	4,685	4,683	2	4,562	97%	121	3%
MULTIDISCIPLINARY	210	210	0	207	99%	3	1%
NEUROSCIENCE & BEHAVIOR	5,197	5,166	31	4,310	83%	856	17%
PHARMACOLOGY & TOXICOLOGY	4,150	4,088	62	3,310	81%	778	19%
PHYSICS	11,174	11,137	37	5,019	45%	6,118	55%
PLANT & ANIMAL SCIENCE	7,319	7,256	63	5,843	81%	1,413	19%
PSYCHIATRY/PSYCHOLOGY	4,127	4,099	28	3,617	88%	482	12%
SOCIAL SCIENCES, GENERAL	8,923	8,702	221	6,914	79%	1,788	21%
SPACE SCIENCE	1,474	1,471	3	1,004	68%	467	32%
Total	148,767	147,429	1,338	94,147	64%	53,282	36%

Finally, in order to analyse the social media attention of each research field, two different figures are shown. First, Figure 1 presents the mean of mentions of these research fields. As can be seen, the Multidisciplinary field (161.84) has the highest mean attention on Twitter, followed by Molecular Biology & Genetics (68.42) and Clinical Medicine (52.74). Those fields with the lowest mean attention in Twitter are Mathematics (2.73), Materials Science (2.99) and Chemistry (3.72). According to Facebook mentions the highest mean attention is for Multidisciplinary (9.38), Clinical Medicine (3.86) and Molecular Biology & Genetics (3.50). The lowest mean attention is for Mathematics (0.07), Engineering (0.20) and Chemistry (0.26). Considering News mentions, the most relevant are Multidisciplinary (21.34), Geosciences (9.88) and Space Sciences (8.89). The lowest mean attention is for Mathematics (0.04), Engineering (0.59) and Computer Science (0.80). Although in these variables some changes can be observed, the main differences are in Patent and Policy mentions. In the case of Patent mentions, the highest mean attention is for Molecular Biology & Genetics (68.42), Multidisciplinary (3.62) and Biology & Biochemistry (3.30), while the lowest are for Space Science (0.01), Economics & Business (0.02) and Social Sciences (0.05). According to the

Policy mentions, the highest mean attention is for Economics & Business (1.83), Multidisciplinary (1.18) and Social Sciences (1.08). The lowest mean attention is for Space Science (0.02), Physics (0.02) and Chemistry (0.02).

Table 2. Description of mentions in Altmetric.com by platform with at least 1% of total share from all the publications indexed in Web of Science that belong to the highly cited papers.

Research Field	*Twitter*	*News*	*Facebook*	*Blog*	*Patent*	*Google*
CLINICAL MEDICINE	827,586	103,186	60,635	21,391	19,746	10,670
MOLECULAR BIOLOGY & GENETICS	312,129	32,743	15,969	11,954	19,004	5,030
SOCIAL SCIENCES, GENERAL	239,355	27,340	11,885	9,528	318	2,878
BIOLOGY & BIOCHEMISTRY	236,759	24,521	11,320	9,248	19,716	5,068
ENVIRONMENT/ECOLOGY	197,255	22,870	8,809	7,995	567	2,468
NEUROSCIENCE & BEHAVIOR	178,812	24,747	14,129	8,903	4,147	4,932
PSYCHIATRY/PSYCHOLOGY	175,135	23,438	10,737	8,097	363	3,552
PLANT & ANIMAL SCIENCE	109,469	10,275	8,433	4,633	2,694	1,583
GEOSCIENCES	101,194	25,506	4,933	7,944	171	2,164
MICROBIOLOGY	68,820	7,565	3,490	2,595	3,646	939
PHYSICS	65,770	13,846	3,396	5,121	9,247	3,305
IMMUNOLOGY	62,675	8,770	5,089	2,106	5,973	788
ECONOMICS & BUSINESS	55,998	5,871	1,797	2,514	43	695
CHEMISTRY	46,294	10,443	3,181	4,363	24,452	1,197
MULTIDISCIPLINARY	33,501	4,417	1,941	1,513	749	636
PHARMACOLOGY & TOXICOLOGY	33,461	3,233	3,971	1,193	7,298	2,152
AGRICULTURAL SCIENCES	27,775	3,679	8,091	1,509	1,102	824
SPACE SCIENCE	26,613	8,928	1,547	2,673	10	1,606
ENGINEERING	18,945	2,185	756	654	3,413	547
COMPUTER SCIENCE	14,400	767	333	438	2,055	386
MATERIALS SCIENCE	13,509	4,173	1,262	1,234	9,286	851
MATHEMATICS	2,672	43	72	217	325	73
Total	2,848,127	368,546	181,776	115,823	134,325	52,344
% Mentions	74.74%	9.67%	4.77%	3.04%	3.52%	1.37%
Mean (SD)	32.52 (161.33)	4.31 (21.57)	2.09 (15.64)	1.46 (6.67)	1.28 (4.38)	0.55 (6.58)

SD: Standard Deviation

Table 3. Spearman correlation coefficients of the different selected variables from the period 2012-2015 Altmetric.com data.

Platforms	*Twitter*	*News*	*Facebook*	*Patent*	*Policy*
Twitter	1	0.526	0.591	-0.051	0.206
News	0.526	1	0.482	0.013	0.203
Facebook	0.591	0.482	1	-0.010	0.165
Patent	-0.051	0.013	-0.010	1	-0.094
Policy	0.206	0.203	0.165	-0.094	1

All the correlations were significant.

Finally, in order to analyse the social media attention of each research field, two different figures are shown. First, Figure 1 presents the mean of mentions of these research fields. As can be seen, the Multidisciplinary field (161.84) has the highest mean attention on Twitter, followed by Molecular Biology & Genetics (68.42) and Clinical Medicine (52.74). Those fields with the lowest mean attention in Twitter are Mathematics (2.73), Materials Science (2.99) and Chemistry (3.72). According to Facebook mentions the highest mean attention is for Multidisciplinary (9.38), Clinical Medicine (3.86) and Molecular Biology & Genetics (3.50). The lowest mean attention is for Mathematics (0.07), Engineering (0.20) and Chemistry (0.26). Considering News mentions, the most relevant are Multidisciplinary (21.34), Geosciences (9.88) and Space Sciences (8.89). The lowest mean attention is for Mathematics (0.04), Engineering (0.59) and Computer Science (0.80). Although in these variables some changes can be observed, the main differences are in Patent and Policy mentions. In the case of Patent mentions, the highest mean attention is for Molecular Biology & Genetics (68.42), Multidisciplinary (3.62) and Biology & Biochemistry (3.30), while the lowest are for Space Science (0.01), Economics & Business (0.02) and Social Sciences (0.05). According to the Policy mentions, the highest mean attention is for Economics & Business (1.83), Multidisciplinary (1.18) and Social Sciences (1.08). The lowest mean attention is for Space Science (0.02), Physics (0.02) and Chemistry (0.02).

Second, Figure 2 shows the percentage of mentions and WoS citations in the different research field. In this sense, the mean of mentions and standard deviations for each social platform is more or less similar (Twitter (3.62± 4.52%), Facebook (4.00± 4.34%), News (3.83± 4.53%) and Patent (3.60± 4.06%)), except to Policy (2.14± 5.03%) that is lower and its standard deviation higher; conversely, the WoS mean citations (4.25± 1.86%) for all the research fields are higher than in social mentions, but with a lower standard deviation. Also, the documents published in the Multidisciplinary field attract the highest attention in the selected platforms, but that does not occur in WoS, where the highest cited field is Molecular Biology & Genetics. In order to better understand these differences between social platforms and WoS, Table 4 shows the mean percentages for the 22 research fields.

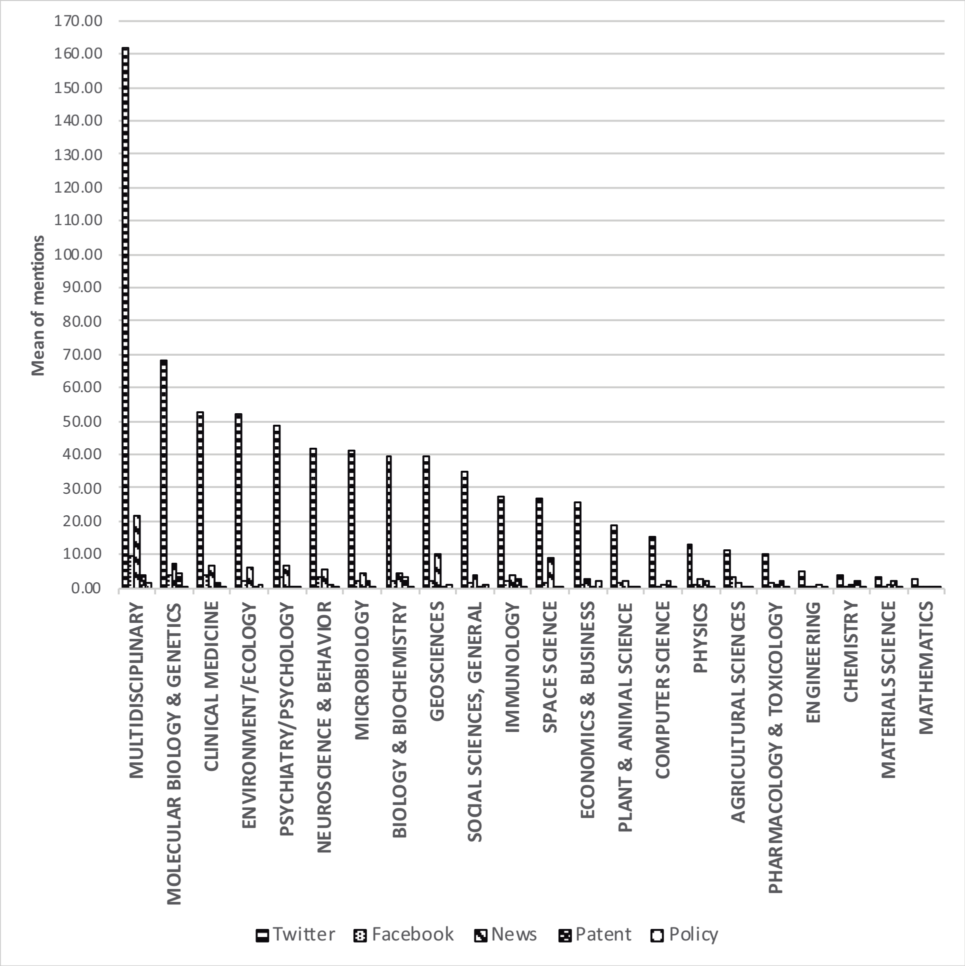


Figure 1. Mean of mentions of the different research field in Twitter, Facebook, News, Patent and Policy documents.

Discussion and conclusions

Once the analysis has been performed, an overview of the social attention of the HCP is discussed. In general, several differences among the research fields were detected. While the mean of citations in WoS seems to be more or less uniform in all the research areas, the social attention unequally focused. Nonetheless, we consider that the results need to be taken with caution. As stated in the Methods section, the percentage of documents without Altmetrics is relatively high (36%); nevertheless, this situation is frequent in this type of analysis (Bornmann & Haunschild, 2018; Didegah, Bowman, & Holmberg, 2018; Moral-Munoz & Cobo, 2018). This data is important in an overall perspective, but it is even more relevant if it is taken into account in each field. Mathematics, Computer Science and Engineering have more than 70% of documents without Altmetrics information. Thus, these documents could be attracting the attention of social media, but they were not measured or the dissemination was performed by different means.

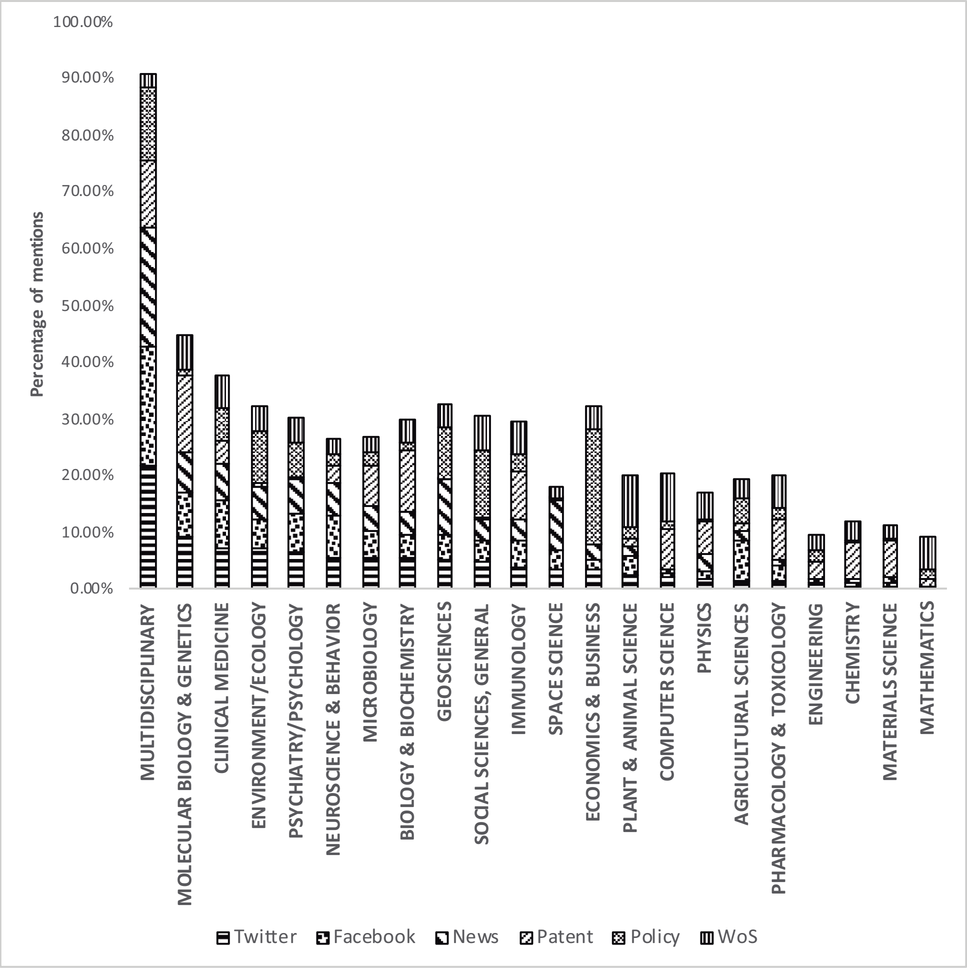


Figure 2. Percentage of mentions and WoS citations of the research fields.

When the total amount of social mentions is considered, Twitter is the platform which the most scientific production is disseminated; followed by the News. In this way, it is well-known the special status of Twitter in the dissemination of relevant news (Banshal et al., 2018; Orellana-Rodriguez & Keane, 2018). Similarly, it has become the place where research consumers (academics, students, clinicians, etc.) converge to report, read, discuss and share new findings. Furthermore, journalists are using this platform to identify potentially highly social relevant news. Thus, it is accepted that Twitter is a way for the scientists to reach a wide popular audience, although it requires a high online engagement and having approximately 2,200 followers at least (Côté & Darling, 2018). Facebook, considered as the most popular social network worldwide with 2,271 millions of users in January 2017 (Statista, 2019), is the third social platform in which the HCP were mentioned. Probably, this is reflecting that scientists are not on Facebook; they prefer to use Twitter as a communication mean. Nonetheless, it is important to highlight that the majority of the Twitter followers of scientists are other scientists (Côté & Darling, 2018). On the contrary, LinkedIn, that is another important social media, has a very low number of mentions. This finding is not in line with previous results (Mas-Bleda, Thelwall, Kousha, & Aguillo, 2014), where it was stated that LinkedIn was the most popular social web for highly cited researches. Consequently, the presence of highly cited researchers on LinkedIn is high, but they do not share their output on this platform.

Table 4. Mean percentages of WoS citations vs. social media attention.

Research Field	*WoS*	*Soc.*
Agricultural Sciences	2.31%	1.49%
Biology & Biochemistry	6.07%	4.18%
Chemistry	6.02%	0.58%
Clinical Medicine	4.37%	6.50%
Computer Science	4.27%	1.54%
Economics & Business	2.73%	2.63%
Engineering	2.73%	0.69%
Environment/Ecology	4.16%	5.94%
Geosciences	3.87%	5.29%
Immunology	6.21%	3.75%
Materials Science	5.63%	0.62%
Mathematics	2.09%	0.37%
Microbiology	4.24%	4.63%
Molecular Biology & Genetics	9.25%	7.75%
Multidisciplinary	8.17%	20.77%
Neuroscience & Behavior	4.95%	5.60%
Pharmacology & Toxicology	3.44%	2.20%
Physics	5.55%	1.77%
Plant & Animal Science	2.59%	1.98%
Psychiatry/Psychology	3.43%	6.40%
Social Sciences, General	2.23%	3.90%
Space Science	5.68%	3.41%

According to the correlation analysis, it was found that only Twitter and Facebook are more correlated; presumably, due to their social nature and the use by a wider group of the population. News are correlated with Twitter and Facebook, such as stated above, journalists are using social platforms as a place where discuss and identify relevant news (Orellana-Rodriguez & Keane, 2018), so this finding is in line with these previous results. In addition, Patent and Policy documents are low correlated to the rest of the platforms. This reflects the question about the linkage between basic research literature and the societal impact, assuming that this impact exists *"when auditable or recorded influence is achieved upon non-academic organization(s) or actor(s) in a sector outside the university sector itself—for instance, by being used by one or more business corporations, government bodies, civil society organizations, media or specialist/professional media organizations or in public debate. As is the case with academic impacts, societal impacts need to be demonstrated rather than assumed. Evidence of external impacts can take the form of references to, citations of or discussion of a person, their work or* *research results"* (Wilsdon et al., 2015). In that way, our findings show a low linkage between the academic and social, though the information provided by Altmetrics has to be taking into account with caution (Bornmann, Haunschild, & Marx, 2016).

Finally, some results about the research field analysis need to be discussed. The first and most important finding is the high social impact of the Multidisciplinary field; with only 207 documents it reaches the highest number of mentions in almost all the platform measured. This is not as surprising when we discover that these documents are published in Nature and Science journals. Thus, the high impact of the documents published in these journals is undisputed. Furthermore, in an overall perspective, the fields attracting the social attentions are those with an output more transferable to the social application, such as Molecular Biology & Genetics, Clinical Medicine or Psychiatry/Psychology. On the other hand, those with the lower application were less mentioned, such as Mathematics, Materials Science and Chemistry. Nonetheless, this does not correspond to the WoS citation relative attention, since the mean percentage for each field is more similar than the social one. Thus this is related to the assumption previously stated in the literature, *"the journals that draw public attention are not the ones that are highly cited except for a small number that receives attention from both the public and academics"* (Banshal et al., 2018).

Although the findings shown in the present work are interesting and present a new overview of the social attention of the HCP, some limitations should be stated. First, due to the characteristics of the Altmetrics, only a percentage of the publications indexed in WoS was analysed; in some research fields, the percentage of documents without Altmetrics is very high. Thus, the results and conclusions of the present study are influenced by the set of documents obtained, due to the methodology drawbacks influential papers could not be included. Second, only a few social platforms were analysed to avoid a text too long and difficult to understand. Third, it should be realised that the output of the different platform needs to be considered in their context. Each platform was designed with a purpose; for example, tweets are designed to be short and very numerous, with an easy way to share. Finally, the intentional mentions by the publisher, the editor of the journal, the author or bots were not analysed. As future research, a study contrasting the HCP with the non-HCP will be performed. This new analysis will discover if there is a different behaviour in social share when a paper is designed as HCP.

The present study shows several findings of the social attention of the HCP. Broadly speaking, the academic, patent and policy attention is not in line with the social attention. Twitter is the main social platform for which the HCP information is disseminated and is related to the share on News. Also, the documents published in journals of a field with a more transferable application to the society obtain higher social attention. Although these are the primary findings of this broad overview, further analysis dividing the documents by periods and deepening in their characteristics is still needed.

Acknowledgments

This work was supported by FEDER funds (TIN2016-75850-R). Authors also want to thank Altmetric.com for facilitating the data access.

References

Banshal, S. K., Basu, A., Singh, V. K., & Muhuri, P. K. (2018). Scientific vs. Public Attention: A Comparison of Top Cited Papers in WoS and Top Papers by Altmetric Score (pp. 81–95). Springer, Singapore.

Bauer, J., Leydesdorff, L., & Bornmann, L. (2016). Highly cited papers in Library and Information Science (LIS): Authors, institutions, and network structures. *Journal of the Association for Information Science and Technology*, *67*(12), 3095–3100.

Bornmann, L., Haunschild, R., & Marx, W. (2016). Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? *Scientometrics*, *109*, 1477–1495.

Bornmann, L., & Williams, R. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics*, *7*(2), 562–574.

Bornmann, L., Bauer, J., & Schlagberger, E. M. (2018). Highly Cited Researchers 2014 and 2015: An investigation of some of the world's most influential scientific minds on the institutional and country level. *COLLNET Journal of Scientometrics and Information Management*, *12*(1), 15–33.

Bornmann, L., & Haunschild, R. (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PLOS ONE*, *13*(5), e0197133.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019.

Côté, I. M., & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops? *FACETS*, *3*(1), 682–694.

Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the differences between citations and altmetrics: An investigation of factors driving altmetrics versus citations for finnish articles. *Journal of the Association for Information Science and Technology*, *69*(6), 832–843.

Docampo, D., & Cram, L. (2019). Highly cited researchers: a moving target. *Scientometrics*.

Glänzel, W. (2008). Seven Myths in Bibliometrics About facts and fiction in quantitative science studies. *Collnet Journal of Scientometrics and Information Management*, *2*(1), 9–17.

Hodge, D. R., & Lacasse, J. R. (2011). Ranking disciplinary journals with the Google Scholar h-index: A new tool for constructing cases for tenure, promotion, and other professional decisions. *Journal of Social Work Education*, *47*(3), 579–596.

Hu, Z., Tian, W., Xu, S., Wang, X., Jiang, L., & Zhang, C. (2018). Why ESI is unreliable in selecting highly cited papers? In *23rd International Conference on Science and Technology Indictors*.

Hu, Z., Tian, W., Xu, S., Zhang, C., & Wang, X. (2018). Four pitfalls in normalizing citation indicators: An investigation of ESI's selection of highly cited papers. *Journal of Informetrics*.

Kolditz, B., Dörhöfer, G., LaMoreaux, J., & Kolditz, O. (2018). Environmental earth sciences—most cited papers: 2015–2016. *Environmental Earth Sciences*, *77*(8), 298.

Li, J. T. (2018). On the advancement of highly cited research in China: An analysis of the highly cited database. *PLoS ONE*, *13*(4).

Mas-Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I. F. (2014). Do highly cited researchers successfully use the social web? *Scientometrics*, *101*(1), 337–356.

Miranda, R., & Garcia-Carpintero, E. (2018). Overcitation and overrepresentation of review papers in the most cited papers. *Journal of Informetrics*, *12*(4), 1015–1030.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. *Journal of the Association for Information Science and Technology*, *66*(9), 1832–1846.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*.

Moral-Munoz, J. A., & Cobo, M. J. (2018). Measuring the online attention of the Rehabilitation Web of Science category : an Altmetrics-based analysis 1. In *23rd International Conference on Science and Technology Indictors*.

Orellana-Rodriguez, C., & Keane, M. T. (2018). Attention to news and its dissemination on Twitter: A survey. *Computer Science Review*, *29*, 74–94.

Priem, J., Piwowar, H. a, Hemminger, B. H., Jason Priem, Heather A. Piwowar, & Bradley H. Hemminger. (2011). Altmetrics in the wild: An exploratory study of impact metrics based on social media. *Metrics 2011: Symposium on Informetric and Scientometric Research. New Orleans, LA, USA*, 1–18.

Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: identifying main topics and actors. *FEMS Microbiology Letters*, *366*(7).

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*(1), 102–117.

Seipel, M. M. O. (2003). Assessing publication for tenure. *Journal of Social Work Education*, *39*(1), 79–88.

Statista. (2019). Global social media ranking 2019.

Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term citation counts? *Journal of Informetrics*, *12*(1), 237–248.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., … Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. *Higher Education Funding Council for England (HEFCE)*. Bristol, Reino Unido: *Higher Education Funding Council for England (HEFCE)*.

Zhang, N., Wan, S., Wang, P., Zhang, P., & Wu, Q. (2018). A bibliometric analysis of highly cited papers in the field of Economics and Business based on the Essential Science Indicators database. *Scientometrics*, *116*(2), 1039–1053.

Industry involvement in biomedical research: authorship, research funding and conflicts of interest

Belén Álvarez-Bornstein[1] and María Bordons[1]

[1] *belen.alvarezchs.csic.es; maria.bordons@cchs.csic.es*
IFS, Centre for Human and Social Sciences (CCHS), Spanish National Research Council (CSIC),
Albasanz 26-28, 28037 Madrid (Spain)

Abstract

Industry plays an important role in biomedical research, especially in clinical and translational research. From a bibliometric perspective, its contribution to research may take different forms, which range from authorship to research funding. Moreover, relationships with industry are increasingly mentioned by authors as potential conflicts of interest (COI), which are very often included in papers under the funding acknowledgment section. This study analyses the contribution of industry in its dual facet of authorship and funder to the scientific publications of Spain-based authors in a biomedical field in Web of Science (WoS) during 2010-2014. Funding acknowledgment data are examined to determine the presence of industry funding and to assess the accuracy of WoS in extracting firms reported as funders, discriminating between funders and entities mentioned due to COI. Frequency of COI disclosure and its main types are described. The influence of a number of variables on the likelihood of papers to present COI is studied. Results show that WoS is not always able to discriminate between funders and companies mentioned due to COI. Industry appears less often as an affiliation of authors than as a source of research funding in this study, but the latter can be overestimated due to the confounding effect of COI.

Introduction

Industry plays an important role in biomedical research. It accounts for almost two thirds of R&D funding in developed countries (OECD, 2015), contributes to major advances in basic and, in particular, in clinical and translational research, and is responsible for a notable share of scientific publications in the biomedical domain.

The involvement of industry in research takes many forms and entails different levels of engagement which ranges from financial support to active participation in the research. Accordingly, from a bibliometric perspective, studying both industry funding and industry authorship can be useful to obtain a more comprehensive view of its contribution to research (Lundberg, Tomson, Lundkvist, Skar & Brommels, 2006).

Approaching the study of industry as a funding source is enhanced at present by the coverage of funding acknowledgments in different databases such as Web of Science or Scopus. Moreover, the fact that the inclusion of funding acknowledgments in publications is becoming mandatory in science increases the interest of this type of study and the significance of the results. However, the analysis of funding data is hampered by their collection in a non-structured section with heterogeneous content comprising financial support but also technical or intellectual assistance as well as conflicts of interest (Álvarez-Bornstein, Morillo & Bordons, 2017). This may lead to errors in the collection or interpretation of funding data, such as confounding funders and companies mentioned due to conflicts of interests in publications, which may distort the results of the studies (Lewison & Sullivan, 2015).

This problem might get worse over the years, as the need to declare potential conflicts of interest (COI), that is, the relationships of authors with entities which might be affected by the research, is increasingly advocated. This is particularly relevant when industry is involved in the research because of its economic interests that may have a negative influence on the objectivity of science and may foster bias in study conclusions. In fact, financial ties of pharmaceutical industry to researchers have been associated with an increased likelihood of reporting results favorable to the intervention being studied (Tungaraza & Poole, 2007; van

Lent, Overbeke & Out, 2013). Disclosure of conflict of interest is thus requested to avoid potential bias and allow readers to form their own opinion about the value of a study.

Previous studies have explored the presence of COI statements in publications in different specialties (e.g.Hakoum et al, 2017), but as far as we know the potential confounding effect of COI on funding data collection has only been analysed in the study of Lewison and Sullivan, who found a decrease in around 40% in the share of papers funded by industry in respiratory diseases after COI removal. Extending the study to other disciplines and including a more detailed analysis of sources of error is needed. Although many professional organisations have developed recommendations about different type of conflict of interest which need to be disclosed (e.g. ICMJE), the lack of a well-established taxonomy hinders this type of study.

In this context, different questions are addressed in this paper. Firstly, we analyse the presence of industry funding and COI in a set of biomedical publications to study whether WoS is able to discriminate correctly among firms mentioned as research funders and those included in a COI statement. Secondly, frequency of COI disclosure and its main types are described while driving factors of COI are explored. Finally, we examine what role (authorship/funding) is played by industry in scientific publications and to what extent funding involvement can be overestimated due to incorrect inclusion of COI.

Methods

Scientific articles by authors based in Spain and published during 2010–2014 were selected from WoS in Cardiac and Cardiovascular Systems (CARD), which is a clinically-oriented domain. The delimitation of this domain was based on the WoS classification of journals into subject categories. Only citable items (articles and reviews) published in English were considered, since FAs are regularly covered by WoS only for items published in this language (Alvarez-Bornstein et al., 2017). A total of 2752 publications were included in the analysis.

Organizations included in the address of authors were normalized and automatically classified into different institutional sectors and papers with an industry affiliation were identified. Concerning funding acknowledgments, they are structured in WoS in three sections: the full text of acknowledgments as it appears in publications (FX), “funding agency” (FA) and “grant number” (GN). In this study, the funding agencies included in the FA section of CARD papers were normalised according to a master file of agencies built at our institution. This master file includes for each agency, its full-normalised name, acronym (if any), institutional sector, type of funding (public, private, mixed) and country.

The full text included in the FX section of publications was manually examined to identify COI and funding agencies and compare them with those included by WoS in the FA section. It was sometimes difficult to determine whether a given funding agency was reported by authors as supporter of the current research or due to previous ties that may lead to COI. To discriminate between both, the guidelines proposed by Lewison & Sullivan (2015) were followed. Detected COI were grouped into seven categories, which include: personal fees, research-support, employment, non-monetary support, patents, stock ownership and drug or equipment supplies. This classification was taken from the literature (Hakoum et al, 2016), after minor adjustments. Multivariable regression analysis was used to assess the predictors of COI presence, that is, the influence of different variables on the likelihood to disclose COI. Main dependent variables included: type of funding (only, public, only private, both), number of funders (1 funder, 2-4, more than 4), foreign funding (0=No, 1=Yes), international collaboration (0=No, 1=Yes), first quartile journal (0=No, 1=Yes) and research level (1=basic, 2=clinical). As a proxy for the basic or clinical nature of the research, we have used the research levels described by Boyack et al (2014), derived from the CHI classification of journals into research levels. It ranges from 1 –most applied– to 4 –most basic–, but they were aggregated into two broad classes: clinical and basic.The statistical package SPSS was used (v.22). It should be noted that the analysis of COI is limited to funded publications, since acknowledgments are collected by WoS only if they include funding information. Finally, the participation of industry as affiliation of authors and as funding source of research was analysed.

Results

A total of 2752 papers in CARD were published by scientists based in Spain during the period 2010-2014. Funding acknowledgments were available in 57% of CARD papers (n=1572) and the government was the main funding source since it was present in 2/3 of the funded publications. Regarding industry, it was acknowledged in almost half of the funded papers (table 1).

Table 1. Distribution of CARD funded papers by type of funding source (FA data)

	No. Papers	*%*
Government	1026	65.27
Industry	710	45.16
NPO	486	30.92
Higher Education	208	13.23
Unknown sector	6	0.38
Total	1572	

NPO= Non-profit organization

The funding text (FX) of papers was manually examined to identify the presence of COI and assess to what extent some of the organizations mentioned due to COI could have been erroneously considered funders and extracted to the FA section by WoS. A total of 710 papers were funded by industry according to agencies collected by WoS in the FA section (table 1), while this figure falls to 570 after our manual revision of data (table 2). In 149 papers (21%) some companies mentioned due to potential conflict of interest were erroneously recorded by WoS as supporters of the current research (false positive data), while no industry funding was recorded in the FA section of 9 papers which acknowledged it in the full text (1%) (false negative data). Accordingly, the share of industry funded papers in CARD declined from 45% to 36% after manual revision of data.

Table 2. Identification of industry funded papers in CARD by type of analysis

	By analysis of FA		
By manual revision of FX	*No industry*	*Industry*	*Total*
No industry	847 (98.90%)	149 (21.00%)	996 (63.60%)
Industry	9 (1.10%)	561 (79.00%)	570 (36.40%)
Total	856 (100%)	710 (100%)	1566* (100%)

*6 papers with unknown institutional sector of funding are not considered

Considering all funded CARD papers, an explicit statement about the presence or absence of COI in the manuscripts was observed in 27% of the cases. Since this figure was surprisingly low, the full text of papers was examined and we noted that COI were not indexed by WoS if they appear in a specific section separated from the acknowledgements. The presence of COI statements rose to 83% after inspection of the full text of funded papers..Specific conflicts of interest were declared in around 36% of the papers, while 46% included a sentence indicating the lack of COI. It should be noted that a higher presence of declared COI were found in industry funded papers (54%) than in the total set of publications (table 3).

Table 3. Distribution of CARD funded papers by presence of COI statement

	All funded papers		Industry funded papers	
	No.Doc.	*%*	*No.Doc.*	*%*
Declared presence of COI	575	36.58	307	53.86
Declared absence of COI	731	46.50	190	33.33
No COI statement	221	14.06	61	10.70
Not identified*	45	2.86	12	2.11
Total	1572	100	570	100

Note: Data obtained by examining the full text of all papers. * Access to the full text was not available

The different types of COI reported by authors and indexed by WoS (n=374 papers) were grouped into seven categories that are shown in table 4. The most frequent cause of conflict was personal fees (in around 77% of the papers), followed by previous research supported by industry (59%).

Table 4. Distribution of CARD funded papers by type of COI

	No.Doc.	*%*
Personal Fees	287	76.74
Consultant	211	56.42
Advisory board	115	30.75
Speaker honoraria	110	29.41
Lecture fee	92	24.60
Honoraria	56	14.97
Speaker bureau	40	10.70
Trials	27	7.22
Educational activities	15	4.01
Royalties	11	2.94
Grant/research support	220	58.82
Employment	90	24.06
Non-monetary support	71	18.98
Stock/shares ownership	50	13.37
Drugs/equipment supplies	24	6.42
Patents	9	2.41

Note: each paper may present more than one type of COI

A binary logistic regression analysis was conducted to assess what variables contribute to explain the presence of COI (Table 5). The most influential variables were type of funding, industry affiliation and number of funders. We can observe that papers with only private support were 7 times more likely to present COI than those with only public support, while those with industry affiliated authors were 8 times more likely to declare presence of COI. Moreover, having more than 4 funders and receiving foreign support increase the likelihood of COI (OR>4 and OR>2, respectively). It should be also noted that papers published in first quartile journals were more likely to present COI (OR=1.9), while basic research was less likely to report conflicts of interest than the clinical one (OR<1).

Table 5. Results of the binary logistic regression analysis for presence of COI

	B	*Wald*	*Exp(B)*
1st Quartile Journal	0.643***	17.306	1.903
Basic research level	-1.249***	34.182	0.287
Type of funding		80.724	
Only private	1.962***	78.577	7.111
Public and private	0.462*	5.247	1.588
(Reference category=only public)			
No. Funders		44.095	
2-4 funders	0.762***	11.861	2.142
>4 funders	1.529***	42.293	4.616
(Reference category= 1 funder)			
Foreign funding	0.925***	34.572	2.522
Industry Affiliation	2.119***	64.384	8.325
Constant	-2.727***	137.336	0.014

**P* < 0.05; ***P* < 0.01; ****P* < 0.001; R^2=0.40; OR = Odds ratio = Exp(B)

Finally, combining data from authorship and funding support, we observe that industry was involved in 24% of CARD papers (668 papers of 2752). Industry contributes more often with funding support (industry in FA) than with active scientists to the research (industry in address of authors). In fact, 85% (n=570) of industry papers received funding from firms, while only in 36% (n=242) an author affiliated to a company participated in the research (Table 6).

Table 6. Distribution of industry papers by type of industry involvement in the research

	No Papers (%)
Industry only in address	98 (14.7%)
Industry only in FA	426 (63.8%)
Industry in FA and address	144 (21.5%)
Total	668 (100%)

Conclusions

This is an on-going research, but some preliminary conclusions can be outlined:

- Algorithms used by Clarivates to extract funding agencies from the FX section are not always able to discriminate correctly between funders of the current research and those included due to previous links of the authors to industry. Around 20% of false positive data are detected (lower share than in the study of Lewison & Sullivan, 2015), which means that industry funding might be overestimated in FA-based studies. Separate disclosure of funding sources and competing interests in papers would be advisable to avoid this type of errors. In fact, WoS mostly confounds COI with funding support when both are not separated or when only COI is disclosed.
- COI are mentioned in 83% of all funded CARD papers and in 87% of the industry funded papers. In fact, competing interests might exist in all papers with industry participation, so COI statements are less prevalent than expected in these cases, in particular because disclosure of conflict of interest (or a statement indicating that there

are not any competing interests) is at present requested by most journals. Moreover, around 5% of funded CARD papers had an industry affiliation but no COI disclosure, which also suggest under-declaration of COI (Darmon et al. 2018).

- COI are more likely to be disclosed in clinical papers, probably because of the greater involvement of industry in this type of research. Moreover, COI are more likely to be disclosed in the most influential journals —Q1 journals—, which suggests it is aligned with good research practices within the scientific community.
- Industry appears more often in CARD publications as source of research funding than as affiliation of authors, so the combination of authorship and funding-based approaches seems to be advisable to obtain a more comprehensive view of the contribution of industry to research.

The results shown correspond to the analysis of CARD –a clinical field-, which will be compared with data from VIROL –a basic field- (still in progress). Differences are expected since both fields differ in their research practices, funding rate and industry involvement..

As for the interest of this research, it should be noted that transparent and clear disclosure of funding sources and potential COI is important to protect research integrity, improve trust in research and enable the development of a wide range of studies with research policy implications.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness (Grant CSO2014 57826 P and predoctoral contract BES 2015 073537)

References

Álvarez-Bornstein, B., Morillo, F., & Bordons, M. (2017). Funding acknowledgments in the Web of Science: completeness and accuracy of collected data. *Scientometrics*, *112*(3), 1793-1812.

Boyack, K.W., Patek, M., Ungar, L.H., Yoon, P. & Klavans, R. (2014). Classification of individual articles from all of science by research level. *Journal of Informetrics*, 8, 1–12.

Darmon, M., Helms, J., De Jong, A., Hjortrup, P. B., Weiss, E., Granholm, A., ... & Azoulay, E. (2018). Time trends in the reporting of conflicts of interest, funding and affiliation with industry in intensive care research: a systematic review. *Intensive care medicine*, 44, 1669-1678.

Hakoum, M. B., Anouti, S., Al-Gibbawi, M., Abou-Jaoude, E. A., Hasbani, D. J., Lopes, L. C., ... & Akl, E. A. (2016). Reporting of financial and non-financial conflicts of interest by authors of systematic reviews: a methodological survey.*BMJ open*, *6*(8), e011997.

Hakoum, M. B., Jouni, N., Abou-Jaoude, E. A., Hasbani, D. J., Abou-Jaoude, E. A., Lopes, L. C., & Guyatt, G. (2017). Authors of clinical trials reported individual and financial conflicts of interest more frequently than institutional and nonfinancial ones: a methodological survey. *Journal of clinical epidemiology*, *87*, 78-86.

International Committee of Medical Journal Editors. ICMJE Form for Disclosure of Potential Conflicts of Interest. http://www.icmje.org/conflicts-of-interest/

Lewison, G., & Sullivan, R. (2015). Conflicts of interest statements on biomedical papers. *Scientometrics*, *102*(3), 2151-2159.

Lundberg, J., Tomson, G., Lundkvist, I., Skar, J., & Brommels, M. (2006). Collaboration uncovered: Exploring the adequacy of measuring university-industry collaboration through co-authorship and funding. *Scientometrics*, *69*(3), 575-589.

OECD (2015). OECD Science, Technology & Industry Scoreboard 2015: Innovation for growth and society. OECD Publishing, Paris. DOI: https://dx.doi.org/10.1787/sti_scoreboard-2015-en

Tungaraza, T., & Poole, R. (2007). Influence of drug company authorship and sponsorship on drug trial outcomes. *The British Journal of Psychiatry*, *191*(1), 82-83.

Van Lent, M., Overbeke, J., & Out, H. J. (2013). Recommendations for a uniform assessment of publication bias related to funding source. *BMC medical research methodology*, *13*(1), 120.

International Register of Academic Book Publishers (IRAP): overview, current state and future challenges

Elea Giménez-Toledo[1], Gunnar Sivertsen[2] and Jorge Mañana-Rodríguez[3]

[1] *elea.gimenez@cchs.csic.es*

Research Group on Scholarly Books (ILIA), Institute of Philosophy (IFS). Spanish National Research Council (CSIC). Madrid, Spain

[2] *gunnar.sivertsen@nifu.no*

Nordic Institute for Studies in Innovation, Research and Education. Oslo, Norway

[3] *jorge.mannana@cchs.csic.es*

Research Group on Scholarly Books (ILIA), Institute of Philosophy (IFS). Spanish National Research Council (CSIC). Madrid, Spain

Abstract

In this contribution, the authors present the background, main underlying concept, data sources, current state and technical and scientific challenges of the International Register of Book Publishers (IRAP) project. This project aims at the creation of a register of scholarly book publishers collecting, normalizing and aggregating different data sources used for evaluation at the national or supra-national level, both for basic research on scholarly books as publication channel and for the provision of aggregated information to all stakeholders involved with scholarly publishing.

Introduction

Looking at the current international trends in research evaluation in the social sciences and humanities (SSH), we observe tendencies to: a) consider full data sources and different types of research results, not only traditional journal publications and metrics; 2) emphasize the societal impact of research; 3) diversify the sources for analyzing outputs and for obtaining indicators (REF, DORA, Leiden, Metric Tide, ENRESSH Manifesto, etc.). All these tendencies have the potential of recognizing the importance of books for scholarly communication in SSH, as well as the diversity of publishers in which humanists and social scientists publish to reach their relevant audiences. The closeness of SSH to the societies and cultures being studied may often imply publishing in local or national languages and with book publishers from the region or country. The diversity of publishers used in SSH can even be seen as requirement for societal impact and for responsible research and innovation. The diversity of scholarly publishers in a country is needed, not only because they publish scientific knowledge which other publishers with a more international profile would not publish, but because they help fulfil the aims of the research itself by communicating with society. These scholarly publishers have an important role in the broader national book market. Given the concentration in the international book market, where a few editorial groups and imprints amass a large portion of the market (*Coufal, 2017),* it seems necessary to facilitate an adequate safeguard of the existing diversity. In this sense, the defense of the publication of scholarly books in the national framework for the aforementioned reasons should also be accompanied by recognition of those publishers and their contribution to research evaluation processes as the select and improve manuscripts before publishing.

The International Register of Academic Book Publishers (IRAP) is being developed as a response to this situation. The register is an initiative within the COST action ENRESSH which has the more general aim of improving the basis and methods for research evaluation in SSH. IRAP in particular is intended to provide structured, precise and quality information on scholarly book publishers, mainly in Europe, but also in United States, Canada and Latin America with the objective of facilitating scientific research, showing the editorial diversity of the different countries This research in progress paper aims to present the state of the art as well as the methodological challenges involved in the development of this project.

International sources for academic books

The two main commercial products providing indicators for scholarly book publishers and individual books are Book Citation Index (Clarivate Analytics) and Scopus Title Expansion Project (Elsevier). The analysis of aspects such as diversity of languages of publication and countries of origin of the publishers revealed some strong biases towards the inclusion of publishers from English-Speaking countries, particularly those that specialize in STEM fields. (Leydesdorff, L., & Felt, U., 2012; Torres Salinas et al., 2014). These findings are consistent with previous analyses of the bibliometric products for journals from the same companies (Moed 2005; Oppenheim and Summers 2008).

Given the relevance of national languages for the SSH, those biases diminish the suitability of the databases for evaluation purposes at the national level in many European countries. As a result, several initiatives have been developed in recent years in several European countries in order to create information systems that allow the provision of reliable information for evaluation purposes. In example, Norway, Finland, Denmark or Belgium (Flanders region) count with Current Research Information Systems (CRIS) which integrate the whole research publication output of the country, thus providing complete data (Sivertsen, 2016). The publication channels (journals or publishers, in example) are then rated in terms of quality by expert panels (Sīle, L. et al. 2017; Sīle, L. et al. 2018)

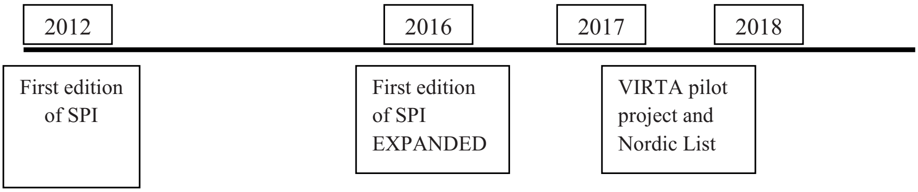
Several other initiatives and approaches have been developed in different countries (Giménez –Toledo et al., 2017; Giménez-Toledo et al., 2018). In Spain, ILIA Research Group (Research Group on Scholarly Books, Spanish National Research Council) has developed *Scholarly Publishers Indicators* (SPI) (Giménez-Toledo et al. 2012); it provides information on the perceived prestige of Spanish and non-Spanish scholarly book publishers, information concerning the manuscript selection processes used by the publishers, the thematic specialization and the presence or absence of a given publisher in five information systems (SPI Expanded, from 2016 onwards). These developments are considered as a reference by the main research evaluation agency in Spain (ANECA).

The development of SPI Expanded has allowed the verification of the potential interest of a register of publishers at the European level, which reflects the presence or absence of the different publishers in various information systems used for evaluation purposes, together with attached information on the quality level that the publisher has in each information system. The Nordic countries have developed a 'Nordic List' merging the respective lists of publication channels from their respective countries through a project funded by NORDFORSK. Finally, there is also a clear interest at the European level in the creation of common infrastructures allowing the convergence in terms of information for research evaluation, as Puuska et al. 2018 (p.1) point out.

International Register of Academic Book Publishers (IRAP)

IRAP is a research proposal and technical development for improving the evaluation of scholarly books while preserving national book industries. Taking into account the described context, the relevance of books in SSH, the need of considering diversity in publishing channels and also the growing research on academic book publishing, a working group of ENRESSH COST action[i] is working on the development of the International Register of Academic Book Publishers. It aims at: a) listing all relevant academic book publishers, it is to say, actually used or to be used by researchers b) providing basic bibliographical information c) offering relevant information for research evaluation purposes such us manuscript selection processes or other useful and transparent information to evaluation agencies, academic institutions, etc.

Figure 1. IRAP history timeline


2012
2016
2017
2018
First edition of SPI
First edition of SPI EXPANDED
VIRTA pilot project and Nordic List

Methodological issues and current state of the Register

SPI Expanded (http://ilia.cchs.csic.es/SPI/expanded_index.html) is the starting point for building up the register. The following table (1) reflects the key features of the Register at its current stage.

Table 1. Key features of the European Register of Scholarly Publishers at its current state.

Feature	*Values*
Number of distinct publishers	5917
Number of different countries of publishers in the Register	110
Number and percentage of University Presses	766 (12.94%)
Number and percentage of publishers with set of ISBN prefixes	2420 (40.89%)

Table 2. Number of distinct publishers in each source.

Database	*Number of different publishers*
BFI (Denmark)	1371
Publication Forum (Finland)	2747
NSD (Norway)	2891
SPI (Spain)	1097
VABB-SHW (Flanders)	134
Book Citation Index	467
Scopus Title Expansion Program	341

Current sources included in the register are: Book Citation Index (2018); Scopus Title Expansion Program (2018); Norwegian NSD (2018) lists; Finnish Publication Forum lists (2018); Danish BFI lists (2018); Spanish Scholarly Publishers Lists (2014); Flemish VABB-SHW lists (2016). The expansion of the register is presently based on the aggregation of information from other existing lists of book publishers at the national level (linked to databases for evaluation purposes). After these expansions, the project has now come to a stage where the scholarly publishers themselves, represented by their international organizations, can be invited to take part in the project. The project is also in contact with ERIH PLUS, the European register of journals in the SSH, which has for a long time planned to include a register of scholarly book publishers.

The construction of the register implies in the first place a work of presentation of the project as well as establishing contact with those responsible of the national information systems (CRIS or equivalent sources), with national organizations in charge of scientific evaluation and with associations of scholarly publishers.

The inclusion of publishers to IRAP through various authorized methods for data importation and crawling implies several phases of technical treatment of the information, but also some challenges from the point of view of research.

Technical challenges

The aggregation of publishers from different sources requires data cleansing and normalization of publishers' names following the common procedures to that effect. This first normalization is followed by a second, manual stage which has the objective of disambiguation of those cases not identified in the first step. In this second step, the official publisher's names available at the Global Register of Publishers (International ISBN Agency) are used.

The depuration of the data from the different source has presented a series of challenges to its reliability

a) Single name and de-duplication

A single name for each publisher is a desirable condition for a register of publishers. Nevertheless, disambiguation and de-duplication of individual names has been one of the main sources of concern in the process of unifying the different sources. On the one hand, the way in which a given publisher is written can take several forms depending on the inclusion of acronyms and the use of common abbreviations for company type. Also, changes in the names of publishers keeping their activity intact are a source of error in the de-duplication process. In those cases, the main source of information for de-duplicating the names have been the use of the Global Register of Publishers (GRP; https://grp.isbn-international.org/), the largest authoritative list of publishers, developed and updated by the International ISBN Agency. It is a challenge for the development of the Register the identification of an optimal process of disambiguation.

b) Imprints and publishing groups

A second type of error sources are those related to the imprints. Many publishing groups have acquired smaller, independent publishers during the course of their business history. Generally, those previously independent publishers (with their set of ISBN prefixes) are included within the publishing group as imprints. With a single name and a series of ISBN prefixes it is possible to find independent publishers, up to a given date and imprints, from the point in time when the publisher was incorporated or merged into a publishing group. The treatment of such instances requires a case by case review of the publishers' history and, apart from time-consuming, the results are not always clear.

c) Co-editions

Co-editions are a further source of error: the co-editing publishers can be kept as independent publishers, the publisher associated to the ISBN prefix can be kept and the rest discarded or co-publications can be discarded beforehand. It remains a challenge to determine which option would be optimal taking into account the different pros and cons of each approach.

Some research challenges

The CRIS systems used in the Nordic countries and Flanders count with complete data on publication in each country. In the case of Norway, Finland and Denmark, the processes and criteria for the classification of publishers are similar. On the other hand, SPI counts with a completely different approach, based on a survey to Spanish scholars on the prestige of both Spanish and foreign publishers. In the case of Flanders, the GRPC provides a source of recognition of individual books but it interpretation in terms of quality or prestige of the publisher is different from the previous ones. Book Citation Index and Scopus provide a completely different set of indicators, based on citation counts and several other information systems include publishers without indicators allowing the categorization or classification of the publishers. It seems clear that the levels, quality labels, citation counts or prestige of the publishers are not comparable. Nevertheless, the presence or absence of the publishers in the databases is driven by an intentional selection process or, in the case of the CRIS-based systems the publishers are rated in a scale according to their quality level and, on the other, the presence of a given publisher in the highest positions in all sources or, by the opposite, in the lower positions can be understood as a potentially useful information (at least in the extreme cases). The option taken until now is to attach the information on the quality of the publisher to its name when available so that, given the proper conditions for its aggregation or further use; there is the possibility of counting with such aggregated data. Furthermore, as pointed out in Mañana-Rodríguez & Pölönen, 2018, the data on the quality 'national' publishers can be imputed into the judgment-based evaluation systems of other countries which do not count with specific information on the quality of foreign, maybe linguistically or culturally distant publishers.

Other relevant point in the development of the register is how to tackle with the information regarding on manuscript selection processes within publishing houses, a critical issue for research evaluation purposes (Giménez-Toledo, Sivertsen & Mañana-Rodríguez). Opening the debate on standards in academic publishing of books entails, among other issues, questioning of peer review and distinguishing the role book editors from the journals editors. A text is not the same published by a publisher or other. The editors provide quality, correction, style or rigor. A scholarly book is part of a publisher and the 'brand' it prints on it (Calasso, 2015). Pointing out the differences between book publishing and journal publishing is a way of breaking down some inertia in research evaluation and science policy. This topic deserves some research for showing different selection practices within publishing houses –apart from peer review- and their relationship with quality or academic recognition. Results from this research might have positive effects in the development of IRAP.

Future expansion of the Register

Different research projects under development are going to offer results on most relevant academic publishers in Colombia (Giménez-Toledo, 2018) and Brazil (Borges de Oliveira, 2018) and other countries in Latin America. It is foreseen to take these results into account for providing information to the Register.

Also, an initial exploration of the information systems used for evaluation in Croatia, Slovenia, Slovakia and the Czech Republic allows concluding that it would be possible to count with structured sets of scholarly publishers from these information systems. On the other hand, further research on the information systems used for evaluation in other European countries such as the UK, France, Italy, Austria or Germany would provide an opportunity for the broadening of the scope of the Register.

References

Borges de Oliveira, Aline (2018). *Los libros en Ciencias Sociales y Humanidades en Brasil: un estudio a partir de los investigadores y las editoriales*. Doctoral Dissertation. Universidad Complutense de Madrid.

Calasso, R. *The Art of the Publisher*. Penguin, 2015.

Coufal, J. (ed.) (2017*). Global Ranking of the publishing industry 2017*. Rüdiger Wischenbart Content and Consulting.

DORA. *San Francisco Declaration on Research Assessment* (2013) http://crln.acrl.org/index.php/crlnews/article/view/9104/9996

ENRESSH (2015). *Memorandum of understanding for the implementation of the COST action European Network for Research Evaluation in the SSH CA15137*. http://enressh.eu/wp-content/uploads/2016/10/CA15137-e.pdf

Giménez-Toledo, E. (2018). *Recognition of academic books in Spanish*. CSIC Research project 201810E125.

Giménez-Toledo, E., Sivertsen, G., & Mañana-Rodríguez, J. (2017). Peer review as a delineation criterion in data sources for the assessment and measurement of scholarly book publishing in social sciences and humanities. In *16th International conference on scientometrics and informetrics*. Wuhan.

Giménez-Toledo, Elea; Mañana-Rodríguez, Jorge & Sivertsen, Gunnar (2017). Scholarly book publishing: Its information sources for evaluation in the social sciences and humanities, *Research Evaluation*, 26, 2, 91–101, https://doi.org/10.1093/reseval

Leydesdorff, L., & Felt, U. (2012). Edited volumes, monographs, and book chapters in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI). *arXiv preprint arXiv*:1204.3717.

Mañana-Rodríguez, J. & Pölonen, J. (2018). Scholarly book publishers' ratings and lists in Finland and Spain: Comparison and assessment of the evaluative potential of merged lists. *ASLIB. Journal of Information Management*, 70, 6, pp. 643-659 https://www.emeraldinsight.com/doi/full/10.1108/AJIM-05-2018-0111

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer Science & Business Media.

Oppenheim, C., & Summers, M. A. (2008). Citation counts and the Research Assessment Exercise, part VI: Unit of assessment 67 (music). *Information Research: An International Electronic Journal*, 13(2).

Research Council of Norway (2017). Evaluation of the Humanities in Norway. Report. O*slo: The Research Council of Norway*. https://www.forskningsradet.no/en/Publications/1220788265688

Research Excellence Framework 2014. *Main panel D criteria* http://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12_2D.pdf

Torres-Salinas, D., Robinson-Garcia, N., Miguel Campanario, J., & Delgado Lopez-Cozar, E. (2014). Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index. *Online Information Review*, 38(1), 24-42.

Wilsdon , J.; Allen, L.; Belfiore, E.; Campbell, P.; Curry, S.; Hill, S.; Jones, R.; Kain, R.; Kerridge, S.; Thelwall, M.; Tinkler, J.; Viney. I.; Hill, J. Wouters, P. and Johnson, B. (2015). *The Metric Tide: The Independent Review of the Role of Metrics in Research Assessment and Management.* doi:10.13140/RG.2.1.4929.1363.

[i] IRAP working group: Gunnar Sivertsen, Vidar Røeggen, Janne Pölönen, Emanuel Kulczycki, Tim Engels, Raf Guns, Elea Giménez, Jorge Mañana, Alesia Zuccala & Kasper Bruun.

Open access journals and the adherence of the elite of Brazilian researchers

Jacqueline Leta[1], Elaine Hipólito dos Santos Costa[2] and Simone Weitzel[3]

[1] *jleta@bioqmed.ufrj.br*
Federal University of Rio de Janeiro, Av. Brigadeiro Trompowisky s/ nº, Prédio do CCS, Bloco B – sala 39,
CEP 21941-590, Rio de Janeiro (Brazil)

[2] *elainebci04@gmail.com*
Federal University of São Paulo, Rua Botucatu, 862, CEP 04023-062, São Paulo (Brazil)

[3] *sweitzel@unirio.br*
Federal University of the State of Rio de Janeiro, Av. Pasteur 458, Prédio do CCCH, sala 418, CEP 22290-240,
Rio de Janeiro (Brazil)

Abstract

The present study aims investigating the adherence of the elite of Brazilian researchers to open access journals as a strategy to publicize their research. The elite was defined as all Brazilian researchers who were receiving in 2016 the most prestigious research fellowship in Brazil. The CNPq productive fellowship 1A is granted to researchers with a very high and continuous performance in scientific publishing and in training human resources as well as a strong visibility inside and outside of Brazil. Information of scientific performance of the elite of Brazilian researchers (n= 1.205) was collected from the Brazilian open directory, Lattes *curriculum*, considering period of 2000-2015. Among the main results, it was found that the largest number of Brazilian articles was published in no-open access journals. Nevertheless, the fraction of articles in open access journals increased from 22.2% in 2000-2003 to 28.6% in 2012-2015 and it was observed an increase in the fraction of researchers with articles in open access journals. The set of results shows that adherence to open access journals by the elite of Brazilian researchers is still low, a worrying scenario considering that this group serves as reference for the rest of the Brazilian scientific community.

Introduction

Since the period known as scientific revolution in the period of 16th – 17th century, communication between researchers has been changing. At that time, books and letters were the main sources for disseminating science knowledge. However, in the middle of the 17th century, scientific journals took on this task, becoming the most important means of registration science knowledge in the following centuries (Meadows, 1974). The success and spread of scientific journals are related to different aspects of the dynamics of scientific communication, including the low cost of publishing in journals, its faster dissemination when compared to books and the insertion of the peer review in the publishing process. Regarding this latter aspect, Zuckerman & Merton (1971) affirmed that process of self-evaluation came to be perceived by scientific community as the mechanism of certification and legitimation of scientific knowledge.

In the 1970s, scientific communication gained its first electronic journal at a time where Internet was still launching. But it was only in the 1990s that many initiatives were organized around the world in order to broaden the number of electronic journals as well as to stimulate researchers to publish in scientific journals in this new format. One of these initiatives was the movement named as Open Access (OA) Movement, which for many authors was born as an alternative against the costly and restricted-access printed journals.

In the last two decades, the OA movement has been structured around some central issues and actions, such as the Budapest Declaration, which states that academic literature should be available online with no costs or other barriers to everyone, including not only the scientists, but scholars, teachers and any other interested person (Budapest, 2002). The Declaration indicates two possible mechanisms to stimulate researchers to join the OA journals: by self-archiving papers at institutional repositories (green road) and by publishing papers in (new) journals committed to the open access ideal (gold road).

It is worth highlighting that the original concept of OA journal, as suggested in the Budapest Declaration, has changed along the last decades, including changes on the aspect of no-payment for authors to publish or for readers to access the articles. In fact, we have witnessed a growth in the number of new OA journals that are supported by article processing charge (APC), especially those that are under the responsibility of commercial publishers. Thus, the expansion of these commercial publishers aimed at the publication of OA journals within different business models such as: immediate open access, hybrid open access, open access late and open promotional access (LAAKSO et al, 2011).

The idea of the article processing charge (APC) as the main source for funding OA journals gained prominence after the publication and diffusion of Finch Report in 2012. This report, entitled "Accessibility, sustainability, excellence: how to expand access to research publications", clearly lists and discusses some recommendations in favour of golden road journals but with APCs, which should be the main strategy for guaranteeing "both effective and sustainable over time, for expanding access to the published findings of research" (Finch, 2012). Initially, APC would display a cost price, as it would cover the primary costs for editing and publishing the on-line journals. Nevertheless, APC values display very high nowadays, what maybe, in contrast to Finch Report, an actually barrier for the consolidation of OA journals, since authors, institutions or governs will have to assume this charge.

An estimate of the OA journals expansion can be obtained from the Directory of Open Access Journals (DOAJ), a repository launched in 2003 with only 300 open access journals. Today, it indexes more than 12.000 titles from almost 130 countries, including Brazil that is among the top-three position in the DOAJ ranking of countries with the highest number of indexed titles: United Kingdom with 1.378, Indonesia with 1.335 and Brazil with 1.242 (DOAJ. 2018).

Brazil's outstanding position in the DOAJ ranking seems to be a result of some national initiatives, as the foundation and expansion of the Scientific Electronic Library Online (SciELO) during the last two decades, as well as some actions led by the Brazilian Institute of Information in Science and Technology (IBICT), such as the introduction of the Electronic Journaling System (a software developed for helping the creation and management of electronic journals) and the launching of the Brazilian Manifesto to Support Open Access to Scientific Information in 2005. These initiatives (and others not described here) widespread open access in Brazil.

Considering the striking presence of Brazilian journals in the DOAJ ranking and, at the same time, the growth of the country's scientific production on an international mainstream (Leta, Thijs, Glänzel, 2013; Sandoval-Romero, Mongeon & Larivière, 2018), we started a large project to investigate to what extend Brazilian scientific community is supporting the OA journal model. The project is based on the premise that publishing in restricted journals (that is, those known as mainstream) or publishing in OA journals would result in different amounts of prestige for researchers or, as conceived by Bourdieu (2004), it would be expressed differently in the accumulation of scientific capital.

Scientific community is strongly structured around a special research grant in Brazil, named CNPq scientific productivity (PQ - the abbreviation in Portuguese) fellowship. The PQ fellowship is granted to a small portion of the whole Brazilian scientific community, that is, those with outstanding levels of scientific and/or technological performance. There are five PQ fellowship categories 1A, 1B, 1C, 1D and 2 granted for active researchers, being the first category considered the elite of Brazilian researchers. In 2016, the number of these PQ fellowships summed 14,342, varying from 1,205 in PQ 1A to 8,037 in PQ 2 (CNPq, 2018).

The present study focuses on Brazilian researchers who received the PQ 1A fellowship in 2016, the highest CNPq - PQ fellowship category. Researchers granted with this fellowship displays a very high and continuous performance in scientific publishing and in training human resources as well as a strong level of collaboration and visibility at the national and international scientific arena. Researchers awarded with this fellowship have some exclusive benefits, such as coordinating research calls with high amount of resources that may be used not only for purchasing equipment but also for supporting students. Such type of benefit increases the gap between the elite of Brazilian researchers (the most top PQ fellowship category, PQ 1A) and the whole Brazilian scientific community, which summed around 200 thousands researchers in 2016 (CNPq, 2016b). In fact, this fellowship acts a sign of distinction (Bourdieu, 2004) and reinforces the hierarchical structure of science in Brazil. Hence, considering the central role of this select group not only as leaders of Brazilian science but also as the group that directly influences the country's science and technology policies, the present paper aims investigating whether they are publishing in OA journals as a strategy to publicize their research. This goal is based in the following research questions: do the elite of Brazilian researchers publish in OA journals? Is this selective and influential group of Brazilian researchers adhering to this new publication format?

It is important to highlight that, for the purposes of this project, OA journals are those whose content is fully available to the Web since its first day of publication, they are catalogued at DOAJ and may or may not be supported by APCs.

Methodology

The information of scientific performance of all 1.205 researchers with the PQ-1A fellowship in 2016 was collected in May 2016 from Lattes *curriculum*. This is a Brazilian open directory created in the 1990’s to compile personal information, academic trajectory, scientific production and other information of all Brazilian scientific community (CNPq, 2018)

Based on the list of names and the webpage of each curriculum, we used the ScriptLattes software (Mena-Chalco & Cesar Junior, 2009) to extract a couple of data from all 1.205 curricula, including: name, sex, area of research, institution of affiliation and details of each publication they published in the period 2000-2015 (coauthors, year, ISSN and source name). It is relevant to underline that only information of documents publish in scientific journals were considered. Simultaneously, a list of the OA journals available at DOAJ was downloaded, including also some additional information, such as ISSN and e-ISSN, country where the journal was edited, whether the journal charges processing article fees, etc.

After cleaning and standardizing the names of the journals and their respective ISSN (or e-ISSN), the list of journals where the 1.205 Brazilian researchers published in the period 2000-2015 was crossed with the list of DOAJ. This was a semi-automatic process and it allowed the classification of each publication of Brazilian researchers as being published in an OA journal, with or without APC, or in a restricted journal.

After all, an excel file was created with personal and academic information as well as with information on scientific publications, especially the indication of published in an open access or in a restricted journal. In order to identify possible changes in the tendencies in publishing of this select group, the data are presented in periods of four years: 2000-2003, 2004-2007, 2008-2011 and 2012-2015. Also, researchers were classified in five groups of adherence to OA journals, 0-20%, 20.1-40%, 40.1-60%, 60.1-80% and 80.1-100%, that are interpreted as very low, low, intermediate, high and very high adherence, respectively.

Results

In order to observe the adherence of the elite of Brazilian researchers to OA journals, the following sessions focus in two main analyses: the proportion of articles in OA journals and the proportion of researchers that publish in OA journals along the studied period.

Number and proportion of articles in OA journals

As shown in table 1, the total number of articles published by the 1.205 Brazilian researchers who received the PQ-1A fellowship, considered here the elite of Brazilian scientific community, increased almost 50% (from 26,363 to 39,067) in the whole period. Such increase was mainly pushed by the 91% of growth of articles published in OA journals in the period (from 5,842 to 11,186).

The higher increase observed in the total number of articles in OA journals led to an increase in the share of these articles when compared to articles published in restricted journals: from 22.2% to 28.6% versus 77.8 % to 71.4%.

Although the number of articles in restricted journals represents almost 2.5 times the number of articles in OA journals in the last period, it is clear that there is a movement towards broadening the adherence to open access journals among this select group of Brazilian researchers.

Table 1. Number and percentage of articles in open access journals and in restricted journals authored by the elite of Brazilian researchers* in four periods.

Period	*Restricted*	*Open#*	*Total*	*Restricted (%)*	*Open (%)*
2000 – 2003	20,521	5,842	26,363	77.8	22.2
2004 – 2007	25,721	8,016	33,737	76.2	23.8
2008 – 2011	28,403	10,193	38,596	73.6	26.4
2012 – 2015	27,881	11,186	39,067	71.4	28.6
2000 – 2015	102,526	35,237	137,763	74.4	25.6

*Includes 1.205 Brazilian researchers that received in 2016 the PQ-1A fellowship, the most prestigious fellowship granted in the country.# Open access journals are defined as those listed in DOAJ.

Although the number of articles in restricted journals represents almost 2.5 times the number of articles in OA journals in the last period, it is clear that there is a movement towards broadening the adherence to OA journals among this select group of Brazilian researchers. Considering this trend, we decided looking closer to the 35,237 articles published in OA journals by Brazilian elite researchers in the whole studied period (2000-2015) to find out whether or not they are supported by article processing charge (APC).

As it can be observed in Figure 1, the studied group publish predominantly articles in OA journals with no APC. In the two first periods, the number of articles in these journals is more than 10-fold the number of articles with APC. Nevertheless, it is clear that, in more recent periods, the number of articles published in OA journals with APC increased substantially. Thus, it is possible that in a near future, the number of articles in OA journals with APC will surpass the number of articles without APC. Such movement is, in a certain way, in accordance with the discussion promoted by Finch Report in which the commitment to promote OA should be to publish in journals with APC.

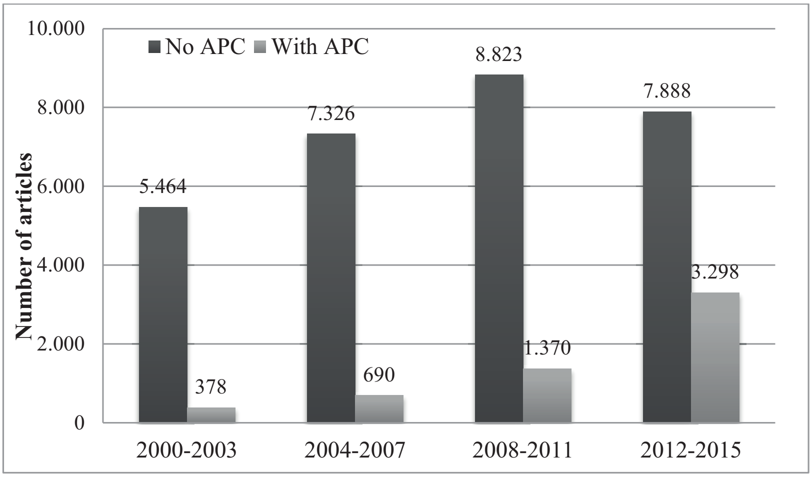


Figure 1. Number of articles in open access journals supported or not by APC (article processing charge) authored by Brazilian researchers* in four periods.

* Includes 1.205 Brazilian researchers that received in 2016 the PQ-1A fellowship, the most prestigious fellowship granted in the country.

In order to better characterize articles published in OA journals, with or without APC, in terms of country where they are edited and their main thematic, we elaborated a list of the top-10 journals with the largest number of articles in both types of journals.

Table 2 presents the top-10 open access journals with no APC and their respective number of articles published by PQ-1A fellowship. The number of articles published in the top-10 journals listed in table 2 sums 1,491 in 2000-2003, 1,719 in 2004-2007, 1,655 in 2008-2011 and 1,057 in 2012-2015, representing 27.3%, 23.5%, 18.8% and 13.4% of the total of each period, respectively. Such decrease in the share indicates that articles are dispersed in a larger number of OA journals in more recent periods.

Looking closer in this list, one first observation is that they are all edited in Brazil and most of them are classified under the fields of agriculture/animal sciences, biological sciences and health sciences, with a single exception, Quimica Nova, which is from Chemistry.

A completely different profile is observed at the top-10 open access journals with APC and the respective number of articles published by the elite of Brazilian researchers (table 3).

Although the amount of number of articles published in these journals is relatively smaller when compared to those of table 2, that is 300 in 2000-2003, 460 in 2004-2007, 629 in 2008-2011 and 1,611 in 2012-2015, they represent higher shares of the total of articles published in OA journals with APC each period, that is, 79.4%, 66.7%, 45.9% and 48.8%, respectively. Such decrease in the share of these journals indicates that articles are dispersed in a larger number of OA journals in more recent periods, but it is important to highlight that articles are still concentrated in a few number of OA journals with APC, a trend that was not observed in the previous analysis (table 2).

Table 2: List of top-10 open access journals with no APC where the elite of Brazilian researchers have published more in four periods.

	2000-2003	*2004-2007*	*2008-2011*	*2012-2015*
1	Brazilian Journal of Animal Science (457)	Brazilian Journal of Animal Science (459)	Brazilian Journal of Animal Science (451)	Pesquisa Veterinária Brasileira (147)
2	Brazilian journal of medical and biological research (199)	Brazilian journal of medical and biological research (213)	Química Nova (179)	Cadernos de Saúde Pública (136)
3	Revista Árvore (124)	Cadernos de Saúde Pública (166)	Cadernos de Saúde Pública (177)	Química Nova (120)
4	Pesquisa Agropecuária Brasileira (114)	Revista Árvore (163)	Revista Brasileira de Ciência do Solo (148)	Revista Brasileira de Psiquiatria (103)
5	Química Nova (111)	Química Nova (158)	Pesquisa Veterinária Brasileira (145)	Revista Brasileira de Zootecnia (100)
6	Cadernos de Saúde Pública (106)	Pesquisa Agropecuária Brasileira (136)	Ciência Rural (124)	Ciência Rural (94)
7	Arquivos de Neuro-Psiquiatria (101)	Ciência Rural (127)	Pesquisa Agropecuária Brasileira (118)	Clinics (93)
8	Ciência e Agrotecnologia (100)	Arquivos de Neuro-Psiquiatria (106)	Clinics (116)	Revista de Saúde Pública (90)
9	Revista Brasileira de Engenharia Agrícola e Ambiental (96)	Revista Brasileira de Engenharia Agrícola e Ambiental (96)	Memórias do Instituto Oswaldo Cruz (104)	Memórias do Instituto Oswaldo Cruz (88)
10	Brazilian Journal of Veterinary Research and Animal Science (83)	Revista de Saúde Pública / Journal of Public Health (95)	Brazilian Journal of Medical and Biological Research (93)	Molecules (86)

* Includes 1.205 Brazilian researchers that received in 2016 the PQ-1A fellowship, the most prestigious fellowship granted in the country. Number of articles in each journal is indicated in parenthesis.

The list of top-10 open access journals with APC shows that the select group of Brazilian researchers tends to publish mostly in journals classified under the fields of biological sciences, health sciences and agriculture. Also, they publish more in national OA journals with APC, especially in the first two periods. In the 2012-2015 periods, Brazilian researchers published in 7 out of the top-10 journals are edited overseas. It is worth noting the contribution of articles published in PLOS' journals, particularly the PlosOne, which articles represent 31% of the total articles published by this group in OA journals with APC in 2012-2015.

In the previous analysis (table 2), we have not observed a trend toward international journals, suggesting that the process of choosing OA journals with APC by Brazilian researchers is in favor of journals with more global visibility. On the other hand, the central role of Brazilian OA journals in both analyses was indeed expected since the country displays one of the highest numbers of OA journal according to DOAJ.

Table 3: List of top-15 open access journals with APC where the elite of Brazilian researchers have published more in four periods.

Item	*2000-2003*	*2004-2007*	*2008-2011*	*2012-2015*
1	Genetics and Molecular Biology on-line (86)	Arquivo Brasileiro de Medicina Veterinária e Zootecnia (145)	Plos One (192)	Plos One (1008)
2	Arquivo Brasileiro de Medicina Veterinária e Zootecnia (62)	Genetics and Molecular Biology on line (111)	Arquivo Brasileiro de Medicina Veterinária e Zootecnia (128)	PLoS Neglected Tropical Diseases (128)
3	Arquivos do Instituto Biológico (28)	Acta Cirúrgica Brasileira (60)	Genetics and Molecular Biology (70)	Arquivo Brasileiro de Medicina Veterinária e Zootecnia (73)
4	Revista da Rede de Enfermagem do Nordeste (24)	Arquivos Brasileiros de Oftalmologia (38)	Genetics and Molecular Biology on-line (55)	Scientific Reports (64)
5	Arquivos Brasileiros de Oftalmologia (21)	Texto & Contexto. Enfermagem (21)	Plos Neglected Tropical Diseases (48)	The Scientific World Journal (62)
6	Acta Cirúrgica Brasileira (20)	Arquivos do Instituto Biológico (20)	Revista da Rede de Enfermagem do Nordeste (33)	BIOMED RES INT (57)
7	Arquivos Brasileiros de Oftalmologia (18)	Biotemas (20)	BMC Genomics (28)	Semina. Ciências Agrárias (56)
8	Texto & Contexto. Enfermagem (15)	BMC Genomics (16)	BMC Microbiology (28)	Semina. Ciências Agrárias (*on-line*) (55)
9	Genetics and Molecular Biology (14)	Genetics and Molecular Biology (16)	Semina. Ciências Agrárias (24)	BMC Genomics (54)
10	Journal of the Brazilian Computer Society (12)	Atmospheric Chemistry and Physics (13)	Arquivos Brasileiros de Oftalmologia (23)	Mediators of Inflammation (54)

* Includes 1.205 Brazilian researchers that received in 2016 the PQ-1A fellowship, the most prestigious fellowship granted in the country. Number of articles in each journal is indicated in parenthesis.

Distribution of researchers that publish in OA journals

As seen in the previous section, the number of articles in OA journals authored by the elite of Brazilian researchers has been increased since 2000, representing almost 29% of total articles published in 2012-2015 period (table 1). This result raised the following question: is such increase related to the adherence of a higher number of researchers that started publishing in OA journals? In order to answer this question, we analysed the distribution of Brazilian researchers according to the number of articles they have published, in restricted or open journals, in each of the four periods of study, as shown in Table 4.

Considering the distribution of researchers by articles published in restricted access journals, it is possible to note that the majority of researchers published 11 or more articles in each period. The most relevant observation, however, is that the distribution of researchers changed very little from one range to another, the exception is the 6-10 articles range, which percentage of researchers reduces along the periods.

A different trend is observed for the distribution of researchers by articles published in OA journals. The largest fraction of researchers is found in the 1-5 articles range, but this fraction was been clearly reduced over time. Such change led to an increase in the number of researchers started publishing more articles in OA journals, especially in 21-50 articles range, where the percentage of researchers has jumped from 4.6% to 11.1%.

Table 4 – Percentage of Brazilian elite researchers* according to number of articles they published in open access journals and in restricted journals in four periods.

	Number of Articles	*2000 - 2003* *%*	*2004 - 2007* *%*	*2008 - 2011* *%*	*2012 – 2015* *%*
Restricted	1 – 5	17.8	14.5	14.1	17.0
	6 -10	22.0	18.2	15.9	15.9
	11 – 20	31.3	27.1	26.3	23.3
	21 – 50	25.1	33.2	34.0	34.2
	51 – 100	3.4	6.1	8.6	8.2
	> 100	0.3	0.9	1.3	1.4
	N. researchers	1,197	1,197	1,192	1,182
Open Acess	1 – 5	64.6	55.4	47.9	44.5
	6 -10	18.7	19.8	20.9	18.9
	11 – 20	11.5	14.8	17.9	20.7
	21 – 50	4.6	9.1	11.1	14.2
	51 – 100	0.7	0.7	2.1	1.6
	> 100	0.0	0.1	0.0	0.1
	N. researchers	898	953	993	1,015

*Includes 1.205 Brazilian researchers that received in 2016 the PQ-1A fellowship, the most prestigious fellowship granted in the country

The data presented in Table 4 indicates that, in fact, more researchers from this selective group are choosing OA journals to diffuse their work. Nevertheless, we questioned whether such change towards OA journals is also observed if we consider the whole set of articles each of these researchers published.

In order to answer this question, we calculated the percentage of articles in OA journals in relation to the total articles each of the 1,205 researchers published in each period. Researchers were classified into five groups of adherence to OA journals: very low (0-20%), low (20.1-40%), intermediate (40.1-60%), high (60.1-80%) and very high (80.1-100%); the first group includes researchers with the lowest fraction of articles in OA journal, while the last group those with the highest fraction.

As shown in Figure 2, researchers with very low adherence comprise the largest group in the four periods. However, it easy to note that this group is been reduced in numbers (from 737 to 520) and in percentage (from 61% to 42%) over time. At the same time, the low and intermediate groups, especially the latter, were enlarged in numbers and percentage, a result that indicates that this select group of researchers is more and more diffusing their research in OA journals.

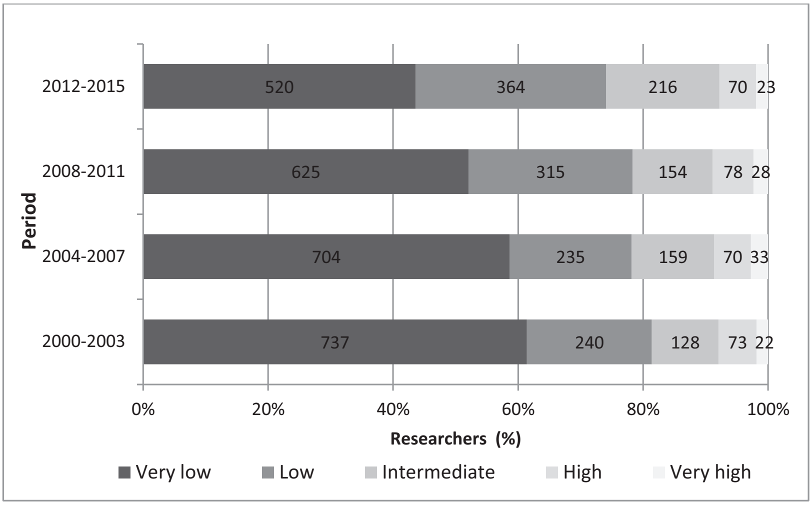


Figure 2: Distribution of Brazilian elite researchers* according to range of adherence to open access journals considering the whole number of articles they published in each period.
(very low = 0 to 20%; low = 20.1% to 40%; Intermediate = 40.1% to 60%; high = 60.1% to 80%; very high = 80.1% to 100%)

Conclusion

Considering the main objective of this study, the set of results presented in previous sections shows that the adherence of the elite of Brazilian researchers is still low, but with a tendency of growing, especially within OA journals with APC. Thus, these results, focused in the adherence of the elite of Brazilian researchers in OA journals, suggest that they are more plural and diverse in choosing the journals.

In spite of this positive movement, the adherence to OA journals among this select group is still far below of that observed in recent literature. Using a combination of sources, Archambault *et al* (2014) have found that around 70% Brazilian scientific production was published in OA journals. One possible explanation for such difference between results found in Archambault's work and results found in this paper includes (a) the extent of sources (different sources versus Lattes only), (b) the period of analysis (2008–2013 versus 2000 – 2015) and concept for OA (wide-ranging versus only DOAJ). In a more recent study, Wang et al (2018) have found that Brazil was among the top-three countries with the highest number of publications in OA journals indexed in WoS in the period from 1990 to 2016; according to the authors, Brazilian share reached 41% in 2014.

It is also noteworthy that increasing adherence to OA journals by the elite of Brazilian researchers seems to be more and more related to the OA model with APC. Such trend is in accordance with the Finch Report guidelines (2012), which points to APCs as a strategy to expand access to the new knowledge published in scientific journals.

The preference for publishing in journals with restricted access as well as the increasing choice for journals with OA supported by APC may be a consequence of a lack of knowledge about the original OA model or even a biased understanding about the quality of OA journals, as pointed in some studies (Nicholas, Huntington & Rowlands, 2005; Rodriguez, 2014). As for Brazilian scientific community, Furnival & Silva-Jerez (2017) surveyed 643 researchers with a position in a Brazilian university discarded the lack of knowledge. The authors found that most of the interviewed display a good level of understanding about the main principles of OA. They declared they are in favour of publishing in OA journals, but whenever possible, they choose higher impact journals. This type of choice supports the notion that scientific capital, expressed here by the impact factor of the journal, as a factor that drives the scientific environment.

Investigating the factors that influence Brazilian researchers is the focus of a study that is in course within our group. We believe that some personal factors (such as age and sex) and academic (such as field and institution) may have an impact in choosing to publish (or not) in OA journals. Also, we are expanding the data analysis, considering the publishing dynamics of Brazilian researchers awarded other PQ fellowship categories as well as researchers without PQ fellowship. In this case, the hypothesis is that the lowest is the prestige of a researcher, the highest is the ratio of papers he/she published in OA journals. Both strategies will allow us to elaborate a better comprehensive view of the adherence to OA journals by the Brazilian scientific community within a theoretical framework that includes some concepts, mainly, accumulation of prestige/scientific capital and power maintenance in science.

By now, the preliminary results shown in this paper suggest that OA journals is not yet used as a strategic means for disseminating scientific knowledge produced by the elite of Brazilian researchers. By one side, it is a surprising finding since Brazil is one of the countries with the largest number of OA journals. By other side, it is an expected finding if we assume that the *status quo* of this selective group is maintained by the journal prestige where they publish, a quality that is not, in general, observed or imputed to OA journals.

Acknowledgments

Authors thanks Dr. Jésus P. Mena-Chalco for the extraction and processing data from Lattes Plataform.

Support

This work is supported by funds from MCTI/CNPq n. 28/2018, Process n. 434146/2018-8, “Comunicação científica brasileira: mapeando fatores que contribuem para a escolha de periódicos de acesso aberto”.

References

ARCHAMBAULT, E., AMYOT, D., DESCHAMPS, P., NICOL, A., PROVENCHER, F., REBOUT, L. & ROBERGE, G. (2014). *Proportion of open access papers published in peer-reviewed journals at the European and World levels-1996-2013*. Montreal: European Commission.. 54 p. (RTD-B6-PP-2011-2: Study to develop a set of indicators to measure open access). Available at: http://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf Access in September, 2017.

BOURDIEU, P. (2004). *Os usos sociais da ciência: por uma sociologia clínica do campo científico*. (Champagne, P. & Landais, E.; trads). São Paulo: Editora UNESP.

BUDAPEST OPEN ACCESS INITIATIVE - BOAI. **Home.** Budapest: ______, 2002. Available at: https://www.budapestopenaccessinitiative.org/read Access in April, 2017.

CNPq. Plataforma Lattes. Curriculo Lattes, 2016. Available at: http://buscatextual.cnpq.br/buscatextual/busca.do?metodo=apresentar Access in January, 2018.

CNPq, Diretorio de Grupos de Pesquisa, 2016b. Available at: http://lattes.cnpq.br/web/dgp/principais-dimensoes Access in January, 2018

DIRECTORY OF OPEN ACCESS JOURNALS - DOAJ. Country of publisher. Disponível em: <https://doaj.org/>. Acesso em: 24 abr. de 2017.

Finch, J. (2012). *Accessibility, Sustainability, Excellence: How to Expand Access to Research Publications*. Available at Research Information Network, http://www.researchinfonet .org/wp-content/uploads/2012/06/Finch-Group-report- FINAL-VERSION.pdf [accessed December 2012].

FURNIVAL, A. C. M.; SILVA-JEREZ, N. S. (2017). Percepções de pesquisadores brasileiros sobre o acesso aberto à literatura científica. *Informação & Sociedade*, v. 27, n. 2, p. 153-166.

LAAKSO. M. et al. (2011). The development of open access journal publishing from 1993 to 2009. *PLOSOne*, v. 6. n. 6.

LETA, J**;** THIJS, B; GLÄNZEL, W. (2013). A macro-level study of science in Brazil: seven years later. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, v. 18, p. 51-66.

MEADOWS J. (1974). *Communication in science*. England: Butterworth & Co. Ltd, 254 p.

MENA-CHALCO, J.P. & CESAR JUNIOR, R.M. (2009). ScriptLattes: an open source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4) : 31-39.

NICHOLAS, D.; HUNTINGTON, P.; ROWLANDS, I. (2011). Open access journal publishing: the views of some of the world's senior authors. *Journal of Documentation*, v. 61, n. 4, p. 497-519.

RODRIGUEZ, J. (2014). Awareness and Attitudes about Open Access Publishing: A Glance at Generational Differences, *The Journal of Academic Librarianship***,** 40, 604–610.

SANDOVAL-ROMERO, V., MONGEON, P., & LARIVIÈRE, V. (2018). Science, technology and innovation policies in Latin-America: fifteen years of scientific output, impact and international collaboration. In *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).

WANG, X., CUI, Y., XU, S., & HU, Z. (2018). The state and evolution of Gold Open Access: A country level analysis. In *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).

ZUCKERMAN, H. Q; MERTON, R. K. (1971) Patterns of evaluation in science: institutionalization, structure and functions of the referee system. *Minerva*, v. 9 n. 1, p. 66-100.

Evaluation Framework for Promoting Gender Equality in Research and Innovation: How to define suitable indicators to evaluate gender equality effects in R&I systems?

Susanne Bührer, Evanthia K. Schmidt, Sybille Reidl, Rachel Palmen, Dora Groo

susanne.buehrer@isi.fraunhofer.de
Fraunhofer ISI, Breslauer Str. 48, DE - 76139 Karlsruhe (Germany)
eks@ps.au.dk
Aarhus University, Bartholins Allé 7, DK - 8000 Aarhus C (Denmark)
sybille.reidl@joanneum.at
Joanneum Research, Sensengasse 1, AT - 1090 Wien (Austria)
rpalmen@uoc.edu
Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss 5, ES - 08860 Castelldefels. (Spain)
dora.groo@nokatud.hu
Association of Hungarian Women in Science, Napraforgo u 17, HU - 1021 Budapest (Hungary

Abstract

Since the topic of "women in research and innovation" has been on the agenda for decades and numerous measures have been implemented at both national and international level to improve the equality of women in the research and innovation systems, it is still unclear under which conditions which measures are most effective. Even less research has been carried out into the effects of better representation of women in terms of (responsible) research and innovation results. Within this paper, an evaluation approach shall be presented, which starts exactly here and uses case studies to show how the concrete implementation of the evaluation model in practice takes place. Furthermore, the results of on the case studies are presented that show how national gender equality measures addressing Higher Education Institutions as well as Research Performing Organisations do not only achieve a better representation of women within these organisations but do also contribute to scientific excellence.

Background and purpose of the study

Despite all efforts undertaken in the past there is no comprehensive and rigorous analytical framework to consider all of the relevant variables in gender equality issues, although there have been a number of European Commission projects such as PRAGES, GENDERA, GenSET, STAGES and GENOVATE, which have explored the gender equality (GE) dimension with different foci. While all these previous studies have illustrated numerous evaluation approaches, concepts, indicators etc. to provide examples of measuring different kinds of impacts, a clear understanding of the mechanisms between different gender equality-related policy initiatives and interventions (inputs) and outputs/results is still not available. In order to address these challenges, EFFORTI (Evaluation Framework for Promoting Gender Equality in Research & Innovation), an EU funded project, aims to clarify the mechanisms between gender equality inputs and the expected results not only on gender equality itself, but also on research and innovation (R&I). The evaluation framework provides the theory and tools for analysing how gender equality-related interventions contribute to the achievement of the three European Research Area’s main objectives on gender equality and how those achievements affect the desired outcomes of (responsible) research and innovation. The uniqueness of the evaluation framework is that it goes beyond conventional research and innovation indicators, taking into account also evaluation dimensions like providing answers to the Grand Challenges and the promotion of Responsible Research and Innovation.

With the rise of the idea of evidence-based policy-making (e.g. Nutley et al. 2002; Solesbury 2001; Sanderson 2002), expectations have grown regarding the use of scientific evidence in policy-making. At the same time, establishing causal relationships between policy interventions and observed changes poses a theoretical challenge as well as empirical and methodological problems. One approach to address these challenges is the theory-based impact evaluation approach (TBIE): In theory-based impact evaluation (TBIE), causality is often defined as a problem of contribution, not attribution. "Why and how" questions are typically being asked instead of "how things would have been without" as counter-factual approaches do. The goal is to answer the "why it works" question by identifying the theory of change ("how things should logically work to produce the desired change") behind the program and assessing its success by comparing theory with actual implementation. The "theories" to be investigated on how gender equality and R&I outcomes interrelate (intervention logics), which in turn link the allocation of resources to the achievement of intended results and finally impacts are still to be developed. These might be complemented by academic theories about public interventions and already existing empirical evidence from former evaluations and impact assessments. The actual results of GE policies will depend both on policy effectiveness and on other context variables. Context factors are organizational structures and cultures, as well as national and regional structures, capabilities and policies. The application of a theory based impact evaluation approach will allow us to take these different levels of influences on policy effectiveness - mechanisms and context - systematically into account. Furthermore, it allows us developing context sensitive and policy specific theories of change.

Methodological Approach

Drawing on already developed and applied indicators in gender equality and R&I research (RIO Observatory, OECD STI Scoreboard etc.), but also on recent studies on RRI indicators (Ravn et al. 2015a, 2015b, European Commission 2015), we carried out a comprehensive desk research as a basis for the collection of a preliminary list of relevant indicators. The team first identified the most relevant indicators according to literature review; clustered these indicators into different categories, dimensions and sub-dimensions, which are based on GE-related literature and smart practice examples implemented in different organisations and contexts; and finally grouped these indicators according to an evaluation logic model. The indicators are differentiated between input, throughput, output, outcome and impact aspects. For each aspect, the indicators are illustrated at micro/individual or team level, meso/organisational level and macro/policy or country level.

The indicators are based on the collection and review of “smart practices” implemented in Europe and beyond. The identification of smart practices was based on an assessment of the practices that are relevant, effective and efficient in the context that they operate in as to their quality of both evaluation and measurement (Kalpazidou Schmidt et al. 2017c). Smart practice examples evaluated measures of different nature and length: some constituted large national programmes with a long-term perspective, while others were of a more limited character. The selection of smart practices was based on the criteria of (1) the quality of the implemented measures, and (2) the impact of the measures. The quality of the measures was assessed based on the parameters of relevance, effectiveness, efficiency, and sustainability of the interventions, while the impact of the measures was assessed in relation to its subjective/objective dimension (Kalpazidou Schmidt & Cacace 2017). Furthermore, we used the existing evidence on the impact of gender diversity on different benefit areas and integrated the respective indicators into the evaluation framework too (for the overview on benefits, see see Bührer / Yorulmaz 2019):

Table 1. Overview on the relation between gender equality and performance

GENDER AND SCIENTIFIC BENEFITS	GENDER AND ECONOMIC BENEFITS
Interdisciplinary & thematic diversity ***Better dissemination of research results and higher share of citations*** ***Social responsiveness and scientific excellence***	***Increased creativity and organizational innovation*** ***Better strategic decision-making & overall competitiveness*** ***Better financial performance*** ***Positive employment effects & job satisfaction*** ***More effective recruiting and retention*** ***Increased organizational attractiveness, brand image and reputation*** ***Better networking and access to customers and markets*** ***Stronger adherence to ethics and rules of conduct***
GENDER AND ENVIRONMENTAL BENEFITS	**GENDER AND SOCIETAL BENEFITS**
More sustainability initiatives ***Higher environmental consciousness in consumption*** ***More eco-innovations***	***Combating gender discrimination through symbolic commitment to equality*** ***Empowerment and confidence*** ***More corporate social responsibility*** ***More supportive and philanthropic behaviour***

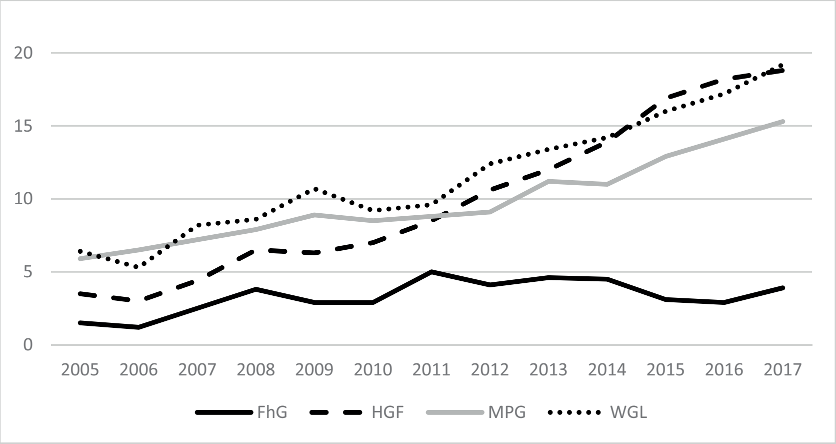
Synthesising the typologies developed by Kalpazidou Schmidt and Cacace (2017) and the fields of action identified by the GENERA project and building on further theoretical and empirical experiences, we developed an intervention typology. Examples of impact stories were developed for a broad spectrum of these intervention types in order to provide examples of the mechanisms regarding intervention intentions and to provide a common framework for understanding the multi-faceted interventions of the cases that will serve as a testing ground for the further development of the tentative evaluation model.

Case Study approach for validation purposes

The EFFORTI intervention logic model forms the conceptual basis for the case study work. The Intervention Logic Model considers inputs, throughputs, and outputs, as well as outcomes and impacts of the former two. The model also aims at showing how, once achieved, these objectives or effects can further affect desired R&I effects such as the number of patents and number of publications and citations, but also new R&I effects, such as providing answers to grand challenges and further promoting RRI. Additionally, the model includes three levels, i.e. team level (research quality, productivity, innovative outputs, and other RRI effects), organisational/ institutional level (workplace quality, recruitment capacity, efficiency, RRI orientation, competitiveness), and country/ system/ policy level (intensity, productivity, ERA orientation, etc.). However, some interventions will most likely overlap between different levels, which was taken into account in the development of the toolbox (EFFORTI Conceptual Evaluation Framework, D3.3, Kalpazidou et al. 2017.8). After having developed a first tentative evaluation framework, a series of case studies is foreseen to validate and further improve the model. Yin (1994.13) defines a case study inquiry as one that *"Investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident."* Therefore, the case study method lends itself to research where contextual factors are highly pertinent to the phenomenon of study (ibid). Case studies as a method have also been used extensively in evaluation research. We used the case study method to inductively build on and validate the evaluation framework. The multiple case study work shed light on those factors and mechanisms that shape and influence the effects of gender equality interventions in R&I on research and innovation outputs. It also attempted to explain how the national/ science system context influences the intervention in terms of the main contextual elements as well as the main agendas, strategies, and policies that shape the intervention.

Case Study Example Germany: The case study examined whether two of the major German flagship programmes to increase the participation of female researchers in the German science system, the "Women Professorship Programme" and the "Pact for Research and Innovation", have actually increased the number of women, especially in leadership positions. In a second step, we analysed whether such an assumed increase influences the publication patterns of authors with German affiliation. The case study was based on literature and desk research as well as bibliometric analysis using Scopus.

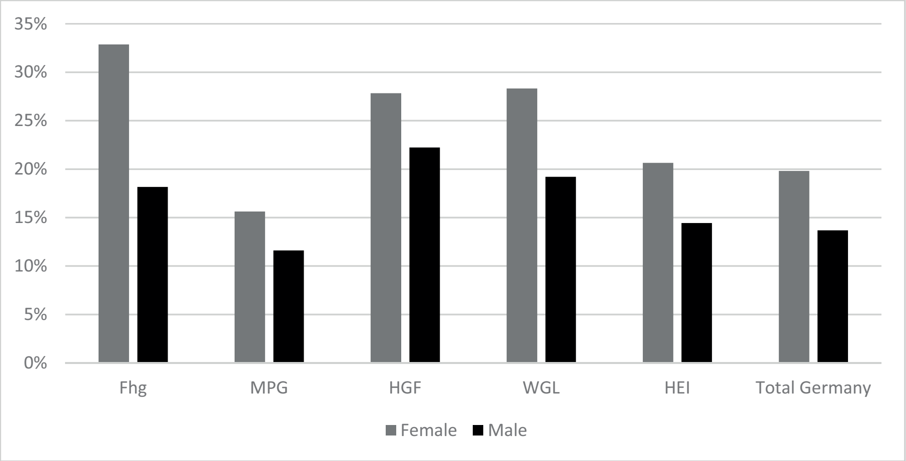
Figure 1. Share of women in leading positions at German non-university research institutions 1992-2016



Source: GWK 2012, GWK 2018

The most important result was that the number of women in research has indeed increased significantly in recent years (see Figure 1) and, accordingly, more women are the (co)authors of scientific publications (see Figure 2). In particular, it can be seen that quality indicators such as citations and excellence rates are high for female authors. This enables us to show that more women in the science system not only bring about a "gain in justice", but also a concrete scientific benefit.

Figure 2. Average annual growth rate in publications by women and men in Germany between 2005 and 2016



Source: Elsevier - Scopus; Fraunhofer ISI calculations.

Discussion and conclusions

Based on a thorough analysis of the relevant knowledge in gender equality, evaluation as well as science and innovation research and the structured analysis of smart practice examples, an evaluation framework has been developed which was then used for the conduction of in total 19 case studies in seven EU countries (Austria, Denmark, France, Germany, Hungary, Spain and Sweden). The case studies cover a broad range of gender equality interventions, from mentoring instruments over structural change approaches up to incentives for integrating gender aspects into research and innovation projects.

Our approach of using a theory-based evaluation framework is appropriate even though it has hardly been possible to measure concrete research and innovation outcomes and impacts of the GE programmes under consideration directly. However, the example presented above show that, not least due to the two big national GE interventions, the role of women academics in the German publication landscape has changed significantly over the past 15 years and there has been a clear increase in the number of (co-)publications by female authors. Furthermore, although the overall number of women has also increased significantly since the introduction of the flagship promotional measures of the Women Professorship Programme and Pact for Research and Innovation, it has not risen to the same extent as women's participation in scientific publications. This means there are clear benefits for Germany in terms of scientific outputs from an increased proportion of women in its scientific workforce.

Our investigation has not yet been able to establish a direct link between the launch of the programmes and the improved representation of women in the different bibliometric indicators. However, we have good reasons to assume that the programmes have at least contributed not only to the higher shares of women within the research performing organisations, but - as a long-term impact - also to improved female publication patterns, especially in terms of citations and excellence rates.

One critique, however, can be that the theory of changes emphasizes differences between male and female researchers and might lead to the promotion of stereotypes. Furthermore, the work with log frames is rather linear and only partly suitable for complex environments, as we are fully aware.

Acknowledgments

The paper is based on the H2020 funded project EFFORTI (Grant agreement No 710470)

References

European Union (2016). *European Innovation Scoreboard 2016*. Brussels.

European Union (Ed.). *Research Innovation Observatory* (https.//rio.jrc.ec.europa.eu/en/stats/key-indicators).

European Commission (Ed.) (2015). Indicators for promoting and monitoring Responsible Research and Innovation. Brussels.

GWK: Joint Science Conference (Gemeinsame Wissenschaftskonferenz). 2012. Pakt für Forschung und Innovation. Monitoring-Bericht 2012. Heft 28. (http://www.gwk-bonn.de/themen/wissenschaftspakte/pakt-fuer-forschung-und-innovation/).

GWK: Joint Science Conference (Gemeinsame Wissenschaftskonferenz). 2018. Pakt für Forschung und Innovation. Monitoring-Bericht 2018. Heft 58. (http://www.gwk-bonn.de/themen/wissenschaftspakte/pakt-fuer-forschung-und-innovation/).

Kalpazidou Schmidt, E. and Cacace, M. (2017). Addressing gender inequality in science. The multifaceted challenge of assessing impact. *Research Evaluation* 2017, 1–13, doi. 10.1093/reseval/rvx003.

Kalpazidou Schmidt, E., Bührer, S., Schraudner, M., Reidl, S. Müller, J., Palmen, R., Haase, S., Graversen, E. K., Holzinger, F., Striebing, C., Groó, D., Klein, S., Rigler, D. and Høg Utoft, E. (2017). *A Conceptual Evaluation Framework for Promoting Gender Equality in Research and Innovation. Toolbox I - A synthesis report. EFFORTI* - Deliverable 3.3.

Nedopil, C., Schauber, C. and Glende, S. (2013): *The art and joy of user integration in AAL-projects.* Brussels.

Nutley, S. Huw D. and Walter, I. (2002). *Evidence Based Policy and Practice: Cross Sector Lessons From the UK*. ESRC UK Centre for Evidence Based Policy and Practice. Working Paper 9. London.

OECD (2014). *Science, Technology and Industry Outlook 2014*, OECD Publishing. Paris.

OECD (2015). *Science, Technology and Industry Scoreboard 2015. Innovation for growth and society*, OECD Publishing. Paris.

Ravn, T., Nielsen, M. W. and Mejlgaard, N. (2015a). *Synthesis report on existing indicators across RRI dimensions*. Progress report D3.1. Monitoring the Evolution and Benefits of Responsible Research and Innovation (MoRRI) (http.//www.technopolis-group.com/morri/).

Ravn, T., Nielsen, M. W. and Mejlgaard, N. (2015b). *Metrics and indicators of Responsible Research and Innovation*. Progress report D3.2. Monitoring the Evolution and Benefits of Responsible Research and Innovation (MoRRI) (http.//www.technopolis-group.com/morri/).

Rommes, E. (2014): *Feminist Interventions in the Design Process*. In: Ernst, W. and Horwath, I. (Eds.): Gender in Science and Technology. Interdisciplinary Approaches. Transcript Verlag, Bielefeld. 41-55.

Sanderson, I. (2002). Evaluation, Policy Learning, and Evidence Based Policy Making, *Public Administration* 80(1), 1-22.

Solesbury, W. (2001). *Evidence Based Policy: Whence it Came and Where it's Going*. In: ESRC UK Centre for Evidence Based Policy and Practice (Ed.). Working Paper 1. London.

Yin, R. (1994). *Case Study Research. Design and Methods*, Second Edition, Sage Publications, London.

Open access challenge at the national level: comprehensive analysis of publication channels used by Finnish researchers in 2016-2017

Janne Pölönen[1], Raf Guns[2], Emanuel Kulczycki[3], Mikael Laakso[4] and Gunnar Sivertsen[5]

[1] *janne.polonen@tsv.fi*
Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki (Finland)

[2] *raf.guns@uantwerpen.be*
University of Antwerp, Faculty of Social Sciences, Centre for R&D Monitoring (ECOOM), Middelheimlaan 1, 2020 Antwerp (Belgium)

[3] *emek@amu.edu.pl*
Adam Mickiewicz University, Scholarly Communication Research Group, Szamarzewskiego 89c, 60-568 Poznań (Poland)

[4] *mikael.laakso@hanken.fi*
Hanken School of economics, Helsinki (Finland)

[5] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education (NIFU), P.O. Box 2815,0608 Tøyen, Oslo (Norway)

Abstract

The purpose of this paper is to provide a comprehensive picture of open access publishing in Finland. Data consists of the complete national peer-reviewed output of 48177 articles and books from 14 Finnish universities in 2016-2017 stored in the VIRTA Publication Information Service. Each publication record contains an indication if it is openly available as Gold or Hybrid OA and/or if it is deposited in OA repository. Using this data, we investigate the share of openly available outputs across fields, as well as journal and book publishing, and analyse the open access status of all 10342 publication channels (journal/series and book publishers) used by Finnish researchers. We also examine the utility of international open access information sources, DOAJ and Bielefeld list for OA journals, and Sherpa/Romeo for self-archiving policies, in estimating the potential for open availability of peer-reviewed outputs, as well as the importance of the largest international commercial publishers in light of these comprehensive national data.

Introduction

In 2016, the European Union member states agreed to "open access to scientific publications as the default option by 2020 and to the best possible re-use of research data as a way to accelerate the transition towards an open science system" (Council of the European Union, 2016). The European Commission supports the transition with a strong open science agenda (European Commission, 2018). Most recently, a group of European research funders known as cOAlition S (which includes Finland's largest research funder, the Academy of Finland) plans to make immediate open access and unrestricted use requirements for all published research funded by the signatories by 2020. This concern, in the first place, journal articles, while a longer transition period is admitted for peer-reviewed book publications.

Finland, like many European countries, is currently developing national strategies and incentives for advancing open access. In 2014-2017, the Ministry of Education and Culture funded a national project, the Open Science and Research Initiative, which set ambitious national targets for the share of openly available publications: 65% in 2017, 75% in 2018 and 100% in 2020 (Ilva, 2017b). According to European Open Science Monitor, the share of OA in Finland is 41.6%, so it ranks 19th out of 36 countries compared. Recently, the Finnish government has approved a new funding model for allocating core-funding annually to universities in 2021-2024. A publication indicator (Pölönen, 2018) will distribute 14 % of the funding, and the publication points based on publication type and channel are multiplied by 1.2

if the peer-reviewed output is openly available (independent of OA mechanism or embargo length). Meanwhile, the Ministry has invested in development of comprehensive national publication data that supports, in addition to the performance-based research funding system (PRFS), monitoring of open access publishing in Finland (Ilva, 2017a).

All countries face the challenge that the vast majority of peer-reviewed outlets used by researchers do not support Gold open access publishing. Many outlets allow individual papers to be made openly available on the publisher website, however, this hybrid OA model is considered unsustainable due to increasing costs and only partial open availability of outputs (Piwowar et al. 2018). It has also been observed that publishing in journals that allow self-archiving (Green OA) does not automatically mean that publications are actually deposited in open access repositories, highlighting a gap between potential and uptake (Laakso, 2014; Björk et al., 2014). Green journals may impose embargoes for the peer-reviewed post-print and publisher version, making them not compliant for example with the Plan S requirements.

Directory of Open Access Journals (DOAJ) and Sherpa/Romeo are the most frequently used information sources to identify Gold and Green OA channels. Nevertheless, even these sources may not provide full coverage of Gold and Green channels. Bielefeld university, for example, provides an ISSN-Matching of Gold OA Journals based – in addition to DOAJ – also to the Directory of Open Access Scholarly Resources (ROAD), PubMed Central (PMC) and Open APC (OAPC) (Wohlgemuth et al. 2016). It is, however, an open question to what extent these existing sources cover the whole variety of publication channels used at the national level, or help estimating the level of open availability of peer-reviewed outputs.

Given that the five largest international commercial publishers account for more than half of the journal output indexed in Web of Science (WoS) (Larivière et al. 2015), most attention at both international and national level is focused on pressuring and/or negotiating with these publishers for open availability of publications. But as Larivière et al. (2015) point out, WoS purports to cover only the most cited international subset of scholarly journals. Especially in the social sciences and humanities (SSH), WoS coverage is seriously wanting due to the importance of national language and book publishing (Kulczycki et al., 2018). In many SSH disciplines, the majority of journal articles are published in national or regional outlets not indexed in WoS (van Leeuwen & Sivertsen 2014; Sivertsen 2016). In addition, up to one half of peer-reviewed outputs in Humanities, and around one-third in the social sciences, are book publications (Engels et al. 2018).

The challenge of implementing and providing open access at a national level has various aspects of which we highlight three. Firstly, analysing what share of national output is published as OA and in how many and what kind of channels. This cannot be easily calculated on the basis of international databases such as WoS or Scopus, or Google Scholar (Martín-Martín, 2018). The implication is that only countries in which current research information systems with full coverage of the SSH publications (Sīle et al, 2018) have been developed can provide an accurate picture of publication patterns and OA publishing in all fields and across publication types. Such an analysis is important as a basis for tailor-made science policy instruments. Secondly, implementing OA at the national level requires infrastructure, tools and resources for open publishing (Sivertsen, 2018). It is an important prerequisite in the ongoing process of flipping journals to the OA model. One possible solution is to use the Open Journal Systems developed by the Public Knowledge Project, to provide a translation of the system into the national language, to provide some training materials, and to ensure resources and create incentives for flipping national journals to the OA model (Ilva, 2018). Another option is to build a national OA platform from the very beginning, as in Croatia (Stojanovski et al, 2009) or in Québec (Larivière and Macaluso, 2011). Thirdly, the challenge of open access at the national level is to provide all mentioned analyses and materials, infrastructure, and platforms also for the peer-reviewed book publications. Scholarly monographs, book chapters and edited volumes play a key role in the social sciences, humanities, and law domains (Montgomery et al, 2018). Thus, not only journal articles but also books should be fully integrated into the OA scholarship.

In this paper we investigate the extent of such challenges by means of a comprehensive analysis of open access publishing in Finland based on complete national publication data. The national information sources remain under-exploited in analysis of open access publishing, and have focused predominantly on journal publishing (Ilva, 2017b; Kronman, 2017; Mikki, 2017; Mikki et al., 2018). Our main research questions are:

1. What is the share of openly available peer-reviewed journal and book publications across fields of science in Finland?
2. How many journals/series and book publishers do Finnish researchers use for publishing peer-reviewed outputs across fields of science, and how large is their share that provides for full, partial or no open availability of Finnish outputs?
3. How large is the share of journals/series that have been identified in VIRTA as OA channels, and what share of these outlets are indexed in DOAJ and Bielefeld list?
4. How large is the share of journals/series that are indexed in Sherpa/Romeo, and does the self-archiving policy influence the share of Finnish outputs in those journals that are openly available?
5. How large is the share of book publishers that are identified in VIRTA as OA channels or have permitted self-archiving?
6. To what extent do the largest international commercial publishers dominate the publishing of Finnish researchers, and are there differences between fields?

Data and Methods

The data consists of unique peer-reviewed outputs published in 2016-2017 that the 14 Finnish universities have reported to the Ministry of Education and Culture and that are stored in the VIRTA publication information service (Sīle et al., 2017; Sīle et al., 2018; Pölönen, 2018). In VIRTA, co-publications of Finnish universities appear as duplicates, however, duplicates are automatically identified on the basis of publication information and indicated in the data. In this study, we use deduplicated publication counts. For each publication, the reporting university has indicated the publication type, OECD field of science, peer review status and open availability. This study includes peer-reviewed articles in journals, books and proceedings, as well as monographs and edited works from all fields of science. For the year 2017 the data collection is not yet entirely complete.

The years 2016 and 2017 have been selected because universities have indicated the open availability of peer-reviewed outputs according to renewed definitions (Ilva, 2017a). Firstly, it is indicated for each output if it is openly available in either Gold or Hybrid OA publication channel. Secondly, it is indicated if the publication is openly available in an OA repository. Information on embargoes or OA licenses, however, is not available in the data. Consequently, it is possible to establish if a peer-reviewed publication is openly available in an OA or Hybrid channel, deposited in a repository, or both. The open availability of a publication can be verified using the URL provided in its metadata. The validation of openly available publications takes place at the universities, and involves both researchers and data collection personnel from the university libraries.

In VIRTA, the publication channel – journal/series or book publisher – of each peer-reviewed output has been identified by matching the publication's bibliographic metadata to the Publication Forum authority list of publication channels. The authority list covers all journals/series and book publishers actually used by researchers affiliated with the 14 Finnish universities. Journals/series include mostly journals but also some book series with ISSN code, as well as some conference proceedings without ISSN. Book publishers mostly have a registered ISBN. For journals/series with ISSN, the Publication Forum channel register contains

the name of the publisher retrieved from the International ISSN Centre. We have complemented the ISSN Centre data with publisher information in the Scopus journal list. It is also indicated if the channel is included in the Directory of Open Access Journals (DOAJ), the Bielefeld list of open access journals, and what the self-archiving policy is according to Sherpa/Romeo.
In 2016-2017, the 14 Finnish universities published a total of 48177 unique peer-reviewed outputs in 10342 publication channels, of which 91.9 % are journals/series and 8.1 % are book publishers (Table 1). 16.5 % of outputs are published with book publishers, while 83.5 % are published in journals/series. Only 62 % of all peer-reviewed outputs are published in journals indexed in Scopus and 52 % in WoS journals (Figure 1). There are, however, large differences between fields in the share of outputs in journals/series, as well as in Scopus and WoS coverage.

Table 1: Number of journals/series and book publishers and their share of outputs by main fields of science

Field of Science	Publication channels			Outputs		
		Journals/ Series	Book publishers		In Journals/ Series	In Book publishers
	N	%	%	N	%	%
Natural sciences	3750	95.3 %	4.7 %	15230	89.7 %	10.3 %
Engineering	1888	91.1 %	8.9 %	6647	81.2 %	18.8 %
Medicine and health	2541	98.4 %	1.6 %	10189	98.5 %	1.5 %
Agriculture and forestry	404	93.3 %	6.7 %	900	95.1 %	4.9 %
Social sciences	3307	89.0 %	11.0 %	10608	72.4 %	27.6 %
Arts & humanities	1782	78.0 %	22.0 %	5920	64.7 %	35.3 %
All fields	10342	91.9 %	8.1 %	48177	83.5 %	16.5 %

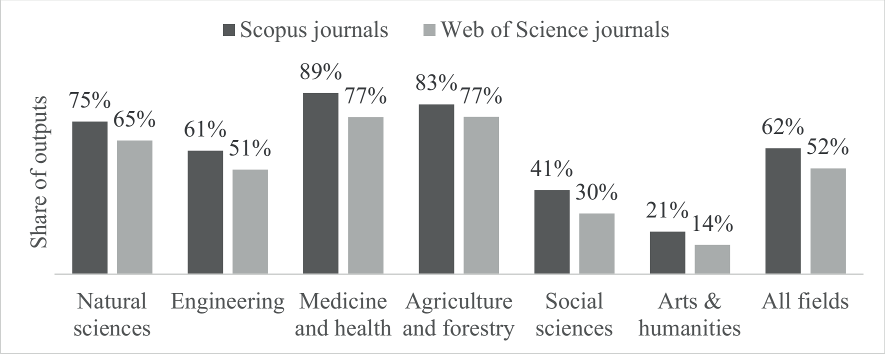


Figure 1: Scopus and WoS coverage of outputs by field of science

Results

Identification of open access status of publication channels based on VIRTA

In VIRTA, there is some evidence of open availability of outputs for one-half of the 10342 publication channels that Finnish researchers have used in 2016-2017 (Table 2). But there is considerable variation in the share of the Finnish outputs that are openly available in different channels. In roughly one-fourth of the channels (24.7 %) all Finnish outputs are openly available, and in one-fourth (25.5 %) of the channels the open availability is only partial. Half (49.8 %) of the publication channels do not seem to have any publications reported as being openly available. This pattern is observed, more or less, in all the main fields, although the share of channels providing no form of open availability is somewhat larger in SSH. This is because open availability is more restricted in the case of book publishers than journal/series.

Table 2: Number of journals/series and book publishers and their share of outputs by main fields of science

Field of science and channel type	Publication Channels (N)	Share of openly available outputs in channel					
		100 %	<100% >=75%	<75% >=50%	<50% >=25%	<25% >0%	0 %
Natural sciences	3750	20.5 %	2.8 %	9.3 %	12.4 %	12.1 %	42.9 %
Engineering	1888	16.4 %	2.8 %	7.5 %	11.5 %	15.7 %	46.2 %
Medicine and health	2541	24.1 %	1.6 %	9.1 %	12.0 %	11.2 %	42.0 %
Agriculture and forestry	404	21.8 %	3.0 %	6.2 %	10.6 %	19.1 %	39.4 %
Social sciences	3307	24.6 %	2.7 %	9.6 %	10.0 %	7.9 %	45.2 %
Arts & humanities	1782	22.6 %	3.3 %	7.8 %	8.1 %	7.2 %	51.0 %
All fields	10342	24.7 %	1.8 %	7.9 %	8.6 %	7.2 %	49.8 %
- Journal/series	9500	25.6 %	1.8 %	8.1 %	9.1 %	7.1 %	48.3 %
- Book publisher	842	14.8 %	1.0 %	6.2 %	3.7 %	8.2 %	66.2 %

Of 9500 journals/series the Finnish researchers used as publication channels, 2074 have at least one peer-reviewed output stored in VIRTA that has been indicated as being openly available in a Gold OA channel (21.8 % of journals/series). In the case of 281 journals/series, outputs are marked as being openly available in both Gold and Hybrid OA channel (3 %), so there is some ambiguity about the OA status of the channel. Outputs from 1137 journals/series have been indicated as being openly available in a Hybrid OA channel (12 %). There are further 1416 journals/series, from which outputs are indicated in VIRTA as being openly available in an OA repository (14.9 %) but not in a Gold or Hybrid channel. For 4592 journals/series used by Finnish researchers we have no indication of any form of open access in VIRTA (48.3 %). The share of journals/series identified as Gold OA channels is smaller for the largest commercial publishers than for the other publishers (Table 3).

Table 3. Type of open access of journals/series by publisher as identified in VIRTA

Publisher	Publication channels	Gold OA channel	Gold or Hybrid channel	Hybrid OA channel	Only self-archiving	No indication of open access
	N	%	%	%	%	%
Elsevier	1373	7.2 %	3.3 %	20.2 %	22.1 %	47.2 %
Springer Nature	605	10.4 %	3.0 %	23.0 %	13.2 %	50.4 %
Wiley-Blackwell	595	8.2 %	2.2 %	18.2 %	16.0 %	55.5 %
Taylor & Francis	553	7.6 %	1.8 %	11.9 %	19.9 %	58.8 %
Sage	273	9.9 %	0.7 %	7.7 %	27.1 %	54.6 %
ACS	46	6.5 %	6.5 %	34.8 %	23.9 %	28.3 %
Other	6055	29.6 %	3.1 %	8.4 %	12.3 %	46.6 %
All journals/series	9500	21.8 %	3.0 %	12.0 %	14.9 %	48.3 %

For the book publishers there is no comprehensive source on OA-status or self-archiving policy, such as DOAJ and Sherpa/Romeo for journals. The VIRTA data indicates, however, that 186 different publishers have at least one output registered as being openly available in a Gold OA channel (22.1 % of the publishers). Outputs from 6 book publishers are indicated as being openly available in both Gold and Hybrid OA channels (0.7 %), so there is ambiguity about the OA status, and 4 book publishers have been identified as Hybrid channels (0.5 %). There are further 89 book publishers, of which outputs have been indicated as being self-archived in an open access repository (10.6 %) but they are not openly available in the publisher website. For 557 book publishers used by Finnish researchers there is no indication of open availability of any outputs (66.2 %). The share of Gold OA channels is about the same for both journals/series and book publishers, however, the availability of Hybrid OA and self-archiving options appear much more limited in the latter case.

There is a considerable difference in share of openly available outputs according to open access status of the channel, as well as according to publications channel type (Figure 2). The share of outputs indicated as being openly available in VIRTA is largest in the identified Gold OA channels (79 %), followed by Hybrid OA channels (31 %) and smallest in journals/series with only self-archived outputs (26 %). The same is observed in case of book publishers, however, the overall share of openly available outputs is much smaller.

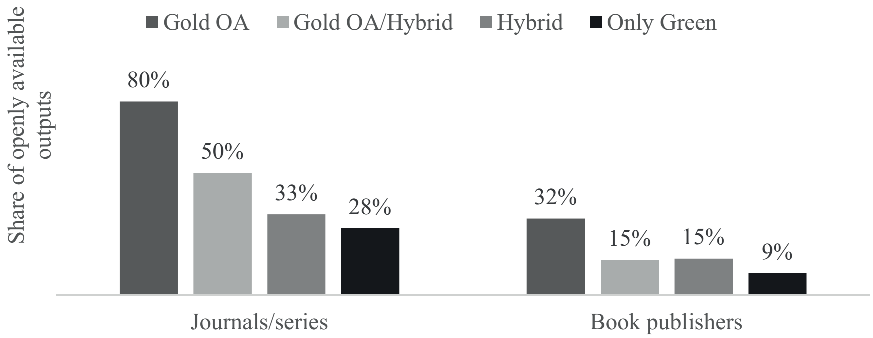


Figure 2. Open Access Status of publication channels as identified in VIRTA and share of openly available outputs.

Comparison of VIRTA based open access status of journals/series with DOAJ and Bielefeld list

Of all 9500 journals/series used by Finnish researchers, 1237 are Gold OA journals indexed in DOAJ with or without a green tick (13 %) (Table 4). Furthermore, 372 journals/series are included in the Bielefeld list of open access journals but are not indexed in DOAJ (3.9 %). Comparison with VIRTA data suggests that inclusion of journal/series in DOAJ and the Bielefeld list is a good predictor of open access, as 96 % outputs from channels in DOAJ and 78 % from channels in Bielefeld are actually indicated in VIRTA as being openly available. For journals/series outside DOAJ and the Bielefeld list, the share of openly available outputs is considerably smaller (25 %), yet as large as 54 % in case of those journals/series indicated as Gold OA channels.

Together, DOAJ and the Bielefeld list cover over 60 % of all journals/series identified as Gold OA channels based on the VIRTA data (including Gold/Hybrid OA journals). Combining all information sources it is possible to identify a total of 2553 potential Gold OA journals, of which 48 % based on DOAJ, 15 % based on the Bielefeld list, and an additional 37 % based on VIRTA (Figure 2). It is noteworthy that 37 % of all potential Gold OA channels are not included in either DOAJ or the Bielefeld list (it has not been possible for us to manually verify their Gold OA status).

Table 4. Comparison of VIRTA based Open Access Status with DOAJ and Bielefeld list, and share of openly available outputs.

	Publication channels	Outputs	Openly available outputs	Openly available outputs
	N	N	N	%
DOAJ	1237	6013	5765	95.9 %
+Bielefeld	372	1249	973	77.9 %
Not in DOAJ or Bielefeld list	7891	32977	8289	25.1 %
All publication channels	9500	40239	15027	37.3 %

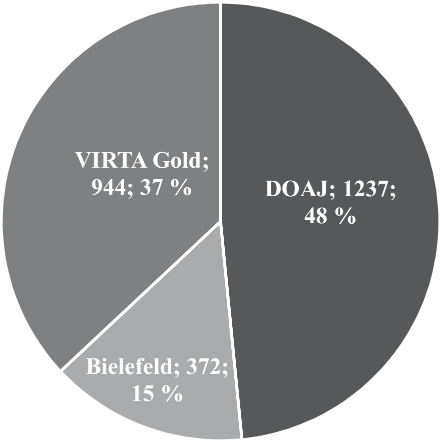


Figure 3. Share of potential Gold OA journals identified based on DOAJ, Bielefeld list and VIRTA

Comparison of Sherpa/Romeo self-archiving policies of journals/series with DOAJ and Bielefeld list and open availability of outputs in VIRTA

Sherpa/Romeo codes indicating the self-archiving policies cover 7537 journals/series (79 % of all journals/series) used by Finnish researchers (Table 5). Sherpa/Romeo includes almost all DOAJ journals (95 %), and a considerable share of Bielefeld listed journals (43 %). Overall, however, the inclusion of journals in Sherpa/Romeo is not a very good predictor of open availability of outputs, the share of which in VIRTA is practically the same as in the case of journals not included in the Sherpa/Romeo service (Figure 3). The share of openly available outputs is much larger for channels included in DOAJ or the Bielefeld list, than for the other channels included in Sherpa/Romeo. Availability of the Gold route clearly has resulted in more complete open availability of outputs than the Green route. The differences in self-archiving policy do not make a great difference, especially if we look at journals/series not in DOAJ or the Bielefeld list.

Table 6. Sherpa/Romeo codes and share of openly available outputs.

Sherpa/Romeo self-archiving policy	Publication channels	
	N	%
Green (publisher version)	5034	53.0 %
Blue (post-print)	361	3.8 %
Yellow (pre-print)	1346	14.2 %
White (none)	267	2.8 %
Gray (unknown)	529	5.6 %
Not in Sherpa/Romeo	1963	20.7 %
All publication channels	9500	100 %

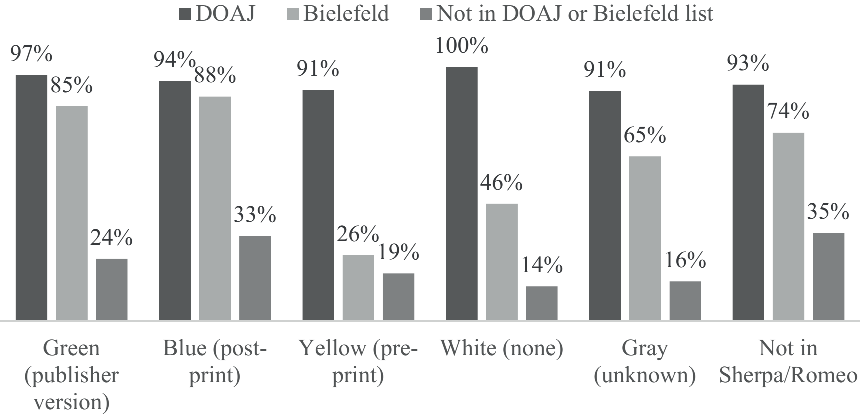


Figure 3. Sherpa/Romeo codes and share of openly available outputs in DOAJ and Bielefeld listed journals

The importance of the largest international commercial publishers and open availability of the outputs across fields

Publication channels owned by Elsevier account for 20.1 % of the 14 Finnish universities' journal outputs in all fields of science counted together (Table 6). Next come Springer Nature (12.8 %), Wiley-Blackwell (9.2 %) and Taylor & Francis (6.8 %). Sage and the American Chemical Society (ACS), which are often also considered among the "big" commercial publishers, account for 2.3 % and 1.9 % respectively. Taken together, these publishers account for 53.1 % of the journal output. This is consonant with studies based on Web of Science data, even though national VIRTA data includes many journals/series not indexed in WoS. If we take into account also peer-reviewed conference articles and book publications, these publishers' joint share of Finish output diminishes to less than half (44.3 %). VIRTA data also suggests that the commercial publishers included in this study are most dominant in Medicine and Agriculture, and least dominant in the social sciences and especially humanities. Thus, our study corroborates the findings of Larivière et al. (2015) concerning the humanities being the field least dominated by the big publishers. In our analysis, however, social sciences is among the least, not the most, dominated fields (this holds true even if we limit analysis to journal articles).

Of all Finnish 2016-2017 peer-reviewed outputs one-third is openly available (33.6 %) and two-thirds are not openly available (66.4 %) (Table 7). The share of openly available outputs is somewhat smaller in case of the large commercial publishers (except Springer Nature) than other publishers. The share of openly available outputs is also larger among journal articles than conference and book publications. Overall, the differences between fields are not great. Nevertheless, natural sciences (39 %) and medicine (37 %) have the largest, while SSH (30 %) and especially engineering (26 %) have smallest share of openly available outputs (Figure 4).

Table 6. The six largest commercial publishers' share of outputs by field of science and publication type

Field and publication type	Outputs	Elsevier	Springer Nature	Wiley-Blackwell	Taylor & Francis	Sage	ACS	Other
	N	%	%	%	%	%	%	%
Natural sciences	15230	17.5 %	17.8 %	8.1 %	2.3 %	0.4 %	3.0 %	50.9 %
Engineering	6647	22.5 %	8.9 %	4.3 %	2.9 %	0.9 %	2.4 %	58.1 %
Medicine and health	10189	19.9 %	17.7 %	13.1 %	6.3 %	2.4 %	0.6 %	40.1 %
Agriculture and forestry	900	27.6 %	12.2 %	10.1 %	5.4 %	0.4 %	0.7 %	43.6 %
Social sciences	10608	8.6 %	8.7 %	3.9 %	14.0 %	4.2 %	0.0 %	60.5 %
Arts & humanities	5920	2.1 %	4.5 %	1.4 %	8.7 %	1.2 %	0.0 %	82.1 %
All fields	48177	14.9 %	12.8 %	6.9 %	6.6 %	1.8 %	1.4 %	55.7 %
- Journal article	34507	20.1 %	12.8 %	9.2 %	6.8 %	2.3 %	1.9 %	46.9 %
- Conference article	6283	2.6 %	9.9 %	0.3 %	0.9 %	0.0 %	0.0 %	86.3 %
- Book publication	7387	1.3 %	15.0 %	1.9 %	10.4 %	0.6 %	0.0 %	70.7 %

Table 7. Type of open availability of outputs by publisher and publication type

Publisher	Outputs	Only publisher service	Publisher service and self-archived	Only self-archived	Not open access
	N	%	%	%	%
Elsevier	7188	5.2 %	8.8 %	11.7 %	74.3 %
Springer Nature	6164	8.4 %	25.5 %	7.0 %	59.2 %
Wiley-Blackwell	3328	5.0 %	9.4 %	8.4 %	77.2 %
Taylor & Francis	3163	3.4 %	6.7 %	11.1 %	78.8 %
Sage	855	3.9 %	6.8 %	14.6 %	74.7 %
ACS	651	1.8 %	5.5 %	9.1 %	83.6 %
Other	26828	13.2 %	16.2 %	8.2 %	62.4 %
All publishers	48177	9.8 %	14.9 %	8.9 %	66.4 %
- Journal article	34507	9.9 %	18.7 %	9.6 %	61.8 %
- Conference article	6283	11.8 %	7.2 %	9.6 %	71.4 %
- Book publication	7387	7.9 %	3.4 %	5.1 %	83.5 %

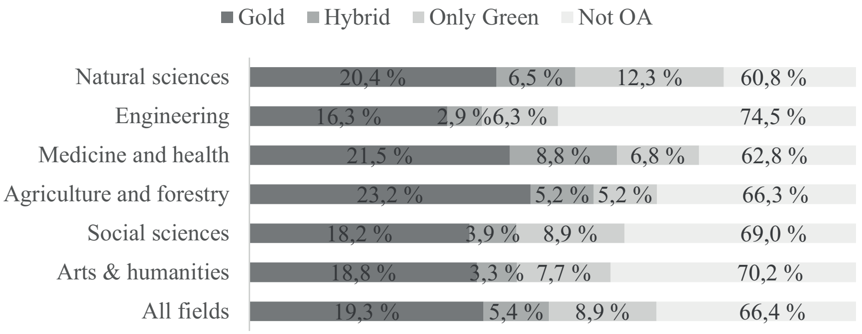


Figure 4. Type of open availability of outputs by field of science

Discussion and conclusions

The international data sources (Web of Science and Scopus), which are most often used for monitoring open access publishing, privilege journal outputs and STEM fields. The national publication data stored in the VIRTA publication information service from the 14 Finnish universities, including 48 177 peer-reviewed outputs from 2016-2017, provides a more complete picture of open access by also including book publications as well as all SSH journal publications. Scopus journals cover only 62 %, and WoS journals 52 %, of all these outputs. Taking all fields and publications types into account, the share of openly available outputs is 34 %. For 25 % of the outputs open availability is provided in a Gold or Hybrid channel, while 9 % are openly available only in repositories. The differences between fields in the share of openly available outputs range from 39 % in the natural sciences to 26 % in engineering.

The Finnish researchers used 10 342 different publication channels as outlets, including 9500 journals/series and 842 book publishers. In 25 % of the channels all Finnish outputs are openly available. In 25 % of the channels, however, the open availability is only partial, and in case of 50 % of the channels no openly available outputs have been reported in VIRTA. It is important to remember that we rely here on universities' self-reported OA status of publications. These results mean that, for Finland to achieve the target of open availability of all peer-reviewed outputs in the near future, around 5000 currently used channels should either be replaced with alternative open access channels or should flip to the required gold or green open access publishing models. Around 2500 channels already provide for open availability of some outputs, hence closing the gap between potential and uptake is the key.

The majority of journals/series used by the Finnish researchers (79 %) have a self-archiving policy registered in Sherpa/Romeo. Analysis of the share of Finnish outputs published in these journals shows that a relatively small share is openly available, irrespective of the self-archiving policy indicated with colour code, unless the outlet also provides open availability via Gold OA (DOAJ-indexed or Bielefeld listed journals). The share of openly available outputs is only slightly larger in the case of Hybrid OA channels than in channels permitting only self-archiving. Our results confirm that there indeed is considerable potential for advancing open availability via Green route (Laakso 2014; Björk et al. 2014). It remains to be seen if open access incentives, such as the extra-weight for openly available publications in the Finnish universities' core funding-model, help to increase the uptake.

As expected, publishing in DOAJ-indexed journals is a good predictor of open availability of outputs. However, only 13 % of the journals/series used by the Finnish researchers are indexed in DOAJ. These account for 15 % of all peer-reviewed journal outputs, and 38 % of all openly available journal outputs (including book publications, the shares are 12 % and 35 % respectively). A total of 944 journals/series identified in VIRTA as OA channels are not covered in DOAJ or Bielefeld list. It has not been possible for us to investigate if these journals/series might meet the DOAJ criteria. Nevertheless, our findings point at considerable gap in the information sources on OA channels. Combining all OA information, it was possible to identify 2553 potential Gold OA journals, of which DOAJ covers 48 % and the Bielefeld list additional 15 %. Our findings suggest that relying on external information sources, such as DOAJ, in the identification of open access publications may not result in complete picture of Gold OA publishing.

We also investigated the importance of large international publishers. Elsevier, Springer Nature, Wiley-Blackwell, Taylor & Francis, Sage and American Chemical Society account for 53 % of the Finnish peer-reviewed journal output, and 44 % of all outputs including conferences and book publications. In all, their dominance appears less pronounced than in analyses using Web of Science data, especially in case of humanities as well as social sciences (Larivière et al., 2015). This means that negotiations with the largest international publishers can provide only partial solution to advancement of open access, which – especially in the SSH – depends on open access publishing models adopted by large variety of relatively small journal and book publishers operating in national context (Ilva, 2018; Late et al., 2018).

The VIRTA publication data provides valuable information on the open availability of peer-reviewed book publications compared to journal articles (conference articles as a group is a mixture of both these publication types). The share of articles in books, monographs and edited works that are openly available is smaller (17 %) than that of journal articles (39 %). Nevertheless, 186 different book publishers (22 % of all publishers used by the Finnish researchers) are identified in VIRTA as Gold OA channels providing for open availability via publisher website at least to some of the outputs. Hybrid and self-archiving options appear, however, more restricted in case of book publishers. Our findings highlight the need for international register of academic/scholarly book publishers that would contain information – like DOAJ – on their peer-review practices, as well as open access status and self-archiving policies.

In all, we conclude that national publication data can provide valuable information on the open availability of peer-reviewed outputs. To enhance comprehensive and comparable monitoring of open access we recommend development of well-structured and comprehensive national and international publication information sources.

References

Bjôrk, B. C., Laakso, M., Welling, P., & Paetau, P. (2014). Anatomy of green open access. *Journal of the American Society for Information Science and Technology*. doi: 10.1002/asi.22963

Council of the European Union (2016). *Outcome of the council meeting, 3470th Council meeting, Competitiveness (Internal Market, Industry, Research and Space) Brussels, 26 and 27 May 2016*. Retrieved January 17, 2018 from: https://www.consilium.europa.eu/media/22779/st09357en16.pdf.

Engels, T., Starčič, A., Kulczycki, E., Pölönen, J. & Sivertsen, G. (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? *Aslib Journal of Information Management*, 70:6 (2018).

European Commission (2018). *OSPP-REC: Open Science Policy Platform Recommendations*. https://ec.europa.eu/research/openscience/pdf/integrated_advice_opspp_recommendations.pdf.

Ilva, J. (2017a). Towards reliable data – counting the Finnish Open Access publications. *Procedia Computer Science* 106, 299–304.

Ilva, J. (2017b). Suomalaisten yliopistojen avoimet julkaisut vuonna 2016 OKM:n julkaisutiedonkeruun tietojen valossa, *Informatiotutkimus* 3–4 (36): 51-69.

Ilva, J. (2018). Looking for commitment: Finnish open access journals, infrastructure and funding. *Insights*, 31, 25.

Kronman, U. (2017). Open Access i SwePub 2010 – 2016. Retrieved January 17, 2018 from: https://web.archive.org/web/20171216091206/http://openaccess.blogg.kb.se/files/2017/12/Open_Access_i_SwePub_2010-2016_v1.pdf

Kulczycki, E., Engels, T., Pölönen, J., Bruun, K., Duskova, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Starčič, A., & Zuccala, A. (2018). Publication patterns in the social sciencesand humanities: evidence from eight European countries, *Scientometrics*, 26.3.2018.

Laakso, M. (2014). Green open access policies of scholarly journal publishers: a study of what, when, and where self-archiving is allowed, *Scientometrics*, 99, 475–494.

Larivière V, Haustein S, Mongeon P (2015). The Oligopoly of Academic Publishers in the Digital Era. *PLoS One*, 10(6): e0127502.

Larivière, Vincent, and Benoit Macaluso. Improving the coverage of social science and humanities researchers' output: The case of the Érudit journal platform. *Journal of the American Society for Information Science and Technology*, 62.12 (2011): 2437-2442.

Late, E., Korkeamäki, L., Pölönen, J. & Syrjämäki, S. (2018). The role of learned societies in scholarly publishing in Finland. In M. Schirone et al. (eds), *Book of Abstracts, Nordic Workshop on Bibliometrics and Research Policy, Borås 7-9 November 2018*.

Martín-Martín, A., Costas, R., van Leeuwen, T. & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis, *Journal of Informetrics*, 12, 819–841.

Mikki, S. (2017). Scholarly Publications beyond Pay-walls. Increased Citation Advantage for Open Publishing. *Scientometrics* 113: 1529.

Mikki, S.; Gjesdal, Ø.L.; Strømme, T.E. (2018). Grades of Openness: Open and Closed Articles in Norway. *Publications*, 6, 46.

Montgomery, L. and Neylon, C. and Ozaygen, A. and Saunders, N. and Pinter, F. (2018). *The Visibility of Open Access Monographs in a European Context: Full Report, The Visibility of Open Access Monographs in a European Context: Full Report*. Curtin University of Technology, Humanities Research and Graduate Studies.

Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. (2017). The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ Preprints* 5:e3119v1

Pölönen, J. (2018). Applications of, and Experiences with, the Norwegian Model in Finland, *Journal of Data and Information Science*, 3(4), 31–44.

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. C. E. (2017). *European Databases and Repositories for Social Sciences and Humanities Research Output*. Antwerp: ECOOM & ENRESSH. Retrieved January 17, 2018 from: https://doi.org/10.6084/m9.figshare.5172322.v2.

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T., Arefievd, P., Duškováe, M., Faurbækf, L., Hollg, A., Kulczycki, E., Nelhans, G., Petr, M., Piskk, M., Soós, S., Stojanovskil, J., Stone, A., Šušolo, J., & Teitelbaump, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: findings from a European survey, *Research Evaluation*, 27(4), 310-322. doi: 10.1093/reseval/rvy016.

Sivertsen, G. (2016). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, 107(2), 357–368.

Sivertsen, G. (2018). Balanced multilingualism in science, *BiD: Textos Universitaris de Biblioteconomia i Documentació*, (40). Doi: http://dx.doi.org/10.1344/BiD2018.40.25.

Stojanovski, J., Jelka, P. & Bojan, M. (2009). The Croatian national open access journal platform. *Learned Publishing* 22.4: 263-273.

van Leeuwen, T. & Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their relationship with research assessment. In *Science, Technology and Innovation Indicators*, Utrecht.

Wohlgemuth, M., Rimmert, C. & Winterhager, M. (2016). *ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA)*. Bielefeld University. doi:10.4119/unibi/2906347.

Knowledge Utilization and Open Science Policies: Noble aims that ensure quality research *or* Ordering discoveries like a pizza?

Julia Heuritsch[1]

[1] *heuritsch@dzhw.eu*
German Centre of Higher Education Research and Science Studies (DZHW), Schützenstraße 6A, 10117 Berlin;
Humboldt Universität zu Berlin, Research Group "Reflexive Metrics", Institut für Sozialwissenschaften,
Unter den Linden 6, 10099 Berlin

Abstract

Open Science has been a rising theme in the landscape of science policy in recent years. The goal is to make research that emerges from publicly funded science to become findable, accessible, interoperable and reusable (FAIR) for use by other researchers. Knowledge utilization policies aim to efficiently make scientific knowledge beneficial for society at large. This paper demonstrates how Astronomy aspires to be open and transparent given their criteria for high research quality, which aim at pushing knowledge forward and clear communication of findings. However, the use of quantitative metrics in research evaluation puts pressure on the researcher, such that taking the extra time for transparent publishing of data and results is difficult, given that astronomers aren't rewarded for the quality of research papers, but rather their quantity. This paper explores the current mode of openness in Astronomy and how incentives due to funding, publication practices and indicators affect this field. The paper concludes with some recommendations on how policies such as making science more "open" have the potential to contribute to scientific quality in Astronomy.

Introduction

Making science more open has been a rising theme for policy stakeholders at international (e.g. European Commission[i] and United Nations[ii]) and national levels (e.g. Netherlands Organisation for Scientific Research; NWO[iii]). The goal is to make research that emerges from publicly funded science to become *findable, accessible, interoperable and reusable (FAIR)* for use by other researchers. Stakeholders like the NWO and European Commission acknowledge that the way science is conducted is fundamentally changing due to the sophisticated digital technologies available. Often these 'open' policies are part of a more general 'knowledge utilization' policy, which aims to efficiently make scientific knowledge beneficial for society at large under the imperative of the *Three Os – Open Innovation, Open Science* and *Open to the World.* The aim of the paper is to give recommendations of what 'open' policies would need to consider in order to support FAIR publication in Astronomy, while at the same time encouraging quality in the knowledge production process. This investigation includes in what way astronomers value openness and how current policies encourage openness and research quality.

New policies may come with new incentives and new ways to measure whether certain goals have been achieved. These may not only come with the aspired effects, but also with "unintended consequences" (drawing back to the notion of "unanticipated consequences of purposive social action"; Merton, 1936) such as quality discrimination. The notion of unintended consequences has been a long-standing debate in the sociology of (e-)valuation as well as in the contexts of notions of "reflexivity", "performativity", self-fulfilling prophecies or “retroaction”. All of these approaches share a common denominator, namely that there seems to be a disjunctive moment between intention, action not only on the micro-level of the individual actor, but also on the level of agglomerations spanning from communities of practice up to society at large. It might therefore be valuable to approach the issue of discourse on a certain topic, in this case “open science” with the notion of what in the most simple formula be termed “feedback” and at the same time the relation between justification and critique that structures part of that discourse.

One aspect of justification and critique becomes apparent in questions how to measure and ensure high-quality of knowledge production in the context of “*Accountability*” and “*transparency*”. Both concepts are closely associated with producing and monitoring metrics (Espeland & Vannebo, 2008). This is because quantification is one means to constitute social entities as things that last and are comparable. As such, the goal of quantification is to enable objectification and to master uncertainty. Through objectification, both a political space and a measuring space, are co-constituted in which things can be compared (Desrosieres, 1998). It permits scrutiny of complex or disparate phenomena in ways that enable judgment (Espeland & Stevens, 2008). Hence, quantification offers a shared language and replaces trust in people with “trust in numbers” (Porter, 1995). Quantification therefore is seen as one mode of de-localization of valuation practices. Quantitative metrics, such as indicators to measure scientific productivity, *commensurate*, which is the act of using numbers to rate and rank, creating a specific type of relationship among objects. Commensuration is one of the most consequential uses of numbers (Espeland & Stevens, 2008), because it turns describing numbers into prescriptive ones. Commensuration attributes meaning to numbers.

Effort or how individual researchers *perform* cannot be monitored efficiently so assessment cannot be based upon it and therefore, a scientist is rewarded and funded for quantitative achievement instead (Rosenberg & Nelson, 1994). The same holds for *usefulness* of science. Not only is it difficult to measure usefulness, but there is also a tension between the ever increasing demand for *societal relevance* (e.g. Bouter, 2008) and the risky nature of basic research. The problem of commensurability of usefulness also becomes apparent, when we observe typical means of addressing usefulness or impact in research evaluation. One of the most notable examples being the narrative form of signifying impact in the Research Excellence Framework (REF). Yet, as elaborated before, there are social forces that adhere to the idea of measurability. Basic research is the human’s endeavour to understand the unknown and as such it is by definition risky (Stephan, 2012). When societal relevance is measured in applicable outputs, and decisions must be made about distribution of researchers amongst research fields, economic pressure to produce such outputs can arise. This may lead to a tension between demonstrating its usefulness to society, on the one hand, and not being able to guarantee that due to the risky nature of research, on the other hand. For research fields that perform mainly basic research, such as Physics and Astronomy, this pressure to justify their societal relevance might be especially high, since results of basic research may have a delay in leading to ground breaking technological developments or theories. More applied sciences, “where the products of research are highly profitable, such as medicine, biotechnology, genetics and military research” (Bourdieu, 2004), face a different economic pressure, the expectation to produce more and better.

Measures that initially may have been designed to describe behaviour can easily be used to judge and control it, due their commensurative character. Measurement intervenes in the social worlds it depicts, as measures are *reactive & performative*; they cause people to think and act differently. Hence, numbers can also exert discipline on those they depict and disciplinary practices define what is appropriate, normal, and to what we should aspire (Espeland & Stevens, 2008). Foucault (1977 & 2003) links statistical practices to "*governmentality*", a term to describe how the government uses numbers to influence citizens so that they fulfil those government's policies. He describes discipline as a mode of modern power that is continuous, diffuse and embedded in everyday routines.

The performative character of indicators leads to the conclusion that policies have constitutive effects on how science is done (Dahler-Larsen, 2014). In other words, indicators and rewards introduced by policies can shape the process how knowledge is produced in science. The fact that indicators commensurate, where all difference is transformed into quantity (Espeland & Stevens, 2008), leads to the argument that their use to assess scientific quality gives rise to an "evaluation gap". This is a term coined by Wouters (2017) to acknowledge a discrepancy between the notions of scientific quality as perceived by researchers of a field and as measured by indicators. In order to meet the targets set by indicators, scientific quality may be sacrificed. This can have "unintended consequences" such as goal displacement, gaming, information overload, questionable authorship practices, unhealthy competition and aversion to risky/innovative projects (Rushforth & De Rijcke, 2015; Laudel & Gläser, 2014).

As outlined above, basic research is under the pressure of demonstrating its societal relevance. While astronomy asks highly fundamental questions, which inspire both scientists and the public at large, Astronomy faces a crisis to demonstrate its usefulness. That is shown by the fact that one can find a large number of flyers and commentaries on the internet and elsewhere explaining why astronomy is important (e.g. Rosenberg et al., 2013). Applications from Astronomy are rather invisible for the public, due to two reasons. First, as outlined above, basic research usually doesn't find immediate societal usefulness. An example is applying the theory of general relativity to enable precision in the Global Positioning System (GPS). Second, uncovering the laws of nature, such as the theory of general relativity, is Astronomy's primary goal, resulting applications are secondary. Nevertheless, technology invented for space exploration often leads to surprising applications on Earth, so-called spin-offs, such as innovations in dental care and breast cancer detection. Hence, there is a tension between needing an output due to economic pressures and doing basic research for its own sake that entails dead ends and no immediate usefulness. Additionally, large and expensive (international) observing instruments involve large collaborations and the use of (open) archives and huge datasets. For these reasons, Astronomy is such an interesting case to study the significance of indicators in its usefulness crisis and their relevance in this tension relationship. To which extent indicators shape the conditions under which astronomers produce knowledge, is what needs to be answered when developing new policies.

Methods

This research[iv] consists of semi-structured interviews[v] and a document analysis. The interview sample was targeted such that the interviewees represent a variety of nations, career-status and research areas. 4 faculty members, 2 postdocs, 1 PhD and 2 Master students from Leiden Observatory (Sterrewacht) were chosen who work in cosmology (2), exoplanets (1) and observational (radio) astronomy in different subfields (6). The Master programme at the Sterrewacht is very research intensive, requiring the student to write two Master theses in total, which is the reason why they are also interesting subjects for this study. In order to shed light on what effects policies have on the field, questions were developed such that an astronomer's definition of quality versus what is measured by indicators can be studied. Next to the meaning of openness and data sharing, topics include career steps, project funding, exposure to assessments, research evaluation, the publication and funding system, different stages of the knowledge production process – from planning, via doing the research to publishing – and the meaning of quality. Each topic was introduced by one overarching question, followed by several potential follow-up questions.

The participating researchers were invited via email and all names are anonymized. All interviews, 80-100 minutes in length, were fully transcribed into electronic form, summarised and coded) according to Grounded Theory. These codes represent themes which emerged by combining sensitivity towards existing literature on constitutive effects of indicator use with insights from the data.

The interview data were complemented with a document analysis of materials collected online or made available via the informants, including CVs of the interviewed researchers, annual reports (1998 to 2015), (self-) evaluation reports of the Dutch Astronomy institutes and their umbrella organisation NOVA. The documents informed the following analysis, however the document analysis itself is not part of this study and can be found in the original report of the project[iv].

In the *Results* section direct quotes of the interviewees will be given between double quotation marks.

Results

The astronomers' stance to openness and effects of current policies

In order to understand what policies related to 'opennes' would mean for Astronomy, we must first ask what data and research results mean to an astronomer. My study found that astronomers generally conduct science for the sake of curiosity and "pushing knowledge forward" (Faculty Member 4). For them, publications are not their priority, but rather a means to publicise their results in order to advance the field in three steps:

> "You have a new science idea. You have asked the question clearly and well, with a well-defined […]. And you have written a paper which demonstrates you have answered that question […]. And you have written it in such way that a non-expert in that field can read it and understand what you have done. […] If it's a crap written paper, then that's crap research – I don't care how brilliant the answer is, if they can't communicate it through a paper or through a presentation, then that's bad research."
> (Faculty Member 4)
>
> "And ahh … if that was not so important [to get papers out] I would probably not bother so much … I mean I would still publish my papers because I – it gives a different motivation to it, right? As a scientist you just want to publish your papers, because you are a scientist and you think this is important for science: 'This is the result, this is what defines the process of science'." (Faculty Member 1)

Rather, it is in the astronomers' interest to share data and results to get knowledge "out to the community" (Faculty Member 1). "To know and understand better" (Postdoc 1) and communicate this knowledge to the community is what makes up an astronomer's intrinsic motivation to conduct research. For an astronomer, the definition of high-quality research is based on this motivation. I found three quality criteria:

- Asking an important question for the sake of understanding the universe better and to push knowledge forward.
- Using clear, verifiable and sound methodology.
- Clear communication of the results so the community can make use of them.

From these criteria it is apparent that astronomers' motivation in doing science is not for some direct societal impact, since they are dedicated to fundamental research which might only find applications decades later (Stephan, 2012).

> "Well, academic quality has always been relatively clear. It has to be verifiable and clear, unbiased etc. But there is these days … a tendency to look at the value of science in terms of economic output, it's called '*valorization'*. And I am totally uninterested in that […]. It is always nice if you find applications that are useful […]. Why not? But that's not why we do it." (Faculty Member 1)

On the other hand, openness in terms of disseminating the knowledge gained does fit within astronomers' values and definition of quality. To understand how 'open' policies could have positive effects on openness in Astronomy, and whether they could at the same time encourage scientific quality, we must first look at the effects of current policies.
This study found factors for extrinsic motivation that drive research in Astronomy as well as the intrinsic factors. Extrinsic motivation arises from what the evaluation system values through its indicators. Astronomers report that *first author publications*, *citation rates* and *number of acquired grants* are what determine their value as a researcher. Since the future career of an astronomer is dependent on these factors, there is a shift from the initial motivation to publish for the sake of disseminating knowledge, to the "need to publish" (e.g. Postdoc 1 & Faculty Member 1). This results in publication pressure and lower quality papers.

> "Your job prospects will depend on this like quantity rate, with which you are publishing." (PhD Candidate)
>
> "It's a system problem I think. Erm, I try to do quality research, but I do feel sometimes that I end up publishing because I have to publish.", "I wish we could just focus on more like quality papers instead of quantity papers." (Postdoc 1)

Observational astronomers are found to be particularly affected by this pressure as compared to theoretical astronomers. On the one hand, they produce data with telescopes which are essential for the knowledge production process in Astronomy. On the other hand, this data is a form of output that is reportedly not valued in evaluations. To produce data, observational astronomers need to compete for limited *observation time* at telescopes and then 'be lucky' to have the right weather conditions. Once granted, observation time does count like received funding in an astronomer's CV. However, non-detections are more common than detections and 90% of non-detections are not publishable. Hence, observational astronomers face the risk to fall through the cracks of metrics in every step of their knowledge production process.

> "So it's essentially [that] negative results are considered as failed research by the community. […] And on that side I disagree. […] So there is always information to be taken from research that is well conducted. Given that the research is using state of the art data, and state of the art methodology, whatever the result is, should be interesting." (Postdoc 2)
>
> "Non-detections. [...] It's just really hard to work with the telescope and I really want to be able to figure it out and do this thing and I think personally I would feel failed if I wouldn't be able to at least … put some limit, that gives a good sort of low sensitivity to it [i.e. finding some implications]." (PhD Candidate)

Current state of openness in Astronomy

The effects of current policies affect not only research quality, but also how *FAIR* astronomers' research data and results are. To demonstrate the current state of openness in Astronomy, output can be divided into three categories: *raw data*, *results* and *negative results*. Policies which advocate for openness and good research quality in Astronomy would have to take these as a starting point.

First, *raw*, unprocessed data is the data to be analysed by the researcher. The most prominent example is telescope data. Most telescopes make the raw telescope data public after a proprietary period to the original observer of one year after obtaining through the archives of the observatories. Additionally, researchers can publish their raw or reduced data through the archive of Centre Donnees Stellaire in Strasbourg or various others, once their paper is accepted. On the one hand, this serves the purpose of openness and gives other researchers the possibility to replicate or conduct their own studies. On the other hand, the relative short proprietary period of one year adds to the publication pressure, as only the first to publish receives the credit.

Second, *results* are written up in publications in the form of scientific articles. In a first step of the data analysis, raw data is processed to reduced data through so-called 'pipelines', which are data reduction codes that clean the data from noise and prepare them for the analysis. Sometimes these reduced data are also published in the publically available observatory archives. Results are the final processed data and the conclusions drawn from them. In order to have impact, astronomers publish in journals, but they usually also upload their papers to the open data repository ArXiv. However, *findable* results does not equal *accessible*, even for fellow researchers. The reasons for non-accessibility are for example that reduction codes are mostly not published or that important steps in calculations are not mentioned in the paper, decreasing transparency and hence the paper's communication value. Interviewees urge for more reader-friendly formats, which would enable, for example, expanding sections, interlinking content, adding simulations/ visualisations and publishing code, as it would be possible with modern technology. Despite this, papers are still written in a style inherited from pre-computer times. As information about methods and analysis gets lost with result oriented papers, written in an out-dated manner, it is more difficult to ensure good communication and replicability of research results. In some cases, mistakes even remain undiscovered.

> "The way that papers are currently being written is perhaps too much tied to the way that papers were published in the past. So they were actual papers in a journal, so they had to be sequential. But this is no more the case, now that we have other ways to … read or get information. We can have, not necessarily interactive things, but, at least content that can be separated into different sources. So you can read on one side about the science of the paper, and on the other side about the technical aspects. And currently the two things are merged into a single file, or work. And even if it's true that you intent to have sections like methodologies and results, so if you are not interested in the methodology, or if you actually want to read about the methodology you can go there or not go there. But people will tend to get take [content] away from the methodology section, because they will consider 'Ah, that's too much […], so let's not mention this or put that into an appendix'. So I think there should be the possibility for authors to be very thorough in explaining the methods and even, that includes the possibility to show code. […] In fact in the Astrophysics community, the skills in programming are fairly low in general. Which is worrysome, because I think there are a lot of bugs running around that are not noticed. And because we can't look at the code we can't say, or see whether this is happening or not." (Postdoc 2)

As mentioned above, raw or reduced data is often published along in various archives. However this process is not standardized and often voluntary. While astronomers cite their (data) sources, interviewees criticise that references are not transitive. For example, a literature review might cite the papers it is referring too, but not the data sources that these papers are based on. This seems hardly *fair* for the producers of the data and doesn't add to transparency of the research process. Since there are no incentives for replicability or for transparency in one's research methods, and since publication pressure is high, astronomers do not have the time to consider whether their results are fully *reusable*:

> "And I think that's very bad, when for example, almost the entire results come from a code which is not publically available. So you cannot look at this code and see … if they are actually doing what they say in the paper. And also if they – sometimes they make a mistake. […] So in the sense, the replicability of the work we do … is not always very high. And in the sense that you can download for yourself, in principle all the raw datasets from a telescope and you can redo everything by yourself. So in this sense, yes it's replicable, but never fully replicable." (Postdoc 2)

Third, conducting fundamental research can naturally lead to a dead end, or to *negative results*. The most common examples in Astronomy are non-detections, which are observations that didn't lead to the predicted detection. They are usually only publishable when the researcher can determine their implication or is able to provide upper or lower limits.

> "[Non-detections are not publishable], unless you have a very good [implication], as in for example the way we sort of explained the upper limits with the non-detection. […] The problem is how to tailor it, right? […] So, yeah, unless you have … like a good way, I mean there is some research that published non-detection – for exoplanets sometimes they publish it when they didn't detect it, because sometimes you sort of predict that it should be there […] And it's an anomaly or something like that […] So there are some ways to publish this, but I think it's very … like 10%. There is a whole 90% that doesn't get published and sometimes, like for example, if you just had bad weather, then it's very difficult, right?" (Postdoc 1)

In some sub-fields of Astronomy, non-detections are more common than detections. Most interviewees are convinced that negative results should "absolutely be publishable" (Faculty Member 4). This is because they are seen as valuable with respect to new knowledge about what does not work. As research is the discovery of the unknown, this kind of information is also essential, "because it either can help [the researchers] discount certain theories, or help them kind of support other theories" (Master Student 2). Hence, astronomers are advocating for the exposure of the negative results as well, so that other researchers don't have to 'reinvent the wheel'.

> "I mean when I was at [the famous institution] we said, we should start a journal on non-detections. Because I am really sure that there are people that have been observing the same objects on and on and in without knowing that other people have already done this. Because nobody published when they don't detect anything." (Postdoc 1)

While efforts like this would be welcomed by astronomers, current metrics do not account for non-detections and hence there are little incentives to invest time in the contribution to such 'non-detection journals' if there are no benefits for one's further career in science.

Conclusion & Recommendations

The results of this study show that Astronomy aspires to be open and transparent. However, this requires support from policy makers as current policies do not provide the incentives to invest valuable time in the publication of data and code. Therefore, the observations above imply five recommendations to policy makers when it comes to knowledge utilization and openness:

1. Goodhart's law which states 'when a measure becomes a target, it ceases to be a good measure', always needs to be kept in mind when establishing new policies. In practice this means, that indicators and rewards have constitutive effects on knowledge production, which need to be accounted for.

2. For the afore-mentioned reasons, *FAIR*ness cannot be implemented if there is no incentive for a change in the way research is published. This may require a change in journal templates and paper-writing style to take advantage of the possibilities offered by modern technology. As one of the astronomer's quality criteria are to communicate results well and transparently, papers would be of higher quality if they included more information on in-between-steps; this could be done through expandable sections that a reader could easily skip if wanted. Incentives for including code would lead to a decrease the propagation of errors and increase *replicability*, *accessibility* & *reusability*. If publications provided an interactive way of delivering visible feedback and updating outdated information, *reusability* would benefit and an active exchange within the scientific community would be fostered.

3. Journals could support the astronomers' aspiration for knowledge utilization in terms of pushing knowledge forward by building on previous research. Providing for transitive referencing would make the sources that pieces of research is based on more transparent. 'Open' policies may encourage the journals to do so and the astronomers to use these opportunities.

4. Advocating for openness and knowledge utilization also means valuing knowledge about what doesn't work. Therefore, policies need to reward the publication of research that led to negative results to ensure that researchers who engaged in those studies receive credit for their work, thereby reducing publication pressure. Especially for observational astronomers in sub-fields where the majority of the detections currently are not publishable, 'open' policies provide a hope for improvement.

5. Unlike Astronomy, applied sciences may naturally be more orientated towards finding economic applications for their research. However, this quote applies for all sciences, and is to be kept in mind when it comes to policies for knowledge utilization:

   > "[These politicians] think that they can direct science. […] – *They think they can order discoveries like you order a pizza*. You. Cannot. Order. A. Discovery. […] You have to work on it, you have to try things, you have to experiment. […] But since science is funded mostly by public funding, we are dependent on the strange conceptions that politicians have on how science works." (Faculty Member 3)

References

Bourdieu, P. (2004), “Science of Science and Reflexivity”, *The University of Chicago Press*, ISBN: 9780226067377 & ISBN: 9780226067384

Lex M. Bouter (2008), Knowledge as Public Property: The Societal Relevance of Scientific Research, OECD, http://www.oecd.org/site/eduimhe08/41203349.pdf

Dahler-Larsen, P. (2014), “Constitutive Effects of Performance Indicators: Getting beyond unintended consequences”, *Public Management Review*, 16:7, p.969-986

Desrosières, A. (1998), “The Politics of Large Numbers – A History of Statistical Reasoning”, *Harvard University Press*, ISBN 9780674009691

Espeland, W.N. & Stevens, M.L. (2008), “A Sociology of Quantification”**,** *European Journal of Sociology*, Volume 49, Issue 03, p. 401 – 436, DOI: 10.1017/S0003975609000150

Espeland, W.N. & Vannebo B. (2008), “Accountability, Quantification, and Law”, *Annual Review of Law and Social Science 3*, p. 21-43

Foucault, M. (1977), “Discipline and Punish: The Birth of the Prison”, London, Allen Lane

Foucault, M. (2003), “The Subject and Power”, in Rabinow Paul and Nicholas Rose, eds., “The Essential Foucault” (New York, *The New Press*, p. 129-144)

Laudel, G. & Gläser, J. (2014), “Beyond breakthrough research: Epistemic properties of research and their consequences for research funding”, *Research Policy 43*, p.1204-1216

Merton, R. (1936), “The unanticipated consequences of purposive social action”, *American Sociological Review,* Vol. 1, No. 6 (Dec., 1936), p. 894-904

Porter, T. (1995), “Trust in numbers”, *Princeton University Press*

Porter, T. M. (1994), “Making Things Quantitative”. *Science in Context, 7:3*, p.389–407, in Dahler-Larsen, P. (2014)

Rosenberg, M. et al. (2013), “Why is astronomy important?”, https://arxiv.org/abs/1311.0508

Rosenberg, N. & Nelson, R. (1994), “American Universities and technical advance in industry”, *Research Policy 32*, p.323-348

Rushforth, A.D. & De Rijcke, S. (2015). “Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in the Netherlands”, *Minerva 53*, p.117-139

Stephan, P. (2012), “How economics shapes science”, *Harvard University Press*

Wouters, P. (2017), “Bridging the Evaluation Gap”, *Engaging Science, Technology, and Society 3*: p.108-118

[i] https://ec.europa.eu/research/openscience/index.cfm
[ii] http://www.unoosa.org/oosa/en/ourwork/psa/schedule/2017/workshop_italy_openuniverse.html

[iii] https://www.nwo.nl/en/policies/knowledge+utilisation

[iv] The complete report of the results of this study can be found on ArXiv (https://arxiv.org/abs/1801.08033), however this paper concentrates on what the findings mean for the latest policy developments regarding the *FAIR* imperative.

[v] Acknowledgements: I would like to thank the interviewees for their willingness to participate in this study and for dedicating research time that is even more valuable in times of publication pressure.

Text Mining to Measure Novelty and Diffusion of Technological Innovation

Sam Arts[1] Jianan Hou[2] Juan Carlos Gomez[3]

[1] *sam.arts@kuleuven.be*
Department of Management, Strategy and Innovation, Faculty of Economics and Business, KU Leuven,
Korte Nieuwstraat 33, 2000 Antwerp, Belgium

[2] *jianan.hou@kuleuven.be*
Department of Management, Strategy and Innovation, Faculty of Economics and Business, KU Leuven,
Korte Nieuwstraat 33, 2000 Antwerp, Belgium

[3] *jc.gomez@ugto.mx*
Department of Electronics Engineering, University of Guanajuato Campus Irapuato-Salamanca,
Carretera Salamanca - Valle de Santiago, Salamanca, Mexico

Abstract

Existing measures of patent novelty and diffusion mainly rely on patent classification or citations. Given that inventive ideas are embedded in the texts of patents, our study provides a new way to assess novelty and diffusion by text mining techniques. As a validation test, we collect a set of patents linked to famous awards such as the Nobel prize. Overall, text-based measures outperform other commonly-used novelty and diffusion metrics.

Introduction

The increasing number of granted patents echoes the prosperity of innovation activities. Nevertheless, the distribution of patent quality is highly skewed as most innovations are categorized as "incremental" (Nelson and Winter, 1982; Henderson and Clark, 1990). To assess the novelty and diffusion of patents, prior studies mainly rely on patent classification or citation information (e.g. Trajtenberg 1990; Fleming, 2001; Dahlin and Behrens, 2005). The validity of traditional measures have been questioned by recent studies (McNamee, 2013; Arts et al., 2018; Kuhn and Thompson, 2019), and one of the most obvious limitations is that neither patent classification nor citation can mirror the technological content of the patent directly.

In this paper, we focus on the technological contents of patents and develop new patent novelty and diffusion measures by text mining. To do so, we collect US patents granted up to 2018 and identify the first occurrence of new word or new word combination to pinpoint the origin of new technology. The reuse frequency of new words and new word combinations are counted as indicator of technology diffusion. To examine the validity of the new measures, we collect a sample of patents awarded by prestigious prizes, such as Nobel Prize and A.M. Turing Award.

Identifying the Origin and Diffusion of New Technologies

We collect titles, abstracts, and claims of US utility patents granted between 1969 and 2018 from the USPTO, the patent claims research dataset (Marco et al., 2016), and PATSTAT. For each patent, we concatenate the title, abstract, and all claims, lowercase the text, tokenize all words, and remove punctuation, words composed of numbers only, one-digit words, words which appear in only one patent, and stop words. Then, we stem the remaining keywords and remove duplicate stemmed keywords from the same patent. Finally, the technical content of each patent is summarized by a collection of unique keywords.

Based on the processed unique words list, we trace the origin of new technologies by identifying the first patent introducing a given new word or new word combination. All patents are sorted by filing date, and keywords from patents filed before 1980 are used to compile the baseline dictionary (Balsmeier et al., 2018). To assess the diffusion of new technology, we count the number of subsequent US patents reusing the given new word or new word combination. Finally, for each patent, we calculate the total number of new words and new word combinations as indicator of novelty and aggregate reuse frequency of all new words and new word combinations as indicator of diffusion.

We calculate several commonly used novelty and diffusion measures and compare their predictive performance with text-based measures. First, we calculate *new subclass comb* as the number of previously uncombined pairs of patent subclasses and weight it by the total number of subsequent patents reusing the focal new subclass combinations to generate the variable *new subclass comb reuse* (Fleming et al., 2007, Arts & Veugelers, 2014). Similarly, we count the number of previously uncombined pairs of cited patents as *new cit comb* and generate *new cit comb reuse* by aggregating the number of future patents reusing the focal new cited patent pairs (Arts and Fleming, 2018). By examining the diversity of cited and citing patents, we calculate *originality* as one minus the Herfindahl index on classes of citied patents, and *generality* as one minus the Herfindahl index on classes of citing patents (Trajtenberg et al., 1997). Finally, we count *forward cit* as the number of citations received by the focal patent within 10 years (Trajtenberg, 1990).

Validation

To assess the predictive ability of text-based measures, we collect a set of patents with arguably high novelty and diffusion from seven prestigious prizes (Carpenter et al., 1981; Arts et al, 2013), namely Nobel Prize, Lasker Award, A. M. Turing Award, National Inventor Hall of Fame, National Medal of Technology and Innovation, Benjamin Franklin Medal, and Bower Award. Given that most awards (except National Inventors Hall of Fame) do not provide the patent number of awarded inventions, we manually match each awarded invention to US patents by the name of laureate, technical description of the awarded invention, year and laureate's affiliation. For each awarded patent, we select one control patent based on text similarity and approximate filing date (Arts et al., 2018).

First, we run t tests to compare the means of the different measures for the award and control patents. Award patents score significantly higher on all measures, except for *originality* and *new cit comb*. *New word comb reuse* shows the strongest discriminating power. Then we run logit regressions to predict the likelihood of being an award patent. All measures are highly significant except *new cit comb*, and *new word comb reuse* strongly dominates other measures in distinguishing awarded patents from control patents.

Conclusion

We develop new text mining techniques to identify the creation and diffusion of new technologies in the population of U.S patents. Whereas prior studies predominantly rely on patent classification or citations, we focus on technical content of patent to measure the technological novelty and impact of a patent. By a validation test, we show that text-based measures outperform traditional measures. We will provide open access to all code and data for all US utility patents granted before May 2018.

References

Arts, S., Appio, F. P., & Van Looy, B. (2013). Inventions shaping technological trajectories: do existing patent indicators provide a comprehensive picture?. Scientometrics, 97(2), 397-419.

Arts, S., & Veugelers, R. (2014). Technology familiarity, recombinant novelty, and breakthrough invention. Industrial and Corporate Change, 24(6), 1215-1246.

Arts, S., Cassiman, B. & Gomez, J. C., (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62-84.

Arts, S. & Fleming, L. (2018). Paradise of novelty-or loss of human capital? Exploring new fields and inventive output, *Organization Science*, 29(6), 989-1236.

Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., Li, G., Luck, S., O'Reagan, D., Yeh, B., Zang, G. & Fleming, L. (2018). Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *Journal of Economics & Management Strategy*, 27(3), 535-553.

Carpenter, M.P., Narin, F. & Woolf, P. (1982). Citation rates to technologically important patents, *World Patent Information*, 3(4), 160-163.

Dahlin, K. B., Behrens. D. M., (2005), When is an invention really radical? defining and measuring technological radicalness, *Research Policy*, 34, 717-737.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1), 117-132.

Fleming, L., Mingo, S. & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success, *Administrative Science Quarterly*, 52(3), 443-475.

Henderson, R. M., Clark, K. B., (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, 35(1), 9–30.

Kuhn, J. M., Thompson, N. C., (2019), How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business,* 26(1), 5-38.

Marco, A. C., Sarnoff, J. D. & deGrazia, C. A. (2016). Patent claims and patent scope, *USPTO Economic Working Paper* No. 2016-04.

McNamee, R. C., (2013), Can’t see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example, *Research Policy*, 42(4), 855-873.

Nelson, R., Winter, S.,(1982), An evolutionary theory of economic change. *MA: Harvard University Press.*

Thompson P & Fox-Kean M. (2005). Patent citations and the geography of knowledge spillovers: a reassessment. *American Economic Review*: 450-460.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics*, 21(1), 172-187.

Trajtenberg, M., Henderson, R. & Jaffe, A. (1997). University versus corporate patents: A window on the Basicness of invention, *Economics of Innovation and New Technology*, 5:1, 19-50.

Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B., (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468-472.

Can the impact of grey literature be assessed? An investigation of UK government publications cited by articles and books

Matthew S. Bickley[1], Kayvan Kousha[2] and Michael Thelwall[3]

[1] *M.Bickley@wlv.ac.uk*

[2] *K.Kousha@wlv.ac.uk*

[3] *M.Thelwall@wlv.ac.uk*

[1,2,3] Statistical Cybermetrics Research Group (SCRG), University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY (United Kingdom)

Abstract

Grey literature encompasses a range of relatively informal textual outputs that are not indexed in citation databases. Although they are usually ignored in research evaluations, it is important to develop methods to assess their impact so that their contributions can be recognised, and successful types of grey literature can be encouraged. This article investigates the extent to which 97,150 UK government publications were cited by Scopus articles and Google Books during 2013-2017 in eleven broad subject areas. A method was used to semi-automatically extract citations to the UK government publications from articles and books with high recall and precision. The results showed that Scopus citations are more common than Google Books citations to UK government publications, especially for older documents, and for those in Healthcare, Education and Science. Since the difference is not huge, both may provide useful grey literature impact data.

Introduction

'Grey Literature' or 'Gray Literature' is a term which describes textual documents that are not published in a standard academic format, such as a book or journal article. The term includes reports, regulations, and policy documents, which are important outputs from many governments and organisations. Fuzzy for many years and still not concrete due to the boundaries between grey literature and non-grey literature varying depending on the situation (IGLWG 1995), the Prague definition of 2010 seems to be now accepted: "*Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body*" (Schöpfel, 2010, p.11). The US Interagency Gray Literature Working Group has given the following alternative definition: "*Foreign or domestic open source material that usually is available through specialized channels and may not enter normal channels or systems of publication, distribution, bibliographic control, or acquisition by booksellers or subscription agents*" (IGLWG, 1995). Hence, grey literature publications can include, but are not limited to, unpublished research, governmental reports, policy statements conference proceedings, and theses or dissertations (GreyNet, 2019, UNE, 2019).

There are many high-profile grey literature repositories, confirming that this is an important document type. The UK government publication repository includes almost 120,000 annual reports, regulations, statistics, or policy documents in different topics (https://www.gov.uk/government/publications). This is a specialised source of grey literature in government policy making. The repository hosts many policy-making papers, such as healthcare reports, which are of high value to society and can be used to improve information on risk factors and how healthcare research is used (Institute of Medicine, 2009).

Other grey literature repositories include those of the World Health Organization (WHO, https://www.who.int/publications/en/), the United Nations (https://digitallibrary.un.org) and the World Bank (http://www.worldbank.org/en/research/brief/publications). Given that large amounts of grey literature have been created by governments and other important organisations, it would be useful to know if they have an impact so that their creators can decide which types of document are worth producing. This article focuses the academic impact as a first step towards this goal.

Citation analysis is commonly used to assess scientific impact of published research. However, there seems to be no practical or standard method to identify grey literature citations. Grey literature publications do not have well-established, centralised and standardised sources, and hence impact indicators are more difficult to calculate.

Google Scholar has been suggested as a good source for monitoring the impact of grey literature (Orduna-Malea, Martín-Martín & López-Cózar, 2017) and dissertations (Kousha & Thelwall, submitted). However, Google Scholar queries cannot be automated on a large scale, except for the facilities of Publish or Perish (Harzing, 2010) and it is therefore not suitable for large scale grey literature evaluations. Web queries have also been proposed for small sets of documents (Wilkinson, Sud, & Thelwall, 2014), but these do not necessarily reflect academic impact.

Given the lack of an accepted solution for determining the academic impact of grey literature, this article proposes and demonstrates two new approaches. First, Scopus (API) cited reference searches can be used to find citations to non-standard academic outputs (Kousha, Thelwall, & Rezaie, 2011) and complex queries can be designed to identify citations to large numbers of documents. Second, the Google Books API can also be used to automatically identify citations to monographs with high accuracy (Kousha & Thelwall, 2015). These strategies are proposed and are important to determine if feasible for grey literature. This paper describes the two new methods in detail and compares their results for 97,150 UK government publications from 2013-2017 across eleven broad subject areas.

Research questions

The underlying goal is to assess if Scopus and Google Books citation searches can be automated for capturing citations to grey literature publications. UK government publications are the focus of the study because the UK government publishes a large number free online, its repository can be crawled, and the authors are familiar with the UK context.

1. Can academic citations to grey literature publications be automatically extracted from Scopus and Google Books on a large scale?
2. Which citation search strategy or indicator is most useful for the impact assessment of UK government publications?
3. Are there disciplinary differences in the answer to the above question?

Methodology

This section describes how the new method was developed through small scale pilot studies.

Data sets

The online repository of documents released by the UK government (held at https://www.gov.uk/government/publications, hereafter: 'the repository') is classified by government-defined policy area and year of release (see Table 3 in the online Appendix (https://figshare.com/s/51a8308bdf43772820b3). This data was collected in July 2018 by a bespoke crawl routine added to the free Webometric Analyst (lexiurl.wlv.ac.uk) software. Each policy area was combined into more general topic areas (Table 1). The most recent five years were chosen to be most relevant for use in this method due to the increase in uploads to the repository at that time. Out of 137,559 documents available, 97,150 (70.6%) are from the years 2013-2017. Each document has a unique URL as well as a title. The URLs were used in subsequent searches to identify citations.

Table 1. All 11 grey literature areas used by combining policy areas as defined in the repository, split by years used, along with total over 2013-2017 (grey and policy areas sorted by largest size).

Grey literature area	*Policy areas merged*	*2013-2017*	*2013*	*2014*	*2015*	*2016*	*2017*
Economics	Business and enterprise; UK economy; Tax and revenue; Employment; Trade and investment; Financial services	21112	2346	5373	4287	4155	4951
Government	Government efficiency, transparency and accountability; Local government; Government spending; Regulation reform; Media and communications	11399	1618	2987	2343	2005	2446
Environment	Environment; Food and farming; Climate change; Wildlife and animal welfare; Rural and countryside	10997	1591	2557	2378	2175	2296
Security	Crime and policing; Law and the justice system; Defence and armed forces; Public safety and emergencies; National security	9729	1096	2308	2030	2028	2267
Housing and travel	Transport; Housing; Planning and building	8995	1281	2028	1766	1574	2346
Healthcare	National Health Service; Public health; Social care	8836	892	1535	1910	2156	2343
International affairs	Borders and immigration; Foreign affairs; International aid and development; Wales; Northern Ireland; Scotland; Europe	8376	1129	1494	2115	1759	1879
Society	Community and society; Children and young people; Welfare; Equality, rights and citizenship; Pensions and ageing society; Consumer rights and issues	6823	994	1500	1604	1381	1344
Education	Schools; Further education and skills; Higher education	6045	597	1171	1295	1497	1485
Science	Energy; Science and innovation	4134	653	979	901	818	783
Leisure	Arts and culture; Sports and leisure	704	141	117	158	116	172
Total		97150	12338	22049	20787	19664	22312

Scopus API citation searches

To find citations to one or more URLs from documents indexed by Scopus, a query of the following form can be used in either the Advanced Search interface or submitted to the Scopus API:

REF("[*search term*]") OR REF("[*search term*]") OR REF("[*search term*]")…

The result is a set of journal articles, magazines, conference papers or books indexed by Scopus that contain a citation in their reference section that matches any [*search term*]. Grey literature titles were not effective as search terms because they were often too short. For example, the UK government report, "Ahead of the curve" has a subtitle of "How UK motorsport technology and innovation can benefit your company". Due to the subtitle not being part of the title, almost exclusively false matches were found in Scopus (1288) when using the article title as the [*search term*]. In comparison, only one match was found when using the URL and omitting https://www., here, REF("gov.uk/government/publications/ahead-of-the-curve"), and this was a correct match. This strategy was not perfect because some URLs can be contained within longer URLs and documents could be cited by title without an URL. Nevertheless, the method can identify citations with high precision. These queries were submitted via the Scopus API to automatically gather the results.

For Scopus API to search the database, a text file for each grey literature area in each year was created. The file contained each query term, listed one per line. Each query included the "REF" part as above, as it is still required to search only the reference sections within Scopus. To match Google Books searches (discussed below), queries without the leading part (www.gov.uk/government/) were used. An example of such a query, using the example above, is:

REF("publications/ahead-of-the-curve")

The list of queries was then input into Scopus API search which automates the search process. In total, 235 query files were produced (47 policy areas per year across 5 years). Results returned are files of all query matches found. After some cleaning and matching, results files were combined into grey literature areas per year (Table 1). The number of matches per policy and grey literature area, and per year can then be calculated, and hence, impact assessed.

Google Books API citation searches

Google Books indexes a substantial fraction of the world's books. The academic books in its collection may contain references to grey literature. The free Google Books API can be queried via Webometric Analyst (WA) for URLs, as in the case of Scopus above. Whilst Scopus only returns a result if an exact match is found within the reference section, Google Books also returns close matches but highlights the matched section in the results returned. WA contains routines to filter out false positives by excluding results that do not contain the original query URL. However, due to the length of the original query on some URLs and imperfections in the Google Books description field (such as additional spaces or text wrapping issues), matches can be missed. Due to this, a second matching method was also used, as described below. All URLs have the form:

gov.uk/government/[*article-title-separated-by-hyphens*]

Here, "gov.uk" is the hostname and "government/[*article-title-separated-by-hyphens*]" is the path. The hostname and first part of the path (gov.uk/government/) are common to all grey literature references within this repository and are therefore useful to match true citations. Nevertheless, text wrapping could cause a problem due to the length of some URLs. If the URL part of the reference were to wrap to more than one line, URLs referenced might change due to the addition of an extra hyphen or a line break, causing a match to be missed. To avoid this issue the hostname and first part of the path (gov.uk/government/) were removed and two Google Books search strategies were formed.

For Google Books API to search the database, a text file for each grey literature area in each year was created. The file contained each query term, listed one per line. Here, each query did not include the "REF" part (as in Scopus), as Google Books does not have the ability to search a reference section specifically. Examples of each search strategy for the example above, to find matches for the document at URL gov.uk/government/publications/ahead-of-the-curve, are:

publications/ahead-of-the-curve
www.gov.uk

The list of queries for each search strategy was then separately input into Google Books API search contained within Webometric Analyst which automates the search process. As before, 235 query files were produced per search strategy (47 policy areas per year across 5 years). Results returned are files of all query matches found, including false positives where a similar match is found. Webometric Analyst also includes further routines to match the original query to the description output for each result, to ascertain true matches.

In pilot studies, comparisons between the two Google Books search strategies were performed to determine if one is inherently more suitable than the other. It was decided that the second search strategy using only the hostname (queries matching only www.gov.uk) was too general, causing matches to general webpages on the UK government website. The first search strategy, although possibly missing some matches due to the length of each query, was more specific and has better precision than the other search strategy.

From this decision, the Scopus search strategy defined above was finalised to be the same as Google Books – so both Scopus and Google Books were searched with the same part of the URL per query. This should help equate precision levels across the separate digital library searches.

After some further cleaning and matching, results files were combined into grey literature areas per year (Table 1). The number of matches per policy and grey literature area, and per year can then be calculated, and hence, impact assessed and compared to Scopus.

Following some pilot studies, some of the highest-ranked documents have very generic URLs. These may be overrepresented in this study as the citation count for the URL may include other URLs within the repository that start with exactly this URL, followed by further phrases.

Manual checking of results is needed, so precision was also calculated due to help remove the inclusion of false positives, estimated from a sample. A random sample of 50 documents in the original data that had at least one citation in Scopus API was extracted and manually searched in Scopus Advanced Search. This was then repeated for a further random sample of 50 with at least one citation in Google Books API and checked manually in Google Books. Precision for each document was calculated by comparing the automated citation count and manual citation count, and the smaller of the two was divided by the larger. The overall precision of each online library was then estimated by taking the geometric mean of the 50 document's precision levels.

Results

Proportions of UK government publications with Scopus or Google Books citations

Since most documents received no citations, the results focus on the proportion cited rather than the average number of citations per document. Other measures of impact exist that can deal with mostly uncited datasets, such as (Equalised) Mean-based Normalised Proportion Cited (MNPC and EMNPC) or Mean Normalised Log-transformed Citation Score (MNLCS) but require a comparison to a world average (Thelwall, 2017). Here, comparisons are between different online libraries across different disciplines, not compared to similar non-grey literature articles.

The results are split by year because comparing the proportion cited between years may be misleading due to the different lengths of time for a document to be cited; older documents with lower impact may report higher than newer document with a higher potential impact.

Comparisons between the original 47 policy areas as defined in the repository between the two search strategies are in the online Appendix (Table 3: https://figshare.com/s/51a8308bdf43772820b3).

The proportion cited from Scopus article path matching are always significantly above the proportion cited from Google Books with article path matching (all lower 95% confidence intervals for Scopus are larger than upper 95% confidence intervals for Google Books article path), across all years and all grey literature areas (55 occasions, 11 areas per year across 5 years) (Figures 1-5).

The more impactful grey literature areas have a proportion cited on Scopus >10% for most years, and some lesser impactful areas still have a proportion cited on Scopus >5% for older years, so a substantial minority have been cited.

As can be seen in figures 1-5, the proportion cited is generally higher in Scopus, and it seems that journals may cite grey literature more often than books. Nevertheless, the difference may be due to different levels of recall for the two search strategies.

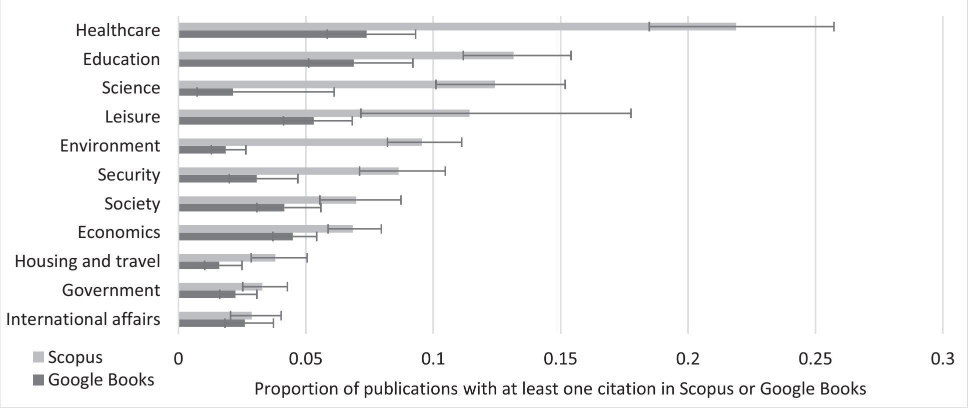


Figure 1. Proportion of UK government publications in 2013 with at least one citation in Scopus or Google Books with 95% confidence interval across 11 areas (Sorted by largest Scopus cited).

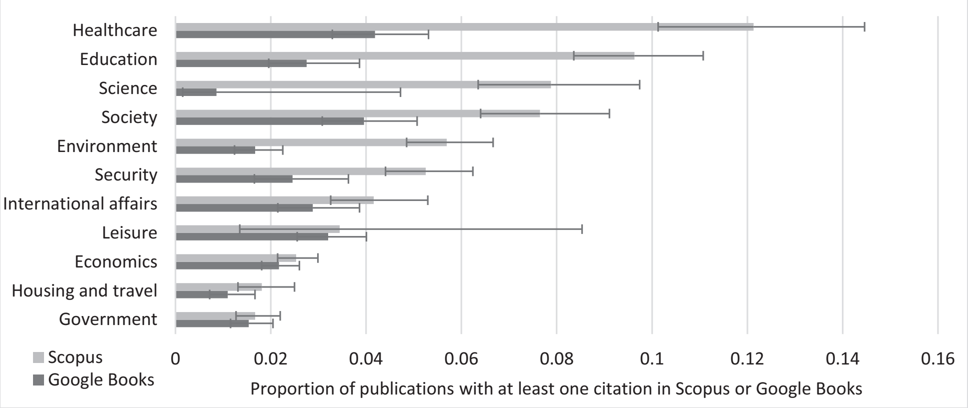


Figure 2. Proportion of UK government publications in 2014 with at least one citation in Scopus or Google Books with 95% confidence interval across 11 areas (Sorted by largest Scopus cited).

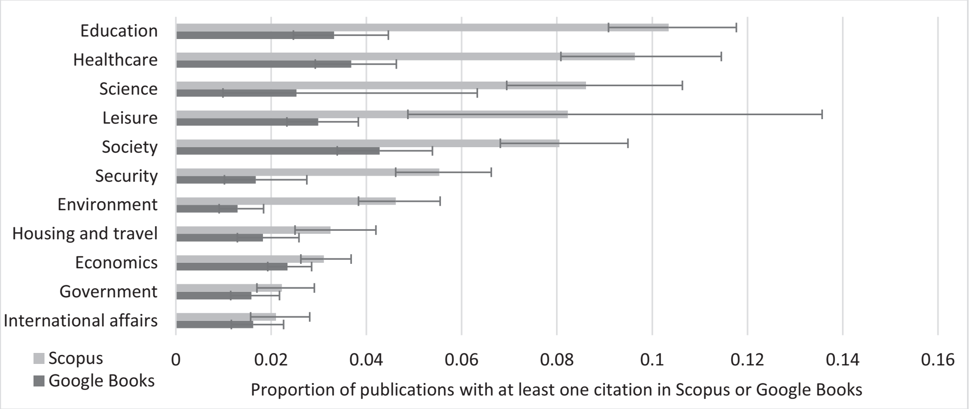


Figure 3. Proportion of UK government publications in 2015 with at least one citation in Scopus or Google Books with 95% confidence interval across 11 areas (Sorted by largest Scopus cited).

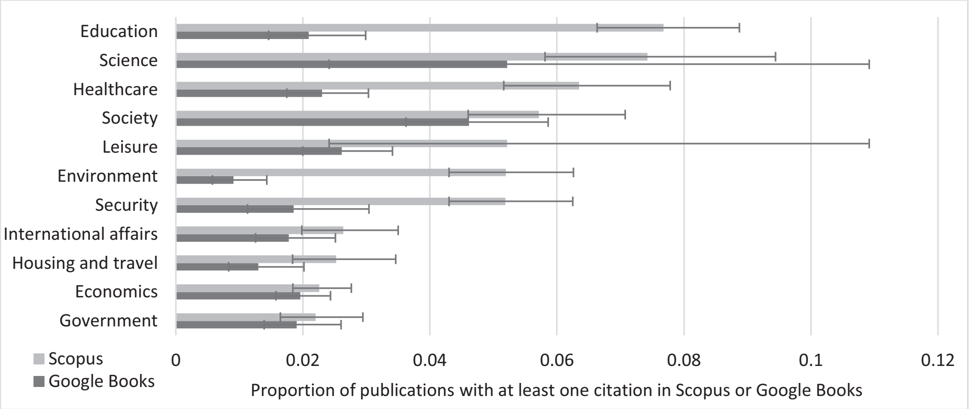


Figure 4. Proportion of UK government publications in 2016 with at least one citation in Scopus or Google Books with 95% confidence interval across 11 areas (Sorted by largest Scopus cited).

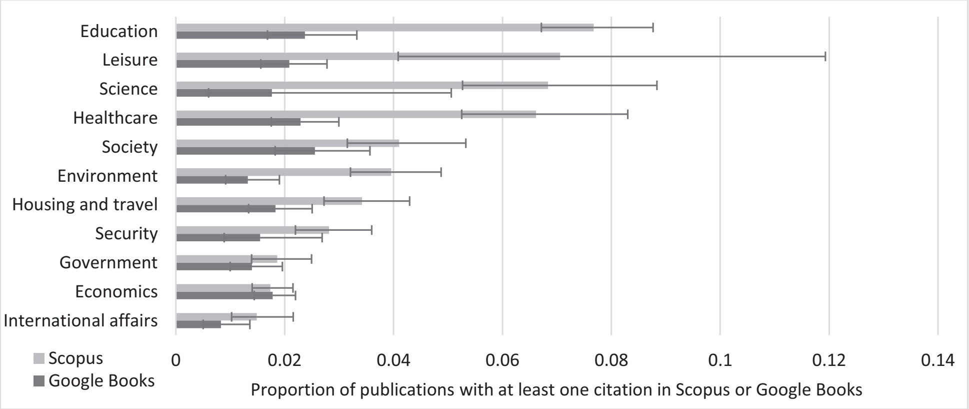


Figure 5. Proportion of UK government publications in 2017 with at least one citation in Scopus or Google Books with 95% confidence interval across 11 areas (Sorted by largest Scopus cited).

Characteristics of the most cited UK government grey literature in Scopus

The top three grey literature areas by proportion cited are Healthcare, Education and Science for each of the years 2013-2016 within Scopus references, and the same three are in the top four in 2017, with Leisure as second most cited, although with a large confidence interval.

In the first two years, Healthcare had the most impact, and has the second, third and fourth highest for 2015-2017 respectively. Education is in the top 2 most impactful grey literature areas; highest in 2013 and 2014, and second in all other years. Science is always third most impactful except for 2016, when it was second. In contrast, the grey literature areas International affairs, Economics and Government regularly finished bottom or near-bottom of the most impactful topics.

The grey literature area Healthcare in 2013 appears to be an anomaly due to its relatively high proportion cited (Scopus 0.22), with no other Scopus measurement above 0.13 for any year. A specific event, such as a national news story or major change in guidelines, may have caused a relative increase in 2013 research citing grey literature.

Table 2 shows the 25 most Scopus-cited grey literature documents across all subject areas. The five most cited grey literature documents in each subject area are shown in the online Appendix (Table 4: https://figshare.com/s/51a8308bdf43772820b3).

Table 2. Top 25 most cited UK government publications as found by Scopus.

Title (in bold) *URL (preceded by gov.uk/government/)*	*Year*	*Policy area*	*Grey literature area*	*Scopus citations*
Prisoners' criminal backgrounds and proven re-offending after release publications/2012	2013	Crime and policing	Security	3933
Housing publications/housing	2016	Tax and revenue	Economics	472
Climate change publications/climate-change	2017	Environment; Food and farming; Wildlife and animal welfare	Environment	333
Costs in disputed applications (PG38) publications/costs	2017	Housing; Business and enterprise	Housing and travel; Economics	260
Mental health and travelling abroad publications/mental-health	2014	Foreign affairs	International affairs	224
Bridges publications/bridges	2015	Government efficiency, transparency and accountability	Government	129
English indices of deprivation 2015 statistics/english-indices-of-deprivation-2015	2015	Community and society	Citizenship	121
Sustainability publications/sustainability	2013	Tax and revenue	Economics	112
NHS reference costs 2012 to 2013 publications/nhs-reference-costs-2012-to-2013	2015	National Health Service	Healthcare	97

NHS reference costs 2014 to 2015 publications/nhs-reference-costs-2014-to-2015	2015	National Health Service	Healthcare	84
NHS reference costs 2013 to 2014 publications/nhs-reference-costs-2013-to-2014	2015	National Health Service	Healthcare	83
Staffing publications/staffing	2017	Government efficiency, transparency and accountability	Government	72
NHS Constitution for England publications/the-nhs-constitution-for-england	2015	National Health Service	Healthcare	65
E-cigarettes: an evidence update publications/e-cigarettes-an-evidence-update	2015	Public health	Healthcare	56
Start active, stay active: report on physical activity in the UK publications/start-active-stay-active-a-report-on-physical-activity-from-the-four-home-countries-chief-medical-officers	2016	National Health Service; Public health	Healthcare	54
Energy consumption in the UK statistics/energy-consumption-in-the-uk	2017	Energy; Climate change	Science; Environment	50
NDNS: results from Years 1 to 4 (combined) statistics/national-diet-and-nutrition-survey-results-from-years-1-to-4-combined-of-the-rolling-programme-for-2008-and-2009-to-2011-and-2012	2017	National Health Service; Public health; Children and young people	Healthcare; Citizenship	46
Facts and figures statistics/facts-and-figures	2014	Business and enterprise	Economics	45
Social media publications/social-media	2015	Wales	International affairs	44
Websites publications/websites	2014	Transport; UK economy	Housing and travel; Economics	44
Open Data Charter publications/open-data-charter	2013	Government efficiency, transparency and accountability	Government	42

The top-ranked documents have generic URLs, such as gov.uk/government/publications/2012 (first in Table 2) and are overrepresented here as this URL does not represent the entire article title, and there are other URLs within the repository that start with this URL (gov.uk/government/publications/2012-user-event-taking-part-survey for example). Following this, URLs such as gov.uk/government/publications/open-data-charter (25th in Table 2) appear to be a generic URL due to words used and length, but will not be as generic as the first one.
For example, documents with citation counts that matched between Scopus API and Scopus Advanced Search had an accuracy of 1 (100%). Those with citation counts of one in either

method and two in the other had an accuracy of 0.5 (50%), and vice versa. This way, each non-agreement results in a fall in accuracy, whether the non-agreement is due to a false positive or a missed match. A combined precision of 0.82 (82%) was estimated for Scopus and 0.71 (71%) for Google Books, each calculated using the geometric mean of 50 text's precision levels.

Excluding these general URLs (Table 2), the themes of the most cited articles (articles with >60 citations) are statistics of an annual report, multiple annual healthcare reports, general healthcare updates/studies and the NHS Constitution. This agrees with the results at the start of this section, showing that healthcare is generally the most cited topic within grey literature. This is possibly due to the importance that current healthcare policy has on relevant practice from medical professionals, teaching within the sector and future policy changes in a publicly transparent field. Furthermore, an example such as "E-cigarettes: an evidence update" is one of the most cited, non-generic URL documents. It is of note due to the rising amount of healthcare research now surrounding the use of electronic cigarettes and derivatives, due to the unknown long-term problems with their use (Callahan-Lyon, 2014).

Another example of time-appropriate research is that of the document "Start active, stay active: report on physical activity in the UK". It has a very specific URL but is relatively highly cited. Physical activity is a useful tool for combatting many issues such as obesity (Bray et al, 2016) and cardiovascular disease (Wilson, Ellison & Cable, 2016), and with an increase of these problems in recent years, it is important to make sure research incorporates all aspects of research, including that of grey literature.

As shown in the online Appendix (Table 4: https://figshare.com/s/51a8308bdf43772820b3), and ignoring the generic URLs cited (as in Table 2), the types of publication within each grey literature area appear to vary. For example, like Healthcare as mentioned in the analysis of Table 2, the grey literature areas Housing and Travel, Science and Security all have highly cited annual reports that would naturally be updated yearly. These may be highly cited as they are updated each year, so the most recent version is always relevant. As new versions are released, old forms may be cited for comparative reasons.

Education, for example, features highly cited articles that centre around unique-to-the-field reasons, namely the National Curriculum. Four of the top five most cited articles are focussed on different subjects or levels within the curriculum, across all ages from school entry to leaving at age 16 or 18. Education is arguably one of the most important areas of research due to the importance of learning from a young age, in addition to the increasing adoption of technology in the classroom at all levels in recent years (Davison & Lazaros, 2015, Domingo & Garganté, 2016). Alongside this, a highly cited article on SEND (Special Educational Needs and Disabilities) is about codes of practice within this area (also classified as a Society grey literature document in this study). This may be due to an increasing focus on inclusion of children with special education needs in the classroom within the regular school lesson (Hornby, 2015, Bryant, Bryant & Smith, 2017).

Discussions and limitations

Using Google Books and matching just the term www.gov.uk in the description field gives more results due to the inclusion of extra spaces and line breaks in the description. However, this is not a good strategy because it introduces extra false matches. Any mention of any governmental page within the description field will cause a match due to all pages starting with the hostname, even if the match is a non-article such as a general webpage. Following, it can be suggested that Google Books article path has a higher precision but likely will miss some matches. Scopus appears to have a balance in terms of higher recall and improved precision compared to Google Books search strategies – a more specific matching term with no major issues found when matching article path to generate results.

From this, Google Books article path has been shown to display a lower proportion cited overall. Although no 'gold standard' to measure online impact within grey literature exists, the results suggest that Scopus API references when matched with the article path part of the URL is likely the best search strategy from those studied here.

To ease the collection and impact assessment of grey literature in future, it may be useful for publishers of these documents to provide their publications with persistent identifiers like DOI.

Limitations

The results are limited using a single case study (UK government publications). The Scopus API requires a paid subscription to use and is limited to 10,000 queries per week. Research of this size may take 10 weeks (n=97,150 for this study), and larger studies may take longer.

Merging of UK government policy areas are somewhat arbitrary for certain areas. Although the policy areas 'schools', 'further education and skills' and 'higher education' form a logical group, others are less intuitive, such as 'food and farming' within 'environment' and the 'housing and travel' grouping. It appears that the more ambiguous groupings were the less impactful, so should not affect the results much, but care should be taken if grouping into grey literature areas. In addition to this, the policy areas defined in the repository used have changed since data was gathered for this study, reducing the number of policy areas. As the total is reduced, it is likely that this may counter some of the problems when defining grouping into grey literature areas.

Several generic URLs were found within the repository that produced many incorrect search matches. This problem needs to be mitigated by data cleaning. The removal of generic URLs may be necessary if studying characteristics of specific documents. Determining which URLs are generic and specific requires manual checking of results, which increases time needed. For the results with extreme citation counts (publications/2012 with 3922 citations, for example), a sample of these matches must be checked to assess the proportion (accuracy and coverage) of false matches to generate an estimate of the total number of correct matches.

Conclusions

In answer to the research questions, a semi-automatic method can be used to identify grey literature publications for both Scopus and Google Books. Although some data collected may need to be cleaned and some text editing required for matching in Webometric Analyst, most steps of the method can be run automatically. From this, the impact of a grey literature article can be gauged using a specific repository. If the repository can be crawled or data can be manually gathered, Scopus can be used to determine how often it has been cited. In addition, the impact of grey literature documents can also be assessed through Google Books.

Scopus appears to be a better measure of impact for grey literature compared to Google Books, at least in terms of generating more matches in addition to a higher level of precision (generated from a random sample of 50 cited documents). Pilot studies showed a larger impact measurement if matching Google Books to a more generic but still suitable matching term. Although recall will be higher, precision would be lost due to the matching term not including any part of the article title (or URL equivalent). Precision and recall are acceptable when using this method for grey literature, as judged for Scopus API, showing clear differences in impact for each grey literature area across all years, when it exists. Google Books suffers with precision if the matching term is too generic, and recall is lower with equivalent matching terms.

Finally, Healthcare, Education and Science seem to be the most cited type of grey literature, at least in terms of UK government documents. Researchers assessing document-based knowledge flows in these areas should include grey literature within their analysis in order to get a more complete picture, who can be assisted by publishers of grey literature by including persistent document identifiers such as DOI.

Appendices

Appendix 1 (Table 3), Appendix 2 (Table 4), and Appendices 3-7 (Figures 6-10), as referred to above, can be found in the online Appendices (https://figshare.com/s/51a8308bdf43772820b3).

References

Bray, G.A., Frühbeck, G., Ryan, D.H. & Wilding, J.P.H. (2016). Management of obesity. *The Lancet*, 387(10031), 1947–1956.

Bryant, D., Bryant, B. & Smith, D. (2017). Teaching Students with Special Needs in Inclusive Classrooms. *ELT Journal*, 71(4), 659.

Callahan-Lyon, P. (2014). Electronic cigarettes: human health effects. *Tobacco Control*, 23(2), 36–40.

Davison, C.B. & Lazaros, E.J. (2015). Adopting Mobile Technology in the Higher Education Classroom. *The Journal of Technology Studies*, 41(1), 30–39.

Domingo, M.G. & Garganté, A.B. (2016). Exploring the use of educational technology in primary education: Teachers' perception of mobile technology learning impacts and applications' use in the classroom. *Computers in Human Behavior*, 56(3), 21–28.

GreyNet International (2019). *Document Types in Grey Literature*. Retrieved January 4, 2019 from: http://www.greynet.org/greysourceindex/documenttypes.html.

Harzing, A. W. K. (2010). *The publish or perish book*. Melbourne: Tarma software research.

Hornby, G. (2015). Inclusive special education: development of a new theory for the education of children with special educational needs and disabilities. *British Journal of Special Education*, 42(3), 234–256.

Institute of Medicine (2009). *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research* (p. 112). Washington, DC: The National Academies Press.

Interagency Gray Literature Working Group (IGLWG, 1995). *Gray Information Functional Plan (GIFP)*, Retrieved January 7, 2019 from: https://apps.dtic.mil/dtic/tr/fulltext/u2/b300928.pdf.

Kousha, K. & Thelwall, M. (2015). An automatic method for extracting citations from Google Books. *Journal of the American Society for Information Science and Technology*. 66(2), 309–320.

Kousha, K. & Thelwall, M. (Submitted). Can Google Scholar and Mendeley help to assess the scholarly impacts of dissertations? *Journal of Informetrics*.

Kousha, K., Thelwall, M. & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, 62(11), 2147–2164.

Orduna-Malea, E., Martín-Martín, A. & López-Cózar, E. D. (2017). Google Scholar and the gray literature: A reply to Bonato's review. arXiv preprint arXiv:1702.03991.

Schöpfel, J. (2010). Towards a Prague Definition of Grey Literature. *Twelfth International Conference on Grey Literature: Transparency in Grey Literature*. Prague, Czech Republic, 6-7 December 2010. Grey Tech Approaches to High Tech Issues, 11-26.

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, 11(1), 128–151.

University of New England (UNE, 2019). *Grey literature*. Retrieved January 4, 2019 from: https://www.une.edu.au/library/support/eskills-plus/research-skills/grey-literature.

Wilkinson, D., Sud, P., & Thelwall, M. (2014). Substance without citation: Evaluating the online impact of grey literature. Scientometrics, 98(2), 797-806.

Wilson, M.G., Ellison, G.M. & Cable, N.T. (2016). Basic science behind the cardiovascular benefits of exercise. *British Journal of Sports Medicine*, 50(2), 93–99.

Exploring the development of science-based nanotechnology

Lili Wang[1] and Zexia Li[2]

[1]*wang@merit.unu.edu*

UNU-MERIT, Maastricht University, Maastricht (The Netherlands)

[2]*lizexia@mail.las.ac.cn*

National Science Library, Chinese Academy of Sciences; Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing, (P.R.China)

Abstract

Scientific research has been acknowledged as an important knowledge resource for developing technologies. Studying science-technology linkages is crucial to help understand the mechanisms of innovation. Using nanotechnology as a case study, this paper investigates what types of scientific research can help improve the impact of technologies. This study uses backward and forward citation analysis, extracted from the Derwent World Patents Index (DWPI). Non-patent citations (NPCs) from each patent are further connected with records indexed in Web of Science, and the forward citations for the cited articles are collected. On the one hand, our results confirm that there is an important contribution from science to technology. High-impact academic research has significantly contributed to the development of high-impact patents. On the other hand, this study also reveals the heterogeneous pattern of patents citing scientific publications, depending on the organizational type, country, and knowledge origin. Compared to those in the U.S., patents developed by Chinese inventors tend to reply on more recent science but with a narrower scientific scope.

Introduction

Over the past decades, the degree to which scientific publications have supported the development of industrial innovations has increased remarkably (Narin et al. 1997; Wang and Li 2018; Branstetter 2005). The involvement of science in technology has not only helped spur new innovations (Mansfield 1991; Beise and Stahl 1999) but also helped to improve the quality of inventions (Sorenson and Fleming 2004; Wang and Li 2018; Branstetter 2005). Despite the rising science and technology (S&T) interactions and the important role science has played in accelerating technologies, as pointed out by many scholars (Meyer-Krahmer and Schmoch 1998; Acosta and Coronado 2003), the mechanisms of S&T linkages are still to a large extent unknown. This is partly due to differences in technological complexity across regions or sectors, and the difficulty of measuring the linkages between science and technology (Mansfield 1991; McMillan et al. 2000; Narin et al. 1997; Acosta and Coronado 2003).

In the science-based fields, such as biotechnology and nanotechnology, it is believed that “academic science might yield the highest economic reward” (Meyer-Krahmer and Schmoch 1998), but yet the communication is insufficient and a better understanding of S&T interactions is crucial. Given the well-recognized sector-specific nature between science and technology, this study focuses on one technological domain, i.e. nanotechnology.

The contributions of this paper are two-fold. On the one hand, we explore how science has contributed to the quality (or impact) of nanotechnologies. On the other hand, special attention is paid to the heterogeneous nature of S&T interactions, depending on the organizational types, countries, and knowledge origins.

Theoretical framework and research questions

The contribution of science to technology can be influenced by various factors, such as the characteristics of the national systems of innovation (Meyer-Krahmer and Schmoch 1998), and the level of technological complexity in a region (Acosta and Coronado 2003). If inventors preferentially cite papers authored in their own country, as suggested by Narin et al. (1997), the pattern of S&T linkages would also depend upon the availability of local scientific knowledge resource. In this study, we aim to link up the characteristics of science with the quality (or impact) of technologies. That is, to explore what types of scientific knowledge are more beneficial to help improve the patent quality (or impact). We focus on a series of scientific characteristics, including quality, newness, scope and country of origin of scientific research.

Quality of science

In the existing literature, the quality of academic research is mainly represented by the evaluation indicators (i.e. ranking) of research departments. However, whether research from higher-quality research departments is really more beneficial to technological development is still a matter of debate. On the one hand, some scholars suggest that the contribution of science to industrial innovation is directly related to the quality (ranking) of the research departments where scientific research has been conducted (Mansfield 1995). It is believed that more reputable research institutes are more likely to produce research that industry is looking for (Tornquist and Kallsen 1994). On the other hand, studies also show that the quality of a research department has no impact on the probability of S&T interactions. Although the ranking of a research department is associated with the percentage of citations received by this institute (Mansfield 1995; Mansfield and Lee 1996), we argue that it is the quality of cited articles, rather than the quality of research institutes, contributing to the technological development. Hence, in this study we move a step further to investigate individual articles cited by patents. The cited articles are linked with information from Web of Science. We use the number of forward citations received by these articles as a proxy for the quality of cited science. By linking up the quality of cited publications with the quality of citing patents, we aim to test whether highly cited articles would help increase the impact (proxied by forward citations) of a citing patent.

Age of scientific results

Recent scientific publications represent the results of up-to-date academic research. However, it takes time for inventors to recognize and utilize scientific results, i.e. there is usually a time lag between the appearance of research output and its application to industry (Adams 1990;

Mansfield 1991). Earlier studies attempted to test the time lag between patents and the cited academic articles. Popp (2017) finds that the probability of patents citing articles peaks 15 years after the article was published, while Finardi (2011) suggests that the time lag between production of scientific knowledge and its technological exploitation is about 3-4 years. As explained by Popp (2017), the time difference is likely caused by the sector-specific feature, for instance, energy research may take longer to progress to a commercialized product. Besides the sector-specific feature, our study takes country-specificity and difference of knowledge origin into consideration. By controlling the country, organization, and knowledge origin variables, we expect that a timely knowledge transfer from science to technology is valuable for generating high-impact technologies.

Scope and variety of knowledge

In addition to the feature of scientific quality mentioned above, the quantity of cited articles is also one of the factors influencing the effect of knowledge transfers from science to technology. With the increasing trend of scientific application into technological development (Narin et al. 1997; Wang and Li 2018; Branstetter 2005), one may wonder whether more references to scientific articles would lead to a better quality patent. What's more, it has been believed that the diversity of knowledge is an important factor facilitating knowledge creation (Liao and Phan 2016). However, by looking at how science and technology are combined in one specific field, Appio et al. (2017) suggest that a high degree of knowledge diversity does not always lead to more impactful inventions. We contend that the patterns of citation (in terms of how many scientific articles to cite or what kind of scientific articles to cite) are country-specific. In studying knowledge transfer, it is important to note the differences between countries of origin (Fernández-Ribas and Shapira 2009).

Knowledge origins

Knowledge resources can be explored by studying the country of origin of references cited by industrial patents. Knowledge documented in an article is known as codified knowledge that can be easily circulated and exchanged. Different from tacit knowledge which is often socially localised, codified knowledge can be transferred easily over large distances (Cohendet and Meyer-Krahmer 2001). In theory, there is an equal chance for inventors to access and cite codified knowledge (e.g. scientific articles) from various geographical locations. However, studies also find that each country's inventions preferentially cite scientific articles from their own country (Narin et al. 1997), or even preferably from closely located institutes (Beise and Stahl, 1999). This indicates that, to some extent, knowledge from home-country might be more relevant than that from foreign countries.

We tend to agree that there is a national bias in the citation patterns for industrial patents (Narin et al. 1997; McMillan et al. 2000). Although the existing studies have been limited to the developed countries, such as the U.S. and Japan, we investigate whether there is also a national bias in the citing patterns in developing countries.

Data collection and methodology

Disciplinary origin is an important factor in influencing the pattern and intensity of knowledge transfer from science to technology (Wang and Li 2018; McMillan et al. 2000). Scientific research may play a more important role in stimulating more complex technologies, while for less complex technologies, scientific elements may not be crucial. In this study we focus on the field of nanotechnology, which is of one the promising and key enabling technologies important for both developed and developing regions (Heinze 2004; Wang et al. 2013; Wang et al. 2019; Coccia and Wang 2015).

The patent data used in this study are collected from the Derwent World Patents Index (DWPI) via the platform Derwent Innovation (previously known as Thomson Innovation). Since 2011, all patent offices worldwide have classified nanotechnology uniformly under the International Patent Classification (IPC) system. The old Y01N system has been transformed to the B82Y category[1]. In the new system, all nanotechnology related patents are classified with an IPC code B82[2]. In total there are 129,123 patent applications related to nanotechnology in the studied period of 2000-2015. For each patent, both backward citations and forward citations are collected. Backward citations include both patent citations and non-patent citations (NPCs). After removing those nano patents without non-patent citations, we obtain 66,105 patents citing non-patent documents. NPCs consist of various types of references, including scientific articles, withdrawn patents, technical manuals, databases, web-based information, news, etc. This study aims to investigate the patents citing scientific articles published in Web of Science (WoS) indexed journals. Based on the DOI and title of the listed non-patent citations in each patent, we identify whether these NPCs are indexed by WoS journals. If they are not, we remove the patents from our sample. This step results in 33,050 nano patents[3].

Based on the information of assignees, the above patents are classified into three organizational types: 1) patents developed by firms, 2) patents developed by research institutes and universities[4], and 3) firm –university collaborated patents. Country codes are extracted based on the addresses of inventors. For each patent, we collect its publication ID, application date, backward citations

[1] http://www.epo.org/news-issues/issues/classification/nanotechnology.html.

[2] There are two sub-categories covered, i.e. B82B and B82Y. The former refers to inventions related to nano-structures formed by manipulations of individual atoms, molecules, or limited collections of atoms or molecules as discrete units; manufacture of treatment thereof. The latter refers to inventions related to specific uses or applications of nano-structures; measurement of analysis of nanostructures; manufacture of treatment of nano-structures (see more at http://www.wipo.int/ipc/itos4ipc/ITSupport_and_download_area/20130101/pdf/scheme/full_ipc/en/b82.pdf).

[3] Scientific references (added by examiners) with a publication year later than the patent publication year are removed.

[4] This type of patent is also called university patents in this paper.

(including number of cited patents and number of cited WoS publications), number of forward citations, and assignee countries. For the cited WoS publications, we collect the information on publication year, ut number[5], number of forward citations, countries of authors.This study aims to explore the impact of scientific research on the quality of technologies. Patent forward citations are collected to represent the impact or quality of the investigated technology. For each nano patent, we construct the following indicators.

Scope of technological background is represented by the number of cited patents (nr_cited_patents), $\sum P_i$. If a great number of patents have been cited by one new invention, this indicates that this new invention was developed based on a wide range of technologies. Otherwise, the scope of technological background is regarded as narrow.

Scope of scientific background is denoted by the number of cited WoS articles (nr_cited_pubs), $\sum S_i$. Similar to the previous indicator, citing more scientific publications indicates a wider scope of scientific background.

Newness of cited scientific articles is measured by the average publication year of the cited WoS articles. Suppose one nano patent cited N WoS scientific articles and the publication year of each article is expressed as YS_i, then the value of newness of cited scientific articles can be calculated by $\frac{\sum_{i=1}^{N} YS_i}{N}$. A lower newness value indicates that this group of cited scientific articles is on average relatively old, while a higher newness value indicates that this patent has been developed based on a group of more recent scientific articles.

If one nano patent cited N WoS scientific articles and the forward citation of each article is FCS_i, then the average forward citation of this group of scientific articles can be measured by $\frac{\sum_{i=1}^{N} FCS_i}{N}$. A higher level of average forward citation implies that a group of high-quality scientific research has been used in developing this studied nano patent. A lower level of average citation implies that this nano patent has been developed based on less impactful (or relatively unknown) scientific articles.

Variety of scientific knowledge resources is measured by the number of author countries of the cited scientific articles. Duplicates have been removed from the list of countries. For instance, if a nano patent cites three WoS articles and the author countries for these three articles are U.S., Netherlands, and U.S., respectively, the value of 'variety of scientific knowledge resources' for this patent is counted as two[6].

[5] This is the accession number, a unique identifying number associated with each record indexed in Web of Science.
[6] The country name of U.S. is counted once.

We use cited WoS articles from the U.S. to represent knowledge flows from advance countries, and cited WoS articles from China to represent knowledge flows from emerging economies. Given that the nature of patents developed by industry is different from that of patents developed by universities (Henderson et al. 1998), we separate firm patents from university patents in most models. Poisson and Negative Binomial are two types of regression used often to model count data. In our case, due to the over-dispersion feature of the outcome variable (i.e. the variance with each sub-group is higher than the mean within each sub-group), we use Negative Binomial regressions.

Results and discussion

Considering the fact that the invention pattern of firms is different from that of research institutes or universities (Wang and Li, 2018), we provide a separate analysis of patents from different organizational types. The scope of technological background (i.e. number of cited patents) varies greatly, ranging from 0 to 2197 for nano patents developed by firms and 0 to 824 for those developed by research institutes and universities. On average, the number of patents cited by firm patents is remarkably higher than that of university patents, with a mean of 30.44 for the former group and 8.75 for the latter. This shows that, as expected, innovation performance carried by firms is more closely related to the market than innovation performance carried by university.

The distribution of cited literature is highly skewed (Popp 2017). Hence we take natural logarithms for the independent variables (except average year of cited WoS publications which is a variable with normal distribution). Given that there is a time lag in receiving citations, it is of importance to take the age of patents into consideration. The likelihood ratio test has been performed by comparing the models without year dummies and those with year dummies. The test shows that the latter group fits significantly better than the former group (prob>chi2=0.0000). Therefore, regressions in this study all include year dummies[7]. To a lesser degree, the national economic environment may also influence the patenting performance in one country. Hence, we use the cluster function in the regression models[8] to capture the differences between countries[9].

Table 1 documents the regression results of three samples: 1) all patents, 2) firm patents and 3) university patents[10]. Given that *Scientific scope* (nr_cited_pub) and *Variety of scientific resources* (nr_all_country) are highly correlated, these two variables are included in separate models.

Technological scope (i.e. number of cited patents) has a significant and positive effect on the dependent variable (i.e. forward citations of nano patents) in all models, including the full

[7] Based on the publication years of patent applications.

[8] The Akaike information criterion (AIC) and Bayesian information criterion (BIC) both suggest that models with country clusters perform better than those without.

[9] Based on the assignee countries.

[10] Firm-university collaborated patents are not included in the sub-samples.

sample (Models 1&2), sub-sample of firm patents (Models 3&4) and sub-sample of university patents (Models 5&6). The second variable, *scientific scope* (i.e. number of cited WoS publications) is positively significant in all three models. This implies that a wider scope of scientific references is likely to lead to a higher patent quality. With regard to the third variable, *newness of scientific knowledge*, there is a slight difference between firm patents and university patents. The coefficient values are higher and more significant for university patents (Models 5&6) than those firm patents (Models 3&4), indicating that citing more recent scientific publications is more beneficial to university inventions. This issue will be further explained later in Table 2 & 3.

Regarding the issue whether the quality of science plays a role, results in Table 1 show that the *quality of cited scientific articles* (forward citations of cited WoS publications) has a positive effect on the quality of citing patents. This positive effect is highly significant for all models at the 0.01 level, indicating that a patent developed based on a group of high-quality scientific research is likely to receive more forward citations itself in the future.

Variety of scientific knowledge resources has a significant coefficient in the full sample models. However, such effect is found to be higher and more significant in the firm patent group than the university patent group. This points out that it is more valuable for firm patents (rather than university patents) to refer to scientific knowledge from various origins. Year dummies are highly significant for all models, which can be explained by the time lag of forward citations.

Table 1: Negative binomial regression results for the full sample

	All patents		Firm patents		University patents	
	(1)	(2)	(3)	(4)	(5)	(6)
Technological scope	0.40334***	0.42616***	0.44894***	0.46629***	0.34772***	0.389198***
	(0.026699)	(0.024569)	(0.022125)	(0.017702)	(0.048329)	(0.047904)
Scientific scope	0.09904***		0.08102***		0.121147***	
	(0.019195)		(0.023472)		(0.034407)	
Newness of scientific knowledge	0.01554***	0.01453***	0.00920**	0.00832**	0.04116***	0.03793***
	(0.003716)	(0.003792)	(0.003822)	(0.003910)	(0.003437)	(0.004351)
Quality of cited scientific articles	0.13186***	0.14154***	0.09527***	0.09935***	0.18546***	0.20526***
	(0.013256)	(0.012240)	(0.015086)	(0.013974)	(0.025328)	(0.026918)
Variety of scientific resources		0.07314*		0.07719**		0.02914
		(0.042004)		(0.035247)		(0.079880)
Constant	-33.0950***	-31.0665***	-20.6358***	-18.8801**	-84.5003***	-77.9982***
	(7.288861)	(7.421608)	(7.640187)	(7.820862)	(6.956649)	(8.678426)
Prob > chi2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Pseudo R2	0.0595	0.0592	0.0504	0.0502	0.0552	0.0547
Observations	29,748	29,748	13,187	13,187	8,863	8,863

Note: 1) The dependent variable is the number of forward citations received by each nano patent. 2) Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. 3) The results on 15 year dummies are not reported in this table.

Following the results in Table 1, we extend the analysis to explore the effect of knowledge resources on patent quality. We use WoS references from the U.S. to represent knowledge flows from advanced countries, and WoS references from China to represent knowledge flows from emerging economies. The knowledge recipients are divided into home and foreign groups. The former stands for the group of patents developed in the same country as the knowledge resource origin, and the later represents the group of patents developed by foreign countries. Considering the aforementioned different nature of inventions developed by different organizational types, firm patents and university patents are tested separately in Table 2 and Table 3.

For patents developed by firms, Table 2 includes four sub-groups, i.e. knowledge flows from the U.S. (sub-groups 1&2) and knowledge flows from China (sub-groups 3&4). Results show that, similar to the aggregated results in Table 1, *technological scope* (i.e. number of cited patents) has a positive and significant effect on the quality of citing patents, irrespective of countries of knowledge resources.

The results on *scientific scope* (i.e. number of cited publications) are mixed, as they depend on the organizational types and knowledge resource countries. A wide scientific scope has a significant positive effect on patents citing scientific articles from the U.S. (Models 1&3). However, for the sample with scientific knowledge flows from China, this variable presents a non-significant effect in the home patent group (i.e. Chinese patents, in Model 3) while a positive and significant effect is presented in the foreign patent group (i.e. patents developed by foreign countries, in Model 4).

In terms of *newness of scientific knowledge*, there is also a noticeable difference between different patent groups. In the sample with knowledge flows from the U.S. (Models 1-4), a more significant and slightly higher coefficient is observed for home patents (Models 1&2) compared to foreign patents (Models 3&4). In the sub-sample for patents receiving knowledge flows from China (Models 5-8), the effect for home patents is almost twice as high as those for non-Chinese patents. That is, more recent scientific research is more beneficial to home patents developed by Chinese inventors (Models 5&6) than patents developed by non-Chinese inventors (Models 7&8). In general, this shows that more recent scientific knowledge is beneficial to patents developed by inventors from the same countries as the knowledge origins. For crossing knowledge flows – i.e. the knowledge resource country is different from the patent inventor country – relatively earlier knowledge is more beneficial. Among all the different groups, Chinese patents citing scientific knowledge from China have the highest coefficients, 0.08295 in Model 5 and 0.08804 in Model 6.

Similar to the results in Table 1, the quality of cited scientific articles is statistically significant and positive in all models in Table 2. This supports our hypothesis that high-quality scientific knowledge can help improve the quality of patents.

The *variety of scientific knowledge resources* has a significant and positive coefficient in sub-groups 1 & 4. This is consistent with the variable of scientific scope. In Model 4 (sub-group 2), this variable is positive but non-significant, in line with the variable of *scientific scope*. This shows that, in most cases, it is more beneficial to have a higher level of *variety of scientific resources* or *scientific scope*. However, for Chinese patents developed based on Chinese science (sub-group 3, Model 6), the *variety of scientific resources* has a negative contribution to the quality of citing patents. This indicates that more scientific publications and a higher diversity of author countries do not lead to a higher quality in the Chinese patents developed based on Chinese science.

Table 2: Negative binomial regression results for the sub-samples: firm patents

	Firm patents with scientific knowledge flows from the U.S.				Firm patents with scientific knowledge flows from China			
	U.S. patents (sub-group 1)		non-U.S. patents (sub-group 2)		Chinese patents (sub-group 3)		Non-Chinese patents (sub-group 4)	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Technological scope	0.39312*** (0.023908)	0.41149*** (0.021991)	0.44993*** (0.032150)	0.46180*** (0.030100)	0.58572*** (0.161350)	0.58165*** (0.153006)	0.48969*** (0.023597)	0.50679*** (0.020617)
Scientific scope	0.11046*** (0.031254)		0.09036* (0.052718)		-0.20815 (0.223325)		0.17332*** (0.053552)	
Newness of scientific knowledge	0.01555*** (0.003836)	0.01458*** (0.003907)	0.01480** (0.006375)	0.01343* (0.007005)	0.08295** (0.034733)	0.08804*** (0.034170)	0.04544*** (0.012396)	0.04160*** (0.011213)
Quality of cited scientific articles	0.05690*** (0.018955)	0.05990*** (0.018951)	0.14378*** (0.027425)	0.14576*** (0.027234)	0.13762* (0.075041)	0.14652** (0.073805)	0.07454*** (0.024545)	0.07336*** (0.022321)
Variety of scientific resources		0.14358*** (0.048648)		0.11753 (0.081973)		-0.40522 (0.274483)		0.27273*** (0.072923)
Constant	-33.284*** (7.690746)	-31.360*** (7.832271)	-32.045** (12.823142)	-29.299** (14.063596)	-168.579** (69.983691)	-178.754*** (68.834802)	-94.023*** (24.987037)	-86.367*** (22.626854)
Prob > chi2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Pseudo R2	0.0493	0.0491	0.0532	0.0531	0.0751	0.0766	0.0819	0.0818
Observations	4,440	4,440	3,652	3,652	204	204	2,617	2,617

Note: 1) The dependent variable is the number of forward citations received by each nano patent. 2) Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. 3) The results on 15 year dummies are not reported in this table.

Table 3 reports the regression results for university patents, with distinguished scientific knowledge flows from the U.S. and China. Most of the regression results in Table 3 are in line with those in the firm-patent sample in Table 2. A major difference is that, compared to Table 2, the coefficients of *technological scope* are not significant in sub-group 3 in Table 3. This indicates that the scope of technological background does not have a significant effect on the quality of Chinese patents developed base on Chinese scientific knowledge. Similarly, on average *variety of scientific resources* in sub-group 3 is also remarkable lower than the mean of this variable in the U.S., i.e. 1.8 v.s. 4.3 for firm patents and 1.5 v.s. 4.5 for university patents. This indicates that, for the group of Chinese patents developed based on Chinese science, *scientific scope* and the *variety of scientific knowledge resources* is considerably low.

Table 3: Negative binomial regression results for the sub-samples_university patents

	University patents with scientific knowledge flows from the U.S.				University patents with scientific knowledge flows from China			
	U.S. patents (sub-group 1)		non-U.S. patents (sub-group2)		Chinese patents (sub-group3)		non-Chinese patents (sub-group4)	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Technological scope	0.38270*** (0.035739)	0.41630*** (0.034223)	0.33098*** (0.096675)	0.33292*** (0.097677)	0.07031 (0.060566)	0.08763 (0.059460)	0.32456*** (0.056700)	0.35075*** (0.059070)
Scientific scope	0.15752*** (0.038941)		0.03266 (0.065184)		0.00224 (0.074884)		0.12225*** (0.030804)	
Newness of scientific knowledge	0.05685*** (0.009586)	0.05096*** (0.009594)	0.05816*** (0.013083)	0.05732*** (0.013389)	0.09070*** (0.014170)	0.08671*** (0.013563)	0.04885*** (0.006532)	0.04389*** (0.007528)
Quality of cited scientific articles	0.20099*** (0.030054)	0.21852*** (0.029949)	0.25996*** (0.033111)	0.26056*** (0.032292)	0.21301*** (0.028349)	0.22466*** (0.028864)	0.28662*** (0.036846)	0.30954*** (0.039172)
Variety of scientific resources		0.151076** (0.059208)		0.045958 (0.107655)		-0.162790 (0.106730)		0.077757 (0.057403)
Constant	-116.724*** (19.2921)	-104.941*** (19.3103)	-119.108*** (26.1910)	-117.437*** (26.7678)	-183.430*** (28.5714)	-175.308*** (27.3465)	-100.770*** (12.9701)	-90.825*** (14.9655)
Prob > chi2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Pseudo R2	0.0507	0.05	0.0396	0.0396	0.0631	0.0635	0.0702	0.0697
Observations	2,563	2,563	2,468	2,468	1,298	1,298	1,760	1,760

Note: 1) The dependent variable is the number of forward citations received by each nano patent. 2) Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. 3) The results on 15 year dummies are not reported in this table.

Conclusions

In order to understand what types of scientific knowledge can help generate impactful technologies, this paper uses nano patents data collected from the Derwent World Patents Index (DWPI) and extracts both patent citations and non-patent citations (NPCs) for each patent. By linking the quality of patents with the characteristics of NPCs (including their information from Web of Science), this study investigates how science can help improve the quality of new technologies.

On the one hand, our study shows that science can help improve the quality of patents in many different ways. This, in accordance with Sorenson and Fleming (2004), emphasizes the importance of S&T linkages. On the other hand, this study stresses that the science-technology linkage patterns can differ across countries. There are also differences between organizations, as well as knowledge origins. This corresponds to the heterogeneous nature of knowledge sources stressed by Audretsch and Link (2019). In our empirical study, we find that more recent scientific knowledge has a more significant contribution to the quality of patents. In most cases, this effect is stronger for university patents than for firm patents. Compared to the U.S., China's patents seem to benefit greatly from recently published scientific research. This might be explained by the fact that, as an emerging country, China's patents focus more on emerging technologies, rather than mature technologies. Hence, more recent science is more relevant in China.

Our results suggest that a wider scientific scope and a high level of variety of scientific resources do not always lead to a higher patent quality. A specific case is that, when Chinese patents are developed based on Chinese science – in which case the scientific scope and variety level is very low – the significant positive contribution is missing. To some extent, in line with Appio et al. (2017), this proves our hypothesis that public science with a larger scope does not always lead to higher-quality inventions. Irrespective of patent types or country origins, our findings show that the quality of cited science has always been crucial in affecting the quality of citing patents. A high-quality scientific base can help lead to a high-quality patent.

References

Acosta, M., & Coronado, D. (2003). Science-technology flows in Spanish regions: An analysis of scientific citations in patents. *Research Policy*, *32*(10), 1783–1803. doi:10.1016/S0048-7333(03)00064-7

Adams, J. D. (1990). Fundamental stocks of knowledge and productivity growth. *Journal of Political Economy*, *98*(4), 673–702.

Appio, F. P., Martini, A., & Fantoni, G. (2017). The light and shade of knowledge recombination : Insights from a general- purpose technology. *Technological Forecasting & Social Change*, *125*(May), 154–165. doi:10.1016/j.techfore.2017.07.018

Audretsch, D. B., & Link, A. N. (2019). *Sources of knowledge and entrepreneurial behavior*. Toronto-Buffalo-London: University of Toronto Press.

Beise, M., & Stahl, H. (1999). Public research and industrial innovations in Germany. *Research Policy*, *28*, 397–422.

Branstetter, L. (2005). Exploring the Link Between Academic Science and Industrial Innovation. *Annales d'Économie et de Statistique*, (79/80), 119–142. doi:10.2307/20777572

Coccia, M., & Wang, L. (2015). Path-breaking directions of nanotechnology-based chemotherapy and molecular cancer therapy. *Technological Forecasting and Social Change*, *94*, 155–169. doi:10.1016/j.techfore.2014.09.007

Cohendet, P., & Meyer-Krahmer, F. (2001). The theoretical and policy implications of knowledge codification. *Research Policy*, *30*(9), 1563–1591. doi:10.1016/S0048-7333(01)00168-8

Fernández-Ribas, A. A., & Shapira, P. (2009). Technological diversity, scientific excellence and the location of inventive activities abroad: The case of nanotechnology. *Journal of Technology Transfer*, *34*(3), 286–303. doi:10.1007/s10961-008-9090-2

Finardi, U. (2011). production of scientific knowledge and its technological exploitation. *Scientometrics*, *89*(1), 37–50. doi:10.1007/s11192-011-0443-5

Heinze, T. (2004). Nanoscience and Nanotechnology in Europe: Analysis of Publications and Patent Applications including Comparisons with the United States. *Nanotechnology Law & Business*, *1*(4), 1–19. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1394375

Henderson, R., Jaffe, A. B., & Trajtenberg, M. (1998). Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965–1988. *Review of Economics and Statistics*, *80*(1), 119–127. doi:10.1162/003465398557221

Liao, Y. C., & Phan, P. H. (2016). Internal capabilities, external structural holes network positions, and knowledge creation. *Journal of Technology Transfer*, *41*(5), 1148–1167. doi:10.1007/s10961-015-9415-x

Mansfield, E. (1991). Academic Research and Industrial Innovation. *Research Policy*, *20*(1), 1–12. doi:10.1016/0048-7333(91)90080-A

Mansfield, E. (1995). Academic research underlying industrial innovations: Sources , characteristics, and financing. *The Review of Economics and Statistics*, *77*(1), 55–65.

Mansfield, E., & Lee, J. (1996). The modern university: Contributor to industrial innovation and recipient of industrial R & D support. *Research Policy*, *25*(7), 1047–1058. doi:10.1016/S0048-7333(96)00893-1

McMillan, G. S., Narin, F., & Deeds, D. L. (2000). An analysis of the critical role of public science in innovation: the case of biotechnology. *Research Policy*, *29*(1), 1–8. doi:10.1016/S0048-7333(99)00030-X

Meyer-Krahmer, F., & Schmoch, U. (1998). Science-based technologies: university–industry interactions in four fields. *Research Policy*, *27*(8), 835–851. doi:10.1016/S0048-7333(98)00094-8

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy*, *26*, 317–330.

Popp, D. (2017). From Science to Technology: The Value of Knowledge From Different Energy Research Institutions. *Research Policy*, *46*, 1580–1594. doi:10.3386/w22573

Sorenson, O., & Fleming, L. (2004). Science and the diffusion of knowledge. *Research Policy*, *33*(10), 1615–1634. doi:10.1016/j.respol.2004.09.008

Tornquist, K. M., & Kallsen, L. A. (1994). Out of the ivory tower: Characteristics of institutions meeting the research needs of industry. *The Journal of Higher Education*, *65*(5), 523–539.

Wang, L., Jacob, J., & Li, Z. (2019). Exploring the spatial dimensions of nanotechnology development in China: the effects of funding and spillovers. *Regional Studies*, *53*(2), 245–260. doi:10.1080/00343404.2018.1457216

Wang, L., & Li, Z. (2018). Knowledge Transfer from Science to Technology—The Case of Nano Medical Device Technologies. *Frontiers in Research Metrics and Analytics*, *3*(11), 1–8. doi:10.3389/frma.2018.00011

Wang, L., Notten, A., & Surpatean, A. (2013). Interdisciplinarity of nano research fields: A keyword mining approach. *Scientometrics*, *94*(3), 877–892. doi:10.1007/s11192-012-0856-9

How well do we evaluate evaluation? An overview of Science, Technology and Innovation Policy Evaluation in Latin America

Adriana Bin[1], Rafaela Andrade[2], Lissa Vasconcellos[3], and Sergio Salles-Filho[4]

[1] *adriana.bin@fca.unicamp.br*
University of Campinas, School of Applied Sciences, Pedro Zaccaria st, 1300, Limeira (Brazil)

[2] *rafamarcelly@gmail.com*
University of Campinas, Geosciences Institute, Science and Technology Department, Carlos Gomes st, 250, Campinas (Brazil)

[3] *lissavasconcellos25@gmail.com*
University of Campinas, Geosciences Institute, Science and Technology Department, Carlos Gomes st, 250, Campinas (Brazil)

[4] *sallesfi@ige.unicamp.br*
University of Campinas, Geosciences Institute, Carlos Gomes st, 250, Campinas (Brazil)

Abstract

Science, technology and innovation (STI) policy evaluation is an increasingly recognized practice around the world. But what is its real role in the formulation, implementation and/or maintenance of these policies? Considering that not only the results but also the quality of evaluations can influence policy design and execution, meta-evaluation, meta-analysis and evaluation synthesis studies are fundamental. This research-in-progress paper aims to present an effort of evaluation of STI policy evaluations in the Latin American context. The research is part of a broader initiative named Science and Innovation Policy Evaluations Repository (SIPER), coordinated by Manchester Institute of Innovation Research (MIoIR). The paper main contributions are: (i) a literature review about the concepts of meta-evaluation, meta-analysis and evaluation synthesis and applications of these tools in STI field; (ii) a framework of indicators to evaluate STI policy evaluations; (iii) an overview of STI policy evaluation in Latin America.

Introduction

The focus of this work is to present an on-going experience of collecting, coding and analyzing Science, Technology and Innovation (STI) Policy Evaluations in Latin America (LA), with emphasis in the evaluation design and methods. In this work, STI policies and programs are understood as any type of public intervention (in partnership or not with non-governmental actors) that promotes scientific and technological development and innovation at international, national, regional or local levels.

The research is part of a broader initiative named Science and Innovation Policy Evaluations Repository (SIPER), coordinated by Manchester Institute of Innovation Research (MIoIR). SIPER is a central source of knowledge on science and innovation policy evaluations. Its aim is twofold: (i) to provide on-line access to a unique collection of policy evaluations, located at a single location; and (ii) to allow policy learning by providing an informed analysis of the database contents that is both searchable by policy makers and other stakeholders and which provides the basis for additional academic analysis.

The paper is organized in three sections: a literature review about meta-evaluation, meta-analysis and evaluation synthesis and its applications in STI field; SIPER methodology and procedures adopted in the Latin American case; and the partial results about STI policy evaluation in Latin America.

Literature review

Meta-evaluation, meta-analysis and evaluation synthesis are not new concepts or instruments. They have been built and implemented since the late 1960s, but there is still ground for research about their operation and use. Jacob and Affodegon (2015) reinforce this point for the specific field of meta-evaluation, arguing that despite its evolution, there are still few studies and many challenges to be faced. For meta-analysis and evaluation synthesis, despite the abundance of studies (especially in the field of health sciences), the challenges of operationalization are equally important.

Meta-evaluation, a concept coined by Michael Scriven in the late 1960s, can be defined as the evaluation of the evaluation (Scriven, 2009; 2010) and is usually oriented to the measurement of the quality of a given evaluation, based on the criteria of validity, credibility, clarity, suitability, utility and generalization (Stufflebeam, 2001). It is, therefore, an evaluation that has the own evaluation as an object of interest (Furtado and Laperrière, 2012).

According to Jacob and Affodegon (2015), this quality measurement is related to the description of how the evaluation was performed (or even how it is being or will be performed) and a judgment based on standards of what can be considered a "good" evaluation. In this sense, evaluating the evaluations systematically assists both the improvement of evaluation techniques, contributing to the evaluators' own work, as well as the view on the robustness and integrity of the evaluation, contributing to its use (Hanssen et al., 2008; Firme and Letichevsky, 2010, Stufflebeam and Coryn, 2014, Jacob and Affodegon, 2015).

Scriven (2009) claims that a partial meta-evaluation is better than none. This means that although ideally a meta-evaluation involves checking the data collected, the evaluation design, analyzes and conclusions, making only one or more parts of this set can contribute in the directions indicated above - to improve the evaluators' work and the subsequent use of recommendations derived from an evaluation. This is a finding related to the very effort required for a complete meta-evaluation. An example in this line can be found in Cahalan and Goodwin (2014). The authors recover data from an evaluation of the Upward Bound US program conducted by a consulting firm (Mathematica) and discuss both the methods and the results achieved by this assessment, consistently indicating the biases of the original assessment.

It is also worth to distinguish the concept of meta-evaluation from two other related concepts: meta-analysis and evaluation synthesis, more oriented to organize and aggregate results of a series of evaluations (Good, 2012). According to Weiss (1998), Stufflebeam (2001) and Edler et al. (2008) the term meta-analysis refers to statistical analyzes that result from the integration of data, especially quantitative, obtained in a set of evaluation studies. In this sense, meta-analysis contributes to summarize the effects of a given policy in certain aspects, from the measurement performed in different evaluations. According to the authors, the great advantage of this instrument is to increase the volume of data available for analysis; the risks associated with it lie precisely in the biases that arise from the objectives and the ways in which different evaluations were carried out, the differences in the quality of these evaluations, as well as the selection of the studies that will compose the meta-analysis. Care and control of these differences are fundamental in the application of this type of instrument.

For its turn, evaluation synthesis, also discussed by Edler et al. (2008), is similar to meta-analysis in the way it is oriented to integrate a set of evaluations. However, evaluation synthesis seeks to answer different questions about the evaluated object, which result from evaluations with different focuses and methodologies, allowing a new interpretation of the different findings. It should be emphasized that the conceptual delimitation between the concepts of meta-analysis and synthesis of evaluations is not consensual since there are many papers that do not use statistical analysis and that call themselves meta-analyses. In this work, it is proposed that these concepts are not used interchangeably.

Despite the differences, there are authors who approach the three concepts, such as Cooksy and Caracelli (2005) and Edler et al. (2008), who understand that meta-evaluation in its original sense can serve as a step in the selection of good evaluations for later conducting an evaluation synthesis or meta-analysis. The meta-evaluation can also serve as a step after the elaboration of an evaluation synthesis or meta-analysis, as a way of analyzing the quality associated to the way of conducting these studies and/or their results.

In order to answer how these efforts apply in STI Policy field, a broad search was made in the Web of Science, Scopus and Scielo bases, combining terms related to STI (science, technology, research, R&D, innovation, policies, programs) with the exact terms of the concepts and tools (meta-evaluation, meta-analysis and synthesis of evaluations), from 2000 to 2018. The refinement of the articles in order to select only the pertinent ones was carried out from three subsequent activities: elimination of duplications, reading of abstracts and reading of articles. At the end, 32 articles were selected.

The articles were published in 26 different journals. All journals had only one article, with exception of Research Evaluation, in which 7 articles were identified. There was a total of 89 different authors, 5 of them with 2 publications: Edna Solomon co-authored with Mehmet Ugur, Erik Arnold, Jari Hyvärinen and Patrik Gustavsson Tingvall. We found 11 meta-analyzes, 11 synthesis of evaluations, 7 meta-evaluations, 2 combinations between meta-evaluation and meta-analysis and 1 combination between meta-evaluation and evaluation synthesis.

There were two main kinds of evaluations used as inputs to these studies. The first one included evaluations of the effects of R&D expenditures and/or research funding in different dimensions, especially in business and economic performance, including evaluations of research and innovation funding agencies and of EU Framework Program. The second group included evaluations of the effects of information and communication technologies (ICTs) in different dimensions, particularly in education.

The studies used as sources of data evaluations exercises mainly described in scientific papers and technical reports. From the methodological point of view, there was variation between the tools. Meta-analysis studies used mainly meta-regression. The evaluation synthesis studies used mainly descriptive statistics and qualitative analysis. The meta-evaluations were also centered on descriptive and multivariate statistics, focusing on the design of evaluations and, secondly on their quality and use. Finally, it is highlighted that two articles used meta-evaluation as a previous step to conduct a meta-analysis or an evaluation synthesis.

From this literature review, it can be concluded that, although they are important tools to understand evaluation results (policy contributions), as well as their quality (evaluation practice), meta-evaluation, meta-analysis and evaluation synthesis studies still have restricted use in the STI field. In general, the papers do not report the use of the results of the researches, although they consider their findings very useful for the formulation of policies.

The research presented in this paper contribute to filling this gap by conducting a comprehensive meta-evaluation study of STI policy evaluation in LA that also serve as a step in the selection of relevant evaluations for later running an evaluation synthesis or meta-analysis.

Methodology

The research follows a three-phase methodology: collect, code and analyze. Collection phase refers to identifying evaluation studies (papers, working papers, evaluation reports) of STI policies in LA countries. As defined by SIPER project methodology, qualified evaluations to be included in the study are those: (i) on science, technology and innovation policy; (ii) evaluating a clearly identifiable, specific program or group of programs; (iii) having a distinguishable methodology; and (iv) providing some sort of evidence. The searches were oriented to 6 Latin American countries: Argentina, Brazil, Chile, Colombia, Mexico and Uruguay.

Collection was done in four main sources: (i) institutional websites (ministries, government agencies and STI-related development agencies); (ii) academic teams, research organizations and companies dedicated to the study or practice of STI evaluation; (iii) journals oriented to STI evaluation; and (iv) world-wide-web in general. These strategies were thought as a way to capture what is available in the selected countries. 153 evaluation documents have been found, divided as follows: Argentina: 26 documents; Brazil: 38 documents; Chile: 35 documents; Colombia: 18 documents; Mexico: 19 documents; Uruguay: 17 documents.

Coding phase was dedicated to characterization of collected documents following SIPER requirements based on a survey, which includes:

(1) Related policy measure characteristics (targets, modality, objectives)

(2) Evaluation characteristics: (2.1) Basic (who conducted; timing, purpose; reference to intervention rationale); (2.2) Topics covered (i.e. outputs, outcomes, additionality, goal attainment, gender issues, degree of satisfaction of stakeholders); (2.3) Design; (2.4) Data collection methods; (2.5) Data analysis methods; and (2.6) Quality Issues.

These topics represent a suitable framework of indicators to evaluate evaluations in STI field, although they miss the satisfaction of the potential users as well as the effective use of evaluation's results. This is for sure a very relevant topic that should be added in a meta-evaluation exercise. Nevertheless, it demands an extra data collection effort from policy makers, since the evaluations themselves generally do not reveal the use that was made from their analysis, conclusions and recommendations.

Up to now, 134 Latin-American evaluations were coded and uploaded in SIPER repository, which also contains 565 evaluations from other countries.

Finally, the last phase comprehends the use of codified information in order to discuss state-of-art of STI evaluation practice in LA.

Findings and discussion

Preliminary results show that STI policy evaluation activity in LA is recent and heterogeneous across countries. Some countries, such as Chile, Argentina and Colombia, have a greater tradition in this field: they started their evaluation efforts before other countries and continue along of the whole period for which the collection occurred. The types of policies that have most been targeted for evaluation are direct funding, such as research grants, subsidies and credit. Direct financing through scholarships and indirect financing, in particular through tax incentives, are also highlighted (Figures 1 and 2).

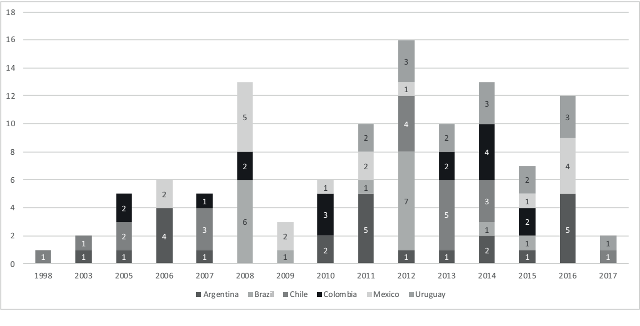


Figure 1: Evaluations by country

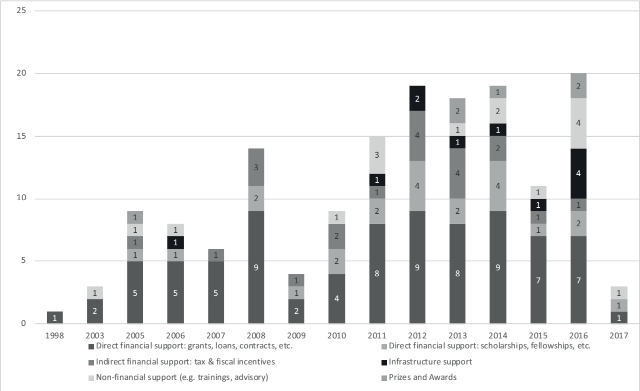


Figure 2: Typology of Policies

Having the national research and innovation systems from these countries as a background, it is possible to assume that the evaluations are in line with research and innovation activities that are being conducted in those nations, with the exception of policies fostering graduation studies through master and doctorate scholarship, that are very common in those countries but not so commonly evaluated.

Evaluations were mostly conducted externally by independent bodies (63%) or by a combination of external contracting by independent bodies and internal staff (27%). Execution by government external teams is minority in the sample. The same happens with non-Latin-American evaluations: 82% were conducted externally, by independent bodies or within government.

Regarding the timing, interim evaluations, that is, those carried out after a certain period of implementation of the policy or program, stand out (116 from 134 evaluations). Ex-post evaluations - those carried out after the end of the implementation period of the policy or program - amount to 12. Finally, 6 ex-ante evaluations were carried out before the implementation of the policy or program. This behavior is similar to non-Latin-American evaluations.

In general, the purpose of these evaluations is a combination of formative and summative (69 documents), followed by only summative (50) or formative (15) evaluations. With only 3 exceptions, the evaluations report the rationale of the evaluated policy or program, either completely or partially. For non-Latin-American evaluations, the ratio for those evaluations that combine formative and summative purposes is even bigger, performing almost 65% from the total, and there were also few cases (19 from 565) that do not report the rationale of the evaluated policy or program.

Quasi-experiments (especially those that use treatment and comparison groups) and non-experiments are the dominant designs, corresponding to 60 and 59 evaluations respectively. There are also combinations between quasi-experimental and non-experimental designs (14 evaluations) and a single evaluation that combines quasi-experiment and an experiment. There is a prominence for non-experimental evaluations in non-Latin-American countries (58% from the total) and also for a combination between non-experimental and experimental evaluations (25% from the total). There are 19 cases using experimental design, solely or in combination with other designs.

Regarding the aspects considered in the evaluation (Figure 3), the focus is on both the results and the impacts, here understood as effects in the long term. Following these aspects, there are the implications for future development of policies / strategies, achievement of objectives, adherence to policy/program (extent to which beneficiaries are attracted), appropriateness to the rationale of the policy, additionality and degree of satisfaction of stakeholders. Little emphasis was placed on mobility, career, networking, gender and minority issues. Differences between Latin-American and non-Latin-American evaluations can be perceived in Figure 3. Appropriateness of rationale, design and goals are aspects more emphasized in those non-Latin-American evaluations when compared to Latin-American ones.

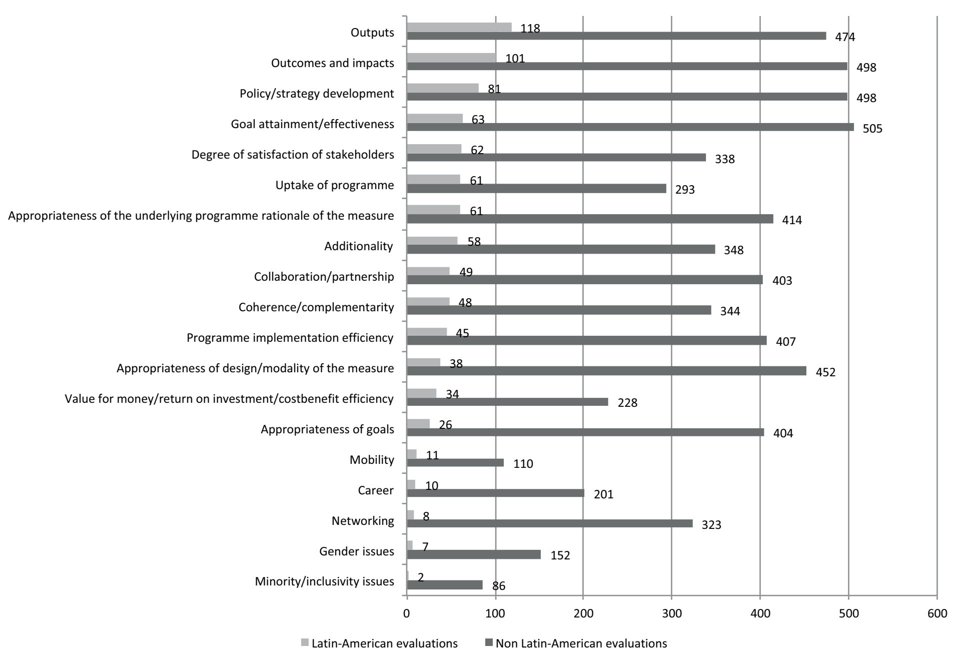


Figure 3: Aspects of the program examined by the evaluations

Of the 118 evaluations that measure policy outputs, only 41 have focused on identifying the quality of these results. Regarding those that measure the long-term impacts of policies (101 evaluations), there is an emphasis on the use of scientific and technological, social and economic indicators. On the other hand, education, competencies and environmental indicators were not highlighted. Of the 56 evaluations that measured economic impact, 28 analyzed cost-benefit or returns on investment. A set of 58 evaluations measures additionality, 29 of which analyze only output additionality, 9 input additionality and 1 behavioral additionality. There is also a combination of input and output additionality (11) and of the three types (5). Additionality is understood here as the difference between the policy/program in terms of investment (in the case of additionality of input), generation of results (in the case of additionality of output) and behavior of beneficiaries.

Regarding the data collection of the evaluations, it is worth noting the use of existing databases (116 evaluations); surveys (64); interviews (52); focal group/workshop/meetings (25); and analysis of scientific publications (19). For non-Latin-American countries, in addition to these tools, evaluations employ peer-reviews, formalized data on intellectual property and site visits.

Considering data collection methods, descriptive statistics (117 evaluations) and econometrics (60 evaluations) are strongly emphasized. Descriptive statistics is also emphasized in non-

Latin-American evaluations, followed by qualitative or quantitative analysis of texts; case-study analysis; input/output, cost/benefit, return on investment analysis, and finally econometric analysis.

Conclusion

The analyzes carried out so far indicate a growing movement towards the institutionalization of STI policy evaluation practices in Latin America, in line with the growing importance of these policies and the perception of their contribution to the countries' economic and social development. However, there are few variations on the methodological designs and indicators used, evidencing the need for substantive advances in this field. Complementary analyzes should be carried out after collection and characterization of all evaluations, seeking to identify the occurrence of a relationship between the analyzed variables, as well as the countries' profile regarding STI evaluation. Further comparisons among STI policy evaluation in LA and other regions and investigating the real use of evaluation results for policy decision-making are also promising developments of this research.

Acknowledgments

We gratefully acknowledge support from São Paulo Research Foundation - FAPESP (Award Number 2016/12258-0) and from National Council for Scientific and Technological Development - CNPq (Award Number 427696/2016-0). We would also like to thank Manchester Institute of Innovation Research (MIoIR) for the opportunity of collaboration.

References

Cahalan, M. & Goodwin, D. (2014). *Setting the Record Straight: Strong Positive Impacts Found from the National Evaluation of Upward Bound. Publication***.** Retrieved February 1, 2019 from: https://files.eric.ed.gov/fulltext/ED555877.pdf.

Cooksy, L. J. & Caracelli, V. J. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evaluation*, 26, 31-42.

Edler, J. et al. (2008). Improving policy understanding by means of secondary analyses of policy evaluation. *Research Evaluation*, 17 (3), 175-186.

Firme, T. P. & Letichevsky, A. C. (2010). O Desenvolvimento da Capacidade de Avaliação no Século XXI: enfrentando o desafio através da meta-avaliação. *Ensaio: Avaliação e Políticas Públicas em Educação,* 2 (5), 180-195.

Furtado, J. P. & Laperrière, H. (2012). Parâmetros e paradigmas em meta-avaliação: uma revisão exploratória e reflexiva. *Ciência & Saúde Coletiva*, 17 (3), 695-705.

Good, B. (2012). Assessing the effects of a collaborative research funding scheme: An approach combining meta-evaluation and evaluation synthesis. *Research Evaluation*, 21, 381-391.

Hanssen, C.E., Lawrenz, F. & Dunet, D. O. (2008). Concurrent Meta-Evaluation A Critique. *American Journal of Evaluation*, 29 (4), 572-582.

Jacob, S. & Affodegon, W. S. (2015). Conducting quality evaluations: four generations of meta-evaluation. *SpazioFilosofico*, 165-175.

Scriven, M. (2009). Meta-Evaluation Revisited. *Journal of MultiDisciplinary Evaluation*, 6 (11), 3-8.

Scriven, M. (2010). *Evaluation evaluations: a meta-evaluation checklist*. Retrieved may 20, 2016 from: http://michaelscriven.info/images/EVALUATING_EVALUATIONS_8.16.11.pdf.

Stufflebeam, D.L. & Coryn, C.L.S. (Ed.). (2014). *Evaluation theory, models, and applications*. San Francisco: Jossey-Bass.

Stufflebeam, D.L. (2001). The Metaevaluation Imperative. *American Journal of Evaluation*, 22 (2), 183-209.

Weiss, C.H. (1998). Have we learned anything new about the use of evaluation? *American Evaluation Association,* 19 (1), 21-33.

Impact of the journals, disciplines, and countries on the citation memory

Jinhuyk Yun[1], Sejung Ahn[2] and June Young Lee[3]

[1] *jinhyuk.yun@kisti.re.kr*
Korea Institute of Science and Technology Information, 66 Hoegiro, Dongdaemun-gu, Seoul, 02456 (Korea)

[2] *sjahn@kisti.re.kr*
Korea Institute of Science and Technology Information, 66 Hoegiro, Dongdaemun-gu, Seoul, 02456 (Korea)

[3] *road2you@kisti.re.kr*
Korea Institute of Science and Technology Information, 66 Hoegiro, Dongdaemun-gu, Seoul, 02456 (Korea)

Abstract

Understanding the evolution of science and technology is essential to comprehend the innovation dynamics of humankind. One crucial in the last finding was the shape of citation distribution, which is following a heavy-tailed distribution with a large disparity. Also, many studies suggested that there is the rich-get-richer effect on the citation that highly cited literature tend to get a new citation easier than other literature; in other words, there is a strong memory of the citation. To understand the mechanism behind the memory, we investigate 21 years of citation evolution through systematic analysis entire citation history of 42,423,644 scientific literature published from 1996 to 2016 contained in SCOPUS. We define a compensated citation measure that separating citation preference of individual article from the influence of the various external agents: i) journals, ii) author countries, and iii) disciplines. From the compensation, we found that the influence of the journal is overwhelming the other two factors. We also developed a generative agent-based model to account for such results, indicating the combination of Matthew effect of the citation and pre-established preference for the journals may result in current status. Our approach sheds light on the unbiased and quantitative understanding of scientific evolution.

Introduction

The progress of science and technology is commonly considered as standing on the shoulders of giants, which means new scientific discoveries are building upon previous researches. A citation is standardized mean to give credit for trail-blazing pioneers. Naturally, the number of citation became a common measure to assess the influence of the scientific outputs, e.g. papers, books and proceedings (Bornmann & Daniel, 2008). One notable finding in the late 20th century was the distribution of paper citation tends to follow a heavy-tailed and highly skewed distribution, such as power-law (Price, 1976; Redner, 1998) and log-normal distributions (Thelwall & Wilson, 2014; Thelwall, 2016). In other words, there are a few but a considerable number of scientific literature that receives an extremely large number of citations compared to ordinary scientific outputs. Researchers have identified substantial factors influencing citation distributions. As the illustrative example, the influence of disciplines (Albarrán *et al.*, 2011) and journals (Yun *et al.*, 2018) have been studied. However, the majority of studies mainly focused on a snapshot of accumulated citation count from the publications, yet hardly considered yearly evolution of the citation count, influenced by various factors.

In this research-in-progress, we have extended our previous study on the influence of the journal for the citation distribution to the influence of disciplines and author countries (Yun *et al*, 2018). For this purpose, we investigate the distributions of three normalized citation counts and compare the degree of memory effect for the normalization citation counts. In particular, we examined the complete 21 years of history for every scientific literature to assess their citation growth over time; thus, we mainly focus on the yearly acquired citation for the scientific literature.

Rescaled measure and impact of the journals, disciplines, and countries.

We begin with the data analysis of the entire SCOPUS CUSTOM XML DATA as of 22 August 2017, which covers from January 1996 to August 2017. This dump includes 42,423,644 academic pieces of literature with title, journal, disciplines of the journals (ASJC; All Science Journal Classification), author information (including author countries), and citation records in XML format. We use the records regardless of the document type to avoid possible bias due to the sampling. We consider the discipline of a paper with a classification scheme of Scimago Journal & Country Rank(SJR) consisting of 309 subject categories refined from the ASJC scheme (Gómez-Núñez *et al.*, 2011) because some of the subject categories in ASJC are barely used (Wang & Waltman, 2016).

In this study, we mainly focus on the impact of journals, disciplines, and author countries by extending of our own previous study on the impact of the journals for the citation distributions (Yun *et al*, 2018) that we proposed the rescaled measures of citation as follows:

$$C_y^*(a) = \frac{C_y(a)}{\sum_{a\in j(a,y_p)} C_y(a)/N\left[j\left(a,y_p\right)\right]}$$

where $C_y(a)$ is the citation count of article a in the cited year y, and $j(a,y_p)$ is the set of articles published in the same journal and published year (y_p) of the article a. Although this measure is primarily defined to compensate influence of the journals for the citation count, the measure is easily extendable for the various factors by altering the normalization group $j\left(a,y_p\right)$ for the normalization factor $\sum_{a\in j(a,y_p)} C_y(a)/N\left[j\left(a,y_p\right)\right]$. For example, if $j(a,y_p)$ is the set of articles published in the same and published year (y_p), and belonging to a specific discipline or author countries, the measure indicates the citation count compensated the influence of discipline or author countries. One should note that one paper can belong to more than one discipline or countries; thus, we take geometric mean for the normalization factors for such cases. In short, we generalize the rescaled measure as follows:

$$C_y^*(a) = \frac{C_y(a)}{\left[\prod_{k=1}^{n_k}\left[\sum_{a\in j_k(a,y_p)} C_y(a)/N\left[j_k\left(a,y_p\right)\right]\right]\right]^{\frac{1}{n_k}}}$$

where $C_y(a)$ is the citation count of article a in the cited year y, and $j_k(a,y_p)$ is the set of articles published in the same discipline (or country) k and published year (y_p) of the article a. A parameter n_k is number of disciplines (or countries) that an article a belonging. The rescaled citation presents the relative citation among the most similar scientific literatures in terms of the age and various factors, *i.e.* journals, disciplines, and author countries.

As the first step, we investigate the best fit model distributions for three normalized and one raw citation count with the Maximum Likelihood Ratio (MLR) methods (Clauset *et al.*, 2009) by testing six candidate distributions: i) power law (PL), ii) power law with an exponential cut-off (PLE), iii) lognormal (LN), iv) lognormal positive (LNP), v) exponential (EXP), stretched exponential (StEX). If the Maximum Likelihood Estimator for a certain distribution is superior to all other distributions, we consider the distribution is the most suitable model distribution for the citation counts based on its own meaning (Clauset *et al.*, 2009). What we observe is the mixture of three distributions for the raw citation with the slight dominance of log-normal for the raw citation, while all three rescaled measure shows the dominance of power-law with an exponential cut-off (Figure 1). Journal normalized citation shows a strong preference of power-law with an exponential cut-off, yet other two normalized measures also show the considerable amounts of log-normal distributions (Figure 1). In other words, discipline and country normalization do not show the universal distribution for entire publication and citation years, meanwhile journal normalization gives the more universal distribution of power law with an exponential cut-off.

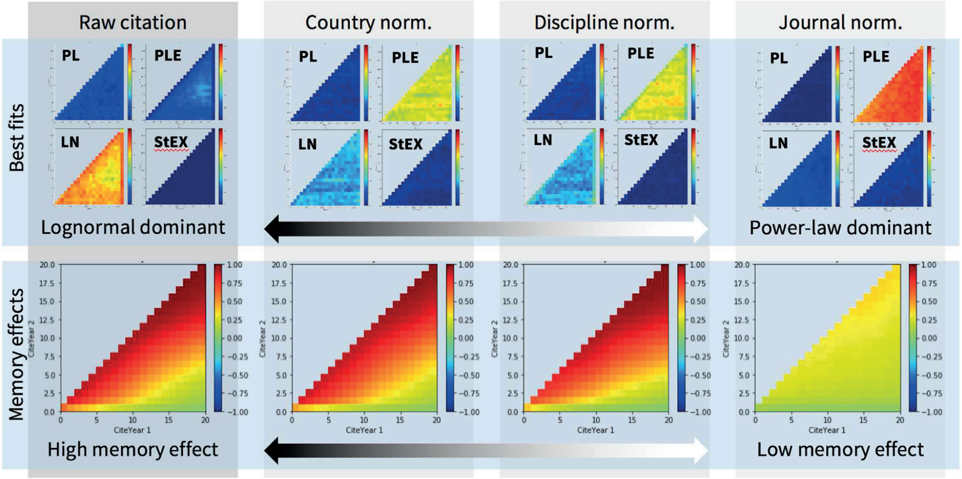
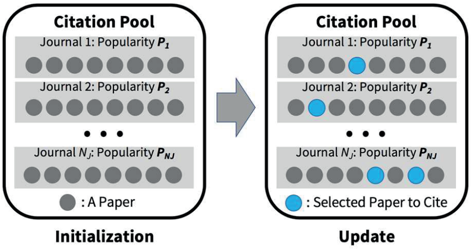


Figure 1. Best fit distributions count for the 309 SJR classification and memory effect of scientific literatures published in 1996 for the raw and normalized citation measures. Results of raw citation and journal normalized citation are taken from our own previous study (Yun *et al,* 2018).

In our own previous study (Yun *et al.*, 2018), we also claimed the existence of strong memory effect of citation, or *the rich-get-richer phenomena*, citation count influenced by the previous citation counts (Borner *et al.*, 2004; Wang 2014). Again, we utilize Pearson's correlation coefficient for different cited years, from lists of the citation count the papers published in the same year. Interestingly, the influence of early citation lasts more than a decade for the raw citation, discipline normalized citation, and author country normalized citation (Figure 1). Although discipline and country normalization also removes the memory effect slightly, the memory effect still exists after normalization. In contrast, journal normalized citation shows insignificant correlations across the cited years (Figure 1). In other words, the influence of the journal is overwhelming the other two factors for the memory effect.



Citation Pool
Journal 1: Popularity P_1
Journal 2: Popularity P_2
• • •
Journal N_J: Popularity P_{NJ}
● : A Paper
Initialization
Citation Pool
Journal 1: Popularity P_1
Journal 2: Popularity P_2
• • •
Journal N_J: Popularity P_{NJ}
● : Selected Paper to Cite
Update

Figure 2. Schematic diagram of the agent-based model.

Agent-based model of citation memory

To prove the dynamics behind the memory effect of the citation, we introduced a simple agent-based model that both popularities of the journals and previous citation history of the paper influence the preference of the new citation. The model begins with a pool of the *N* papers that can be cited in the simulation (Figure 2). There are also N_j journals that papers are belonging to. Journals have their own pre-established popularity P_j, which is randomly drawn from the log-normal distribution of mean μ and standard deviation σ. In our model, we considered the author's citation is driven by the popularity of the paper. Many factors may affect the choice for the citation, but we assumed three: i) popularity inherited by the journal, ii) popularity due to the number of the previous citation of the paper, and iii) decaying of the influence of the previous citation due to the aging. For every step, a new paper cites k articles in the citation pool with a probability $Prob(i) \propto P(i) \times C^*(i;t)$ (see Figure 2). Here, $P(i)$ is the popularity of the journal paper i belonging to. $C^*(i;t)$ indicates the sum of the citation of paper i with decaying due to the aging as follows:

$$C^*(i;t) = 1 + \sum\nolimits_{c \in C(i)} \frac{\exp\left(-\frac{a(c;t)}{T}\right)}{k}$$

where, $C(i)$ is the set of citations that paper i received since the onset of the system, $a(c;t)$ is the age of the citation c at time t, T is temporal decaying rate for the citation history, and k is the number of reference for a new article. For the simplification, we fix the parameters $P(i)$, k, and T as the constant. We also perform the similar journal normalization for the model as we considering a year in the model as the *N* steps of the simulation, which is same as the number of the papers in the citation pool.

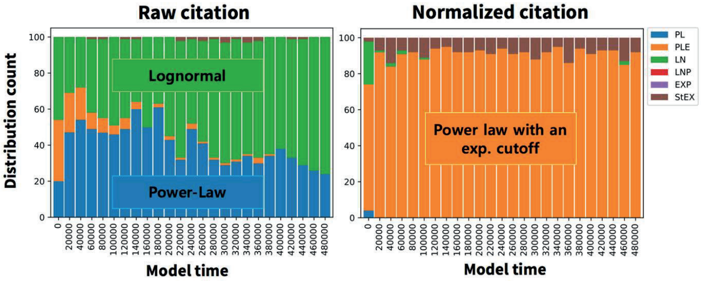


Figure 3. The best fit distributions of the agent-based model. The simulation is performed under following parameters: number of papers *N* = 20000, number of journals N_j = 500, number of ensembles *e* = 100; popularity distribution parameters mean $\boldsymbol{\mu = 0.0}$, standard deviation $\boldsymbol{\sigma = 1.6}$, decay parameter $\boldsymbol{T = 3 \times 10^4}$. We consider model year is simulation time of 20000, which is same as the number of papers in the system

The results of the agent-based model are consistent with the data analysis. The best fit distribution for the raw citation is lognormal with a considerable amount of the power-law (Figure 3; compare with Figure 1). As the model time goes by, the model shows a stronger log-normal prevalence. In contrast to the raw citation, most (journal) normalized citation is observed as the power law with an exponential cut-off (Figure 3; compare with Figure 1). Although our model disagrees with the second popular distribution of the raw citation (power-law with an exponential cut-off for the empirical data and the power-law for the model data), the model successfully reproduces the change of best fit distributions for the journal normalization.

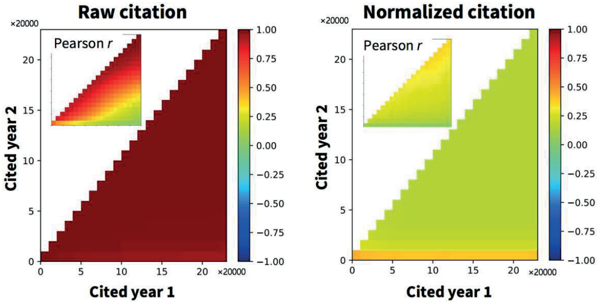


Figure 4. Memory effect of the agent-based model (large panels) and empirical data (insets). For both cases, we perform journal normalization. The simulation is performed under following parameters: number of papers $N = 20000$, number of journals $N_j = 500$, number of ensembles $e = 100$; popularity distribution parameters mean $\mu = 0.0$, standard deviation $\sigma = 1.6$, decay parameter $T = 3 \times 10^4$. We consider model year is simulation time of 20000, which is same as the number of papers in the system

In addition to the best fit distributions for the raw and normalized citation count, our model also reproduces the trend of memory effect for the yearly acquired citation reported in Figure 1. As shown in Figure 4, the interrelationship between the number of citation between two different years produces a similar result. For both model and data, we observed a significant correlation between across the citation years, which is varnished by the journal normalization. Although the model failed to imitate the decay of correlation with more than 10 years of time differences, it successfully demonstrates the result of the normalized citation (compare Figure 1 and 4).

Discussion

In this research-in-progress, we explored the influence of the journals, disciplines, and countries on the citation through a massive history of metadata in SCOPUS over the past two decades. We show a strong memory effect can be compensated by the journal normalization, yet discipline and country normalization do not give much impact on the memory effect. We suggest that in-depth analysis of other factors that influence the number of citations, e.g. impact of authors and institutes may be promised to enhance the impact of our approach, yet left for the further research due to the difficulty of the disambiguation. We hope to refine our agent-based model with more factors can improve the quality of the model, especially reproducing the decaying of memory more than 10 years of time differences. Going one step forward, if data-driven analysis accompanied by the proper model, it provides the evidence to understand the evolution of the knowledge formation, as we hope.

Acknowledgments

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government through Grant No. NRF-2017R1E1A1A03070975 (J.Y.; S.A.) and the Korea Institute of Science and Technology Information. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, 88(2), 385-397.

Börner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences, 101(suppl 1), 5266-5273.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of documentation*, 64(1), 45-80.

Clauset A., Shalizi C.R., Newman M.E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4), 661–703.

Gómez-Núñez, A. J., Vargas-Quesada, B., de Moya-Anegón, F., & Glänzel, W. (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, 89(3), 741.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5), 292-306.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2), 131-134.

Redner, S. (2005). Citation statistics from 110 years of physical review. Physics today, 58(6), 49-54.

Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, 8(4), 824-839.

Thelwall, M. (2016). Citation count distributions for large monodisciplinary journals. *Journal of Informetrics*, 10(3), 863-874.

Wang, J. (2014). Unpacking the Matthew effect in citations. *Journal of Informetrics*, 8(2), 329-339.

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, 10(2), 347-364.

Yun, J., Ahn, S., & Lee, J. Y. (2018). On the Heterogeneous Distributions in Paper Citations. *arXiv preprint* arXiv:1810.08809.

Exploring Barriers to Interdisciplinary Research

Daniele Rotolo[1] and Michael M. Hopkins[2]

[1] *d.rotolo@sussex.ac.uk*
SPRU – Science Policy Research Unit, Business School, University of Sussex, Brighton (United Kingdom)

[2] *m.m.hopkins@sussex.ac.uk*
SPRU – Science Policy Research Unit, Business School, University of Sussex, Brighton (United Kingdom)

Abstract

This paper aims to increase our understanding of which barriers may hinder researchers from undertaking interdisciplinary research (IDR). IDR is research that builds on a set of theories, concepts, tools, data, and methods that are not available within a single discipline or specialty. IDR is a common policy priority because it is expected to generate knowledge to address complex societal problems such as climate change, sustainability, population ageing, etc. In order to understand the perceived barriers to undertaking IDR, we surveyed three groups of stakeholders in the UK Higher Education (HE) system: (i) researchers, (ii) managers in HE institutions, and (iii) managers in research funding organisations. We first defined three samples of potential respondents: a sample of 16,625 researchers based in UK HE institutions from 2013 to 2015 (this sample was stratified by regions at NUTS-1 level); a sample of 1,080 managers in 15 universities representing a cross section of UK HE institutions; and a sample of 539 managers in research funding organisations. The survey response rates were about 13%, 33%, 17% in the case of researchers (i.e. 2,183 responses), managers in HE institutions (i.e. 367 responses), and managers in research funding organisations (i.e. 94 responses), respectively. The survey provides evidence of discrepancies of the perceptions of barriers to IDR around four main areas (i) collaboration, (ii) career, (iii) evaluation, and (iv) funding across stakeholder groups and research areas. This has policy and managerial implications for the ease with which these can be addressed.

Introduction and related literature

Interdisciplinary research (IDR) has attracted widespread attention from researchers, policy makers, and funders (e.g. Horton, 2017; National Academy of Sciences, 2005; Nurse, 2015). Strong expectations exist about the role of IDR in providing solutions to address the most complex challenges in climate change, sustainability, and global health, which single disciplines may not be able to address. For example, there is increasing consensus among policy makers that addressing major issues such as antimicrobial resistance requires a global and interdisciplinary response (Syed, Ducrotoy, & Bachmann, 2016).

This has resulted in growing pressures on research organizations to steer their 'monodisciplinary' research trajectories towards interdisciplinary approaches (Spelt, Biemans, Tobi, Luning, & Mulder, 2009). A large variety of actors is involved in this process: from funders – focussed on the design of policy instruments to support an efficient allocation of funding as well as to create a 'protected space' for researchers to undertake IDR – to research managers/administrators in Higher Education (HE) institutions – facing resource distribution and research evaluation challenges – and the research teams actively involved in IDR. Actions and behaviours of these groups towards IDR are likely to be shaped by the perception that these groups have of which factors may incentivise or hinder researchers to undertake IDR efforts. In this regard, our understanding remains non-systematic, thus posing considerable challenges on devising organisational and policy mechanisms capable of addressing barriers to IDR at different levels.

The paper attempts to address the following question: how do barriers to IDR are differently by different groups of stakeholders involved in the research process? To address this research question, we build on a large-scale UK-based survey analysis involving researchers working in different disciplines, managers in HE institutions, and managers in funding organizations.

Academic and policy research on IDR has been extensive (e.g. Frodeman, Klein, & Mitcham, 2010; Wagner et al., 2011). This includes conceptual and theoretical efforts on defining IDR and building typologies of IDR (e.g. Barry, Born, & Weszkalnys, 2008; Huutoniemi, Klein, Bruun, & Hukkinen, 2010; Lattuca, 2001; Miller, 1982), and studies focused on the development of methodological approaches and indicators to evaluate IDR. Pioneering quantitative approaches have extensively built on the analysis of text mining and citation patterns in bibliometric data (e.g. Bromham, Dinnage, & Hua, 2016; Fox, 1999; Leydesdorff & Rafols, 2011; Porter & Rafols, 2009; van Raan & van Leeuwen, 2002), while more recent efforts have started to capitalise on the increasing access to textual sources and advancements of text-mining techniques. Researchers have also argued that statistics are necessary, but not sufficient to capture the complexity of research process (Klein, 2008). For this reason, qualitative measures of IDR have been also proposed. These include measures of IDR collaboration based on participants' self-assessments (Hall et al., 2008; Mâsse et al., 2008).

Evidence has been provided that IDR and research impact are positively related. For example, the relationship between IDR, as measured on the basis of indicators derived from the list of cited references in publications, and publication citation count has been found to be positive. According to Uzzi et al. (2013), publications with 'unusual combinations' of prior works (which can be conceived as a proxy of IDR) are twice as likely to be highly cited than publications with 'conventional combinations' of references. Chen et al. (2015) found that top 1% most cited papers are those with the highest levels of IDR (as assessed on the basis of the Simpson Index of the disciplinary classification of the journals of the cited referenced). Yegros-Yegros et al. (2015) found also a positive relationship between IDR (as assessed on the basis of the disciplinary diversity of publication references, i.e. variety, balance and disparity of the WoS categories of cited references) and citation count. Yet, this relationship seems to be curvilinear (inverted U-shape). Similar findings were also reported by Wang et al. (2015).

Evidence of the positive relationship between IDR and research impact have been also found in the impact cases submitted for the 2014 UK Research Excellence Framework (REF). The text-mining analysis of the full-text of such cases revealed that two-thirds of REF Impact Cases could be classified in two or more disciplinary areas, thus they could be considered to be at least multidisciplinary (Kings College London & Digital Science, 2015).

Although these conceptual and methodological research efforts have remarkably increased our understanding of interdisciplinarity, considerable challenges remain in evaluating IDR (Klein, 2006; McLeish & Strang, 2016).

We aim to contribute to this literature by providing a more systematic understanding of which barriers hinder IDR according to the perspective of different actors involved in the research process. It is worth noting that there is a lack of consensus around the definition of IDR (Lau & Pasquini, 2004). A variety of terms have been associated with the concept of IDR: (i) interdisciplinarity as including research approaches that "[…] integrate separate disciplinary data, methods, tools, concepts, and theories in order to create a holistic view or common understanding of a complex issue, question, or problem" (Wagner et al., 2011); (ii) multidisciplinarity as when "Theory, methods, and interpretive standards of the different disciplines are employed. Interpretation of the results from different disciplines typically occurs post hoc, often from the perspective of one discipline that may emerge as dominant within the project" (Rossini & Porter, 1979); (iii) trans- disciplinarity as "Trans-sector, problem-oriented research involving a wider range of stakeholders in society" (Klein, 2008); and (iv) cross-disciplinarity, often used to describe the previous three research modalities.

The complexity of defining what is a 'discipline' from the conceptual and empirical points of views and how individual researchers may differently perceive the boundary of their discipline are at the core of this debate (e.g. Huutoniemi et al., 2010). This paper does not aim to enter into the philosophical and epistemological debate on IDR. For the purpose of our empirical analysis, we define IDR *as research activities that cross the boundary of a single discipline or specialty*.

Data and Methods

We conducted a survey analysis targeted to three main groups of stakeholders in the HE system: (i) researchers, (ii) managers in HE institutions, and (iii) managers in funding organizations. Our empirical analysis focuses on the HE system in the United Kingdom.

We designed a 20-minute survey on the basis of a literature review and input received from three workshops with stakeholders. These workshops were facilitated by the Higher Education Funding Council for England (HEFCE) and involved representatives of researchers (21 participants), managers in HE institutions (15 participants), and managers in funding organizations (9 participants) – for more details about the organisation of the workshops, please refer to Davé et al. (2016). The survey was structured along four groups of barriers – i.e. career, funding, collaboration, and outcomes – and was opened to respondents from March to May 2016.[1] For comparative purposes, most of the survey questions were shared across the three groups of stakeholders. However, the survey was also adapted to each group to explore barriers to IDR that could not be explore in all the groups.

To identify our target populations of researchers, managers in HE institutions, and managers in funding organizations, we developed three different empirical strategies. These strategies are described below and summarized in Table 1.

In the case of researchers, our target population is represented by all research-active individuals employed in UK HE institutions. We assumed a researcher in a UK HE institution to be research-active if the researcher was corresponding author of at least one publication from 2013 to 2015, i.e. three years before the survey opened. We used publication data from Web of Science to define a frame population that is as close as possible to the target population (de Leeuw, Hox, & Dillman, 2008; Heeringa, West, & Berglund, 2010). More precisely, we identified all publication records from 2013 to 2015 including at least one author based in the UK as reported in authors' affiliation addresses. This led to an initial sample of 566,957 UK publications of which 219,182 publications (39%) had corresponding authors' email addresses ending with the ".ac.uk" extension. From this sample, we extracted 109,698 distinct email addresses of UK corresponding authors. These email addresses were then matched with a list 164 UK HE institutions (e.g. "sussex.ac.uk" was matched with "University of Sussex"). The NUTS-1 level code of each HE institution was then identified and used to draw a sample of 16,625 email addresses/researchers stratified by region.

In the case of managers in HE institutions, our target population included all individuals responsible for managing research at UK HE institutions. We define our frame population as all individuals responsible for managing research as reported on UK HE institutions' websites. These included roles such as Pro-Vice Chancellor, Head of Department, Head of Faculty, Head of School, Head of Doctoral Studies of Department, Head of Programme, Head of a Unit of Assessment (REF2014), etc. We selected a sample 15 UK HE institutions to include institutions

[1] The survey also explored incentives for stakeholders to undertake IDR. The results of this analysis are not presented in this paper.

from all UK nations, small and large HEIs, and specialized and non-specialized institutions.[2] The websites of these organisations were searched to obtain contact details for research managers from across the whole of these targeted institutions. The list of identified managers and associated email addresses was sent senior managers in the sampled institutions to ensure accuracy. This approach enabled us to build a sample of 1,080 managers.

Table 1. Sampling strategy.

Survey	**Researchers in HE institutions**	**Managers in HE institutions**	**Managers in funding organisations**
Target population	Research-active individuals employed UK HE institutions	Individuals responsible for managing research at UK HE institutions	Staff at UK-based funders involved in the allocation of funding resources to HE institutions
Frame population	Corresponding authors of publications from 2013 to 2015 employed in UK HE institutions: (i) 566,957 UK publications in WoS (ii) 109,698 email addresses with the "ac.uk" extension (iii) Matching between email extensions and 164 UK HE institutions	Individuals responsible for managing research as reported on UK HE institutions' websites: (i) HE institutions' websites to compile a contact list (ii) Senior managers' inspection of contact list	Staff at UK-based funders involved in the allocation of funding resources to HE institutions (i) RCUK funding bodies' website: AHRC, BBSRC, EPSRC, MRC, NERC, HEFCE, HEFCW, SFC, DEL (ii) Funders that issued at least three calls (+£1,000) as listed in Research Professional in the 365 days to May 10 2016
N	105,839 (96% of UK HE institutions)	Unknown	118 (RCUK funding bodies) 421 (Other funders)
Sample population	Region stratification (NUTS-1)	15 HE institutions (representative of region, size, specialisation)	All identified
N	16,625	1,080	118 (RCUK) 421 (Other funders)
Responses	2,183	367	27 (RCUK) 67 (Other funders)
Response rate	13.1%	32.6%	22.9% (RCUK) 15.9% (Other funders) *[Average 17.4%]*

Source: Authors' elaboration.

[2] We defined a HE institution to be "large" if more than 600 Full-Time-Equivalent (FTE) researchers were submitted to the 2014 UK REF 2014, "small" otherwise. We considered specialised HE institutions those focused on a specific research area (e.g. Colleges of Arts, Veterinary Colleges).

In the case of managers in funding organizations, our target population includes staff at UK-based funders involved in the allocation of funding resources to HE institutions. To identify our frame population, we first collected the email addresses of managers in RCUK funding bodies (i.e. AHRC, BBSRC, EPSRC, MRC, NERC, HEFCE, HEFCW, SFC, DEL) as reported on the website of these organisations. This led to a first sample of 118 managers in RCUK funding bodies. We then queried the ‘Research Professional’ platform to identify all calls for funding with budgets of more than £1,000 in the 365 days to 10th May 2016. We extracted the contact details reported in these calls, thus identifying an additional sample of 421 managers in other UK-based funding organisations (e.g. charities, foundations).

The survey response rates were about 13%, 33%, 23%, and 16% in the case of researchers (i.e. 2,183 responses), managers in HE institutions (i.e. 367 responses), managers in RCUK funding bodies (i.e. 27 responses), and managers in other funding organizations (i.e. 67 responses), respectively. We perform a survey analysis including, when population data were available, base weights, non-response weights, and post stratification weights to correct for discipline, career stage, and gender bias that may affect the sample of respondents (de Leeuw et al., 2008). The weighting approach is summarized in see Table 2.

Table 2. Response weighting strategy.

Survey	Researchers in HE institutions	Managers in HE institutions	Managers in funding organisations
Base weight $(w_{B_{h,i}})$	Region (NUTS-1)	Nation	-
Non-response weight$(w_{NR_{c,i}})$	Region (NUTS-1)	HE institution	-
Post-stratification $(w_{PS_{l,i}})$	Contract (role) Gender	- (no population data available)	- (no population data available)
Total weight (w_i)	$w_i = w_{B_{h,i}} \times w_{NR_{c,i}} \times w_{PS_{l,i}}$ $w_{B_{h,i}} = \frac{N_h}{n_h}$, inverse probability of selection in stratum *h* (region) $w_{NR_{c,i}} = \frac{\sum_{i=1}^{n_{rc}} w_{B_{h,i}}}{\sum_{i=1}^{n_c} w_{B_{h,i}}}$, inverse probability of responding in cell *c* (region) $w_{PS_{l,i}} = \frac{N_l}{\sum_{i=1}^{n_{rl}} w_{B_{h,i}} \times w_{NR_{c,i}}}$, inverse probability of responding in post-stratum *l* (contract-region) N_h population size of stratum *h* (region) n_h sample size of stratum *h* (region) n_c sample size of cell c (region) n_{rc} response of cell c (region) N_l population size of post-stratum *l* (contract-gender) n_{rl} response of post-stratum *l* (contract-gender)		

Source: Authors’ elaboration.

Results

Preliminary analyses revealed the presence of alignment and misalignment of how barriers to IDR are differently perceived by stakeholder. These analyses are summarized in Table 3.

First, in terms of career barriers to IDR, there is agreement across the three groups of stakeholders that IDR requires a combination of strong disciplinary training followed by training in other areas. There is agreement, but less strong support, on the need of additional training to undertake IDR. Together, these findings indicate a potential barrier to entry greater for IDR researchers compared to mono-disciplinary researchers.

Managers in HE institutions and funding organisations are also more in agreement that researchers' peers see IDR as less rigorous – although researchers indicate less concern on this point. Managers in funding organisations more often perceive that promotion and tenure policies adversely affect those engaged in IDR, than researchers or managers.

Second, in terms of collaboration barriers to IDR, respondents from all three groups agreed that collaboration involving IDR was more challenging due to communication difficulties in IDR teams. There was a more neutral response to the suggestion that IDR required co-location of researchers. Although managers in HE institutions were more likely to favour this proposal than managers in funding organisations. Managers also agreed that IDR was complex to support and needed more resources, while funders agreed on this to a lesser extent (and not significantly less). Researchers were not asked these questions.

Third, in terms of barriers, a considerable proportion of respondents across stakeholder groups perceived that IDR is less likely to be funded than monodisciplinary research: 48% of respondents in the case of researchers, 44% in the case of managers in HE institutions, and 41% in the case of managers in funding organisations. The strength of this perception is strongest in researchers and managers in HE institutions, and significantly less so (but still in agreement) in the case of funders. There was considerable support for the suggestion that this is caused by monodisciplinary perspective of the reviewers. This view was significantly more supported by researchers and their managers than in the case of funders.

There was a significant difference in perception between researchers and managers in HE institutions perceiving the IDR to be seen as more 'risky' during peer review, compared with funders who agreed less with this view. The three stakeholder groups were more neutral on the suggestion that difficulties in producing IDR proposals or perceptions of IDR being of lower quality were to blame for lower funding success.

Finally, in terms of barriers to IDR related to research outcomes, stakeholders agreed that it was more difficult to publish the outcomes of IDR in leading journals, with HE managers significantly more likely to support this view than researchers or funders. Researchers and HE managers also agree significantly more than funders that IDR outputs take longer to produce. There is also agreement across groups (although slightly less in the case of funders) that research evaluation processes undervalue IDR.

Discussion and conclusions

The results of the survey provide evidence that researchers, research managers in HE institutions, and research funders in the UK all recognise a number of barriers to IDR in the UK, including career entry barriers around training, lower funding success for IDR, and IDR being undervalued in research evaluation processes. However, even where there was agreement, funders are sometimes less convinced than researchers and HE managers of the existence of particular barriers. This suggests a need for more close alignment between these groups in order to best address the challenges around IDR.

Table 3. Perception of barriers to IDR across groups of stakeholders [Likert scale from 1 (Strongly disagree/Not at all influential) to 5 (Strongly agree/Extremely influential)].

		Researchers in HE institutions			**Managers in HE institutions**			**Managers in funding organisations**			Rank test (Kruskal-Wallis)
	Barrier	Mean (weighted)	Std. Dev. (weighted)	Obs.	Mean (weighted)	Std. Dev. (weighted)	Obs.	Mean (weighted)	Std. Dev. (weighted)	Obs.	
Career	• A strong disciplinary training is required to undertake IDR	3.81	0.99	2023	3.97	0.91	358	3.93	0.86	85	*
	• Additional training is required to undertake IDR	3.55	1.09	2036	3.47	0.97	357	3.36	0.94	84	
	• Peers in researchers' core discipline(s) consider IDR less rigorous	3.17	1.23	1987	3.50	1.09	354	3.38	0.98	72	***
	• Promotion/tenure policies discourage IDR	3.17	1.31	1959	2.90	1.13	359	3.32	1.14	59	***
Collaboration	• Communication in IDR teams is a challenge	3.56	1.03	2037	3.65	0.98	365	3.4	0.98	85	
	• IDR requires co-location of researchers	3.11	1.08	1986	3.21	1.05	354	2.76	0.95	82	***
	• Supporting IDR is complex	-	-	-	3.99	0.86	362	3.53	1.06	90	
	• IDR is more likely to involve non-academic partners	2.96	1.03	1967	3.16	0.91	359	3.01	0.9	86	***
	• Finding appropriate partners for IDR is a challenge	3.40	1.09	2037	3.56	0.94	361	3.66	0.84	88	**
	• IDR requires more institutional resources	-	-	-	3.63	1.02	358	3.22	1.04	72	
Funding	• IDR is less likely to be funded because of the disciplinary focus of funding	4.18	0.86	816	3.95	0.96	135	3.72	1.02	32	**
	• Producing strong IDR research proposals is a challenge	2.79	1.23	813	2.94	1.18	133	3.15	1.15	33	
	• IDR may be considered of a lower quality	3.21	1.17	791	3.3	1.10	135	2.87	1.12	31	
	• Reviewers' monodisciplinary view makes IDR less likely to be funded	4.44	0.76	821	4.52	0.66	134	3.91	0.93	32	***
	• IDR may be considered more risky	3.48	1.07	804	3.58	1.01	135	3.06	1.03	33	*
Outcome	• IDR takes more time to produce outcomes	3.77	0.96	2009	3.65	0.98	352	3.17	0.88	81	***
	• Research evaluation undervalues IDR	3.71	1.09	1745	3.58	1.09	341	3.48	1.13	60	
	• IDR is less likely to be published in top-tier disciplinary journals	3.46	1.17	1973	3.71	1.1	353	3.49	1.04	73	***

Note: *** p <0.001, ** p<0.01, * p<0.05.
Source: Authors' elaboration.

Acknowledgments

This paper builds on data collected for a report by Technopolis and SPRU (University of Sussex), for the Higher Education Funding Council for England (HEFCE) in 2016. This report is titled "Landscape Review of Interdisciplinary Research in the UK" and it is available at http://webarchive.nationalarchives.gov.uk/20180322111256/http://www.hefce.ac.uk/pubs/rereports/year/2016/interdis/). We thank Web of Science (Clarivate Analytics) for providing the raw data of the publication records involving UK authors. The authors are grateful for the time generously given by respondents who completed the survey, as well as to the staff at HEIs that helped to identify participants. We are also grateful to colleagues working on the report for assistance in building the dataset and helping to improve the questionnaire. Finally, we are grateful to the project advisory board members. The views reported here are those of the authors, and do not necessarily reflect the views of those who commissioned or advised on this research project.

References

Barry, A., Born, G., & Weszkalnys, G. (2008). Logics of interdisciplinarity. *Economy and Society*, *37*(1), 20–49. https://doi.org/10.1080/03085140701760841

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature*. https://doi.org/10.1038/nature18315

Chen, S., Arsenault, C., & Larivière, V. (2015). Are top-cited papers more interdisciplinary? *Journal of Informetrics*, *9*(4), 1034–1046. https://doi.org/10.1016/j.joi.2015.09.003

Davé, A., Hopkins, M. M., Hutton, J., Krčál, A., Kolarz, P., Martin, B., … Stirling, A. (2016). *Interdisciplinary research: Landscape and environment*. Technopolis (Brighton, UK) and SPRU-University of Sussex (Brighton, UK) for HEFCE.

de Leeuw, E., Hox, J., & Dillman, D. (2008). *International Handbook of Survey Methodology*. *International Handbook of Survey Methodology* (Vol. 21). London, UK: Routledge. https://doi.org/10.4324/9780203843123

Fox, M. F. (1999). Gender, Hierarchy, and Science. In J. S. Chafetz (Ed.), *Handbook of the Sociology of Gender* (pp. 441–457). New York, NY: Kluwer.

Frodeman, R., Klein, J. T., & Mitcham, C. (2010). *The Oxford Handbook of Interdisciplinarity*. *2010*. https://doi.org/10.1093/oxfordhb/9780198733522.001.0001

Hall, K. L., Stokols, D., Moser, R. P., Taylor, B. K., Thornquist, M. D., Nebeling, L. C., … Jeffery, R. W. (2008). The collaboration readiness of transdisciplinary research teams and centers. Findings from the National Cancer Institute's TREC Year-One evaluation ttudy. *American Journal of Preventive Medicine*, *35*(2 SUPPL.), 161–172. https://doi.org/10.1016/j.amepre.2008.03.035

Heeringa, S. G., West, B. T., & Berglund, P. a. (2010). *Applied Survey Data Analysis*. Boca Raton, FL, United States: Taylor & Francis Group. https://doi.org/10.1201/9781420080674

Horton, R. (2017). Offline: The unspoken dangers facing UK medical science. *The Lancet*, *390*(10113), 2616. https://doi.org/https://doi.org/10.1016/S0140-6736(17)33299-3

Huutoniemi, K., Klein, J. T., Bruun, H., & Hukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy*, *39*(1), 79–88. https://doi.org/10.1016/j.respol.2009.09.011

Kings College London, & Digital Science. (2015). *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies King's*. London, UK.

Klein, J. T. (2006). Afterword: The emergent literature on interdisciplinary and transdisciplinary research evaluation. *Research Evaluation*. https://doi.org/10.3152/147154406781776011

Klein, J. T. (2008). Evaluation of interdisciplinary and transdisciplinary research. A literature review. *American Journal of Preventive Medicine*, *35*(2 SUPPL.), S116–S123. https://doi.org/10.1016/j.amepre.2008.05.010

Lattuca, L. R. (2001). *Creating Interdisciplinarity: Interdisciplinary Research and Teaching among College and University Faculty*. Nashville: Vanderbilt University Press.

Lau, L., & Pasquini, M. W. (2004). Meeting grounds: perceiving and defining interdisciplinarity across the arts, social sciences and sciences. *Interdisciplinary Science Reviews*, *29*(1), 49–64. https://doi.org/10.1179/030801804225012437

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87–100. https://doi.org/10.1016/j.joi.2010.09.002

Mâsse, L. C., Moser, R. P., Stokols, D., Taylor, B. K., Marcus, S. E., Morgan, G. D., … Trochim, W. M. (2008). Measuring Collaboration and Transdisciplinary Integration in Team Science. *American Journal of Preventive Medicine*. https://doi.org/10.1016/j.amepre.2008.05.020

McLeish, T., & Strang, V. (2016). Evaluating interdisciplinary research: The elephant in the peer-reviewers' room. *Palgrave Communications*, *2*(8), 16055 EP. https://doi.org/10.1057/palcomms.2016.55

Miller, R. C. (1982). Varieties of interdisciplinary approaches in the social sciences: A 1981 overview. *Issues in Integrative Studies*, *1*, 1–37.

National Academy of Sciences. (2005). *Facilitating Interdisciplinary Research*. *Public Policy*. WASHINGTON, D.C.: National Academies Press. https://doi.org/10.1007/s00103-011-1362-6

Nurse, P. (2015). *Ensuring a successful UK research endeavour - A Review of the UK Research Councils*. London, UK.

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745. https://doi.org/10.1007/s11192-008-2197-2

Rossini, F. A., & Porter, A. L. (1979). Frameworks for integrating interdisciplinary research. *Research Policy*, *8*(1), 70–79. https://doi.org/10.1016/0048-7333(79)90030-1

Spelt, E. J. H., Biemans, H. J. A., Tobi, H., Luning, P. A., & Mulder, M. (2009). Teaching and learning in interdisciplinary higher education: A systematic review. *Educational Psychology Review*, *21*(4), 365–378. https://doi.org/10.1007/s10648-009-9113-z

Syed, S. N., Ducrotoy, M. J., & Bachmann, T. T. (2016). Antimicrobial resistance diagnostics: time to call in the young? *The Lancet Infectious Diseases*, *16*(5), 519–521. https://doi.org/https://doi.org/10.1016/S1473-3099(16)30011-1

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474

van Raan, A. F. J., & van Leeuwen, T. . (2002). Assessment of the scientific basis of interdisciplinary, applied research: Application of bibliometric methods in Nutrition and Food Research. *Research Policy*, *31*(4), 611–632. https://doi.org/https://doi.org/10.1016/S0048-7333(01)00129-9

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., … Borner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. Journal of Informetrics. *Journal of Informetrics*, *165*(1), 14–26.

Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE*, *10*(5). https://doi.org/10.1371/journal.pone.0127298

Yegros-Yegros, A., Rafols, I., & D'Este, P. (2015). Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PLoS ONE*, *10*(8), e0135095. https://doi.org/10.1371/journal.pone.0135095

Have you read this? An empirical comparison of the British REF peer review and the Italian VQR bibliometric algorithm

Daniele Checchi[1], Alberto Ciolfi[1], Gianni De Fraja[2], Irene Mazzotta[1], Stefano Verzillo[3]

[1] *daniele.checchi@anvur.it; alberto.ciolfi@anvur.it; irene.mazzotta@anvur.it*
ANVUR, Via Ippolito Nievo, 35 00153 Rome (Italy)

[2] *Gianni.Defraja@nottingham.ac.uk*
University of Nottingham, School of Economics University Park, NG7 2RD, UK; Università di Roma "Tor Vergata", DEF, Via Columbia 2, I-00133 Rome (Italy)

[3] *stefano.verzillo@ec.europa.eu*
European Commission, JRC and University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan (Italy)

Abstract

This paper determines the ranking of the publications units of assessment which were submitted to the UK research evaluation carried out in 2014, the REF, which would have been obtained if their submission had been evaluated with the bibliometric algorithm used by the Italian evaluation agency, ANVUR, for its evaluation of the research of Italian universities.

Introduction

The week before Christmas 2014, university offices were abuzz with discussions and dissections of the freshly published results of 2014 "Research Excellence Framework" (REF), the official evaluation of research conducted by UK academic institutions in the period 2008-13. Since the initial dummy run held in 1986, the raison d'être of this peer review based exercise is twofold: to ensure accountability for the taxpayer's investment in academic research and persuading the public of its benefits; to form the basis for the allocation of the annual "block" budget for research, around one quarter of all the funds to higher education institutions (HEIs).

Following the 2008 exercise, the funding agency run a pilot study with a view to replace the expensive peer review with a bibliometric-based assessment, but concluded that "bibliometrics are not sufficiently robust … to replace expert review in the REF" (HEFCE, 2009) and so the 2014 exercise continued to rely on peer evaluation, with an estimated overall cost of approximately £246m (Farla and Simmonds, 2015), comparable to the annual budget of a medium size university, and dividing up at £4000 per academic assessed. The exercise planned for 2021 will also be conducted via peer review, even when several other countries do adopt a bibliometric evaluation, as highlighted in Wang, Vuolanto, and Muhonen (2014)'s survey. The question of the closeness between a peer review and a bibliometric approach is addressed by Bertocchi, Gambardella, Jappelli, Nappi, and Peracchi (2015) in a report on the working method of the economics and management assessment panel in the Italian eValuation of the Quality of the Research (VQR) 2004-10 assessment, which randomly selected some of the journal articles assigned to bibliometric evaluation also to be peer reviewed, precisely to assess to correspondence between the two methods (see also Baccini and De Nicolao (2016) and the reply, Bertocchi, Gambardella, Jappelli, Nappi, and Peracchi (2016)). Mryglod, Kenna, Holovatch, and Berche (2015) assess the correlation between the score and the rank obtained by each institution with the corresponding "departmental h-index" (Hirsch, 2010). The latter paper examines a broader range of research areas than Bertocchi, Gambardella, Jappelli, Nappi, and Peracchi (2015), and reports good correlations in the various subject areas, between 0.36 and 0.89. However, it uses a different set of articles from those evaluated by the REF panels, and indeed, as we explain below, it includes articles written by academics who were not submitted as part of the group evaluated by the relevant REF panel. In the same vein, Harzing (2017) has shown that ranking UK departments according to the "departmental h-index" correlates to the REF power ranking at 0.97.

In detail, we assess the papers which were submitted to the REF, and are included in the Scopus database, using the bibliometric criteria which ANVUR, the Italian evaluation agency, used to assess the outputs published from 2011 to 2014 submitted for the VQR.

We stress at the outset some limitations of the exercise (illustrated in Table 3), which makes its contribution more a template for more analysis than policy advice: there are specific differences between the two evaluations, and we did not adjust the algorithm to account for these. Even then, we find a remarkable correspondence between the methods: in the 18 REF research areas where at least 75% of the outputs submitted to the REF could be evaluated bibliometrically, the average correlation between REF peer review score and the corresponding measure calculated with the ANVUR algorithm is 0.81, and the average rank correlation is 0.76: for the full sample, the figures are 0.63 and 0.6. Correlation is very much higher for other measures of departmental research quality, which consider the size of the unit as well as its average quality: of particular interest to policy makers is the correlation in the funding that would be attributed by the two methods, which stands at 0.995 when the departments with at least 75% of the outputs could be evaluated bibliometrically, and at 0.986 for the whole sample. Moreover, we show that, had the annual funding to institutions been allocated following the ANVUR assessment methods, the outcome would have differed relatively little. The summary result of the correlation in the institutional funding is most striking: if the output submitted had been evaluated with the bibliometric algorithm used in the VQR, with peer review assessing the rest of the institutional submission, the correlation between the actual funding assigned to each institution and the funding it would have received if calculated with the VQR score would have exceeded 0.9997, and hence the difference in funding would have been minuscule.

We close the paper with a simple attempt to uncover association between the closeness of the measure and other institutional variables, finding very little systematic variation.

The 2014 Research Excellence Framework (REF2014)

The REF2014 evaluated the research conducted by 52,000 academic researchers associated to 1911 units of assessment in 154 HEIs in UK. The assessment was carried out by over 1000 experts in 36 panels, one in each area of research, in turn grouped into four "main panels". The full list is in Table 5 below. It may therefore be useful to fix the terminology: we denote as "subject areas" the 350 subject categories in Scopus (finest classification of topics); we will then denote as "VQR research areas" and "REF research areas" the groups of subject areas which were assessed by the 16 VQR individual panels (known as GEV) and the 36 REF panels. In the formal analysis we index with h the subject areas and with i the research areas.

Panels assessed three main dimensions of an institution's activity. (i) individual research outputs consisting, for each member of staff submitted, of four outputs published in the reference period 2008-2013; (ii) the research environment, as described in words by each institution; (iii) the impact of research on the wider society, in terms of knowledge transfer and/or public engagement, as evidenced in case-study reports, numbering one per every eight researchers.

The panels determined the percentage for each of the three dimensions of the activities of each submission to be assigned to the five quality categories, ranging from the best, 4-stars "quality that is world-leading in terms of originality, significance and rigour" to the worst, 0-stars "quality that falls below the standard of nationally recognised work". On Thursday 18 December 2014, the panels' assessments was made public, together with the aggregate profile, obtained as a weighted average of the outputs, environment, and impact components, with the weights 0.65, 0.15, 0.2.

The unit of assessment (UoA) is the group of researchers submitted to a given panel: there was no requirement that all the academics submitted to the unit should be all part of an institutional group, such as a department, a school or an institute. There were also members of one department being submitted as part of a different UoAs. To lighten the exposition we refer as department or unit, the group of academics which an institution submitted for assessment to a specific UoA, but it must be kept in mind that, for example, health economists, behavioural economists, econometricians, political economists, development economists, all working in their economics department were submitted to the "Public Health", "Psychology", "Mathematical Sciences", "Politics and International Studies", "Anthropology and Development Studies" panels, respectively. And indeed, many institutions submitted the entire department of economics to the "Business and Management Studies" panel. The decisions regarding submissions were taken usually at institutional level with the attempt to improve the overall result. With no obligation either to submit all departments for evaluation or to submit all the academic members of each department submitted, HEIs took different approaches to the decision whether or not to submit a researcher at all, some leaving out weaker researchers, other including every academic on payroll. These considerations suggest a loose correspondence between units of assessment and departments which moreover is unlikely to be orthogonal to the quality of the research output and casts doubts on the possibility of extending to all disciplines the approach of drawing on departmental information to map the outcome of the REF taken by Mryglod, Kenna, Holovatch, and Berche (2015) and Harzing (2017).

Outputs can be submitted by an institution as long as the author is employed by that institution on the REF census date, 31st October 2013, irrespectively of where the author was when the paper was written or published.

The environment component is a written submission describing the achievements of the academic department, together with data on research grant income and PhD completions. Impact is assessed by considering written 'case studies', one for every eight academics submitted, accompanied by supporting evidence which shows how the research of the department has brought benefits outside of academia. Unlike output, impact is attributed to the institution where it was carried out, irrespective of which institution is currently employing the researcher responsible for it at the census date. The measures of environment and impact have no correspondence in the VQR and cannot be the object of a bibliometric approach, so we limit our comparison to the output component of the REF.

Table 1. Summary statistics and cross correlations of REF performance by component[1].

	GPA Score	*GPA Outputs*	*GPA Environment*	*GPA Impact*	*Mean*	*St. Dev*
GPA Score	1				2.82	0.433
GPA Outputs	0.93***	1			2.76	0.369
GPA Environment	0.883***	0.71***	1		2.88	0.751
GPA Impact	0.826***	0.578***	0.726***	1	2.98	0.689

Unlike its Italian counterpart, the UK funding agency does not present a single score. Commentators and the public have therefore stepped in, variously aggregating the profiles into single numbers so as to draw ranking of UoA and institutions in national league tables. The most commonly used are the grade point average, GPA, and the research power, RP

[1] Sample size = 1828 departments submitted to REF 2014. For explanation of the components, see main text. *** denotes significance at 1% level.

(Forster, 2015). GPA is calculated as a weighted average of the scores, with the proportion in each category as weight: the GPA of department i's in institution k is calculated simply as:

$$GPA_{ik}^{REF} = \sum_{s=0}^{4} \pi_{ik}^{s} s$$

where π_{ik}^{s} is the proportion of the activity of department i's in institution k which was assessed to be of s star quality. Table 1 shows that the correlation between the three components is high, but not so much as to make it meaningless to assess the three components separately.

The RP is simply the product of the GPA by the number of staff submitted:

$$RP_{ik}^{REF} = n_{ik} \times \sum_{s=0}^{4} \pi_{ik}^{s} s$$

where n_{ik} denotes the number of full–time equivalent researchers submitted by institution k to panel i. Thus, GPA measures the average quality, without reference to the size of the UoA, which is instead taken into account by the RP measure: excluding a relatively weak member of staff would definitely increase the GPA and reduce RP.

While less prominent in the media, the government does determine a funding score formula, FS, which is used to calculate how to allocate the overall "quality related" funding made available to the sector in each year. Universities are free to spend this funding as they wish, with no link to projects or even disciplines.

When designing the funding formula, the government intended to provide incentives towards high quality research, and so it gave high weight to 4* output, specifically four times higher than the weight given to 3* output, and *no weight* to output judged less than 3*. With the above notation, an institution's funding in year t until the following evaluation exercise is given by

$$FS_{ikt}^{REF} = \Phi_t \times \Gamma_i \times (4\pi_{ik}^{4} + \pi_{ik}^{3}) \times n_{ik}$$

where Φ_t is the coefficient (in the jargon the "QR unit funding"), which varies from year to year, and depends on the overall public funding for universities, and Γ_i is a research area specific weight which takes value 1.6 for STEM subjects, UoAs 1-15, 1.3 for intermediate cost research areas such as geography, architecture, sport sciences, design, music, UoAs 16, 17, 26, 34, and 35, and 1 for all other research areas.

Table 2. Correlation between possible measures of performance[2].

	GPA Score	*Research Power (RP)*	*Funding Score (FS)*	*Mean*	*St. Dev*
GPA Score	1			2.82	0.433
Research Power (RP)	0.377***	1		79.62	93.11
Funding Score (FS)	0.508***	0.978***	1	38.197	50.964

Table 2 shows the correlation between these measures, indicating that the size based ones, RP and FS, are fairly close but both rather different from the GPA; the correlation between the number of academics submitted, n, and the GPA score is 0.433, indicating that the low correlation between GPA and RP may be due to institutions pursuing different strategies, some preferring selecting only their best performers, others pursuing the funding associated with larger submissions.

Our main aim is to determine degree of similarity between the REF peer review and the Italian bibliometric measurement. To do so, we calculate the quality scores of the output

[2] Sample size = 1828 departments submitted to REF 2014. For explanation of the measures see main text. *** denotes significance at 1% level.

component of the research activity of the UK institutions that would have resulted if the REF assessment of the outputs had been carried out using the algorithm that was used to assess the quality of the research of Italian institutions. We stress that we do not attempt to perform a comparison between Italian and British institutions. Given the many differences between the set of rules used in the two assessment methods, as illustrated in Table 3, this seems unlikely.

Table 3. Differences between the VQR (Italy) and the REF (UK)[3].

	REF	*VQR*
All researchers submitted	NO	YES
Portability of output	YES	YES
Weight of output in assessment	65%	80%
Period of evaluation	2008-13	2011-14
Census date	31 October 2013	30 November 2014
Number of outputs per person	4	2
Expert panel	YES	YES
Peer Review	YES	depending on VQR subject area
Bibliometric indicators	available: use at the discretion of the panel	must be used for STEM research areas
Peer review by	panel members or other panels	panel members and external reviewers
Overall funding to research area	depending on evaluation	pre-determined *
Funding attributed to	institutions only	both institutions and departments**
Entity assessed	department/unit	individual output

Differences between the results of the two assessment methods could spring from two sources. One the one hand there could be structural differences between the methods, which would be the case if a substantial fraction of the highly cited papers published in prestigious journals were, rightly or wrongly, considered to be of poor quality by the peer reviewers, or vice versa, if peer review assessed as being of top quality many papers published in obscure journals and with low citation counts. On the other hand, there might be systematic difference in the submission strategy of different institutions: for example large institutions may be able to devote more resources to assess internally the quality of each output submitted, while smaller ones having to rely on bibliometric algorithm to select the papers and the academic to submit for evaluation. Of course, a similarity between the VQR bibliometric and the REF peer review assessment could emerge if they did *in general* yield different results, but in the specific case of the 2014 REF, these various factors cancelled each other out. Thus, the nature of our paper can only be suggestive, even though, compared to some of the existing literature, it covers the whole of the research carried out in the UK.

The VQR bibliometric algorithm.

The VQR algorithm identifies a paper by four parameters: (i) the year of publication, $t = 1,. . ., T$, (ii) the subject area, indexed by h, (iii) the number of citations at the census date, and (iv) the journal where it was published. The last two parameters are both turned into a number in [0, 1] by normalising their position in an appropriate distribution. The algorithm computes the

[3] Summary comparison between the VQR and the REF, see text for more details. Information obtained from www.ref.ac.uk/2014 (REF) and http: //www.anvur.it/attivita/vqr/vqr-2011-2014/ (VQR).

* The amount allocated to all the submissions in a given VQR research area is independent of the evaluations given by the VQR panel to the institutions in that research area.

** The round of annual funding is allocated to institutions, but a subsequent law awarded a numbers of posts directly to departments, partly on the basis of their VQR score.

distribution of the citations obtained by all the articles published in research area h in year t; let this be denoted by $\Phi^{C}_{ht}(n) \in [0,1]$. That is, $\Phi^{C}_{ht}(n) \in [0,1]$ is the proportion of papers published in research area h in year t that have obtained n citations or less. Similarly for journals, where the relevant measure is the journal impact metric: $\Phi^{J}_{ht}(n) \in [0,1]$ is the proportion of journals included in the Scopus database as pertaining to research area h that, in year t, had impact metric at most x.

In order to do so, it is therefore necessary to know the world distribution of citations and impact metrics at the earliest available date after the REF census date. We purchased from Scopus bibliometric information (namely the number of citations and the SCImago Journal Rank) on 1/1/2015, for each of the papers submitted to the REF; given the suggestive nature of the exercise, we opted to use data made available by ANVUR, which included these distributions on 1/1/2017. This might generate a measurement error, which however is systematic only to the extent that there are different trends in the citations patterns and the impact metrics of the journals where certain institutions are more inclined to publish.

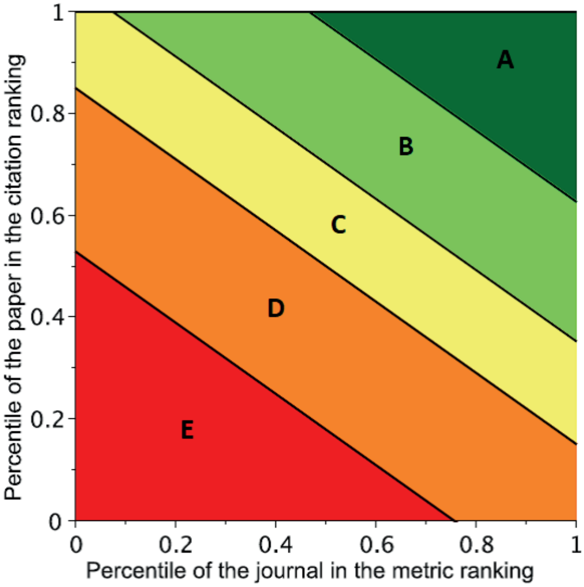


Figure 1. Allocations of products to quality classes.

In the next step of the procedure[4], the unit square $[0,1]^2 \subseteq \mathbb{R}^2$ is divided into five subsets as shown in Figure 1 by four parallel downward sloping straight lines, in such a way that the dark green (A) area[5] is 0.1, the light green (B) and yellow (C) areas are both 0.2, the orange (D) area is 0.3, and the red (E) area is 0.2. Simple computations determine the boundary lines; these are given by $y = a_{it} - b_{it}x$, where a_{it} is the solution in a, for σ = 0.1, 0.3, 0.5, 0.8, of:

$$1 - \max\left\{0, \frac{a-1}{b_{it}}\right\} - \int_{\min\left\{1, \frac{a}{b_{it}}\right\}}^{\max\left\{0, \frac{a-1}{b_{it}}\right\}} (a - b_{it}x)\, dx = \sigma.$$

[4] The procedure is described in detail in Anfossi, Ciolfi, Costa, Parisi, and Benedetto (2016).

[5] The normalisation with the percentiles ensures that the distribution is uniform in the unit square.

The solution is given by

$$a_{it}(\sigma, b_{it}) = \begin{cases} 1 + \frac{b_{it}}{2} - \sigma & \text{if } \sigma \leq \frac{b_{it}}{2} \\ 1 - \sqrt{2\sigma b_{it}} + b_{it}(1 - x) & \text{if } \frac{b_{it}}{2} < \sigma \leq 1 - \frac{b_{it}}{2} \\ \sqrt{2b_{it}(1 - \sigma)} & \text{if } \sigma > 1 - \frac{b_{it}}{2} \end{cases}$$

In the previous equation b_{it} is the slope used to assess outputs in the VQR research area i in year t: it is chosen subjectively by each panel, to reflect the trade-off between visibility of an article and prestige of the publishing journal, and the manner in which it changes with time.
In order to account for the different citation patterns and the fact that more recent papers have less opportunity to collect citations, the slopes separating the areas in Figure 1 increased in absolute value with the year of publication so as to reduce the importance of citation for younger articles.

Table 4. Slopes of trade-offs between citations and impact factor[6].

	VQR				*REF*		
Research Areas	2011	2012	2013	2014	2008-11	2012	2013
Computer Science	1	1.25	1.5	1.75	1	1.25	1.5
Mathematics	depending on subarea				1.1	1.4	1.7
Physics	.4	.6	.9	1.5	.4	.6	.9
Chemistry	.4	.6	.8	1.2	.4	.6	.8
Earth Sciences	.4	.6	.9	1.5	.4	.6	.9
Biology	.4	.6	.8	1.2	.4	.6	.8
Medicine	.4	.6	.8	1.2	.4	.6	.8
Agricul. and Vet. Sciences	.7	.9	1.5	2	.7	.9	1.5
Architecture	.6	.9	1.5	2	.7	.9	1.5
Civil Engineering	.7	.9	1.5	2	.7	.9	1.5
Ind. and Inform. Engineering	.4	.6	.9	1.5	.4	.6	.9
Psychology	.4	.6	1	1.5	.4	.6	1

Table 4 reports the slopes that were used in the VQR, and those that we have used to obtain the score for each of the articles we have assessed. The overlap between the REF and the VQR is such that we could use the VQR slopes only for the years 2011-2013. For the other years, we deliberately chose to reduce our degrees of freedom by setting the slopes outside the overlap period to be the same as at its beginning.[7] The first four columns report the coefficients used in the VQR, the last three those we have used to compute the scores of papers submitted to the REF. The score assigned to an article published in a journal included in subject area h in year t depends on the number of citations that it received relative to the world distribution of citation for articles published in subject area h in year t, and on the impact metric of the journal where it was published, again relative to the distribution of the impact metrics of journals in subject area h in year t. In detail, consider an article which was in percentile p^C of the world distribution of citation for articles published in subject area h in year t, published in a journal whose impact metric placed it in percentile p^J of the corresponding world distribution of journals' impact metrics. Then, this article's score is

[6] The slopes of the lines in Figure 1, for different VQR research areas and different years.

[7] The VQR measured citations accumulated up to 2015 of articles published in the 2011-14 period, while the REF looked at 2015 citations of articles published in the 2008-13 period. As a consequence, the REF articles had longer to be cited, and this is why we disregard the slopes used by the Italian VQR in the final year. Moreover, Italian researchers chose the panel to which they submitted their paper, without knowing in advance the slopes which the panel would adopt; in the case of REF, we have relied on the subject area of the publishing journal, which had a correspondence into the Italian Panels reconstructed by Scopus.

given by

$$s_{VQR} = \begin{cases} 1 & \text{if} \quad p^C \geqslant a_{it}(0.1, b_{it}) - b_{it}P^J \\ 0.7 & \text{if} \quad a_{it}(0.1, b_{it}) - b_{it}P^J > p^C \geqslant a_{it}(0.3, b_{it}) - b_{it}P^J \\ 0.4 & \text{if} \quad a_{it}(0.3, b_{it}) - b_{it}P^J > p^C \geqslant a_{it}(0.5, b_{it}) - b_{it}P^J \\ 0.1 & \text{if} \quad a_{it}(0.5, b_{it}) - b_{it}P^J > p^C \geqslant a_{it}(0.8, b_{it}) - b_{it}P^J \\ 0 & \text{if} \quad p^C < a_{it}(0.8, b_{it}) - b_{it}P^J \end{cases},$$

where, in each row, the dependence of a_{it} on σ and b_{it} derived in previous equation is made explicit.

An article is considered as "excellent" (score 1) if it corresponds to the best 10% in the world joint distribution of citations and journal metric; it is assessed as "good" (score 0.7), if it falls within 10% and 30%; it is considered "fair" (score 0.4), if it falls within 30% and 50% and as "acceptable" (score 0.1), if it falls within 50% and 80% of the world distribution. The remaining papers are labelled as "limited", and receive a score of 0.

Approximately 70% of the outputs submitted to REF are published in journals which the VQR had allocated to one or more VQR research areas. We allocated the remaining ones to close VQR research areas by exploiting information on the frequency of publications in journals of a given Scopus subject areas by the academics submitted to a VQR research area. The entire allocation procedure was such that around 46% of the outputs submitted to the REF and contained in Scopus was published in journals which are associated to multiple VQR research areas. Depending on where they fall in the version of Figure 1 of each VQR research area, a given output could have different values of these scores: when this happened, we assessed the given output in all the selected VQR research areas, and then chose the highest evaluation score.[8] After each output was assigned to the corresponding class, the score could be aggregated by averaging or adding up all the scores for each article submitted by members of each unit assessed (department, faculty or university). The corresponding score for each institution *i* evaluated according to the VQR algorithm is given by:

$$GPA_{ik}^{VQR} = 4\pi_{ik}^{1} + 3\pi_{ik}^{0.7} + 2\pi_{ik}^{0.4} + \pi_{ik}^{0.1}$$

where π_{ik}^{s} is the proportion of the articles of institution *i* published in research area k to which the algorithm assigned a score $s_{VQR} = s, s = 1, 0.7, 0.4, 0.1$. Note of course that $\sum_s \pi_{ik}^{s} \leq 11$, but it can be strictly less than 1, as some output may score zero. In previous equation, we calculate the GPA with the weight vector (4, 3, 2, 1, 0) used in the REF, rather that the VQR weight vector, which was (1, 0.7, 0.4, 0.1, 0). The overall correlation between the measures, at 0.998, is very high.

The data

The outputs submitted to the REF can be downloaded as Excel files from www.ref.ac.uk/2014. The total number of outputs assessed is 190962, with 81.09% of the total (154854) journal articles, the remainder consists mainly of chapters in books (7.5%) and books (5.4%).

[8] This is equivalent to assume that the institutions knew in advance the assessment criteria of the potential panels, and would submit each paper to the unit of assessment giving that paper the highest evaluation: again, we have no reason to think that papers with different areas would be systematically concentrated in certain institutions.

Table 5. Summary statistics of the paper submitted to REF 2014.

No.	Unit of Assessment	number of institutions	product assessed with VQR algorithm	% of total REF submissions.	% assessed by REF as			
					4*	3*	2*	1*
	Main Panel A	121	48356	94.44	37	44	17	1
1	Clinical Medicine	31	13400	97.34	39	44	15	1
2	Public Health	32	4881	93.26	39	41	17	3
3	Allied Health Professions	82	10358	93.33	31	50	17	1
4	Psychology	81	9126	97.04	38	40	19	2
5	Biological Sciences	44	8608	98.18	37	46	15	1
6	Agriculture and Veterinary Science	29	3919	96.61	35	41	20	3
	Main Panel B	105	44830	89.11	26	57	15	2
7	Environmental Sciences	44	5184	96.53	24	59	15	2
8	Chemistry	37	4698	98.47	28	63	9	0
9	Physics	41	6446	97.91	28	60	11	1
10	Mathematics	53	6994	90.65	29	55	15	1
11	Computer Science	89	7651	67.39	26	44	24	5
12	Chemical and Manuf. Engineering	22	4143	95.73	25	57	17	1
13	Electrical Engineering	32	4025	96.77	25	62	11	2
14	Civil Engineering	14	1384	92.41	24	56	16	3
15	General Engineering	62	8679	95.09	26	56	16	2
	Main Panel C	124	36432	67.61	27	42	26	4
16	Architecture	43	3781	66.81	29	40	25	6
17	Geography and Archaeology	58	6017	76.32	27	42	26	5
18	Economics and Econometrics	28	2600	86.88	30	48	19	2
19	Business and Management Studies	98	12202	89.08	26	43	26	4
20	Law	65	5522	30.21	27	46	23	4
21	Politics and International Studies	55	4365	60.34	28	40	26	6
22	Social Work and Social Policy	62	4784	64.61	27	42	25	5
23	Sociology	29	2630	64.9	27	45	26	2
24	Anthropology and Develop. Studies	21	2013	57.68	27	42	26	4
25	Education	75	5519	65.43	30	36	26	7
26	Sport Sciences, Leisure and Tourism	50	2757	83.9	25	41	27	6
	Main Panel D	138	9850	25.55	30	41	24	4
27	Area Studies	22	1724	40.55	28	42	25	5
28	Modern Languages and Linguistics	47	4932	27.58	30	42	23	4
29	English Language and Literature	86	6923	19.2	33	41	22	4
30	History	81	6431	31.27	31	44	23	2
31	Classics	22	1386	12.77	34	42	22	2
32	Philosophy	39	2173	46.71	31	42	24	3
33	Theology and Religious Studies	31	1558	20.54	28	40	27	5
34	Art and Design	71	6321	15.57	26	42	25	6
35	Music, Drama and Dance	72	4246	16.77	29	39	24	6
36	Media Studies	69	3517	35.34	29	38	24	8
	Total	154	139468	64.20	19	45	29	5

For each output, the file contains the type of output, the institution that submitted it, the unit of assessment it was submitted to, as well as the DOI, the publication year, the number of co-authors, the title the place of publication and so on. The outputs are distributed evenly in the six years covered by the REF, with the exception of 230 outputs, which have 2007 as publication date.

Scopus returned the required data for 139847 journal articles, the remaining submissions having being published in outlets not covered by Scopus (books, editorials, notes, etc.). In addition, a handful of other products could not be evaluated, for various reasons (301 were of a type not considered by the VQR algorithm, 61 were allocated in the REF published data to an anonymised UoA, and 17 had missing data). The final tally of outputs we assessed was thus 139468. Table 5 presents summary statistics of the output data, showing the research areas where the typical publication outlet are refereed journals.

Results

Our main results are reported in Table 6. Column (1) reports the correlation between the individual GPA scores calculated for the outputs of the various institutions which submitted to the corresponding UoA using the VQR algorithm, and the scores awarded to these units by the REF expert panel. Column (2) reports the rank correlation between these sets of scores. These two sets of correlations are themselves highly correlated (0.973). Both the correlations between values and the rank correlations are positive, and many are very high. GPA scores are averages, and so are independent of the number of academics submitted. Columns (3) and (4) reports the correlations in RP, while columns (5) and (6) reports the correlations in the FS measure, the funding attributed to each unit submitted. In column (5), in the majority REF research areas this correlation exceeds 0.99 with the lowest value at 0.913, for “Music Drama and Dance”. This is extremely high, considering that for this area we could assess less than 17% of the outputs. The weighted average across REF research areas (with weights the output submitted to the REF) is 0.989. The very high values of the correlations even for REF subject areas where relatively few outputs where in Scopus journals can be explained with a correlation between the quality of the outputs submitted to journals and the quality of the books and other forms of outputs in these REF research areas.

Table 6. Correlation in the measures and the rankings[9].

	(1) *Corr* *GPA*	*(2)* *Spearman* *GPA*	*(3)* *Corr* *RP*	*(4)* *Spearman* *RP*	*(5)* *Corr* *FS*	*(6)* *Spearman* *FS*
Chemistry (8)	0.857	0.788	0.987	0.975	0.995	0.993
Biology (5)	0.884	0.747	0.989	0.972	0.998	0.993
Physics (9)	0.896	0.828	0.992	0.977	0.998	0.993
Medicine (1)	0.753	0.811	0.988	0.994	0.999	0.997
Psychology (4)	0.847	0.875	0.984	0.963	0.998	0.99
Elect. Engineering (13)	0.825	0.808	0.976	0.956	0.993	0.988
Agriculture (6)	0.777	0.691	0.977	0.975	0.996	0.993
Environment (7)	0.794	0.763	0.98	0.983	0.996	0.991
Chem. Engineering (12)	0.690	0.613	0.972	0.943	0.991	0.985
General Engineering (15)	0.785	0.78	0.965	0.952	0.994	0.989
Health Professions (3)	0.82	0.800	0.979	0.969	0.996	0.991
Public Health (2)	0.909	0.761	0.994	0.947	0.999	0.995
Civil Engineering (14)	0.832	0.846	0.93	0.951	0.991	0.991
Mathematics (10)	0.779	0.68	0.987	0.965	0.998	0.993
Management (19)	0.818	0.852	0.985	0.969	0.996	0.996
Economics (18)	0.899	0.880	0.987	0.917	0.996	0.973
Sport Sciences (26)	0.522	0.467	0.899	0.807	0.985	0.963
Geography (17)	0.834	0.777	0.954	0.954	0.994	0.988
Computing (11)	0.758	0.665	0.933	0.909	0.989	0.979
Architecture (16)	0.624	0.600	0.95	0.859	0.993	0.982
Education (25)	0.565	0.575	0.966	0.819	0.996	0.981
Sociology (23)	0.542	0.46	0.904	0.933	0.983	0.988
Social Work (22)	0.649	0.638	0.907	0.837	0.987	0.980
Politics (21)	0.666	0.646	0.957	0.907	0.994	0.982
Anthr. & Development (24)	0.308	0.381	0.844	0.836	0.982	0.990

[9] Comparison between the score and the rank obtained using the VQR algorithm and the actual REF score. The horizontal line divides between UoAs where the fraction of products assessed is above 75% and UoAs where the same fraction was below. The number in brackets after the UoA’s name is the UoA’s number. Pairwise correlations between each pair are respectively: 0.973***, 0.778*** and 0.903***.

Philosophy (32)	0.557	0.521	0.978	0.944	0.988	0.978
Area Studies (27)	0.357	0.299	0.89	0.782	0.974	0.928
Media Studies (36)	0.443	0.495	0.813	0.788	0.962	0.952
History (30)	0.623	0.623	0.967	0.914	0.995	0.984
Law (20)	0.612	0.598	0.896	0.861	0.987	0.976
Modern Languages (28)	0.001	0.066	0.812	0.715	0.964	0.945
Theology (33)	0.400	0.369	0.742	0.686	0.967	0.939
English (29)	0.289	0.234	0.868	0.800	0.967	0.958
Music (35)	0.136	0.142	0.586	0.487	0.913	0.874
Arts (34)	0.211	0.308	0.836	0.664	0.96	0.902
Classics (31)	0.345	0.336	0.899	0.684	0.979	0.852

The results for the rank correlation are less extreme, due to the fact that many scores are very tightly bunched, and so small measurement errors change little in the absolute scores, but may have large impact in the ranking. Given that the aim of the UK exercise is to assess research, not rank institutions, this is the less relevant of the two correlation measures.

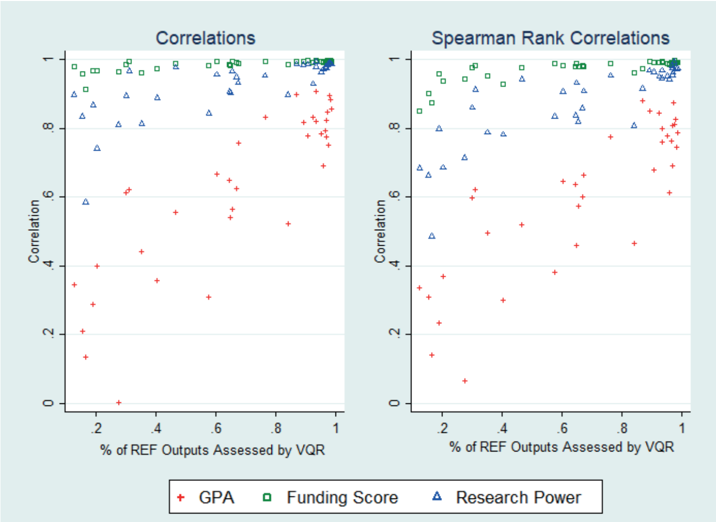


Figure 2. Correlations between performance scores.

Figure 2 illustrates the correlations and the rank correlations in the various units of assessment according to the various measures we have considered. The high correlation in institutional funding is a consequence of the high correlation between the scores, illustrated by the square dots.

Concluding remarks

We have performed an exercise to compare the outcome of the assessment of the research of the 2014 REF with the outcome that would have resulted had the publications which were included in submissions been evaluated, when possible, using the VQR bibliometric algorithm used in the Italian corresponding exercise. While we are keenly aware of the rough and approximate nature of our analysis, we find the closeness of the outcome, especially when comparing size sensitive measures, strongly suggestive that the method could be used to assess the publications at least for the research areas where the main outlet are refereed journals.

These results point to the similarity between peer review and bibliometric criteria (citations and impact factors) when the main outlets are articles published in journals contained in large bibliometric databases. An easy interpretation could be that British peer reviewers were influenced by citation data, in a sort of "informed peer review" (and this was not prohibited by REF rules). A more complex interpretation points to the existence of unobservable quality (creativity, methodological rigour, knowledge of the literature) which can be measured by direct reading (as the reviewers were expected to do) or can be sensed by other scholars, reporting their appreciation by citing the paper. In both cases, the scores of the articles would be correlated, leading to outcomes similar to what we have obtained.

The nature of the research output might be affected by the manner in which it is measured, in a coarse macroscopic version of the Heisenberg Uncertainty Principle. A statement that only journal articles will be considered worthwhile output for assessment would obviously direct academics to try to publish mainly in these outlets. This effect could be particularly strong for early career researchers, many of whose outputs were submitted in the form of working papers, and who might decide or be persuaded to submit their work to less prestigious journals, rather than risk being unable to submit outputs which the rules deem of lower quality.

Acknowledgments

We would like to thank ANVUR for financially supporting this research project. All opinions expressed in the paper are personal and do not involve the Agency. The information and views set out in this paper are those of the authors and do not reflect the official opinion of the European Union, where one of the author is employed. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

References

Anfossi, A., Ciolfi A., Costa F., Parisi G., and Benedetto S. (2016). Large-scale assessment of research outputs through a weighted combination of bibliometric indicators. *Scientometrics*, 107(2), 671–683.

Baccini, A. & De Nicolao G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, 108(3), 1651–1671.

Bertocchi, G., Gambardella A., Jappelli T., Nappi C.A. & Peracchi F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy*, 44(2), 451–466.

Bertocchi, G., Gambardella A., Jappelli T., Nappi C.A. & Peracchi F. (2016): "Comment to: Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise," *Scientometrics*, 108, 349–353.

De Fraja, G., Facchini G. & Gathergood J. (2016). How Much Is That Star in the Window? Professorial Salaries and Research Performance in UK Universities. Discussion Paper 11638, CEPR Discussion Paper.

Farla, k. & Simmonds P. (2015). REF Accountability Review: Costs, benefits and burden. Discussion paper, Technopolis Group.

Forster, J. (2015). Report from the RSS Working Group on Research Excellence Framework (REF) League Tables. Discussion paper, Royal Statistical Society, London, UK.

Harzing, A.W. (2017). Running the REF on a rainy Sunday afternoon: Do metrics match peer review? from: www.harzing.com

HEFCE (2009). Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework. Discussion paper, Higher Education Funding Council for England, London UK.

Hirsch, J. E. (2010). An Index to Quantify an Individual's Scientific Research Output that Takes into Account the Effect of Multiple Coauthorship. *Scientometrics*, 85, 741–754.

Mryglod, O., Kenna R., Holovatch Y. & Berche B. (2015). Predicting results of the research excellence framework using departmental h-index: revisited. *Scientometrics*, 104(3), 1013–1017.

Wang, L., Vuolanto P. & Muhonen R. (2014). Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives. Discussion paper.

The Maturity of Scientific Research Problems：A Method to Identify the Subsequent Influence of New Published Papers

Haiyan Wang[1], Zheng Ma[2] and Zhenglu Yu[3]

[1] *wanghaiyan@istic.ac.cn*
Institute of Scientific and Technical Information of China, No.15 Fuxing Road, Haidian District, 100038, Beijing(China)

[2] *mazheng@istic.ac.cn*
Institute of Scientific and Technical Information of China, No.15 Fuxing Road, Haidian District, 100038, Beijing(China)

[3]*luluyu@istic.ac.cn*
Institute of Scientific and Technical Information of China, No.15 Fuxing Road, Haidian District, 100038, Beijing(China)

Abstract

The follow-up academic high-impact method of scientific papers is a future-oriented predictive analysis method. This paper uses a new measure of scientific research problem maturity to identify the subsequent influence of new published papers. This paper chooses a special type of highly cited paper, which is the case of delayed recognition of papers. Through the scientific problem maturity method proposed in this paper, it is verified that the relationship and characteristics of scientific research and patent in the process of high citation for the delayed recognition paper reaches after several years of publication is in line with the actual situation. Experiments show that the scientific research problem maturity method, which complement other index or methods, become an effective method to predict the subsequent high impact of newly published papers.

Keywords

High impact paper, scientific research problem maturity, patent growth rate, paper growth rate, academic influence

Articles of Science and Technology are a kind of carrier of knowledge and issues. In face of more and more large amount of latest published articles, how could we identify a new article is going to have high impact in the future? It has been an important research direction on the field of identification of articles with high-impact in the future and academic trend analysis, and meanwhile, has been a focus of field study researchers, valuators and researchers of science policy.

The theory and method of identification for articles with high-impact in the future is an emerging hybrid FTA method. The high-impact scientific papers presented in this paper refer to those scientific research papers that have made significant impact on the development of the subject and the exchanges in the scientific community.

This paper presents a new measurement method namely scientific question readiness level to identify future impact strength of articles. The approach for study readiness level is constructed with an example of delayed recognition papers in the information security field, analyzes and verifies the situation of articles with future high-impact in this field. Delayed recognition papers refer to those papers initially unnoticed or unused， but then was considered very important, whose definition based on times cited in several years after publication. According to experimental study, the approach for study readiness level, is able to reflect the level of high-impact in the future, by means of complementing with other methods is efficient method to predict high-impact strength in the future of articles.

Introduction

At present, the research on the identification methods of high-impact scientific papers in the future is concentrated in three aspects. The indirect indicators are used to predict the subsequent citations of scientific papers, and the social network analysis methods are used to identify key position papers and semantic judgments.

Identification method based on indirect indicators of refracting high impact

If the first-hand evidence or indicators are not available, such as the newly published scientific papers have no or only low citation frequency and number of downloads, indirect metrics are needed. Variable analysis is the method to predict future citations by the characterization data of the paper.

Perneger and Brody think that the number of early hits is a potentially useful measure of the scientific value of published medical research papers (Perneger,2004; Brody,2006). Publication of hit counts by online journals should be encouraged. The paper's scientific value also leads to citation by other researchers. Earlier web usage statistics could be the predictors of later citation impact. Levitt and Thelwall(2008) think that for highly cited articles the prediction of citation ranking of from the sum of citations during their first six years was less accurate than prediction using the sum of the citations for only the fifth and sixth year. Adams (2005) also thinks early citation counts correlate with accumulated impact.Higher citations were predicted by indexing in numerous databases; number of authors; abstraction in synoptic journals; clinical relevance scores; number of cited references; and original, multicentre, and therapy articles from journals with a greater proportion of articles abstracted. Walters(2006) Predicted subsequent citations to articles published in twelve crime-psychology journals, the Author impact is more important than journal impact. Some researchers have carried out the author's reputation and the reputation of the journal on the high citation of the paper. Dalen and Henkens(2001,2005) thinks By means of a citation analysis, several quantifiable characteristics of the articles (characteristics with respect to authors, visibility, content and journals) are strongly related to their subsequent impact in the social sciences. They points out that the reputation of journals plays an overriding role in gaining attention in science.

A study by Malissa A. Schilling and Elad Green(2011) found that several factors, such as search scope or search breadth, search depth, and atypical connections between different research areas, can increase the impact of the paper. The above conclusions are still true even if the academic experience of the author and the success factor of the previously published article are controlled. Researchers such as Boyack (2005) developed a method based on journal importance, reference importance, and author reputation to predict the importance of current papers. According to their report, among the three important variables, journal influence has a strong correlation with the discipline. They suggest that these important factors can be used to rank papers without waiting to accumulate the number of citations. This implies a hypothesis that a paper can obtain more citations by citing more high-cited articles. Some researchers have also paid attention to the prediction of the subsequent citation of the paper based on the relevant indicators of the authors based on different research fields. Kostoff (2007) found that highly cited papers tend to have more co-authors, cite more literature, and write longer summaries and pages. The study by Skilton (2009) found that in the natural sciences, highly cited co-authors and authors tend to be highly cited if they have a multidisciplinary background or come from different regions. Bornmann and Daniel(2006), Tijssen, Visser, & van Leeuwen (2002) conducted research on the citation behavior, paper type, etc.

Identification research based on social network analysis method

The social network analysis method uses the topological indicators in the citation network to predict future citation behavior. Professor Chen Chaomei(2011) studied and established the theory of discriminating scientific turning points through social network methods. N. Shibata, Y. Kajikawa, Y. Takeda (2009)analyzed the correlation between the three centrality indicators in social networks and the number of future citations in two areas. According to the research, although there are only a few citations, the papers with high mediation degree tend to get more citations in the future because those papers connected the knowledge of different fields.

Recognition method based on semantic judgment

In addition to these indicators, will future high-impact papers be identified by other signals? Some scholars have discussed the semantic properties of the paper. Some researchers have proposed evaluation indicators for the quality of the paper. For example, Van Dalen (2005) proposed two indicators, impact and speed, to judge the quality of the paper in 2005, but these two indicators still depend on the calculation of the cited frequency and the first cited time. Related follow-up experiments (wang,MY ,2011,2012)demonstrate that predictions are more relevant to the characteristics of the paper, rather than relying on data sets and classifiers. In terms of combining semantic analysis, many researchers pay more attention to the keywords of the papers, and hope to distinguish the themes and connections of different papers from the perspective of conceptual understanding (Yufeng Zhang,2011). Among them, researchers have proposed the title semantic information, citation link relationship and order as the three dimensions of the frontier of detection research (Luxio-Aroas D and Leydesdorff L,2009). Some researchers combined the method of word co-occurrence analysis to solve the problem of automatic recognition and judgment of word semantics. In recent years, more scholars have used ontology to improve the effect of semantic analysis, and some researchers have conducted related research on the ontology of academic resources (Hao Wang and Xinning Su,2011).

At present, the research on the high influence of the paper in the future is mainly to use the external measurement index of the paper as a surrogate index to measure the future cited trend of the research paper. And it weakens the quality of the paper itself, which is the key feature of whether the paper is cited. Moreover, the existing research only predicts that the future citation frequency of the paper is still slightly single from the perspective of the characteristics of the paper itself. This study adopts the method of judging the maturity of scientific research in scientific papers, not only taking into account the paper indicators, but also introducing patent indicators, reflecting the interaction between scientific research and technology. This method analyzes the characteristics of the maturity of the research problems proposed by the delayed recognition paper, and attempts to propose a method to measure the future high impact of the paper.

Method

The maturity of the research problems

The maturity of the research problems refers to a study that may be at different stages of the research life cycle. It is based on the concept of scientific research life cycle to select the two indicators of scientific paper growth rate and patent growth rate to represent the development stage of scientific research. This embodies the interaction between scientific research and technology – scientific research is transformed into technology, and the development of technology promotes scientific research.

The development of basic research can be reflected by the growth rate of scientific papers, and the development of technology can be expressed by the growth rate of patents. The growth rate of scientific papers is normally distributed, and the growth rate of patents also has the characteristics of normal distribution. Because the production time of scientific papers and patents is inconsistent, the normal distribution curve of patent growth rate lags behind a period of time t[0-n] relative to the normal distribution curve of scientific papers. t depends on the speed at which scientific research problems embodied in scientific papers in the field are translated into patents.

Therefore, this study selects the two indicators of scientific paper growth rate and patent growth rate to construct the research life cycle graph (Fig. 1,2,3). For a specific scientific paper, the stage of the scientific research life cycle of the research problem can be judged according to the position of the two indicators (the growth rate of the paper and the patent growth rate) projected onto the coordinate map.

Usually, the number of scientific papers increases from slow to rapid growth under the stage from the germination to the rapid growth of basic research. Basic research papers on scientific research issues are produced at this stage. At this time, the research activities are mainly distributed among a small number of scientific research groups, and their concentration is relatively high. Furthermore, when scientific research issues have made breakthroughs or caused widespread concern, related scientific papers will increase rapidly, and the growth rate of papers will rise rapidly. Therefore, the basic research papers and related papers of breakthrough research of this scientific research are concentrated in this stage. It’s very important because it covers the basic theories and methods of the research, the breakthrough development or the theory and method of important turning points.

When the technology is in the mature stage, important basic inventions are born. As more and more R&D companies enter the field, basic research is becoming more sophisticated, and breakthrough technologies are constantly emerging. At this time, important basic patent technologies and breakthrough technologies are concentrated at this stage, and these technologies are bound to become technological prototypes for future development technologies. The development of technology has promoted the study of scientific issues. The scientific papers of technology mapping have embodied important technological developments. It is foreseeable that with the in-depth study of scientific issues and the continuous development of technology, its influence is constantly improving.

During the recession, the number of papers slowly fell back to a very small number, and the growth rate of the paper dropped rapidly and tended to zero. At the same time, due to the lack of sustained support and breakthroughs in basic research, the development of technology has entered a period of decline. The number of patent applications per year shows a negative growth, and the absolute value of negative growth rates continues to rise to a high point. At this time, the growth rate of the paper and the growth rate of patents are both negative growth states, indicating that the scientific research problem has entered a recession stage.

After entering the revival stage of basic research, the number of papers represented basic problems have made breakthrough, and the corresponding papers quantity has slowly increased. It has begun to enter a new round of germination. At this point, the negative growth rate of patent growth has declined, and the impact of breakthrough progress on patent conversion has not yet been demonstrated. The positive growth rate of the paper has shown a positive slow increase, indicating that the basic research in this stage has entered the recovery phase.

Therefore, based on the above analysis, we think that the development stage of scientific problems can be characterized by the growth rate of scientific papers and the growth rate of patents. If the research problems are in the stage of germination to the rapid growth of basic research or in the mature stage, it is considered that the scientific paper has a high probability of high influence in the future.

scientific research problem maturity development hypothesis

According to the above analysis, the development process of a field is divided into four stages, namely, the rapid growth stage of basic research, the mature stage of technology development, the stage of decline, and the revival stage of basic research.

We have designed a coordinates of the scientific research problem maturity judgment of the scientific papers (Fig. 1, Fig. 2, Fig. 3). Among them, Fig. 1 simulates the annual growth of the number of papers and the process of increasing the number of patents. Figure 2 simulates the growth of the papers. The rate and the growth rate of the patent growth rate, Figure 3 shows the maturity graph of the scientific research papers based on the paper growth rate-patent growth rate.

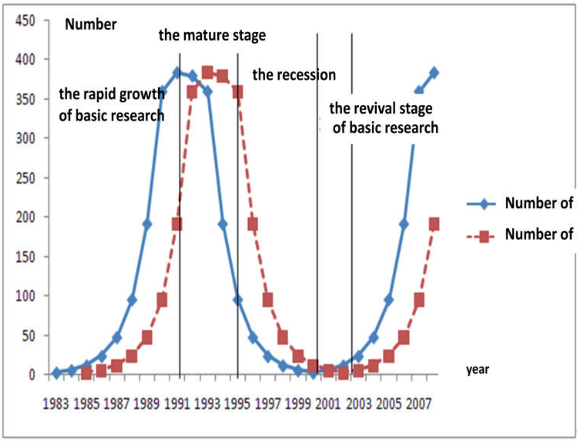


Figure 1 Schematic diagram of the growth of the paper and the growth of the patent

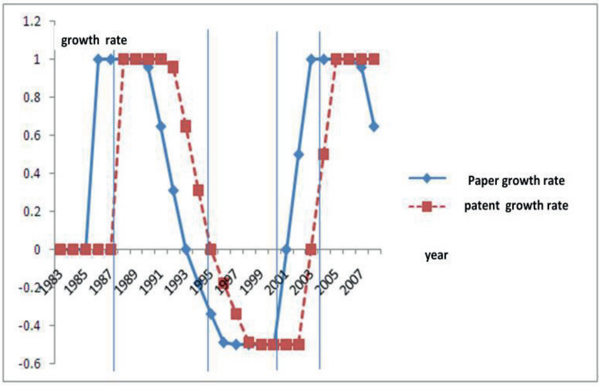


Figure 2 Schematic diagram of the growth rate of the paper and the growth rate of the patent

This study simulates the development process that papers and patents change from growth to recession through data. It builds a model of the relationship between the growth rate of papers and the growth rate of patents (Fig. 3). In Figure 3, the X-axis represents the patent growth rate, indicating the growth rate of the patent for a certain period; the Y-axis represents the growth rate of the paper, indicating the growth rate of the paper for a certain period. The diameter of the circle is determined by the ratio of the number of papers at that time point to the number of patents. The ratio of the number of papers and the number of patents at each time node can be judged initially.

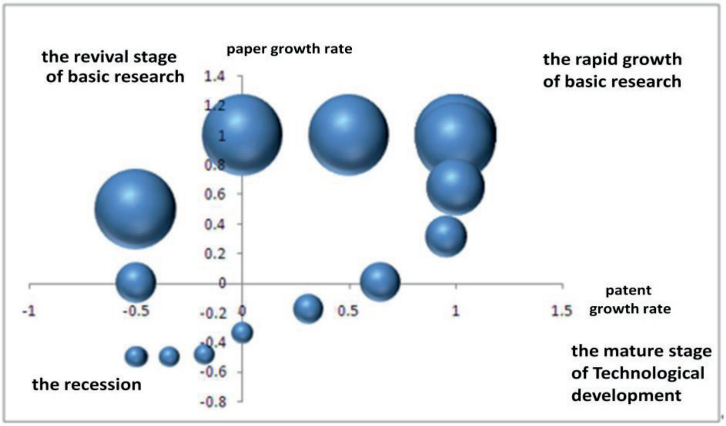


Figure 3 Science and technology paper research problem maturity graph

The various indicators involved in Figure 3 are explained below:

The growth rate of the paper refers to the change in the number of papers published at this time point and the number of papers published at the previous time point, also known as the growth rate.

$$\Delta Ai = \frac{A_i - A_{i-1}}{A_{i-1}} \times 100\%$$

$$= \left[\frac{A_i}{A_{i-1}} - 1 \right] \times 100\%$$

Ai is the number of papers published in the current period;
i=n（n=1,2……,n）

The patent growth rate refers to the change in the number of patents at this time point and the number of patents at the previous time point, also known as the growth rate.

$$\Delta P_j = \frac{P_j - P_{j-1}}{P_{j-1}} \times 100\%$$

$$= \left[\frac{P_j}{P_{j-1}} - 1 \right] \times 100\%$$

Pj is the number of patents in the current period
j=m（m=1,2……,m）

Round area:

$$S = \frac{1}{2} \pi \left[\frac{A_{i=n}}{P_{j=m}} \right]^2$$

$$R = \frac{A_{i=n}}{P_{j=m}}$$

S represents the size of the circle projected into the coordinate map， and is represented by the area;
R represents the diameter of the circle projected onto the coordinate map.
Ai is the number of papers published in the current period;
Pj is the number of patents in the current period;
i=n（n=1,2……,n）
j=m（m=1,2……,m）

Finding

Data and Domain

In recent years, due to the important academic and application value of information security Technology, countries have paid great attention. Research on information security technology has made substantial progress in both technical research and information industry. Therefore, this study selected the scientific and technical literature in the field of "information security technology" as the material for statistical analysis.

The domain data comes from the database in the Web of Science platform, including the core set SCI-EXPANDED, CPCI-S, based on the 12 types of subdivision techniques of the technical system defined in this study. The qualified document types are ARTICLE, REVIEW, PROCEEDINGS PAPER. A total of 14,265 valid search results published in 1995-2004 were obtained, which constitute the data set for the analysis of the Delayed recognition papers in this experimental study.

Considering the difference in the fields, the specific measurement standard for identifying delayed recognition papers in this paper is set as follows: the first three years after the publication of the paper, there is no citation frequency, and in the following years, the frequency of citation increases faster within the statistical time window(Haiyan Wang,2015). The total citation of the paper is the first 30% of the citations of the paper published in the same year. In the field of information security technology, the papers published from 1995 to 2004 were obtained. According to the criteria of the top 30% of the papers published in the same year, we got 41 papers. This article will judge the maturity of these 41 articles separately.

scientific research problems maturity judgment ideas

Firstly, determine the 12 types of technical directions in the field of information security technology, search for the number of papers and patents in the same technical direction and t in the corresponding year; secondly calculate the annual paper growth rate and patent growth rate; then draw maps: the number of papers - the number of patent annual growth process map, paper growth rate-patent growth rate map; Finally, a according to the position of the paper growth rate and the patent growth rate data point projected in the research problem maturity quadrant, we can judge the subdivision technology research problem maturity level, and then predict its development trend.

The idea of analyzing a single paper is as follows. Firstly, determine the research topic of a single paper, and determine the technical direction category of the paper by calculating topic similarity, then judge the development trends of the research problem based on the maturity of the technical direction; Secondly, according to the research theme of the paper, set the search terms, retrieve the number of the same subject scientific papers and the number of patents, calculate the annual paper growth rate and patent growth rate; then draw the annual growth process diagram of the paper - patent, the paper growth rate-the patent growth rate maturity quadrant map; finally, according to the position of the paper growth rate and the patent growth rate data point projected in the research problem maturity quadrant, we can judge the subdivision technology research problem maturity level, and then predict its development trend.

Scientific research problem maturity judgment process

This section analyzes the Delayed recognition papers separately. The following is the analysis process and conclusions of one of the papers.

Paper 1, entitled "An efficient remote use authentication scheme using smart cards", published in the "IEEE TRANSACTIONS ON CONSUMER ELECTRONICS" journal in 2000. This paper discusses a smart card-based remote authentication scheme.

After studying, the paper belongs to the cross-technical direction of identification and authentication technology and cryptography. To determine the technical direction of each thesis's research topic, it can be judged by co-occurrence of the topic words and co-citation of the literature. Because the number of the Delayed recognition papers involved in this study is small, the method of judging the technical direction of each article is adopted. This study separately analyzed the maturity of the cross-technical direction of identification and authentication technology and cryptography.

Technical direction 1: Identification and authentication technology.

According to the above search scheme, 132,718 papers and 25634 patents were obtained. Based on this data, draw the maturity map of the scientific research papers (Fig. 4).

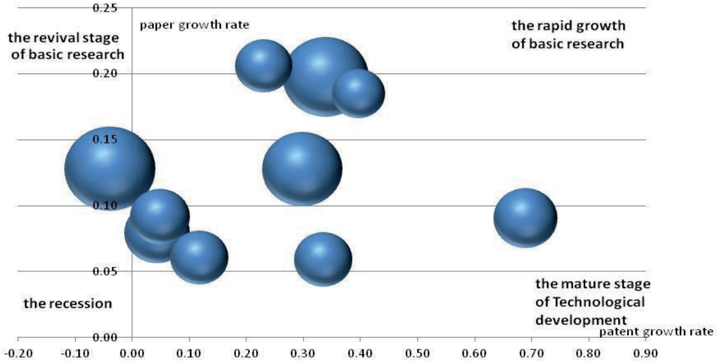


Figure 4 Technical direction 1 scientific paper research problem maturity graph

Technical direction 2: password technology.

According to the above search scheme, 10851 papers and 9075 patents were obtained. Based on this data, draw the maturity map of the scientific research papers (Fig. 5).

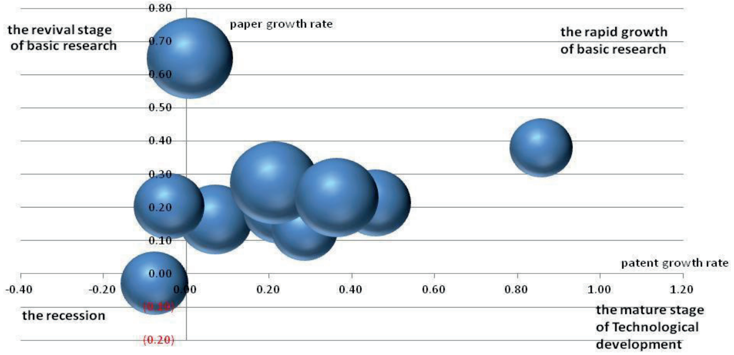


Figure 5 Technical direction 2 Scientific paper research problem maturity graph

For the Delayed recognition article : smart cards, the design search is as follows:

Paper search:(ts=("identification" or "authentication" or "access control") or ts=("cryptanalysis" or "encryption" or "cryptography" or "password")) and ts=(information or network or cyber or software or computer or wireless or communication or data security) and ts=(“smart card”)

Patent search:((TI=("identification" or "authentication" or "access control") or TI=("cryptanalysis" or "encryption" or "cryptography" or "password")) and TI=(information or network or cyber or software or computer or wireless or communication or data security) and …

According to the above search scheme, 460 papers and 203 patents were obtained. Based on this data, draw the maturity map of the scientific research papers (Fig. 6).

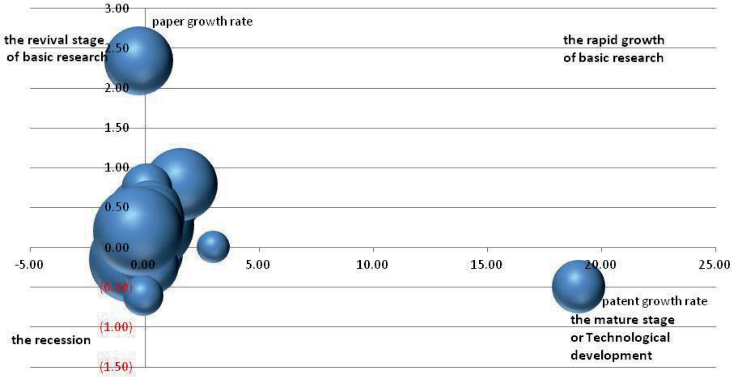


Figure 6 Science paper research problem maturity graph

For Smart Card applications and trends in this research topic:

The data points of the paper growth rate and patent growth rate of the smart card theme are projected in the first quadrant of the research problem maturity model—the rapid growth stage of basic research, the third quadrant—the mature stage of technology development, the fourth quadrant—the revival of basic research stage. According to the definition of this model, the theme has experienced the germination, growth, maturity and revival stage of basic research after new technology improvement.

In the period of rapid growth of basic research and the revival of basic research, it can predict its future high impact - becoming a highly cited paper, it is in line with the actual situation of the Delayed recognition papers. Taking the smart card thesis as an example, according to the performance of the number of papers, the smart card technology experienced a rise, a turn, and then rise, the development of technology, stagnation and continuous change, then making the field undergo a renaissance stage. This revival phase is shown in the fourth quadrant in the maturity graph.

Discussion and Conclusions

It can be seen from the analysis of the existing research that there are still some shortcomings in the identification of subsequent high-impact papers, and the recognition angle or method is slightly single. The follow-up academic high-impact method of scientific papers is a future-oriented predictive analysis method. This paper uses a new measure of scientific research problem maturity to identify the subsequent influence of new published papers. This paper chooses a special type of highly cited paper, which is the case of delayed recognition of papers. Through the scientific problem maturity method proposed in this paper, it is verified that the relationship and characteristics of scientific research and patent in the process of high citation for the delayed recognition paper reaches after several years of publication is in line with the actual situation. Experiments show that the scientific research problem maturity method, which complement other index or methods, become an effective method to predict the subsequent high impact of newly published papers.

At the same time, we also see that this study verifies the development of a single paper through the development of domain themes, and it is aimed at the macro or meso level, applied to the discipline planning of research institutions and the development support of the national science and technology output stage. However, this method cannot directly face micro-evaluation, such as the application of personal scientific evaluation. If it needs to be applied to personal scientific evaluation, other auxiliary indicators should be introduced.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant Nos.71373252). China Postdoctoral Science Foundation Special Funding Project (No. 2015T80119). Institute of Scientific and Technical Information of China (Project Number: ZD2018-18)

References

Bornmann, L., Daniel, H.-D.(2006). What do citation counts measure? A review of studies on citing behavior . *Journal of Documentation*, 1, 45-80.

Boyack, K.W., Klavans, R., Ingwersen, P., Larsen, B. (2005).Predicting the importance of current papers. Ingwersen, P, Larsen, B ISSI 2005: *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics* (ISSI 2005).(pp.335-342). Stockholm, Sweden:ISSI

Brody,T.,Harnad,S.,&Carr,L.Earlier.(2006). Web usage statistics as predictors of later citation impact. *Journal of the American Association for Information Science and Technology,* 8,1060-1072.

Chaomei Chen.(2011).Turning Points—The Nature of Creativity(pp.116-132).Beijing：*Higher Education Press.*

Haiyan Wang, Zheng Ma, Yuntao Pan , Yishan Wu. (2015).Study on citation curves of highly-cited articles with sleeping beauties articles. *Library and Information Service*, 16, 83-89.

Hao Wang, Xinning Su.(2010). Research on construction and applications of the CSSCI academic esources networks model based on ontology. *Journal of the China Society for Scientific and Technical Information*,6,658-667.

J. Adams.(2005). Early citation counts correlate with accumulated impact. *Scientometrics*,3,567-581.

Kostoff, R.(2007). The difference between highly and poorly cited medical articles in the journal Lancet. *Scientometrics*, 3,513-520.

Levitt, J., &Thelwall, M.(2008).Patterns of annual citation of highly cited articles and the prediction of their citation ranking: A comparison across subjects. *Scientometrics*,1,41-60.

Lucio-Aroas D，Leydesdorff L. (2009).An Indicator of Research Front Activity: Measuring Intellectual Organization as Uncertainty Reduction in Document Sets. *Journal of the American Society for Information Science and Technology*, 12,2488-2498.

Malissa A. Schilling，Elad Green.(2011) .Recombinant search and breakthrough idea-generation: An analysis of high impact papers in the social sciences.Research Policy,10,1321-1331

N. Shibata, Y. Kajikawa, Y. Takeda, I. Sakata, K. Matsushima. (2009).Early Detection of Innovations from Citation Networks.*2009 IEEE International Conference on Industrial Engineering and Engineering Management* (IEEM 2009),(pp.54-58). Hongkong , China:IEEM 2009.

Perneger, TV. (2004).Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ. *British Medical Journal*,7465,546-547.

Skilton, P.(2009). Does the human capital of teams of natural science authors predict citation frequency? *Scientometrics*, 3,525-542.

Tijssen, RJW, Visser, MS, van Leeuwen, TN.(2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*,3,381-397.

Van Dalen, HP, Henkens, K.(2001). What makes a scientific article influential? The case of demographers .*Scientometrics*, 3,455-482

Van Dalen, HP, Henkens, K.(2005). Signals in science-on the importance of signalling in gaining attention in science. *Scientometrics*,2,209-233.

Walters, GD.(2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics*, 3, 499-510.

Wang, M.Y., Yu, G., Xu， J.H., Yu,DR.(2012). Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics*, 4,586-599.

Wang,M.Y.,Yu,G.,Yu,DR.(2011). Mining typical features for highly cited papers. *Scientometrics*, 3,695-706.

Yufeng Zhang, Jiaojie Cai.(2011). Research on user interest ontology learning based on web mining technology. *Journal of the China society for scientific and technical information,* 4,369-373.

Disciplinary Variations in Altmetric Coverage of Scholarly Articles

Sumit Kumar Banshal[1], Vivek Kumar Singh[2], Pranab K. Muhuri[3] and Philipp Mayr[4]

[1] sumitbanshal06@gmail.com
Department of Computer Science, South Asian University, New Delhi (India)

[2] vivek@bhu.ac.in
Department of Computer Science, Banaras Hindu University, Varanasi (India)

[3]pranabmuhuri@gmail.com
Department of Computer Science, South Asian University, New Delhi (India)

[4] philipp.mayr-schlegel@gesis.org
GESIS Leibniz Institute for Social Sciences, Cologne (Germany)

Abstract

The popular social media platforms are now making it possible for scholarly articles to be shared rapidly in different forms, which in turn can significantly improve the visibility and reach of articles. Many authors are now utilizing the social media platforms to disseminate their scholarly articles (often as pre- or post- prints) beyond the paywalls of journals. It is however not very well established if the level of social media coverage and attention of scholarly articles is same across all research disciplines or there exist discipline-wise variations. This paper aims to explore the disciplinary variations in coverage and altmetric attention by analyzing a significantly large amount of data from Web of Science and Altmetric.com. Results obtained show interesting patterns. Medical Sciences and Biology are found to account for more than 50% of all instances in Altmetrics. In terms of coverage, disciplines like Biology, Medical Science and Multidisciplinary Sciences have more than 60% of their articles covered in Altmetrics, whereas disciplines like Engineering, Mathematics and Material Science have less than 25% of their articles covered in Altmetrics. The coverage percentages further vary across different altmetric platforms, with Twitter and Mendeley having much higher overall coverage than Facebook and News. Disciplinary variations in coverage are also found in different altmetric platforms, with variations as large as 7.5% for Engineering discipline to 55.7% for Multidisciplinary in Twitter. The paper also looks into the possible role of source of publication in altmetric coverage level of articles. Interestingly, some journals are found to have a higher altmetric coverage in comparison to the average altmetric coverage level of that discipline.

Introduction

The rapid growth of the Internet and social media has not only transformed the businesses, organizations and society, but has also changed the entire process of scholarly information processing, including article storage, access and dissemination. Not only research articles are being stored in and accessed from online digital libraries, but they are also disseminated through different social media platforms. Scholarly articles are now disseminated and shared on different social media platforms such as ResearchGate, Twitter, Facebook etc. There are some other popular platforms dedicated mainly to dissemination and sharing of academic texts, such as Academia and Mendeley. These academic networks provide wide-range of facilities which can be useful for academics (Gruzd & Goertzen, 2013). Several studies (such as by Williams & Woodacre, 2016) have found these types of academic social networks very informative and relevant for quantitative characterization in research assessments. This social media phenomenon of scholarly articles has become so popular that now a new range of metric have been designed, called alternative metric, to measure the interaction of social media with scholarly information processing (Priem, 2014; Priem & Hemminger, 2010).

Altmetrics is now an interesting area of research where researchers try to analyze the social media coverage and consumption of scholarly articles; and sometimes Altmetric values are even able to predict future citations. However, most of the attention of research in the area has so far been concentrated on measuring correlations and interactions among social media transactions and citation behaviour of scholarly articles. Relatively lesser attention has been paid on measuring disciplinary variations in social media coverage and usage of scholarly articles. This paper tries to address the issue through a comprehensive study involving large amount of data collected from Web of Science and corresponding values from Altmetrics. The main objective is to find out if there exist discipline-wise variations in social media coverage and consumption patterns of the scholarly articles. The data for different platforms (namely Twitter, Facebook, News and Mendeley) is analysed computationally for the purpose. Statistics for some highly visible journals in Social Media and journals with high impact factor are also analysed to understand the role of source of publication and disciplinary variations.

Related Work

There has been some attention of researchers on understanding and analyzing the relationship of social media and scholarly information systems. Some of these studies (Priem, 2014; Haustein et al., 2014; Thelwall & Kousha, 2015; Sugimoto et al., 2017) tried to understand and demonstrate if social media platforms can be used (or not) as a tool to attract more attention towards a published work. Few others (Shema, Bar-Ilan & Thelwall, 2014; Thelwall, 2016; Peters et al., 2016) tried to see if Altmetrics could correlate with citations, with few (Costas, Zahedi & Wouters, 2015a) going to the extent to see if it can complement citations or not. There have also been studies that tried to predict early citations from different platforms of social media, such as Mendeley (Thelwall, 2018), ResearchGate & Google Scholar (Thelwall & Kousha, 2017 a), altmetric.com (Thelwall & Nevill, 2018), and CiteULike bookmarks (Sotudeh, Mazarei & Mirzabeigi, 2015) etc. Country specific Altmetric studies has also been done, such as for India (Banshal et al., 2018) and China (Wang et al., 2016) etc.

Discipline-specific studies of understanding Altmetric coverage and impact have been done by some researchers. For example, a study by Bar-Ilan (2014) mapped astrophysics research output with Mendeley readership behavior using Scopus and Arxiv. Another study (Sotudeh, Mazarei & Mirzabeigi, 2015) analyzed the correlation of research impact and CitedULike bookmarks in Library & Information Science discipline. Few other such efforts are analyzing the relationship between traditional and alternative matrices in psychology literature for the period of 2010-2012 (Vogl, Scherndl & Ku, 2018); online media presence of Swedish articles in humanities in the year 2012 (Hammarfelt, 2014); and evaluation of the impact of Altmetrics in social sciences and humanities research published by Taiwan based researchers (Chen et al., 2015). In a recent work (Htoo & Na, 2017) worked towards alternative metrics across various disciplines of Social Sciences and visualized the significance of ten selected indicators on nine disciplines of social science. However, there has been relatively less attention on understanding disciplinary variations in altmetric coverage of scholarly articles.

The only past works found on Altmetrics with focus on disciplinary analysis are as follows: Holmberg & Thelwall (2014) conducted a study on data from Twitter of ten selected disciplines to map their coverage and frequencies in twitter. Authors here selected ten different disciplines based on their publication size and pattern variations to represent variations in publishing scholarly communication. Similar to this work, authors selected ten disciplines of social sciences and humanities from web of science subject areas to correlate with Mendeley readership (Mohammadi & Thelwall, 2014). However, analyzing overall disciplinary variations in coverage was not the main objective of the paper. Another work (Zahedi, Costas & Wouters,

2014) used a multi-disciplinary approach on different online and social networks to assess coverage and distribution of randomly selected 20,000 articles published between 2005 and 2011. This approach also outlined alternative metrics into seven different broader areas of research. These seven broader areas are classified based on high level classification which classified the research areas as 'Natural Sciences', 'Engineering Sciences' etc. In another related work (Costas, Zahedi & Wouters, 2015b), authors tried to understand the thematic orientation of publications mentioned on social media. However, this paper only uses altmetric data and tries to understand the distribution of the data into various high-level disciplines. It does not measure coverage levels of different disciplines in altmetric. Another work (Ortega, 2015) used thousands of Spanish researchers' profiles to explore the disciplinary behavioral patterns in three online media, namely ResearchGate, Academia & Mendeley. Furthermore, in this analysis the scholarly articles are classified into eight broader areas to visualize the presence and coverage of same discipline's researchers across different platforms. In one relatively recent work (Thelwall & Kousha, 2017b) scholarly communications shared in ResearchGate is being classified into 27 Scopus categories of subject area where the Scopus subject areas are used as it is defined to classify the articles and understand the disciplinary variations.

The present work has focused objective of analyzing coverage levels of articles from different disciplines in Altmetrics. It uses a large amount of data from Web of Science (about 1.4 million records to be precise) and their corresponding entries in Social media platforms. The objective is to understand whether research articles from all disciplines get equal coverage in social media platforms or not. Data is categorized into 14 different well-identified broader research areas/ disciplines and variations in altmetric coverage across these disciplines are identified. Unlike, previous studies the present work mainly tries to analyze altmetric coverage levels of different disciplines and not the disciplinary distribution of altmetric data, pursued by many of the previous studies. Further, a journal-based analysis is also done to understand the disciplinary variation and its impact on altmetric coverage.

Data

The data is obtained from two sources: Web of Science and Altmetric.com. First of all the publication records for the year 2016 are downloaded from Web of Science. The data download process is performed during 1-10th December, 2018. A total number of 2,528,868 publication records are found for the year 2016. This data is then scanned for DOI entries and those records that do not have DOI are removed. This process reduces the data to 1,460,124 records. The second step in data collection involved collecting altmetric data for the 1,460,124 publication records of Web of Science with DOIs. For this purpose, the popular portal altmetric.com was accessed. In altmetric.com, 18 different types of mentions and stats are provided. These comprise of different social network mentions and reads. Out of the 1,460,124 records, a total of 681,274 publication records are found indexed in altmetric.com. Out of these 650,009 records are found with at least one kind of statistics. This corresponds roughly to 45% of data collected from Web of Science having DOI. Though altmetric.com captures statistics from various social platforms; platforms like Twitter, Facebook, News and Mendeley are found to be more popular. We have, therefore, used the altmetric data for these four platforms. The analysis also involves Impact Factor data for different journals, which is collected from Web of Science Reports.

Disciplinary Tagging

To understand disciplinary variations in altmetric coverage, it is necessary to tag each publication record with at least one specific discipline. For this task, each publication record in the dataset is classified into one of the 14 broad research disciplines, as proposed in an earlier work (Rupika et al., 2016). This tagging is done by using Web of Science Category (WC) field information. One record can be tagged with multiple disciplines of research based on its WC entries. These 14 broader disciplines are as follows: Agriculture (AGR), Art & Humanities (AH), Biology (BIO), Chemistry (CHEM), Engineering (ENG), Environment Science (ENV), Geology (GEO), Information Sciences (INF), Material Science (MAR), Mathematics (MAT), Medical Science (MED), Multidisciplinary (MUL), Physics (PHY) and Social Science (SS). Thus the 255-category division of articles in Web of Science is reduced to these 14 broader disciplines and each publication record is tagged with one (or more) broad disciplines. All further analysis on disciplinary variations in altmetric coverage are done across these 14 broad disciplines. The articles are grouped discipline-wise and analytical results are obtained accordingly.

Disciplinary Distribution and Coverage

The first point of analysis was to find out disciplinary distribution of articles in Web of Science & Almetrics and to see if disciplines are distributed in same proportions in Web of Science & Altmetrics. **Figure 1** presents the discipline-wise distribution of research output in Web of Science and altmetric.com. We can observe that some disciplines with higher proportionate distribution in Web of Science have relatively lesser proportion of presence in altmetric.com. In contrast, the Medical Science (MED) discipline accounts for about 30.2% proportion of output in Web of Science whereas in altmetric.com, it accounts for more than 41% of articles covered. Thus, this discipline is over covered in Altmetrics. For many other disciplines, proportionate contribution in Web of Science and altmetrics.com are different. For example, in Web of Science, PHY has the second most published output with contribution of 13.8% whereas its proportionate contribution in altmetrics.com is 7.9%. In Altmetrics, Social Sciences (SS) has the second highest proportionate contribution with a share of 15.4% followed by Biology with a share of 11.9%. Thus, MED and BIO disciplines are more visible in altmetric coverage, being covered in proportion higher than their proportion of published articles indexed in Web of Science. Disciplines like PHY, MAR, MAT, INF and ENG are proportionately less covered in Altmetrics. These results indicate that difference in altmetric coverage proportion of different disciplines are likely.

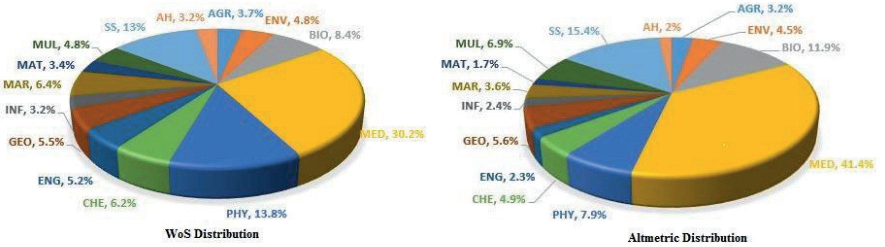


Figure 1. Discipline-wise Article Distribution in Web of Science (WoS) and Altmetric.com

The second point of analysis was to look at each discipline and find out its coverage level in Altmetrics. For this purpose, the Web of Science article counts for different disciplines is taken and altmetric.com is searched to see if they are covered in Altmetrics. This is done through an article-wise lookup in altmetric.com, for each article in Web of Science. **Table 1** shows the counts of articles of different disciplines that are indexed in Web of Science, number of articles found in altmetrics.com, and the altmetric coverage percentage for each of the 14 disciplines. It is observed that there is a significant difference in altmetric coverage percentages. For example, MUL, MED and BIO disciplines have a coverage percentage above 60%, which shows that out of all publications from these disciplines in Web of Science, more than 60% are found covered in Altmetrics. Articles from SS discipline have a coverage percentage of more than 50%. Interestingly, disciplines like ENG, MAT and MAR have less than 25% of their articles covered in Altmetrics. There is, therefore, a clear disciplinary variation in altmetric coverage of articles.

Table 1. Discipline-wise data for altmetric coverage of articles indexed in Web of Science (WoS)

Discipline	*Articles in WoS*	*Altmetric Presence*	*Coverage Percentage*
AGR	53749	21068	39.2
AH	47186	12871	27.3
BIO	123180	77259	62.7
CHE	90959	31670	34.8
ENG	75834	14737	19.4
ENV	69709	29194	41.9
GEO	80477	36420	45.3
INF	46438	15568	33.5
MAR	94117	23571	25
MAT	49385	10865	22
MED	**441032**	**268830**	61
MUL	69445	44778	**64.5**
PHY	201373	51569	25.6
SS	189835	100029	52.7

The third point of analysis was to find out if the discipline-wise coverage patterns are same across different altmetric platforms or if they vary significantly. For this purpose, coverage patterns across four different altmetric platforms, namely Twitter, Facebook, News and Mendeley are identified. **Table 2** shows the data for coverage of articles indexed in Web of Science in different altmetric platforms, corresponding to different disciplines. It is observed that, Mendeley and Twitter have in general higher coverage percentage for most of the disciplines as compared to News and Facebook. Thus, Mendeley and Twitter appear to be more popular altmetric platforms. It can also be observed from the table that there are noticeable disciplinary differences in altmetric coverage of articles. For example, articles from MUL discipline have highest presence in both Twitter & Mendeley with 55.7% and 63.6% coverage followed by BIO with 54.6% & 62.1% coverage. But articles from disciplines like ENG and INF have less coverage in Twitter (7.5% and 9.3%, respectively) and Mendeley (18.6% and 30.5%, respectively). In case of Facebook and News, coverage levels are low, with highest coverage being for MUL discipline of 17.8% followed by MED discipline of 13.7%. In News platform, coverage levels are further low with highest being 13% for MUL discipline followed by 7.5% for MED discipline. Interestingly, articles from ENG discipline have low coverage (7.5% in Twitter; 1.4% in Facebook; 18.6% in Mendeley and 0.5% in News) across all the platforms.

In terms of variations across disciplines, Twitter has the largest variation in coverage ranging from low of 7.5% for ENG to 55.7% for MUL. The variation range is in Mendeley is from 18.6% for ENG to 63.6% for MUL discipline, almost similar as Twitter. Facebook has variation in coverage percentage ranging from 1.3% for INF to 17.8% for MUL discipline. Thus, it is clearly observed that there exist disciplinary variations in altmetric coverage of articles, which varies further across different altmetric platforms. It may also be interesting to see if these variations can be attributed mainly to disciplines or if there are other factors such as the source of publication (journal), which play an important role.

Table 2. Discipline-wise Coverage across Different Platforms of Articles indexed in Web of Science (WoS)

Discipline	*Articles in WoS*	*Twitter*		*Facebook*		*News Mention*		*Mendeley*	
		#of Articles	*Coverage Percentage*	*#of Articles*	*Coverage Percentage*	*#of Articles*	*Coverage Percentage*	*#of Articles*	*Coverage Percentage*
AGR	53,749	16,132	30	4,406	8.2	1,468	2.7	20,784	38.7
AH	47,186	8,690	18.4	2,025	4.3	350	0.7	10,763	22.8
BIO	123,180	67,281	54.6	16,850	13.7	9,006	7.3	76,480	62.1
CHE	90,959	24,733	27.2	4,673	5.1	2,332	2.6	31,331	34.4
ENG	75,834	5,663	7.5	1,067	1.4	355	0.5	14,128	18.6
ENV	69,709	22,196	31.8	4,722	6.8	2,219	3.2	28,961	41.5
GEO	80,477	26,873	33.4	5,445	6.8	3,599	4.5	35,902	44.6
INF	46,438	4,330	9.3	583	1.3	373	0.8	14,151	30.5
MAR	94,117	15,096	16	2,508	2.7	1,674	1.8	23,280	24.7
MAT	49,385	5,773	11.7	792	1.6	618	1.3	9,777	19.8
MED	**441,032**	**224,132**	50.8	**70,401**	16	33,021	7.5	**264,405**	60
MUL	69,445	38,675	**55.7**	12,371	**17.8**	9,021	**13**	44,194	**63.6**
PHY	201,373	33,571	16.7	5,973	3	3,908	1.9	50,031	24.8
SS	189,835	78,799	41.5	24,557	12.9	9,258	4.9	96,180	50.7

Analysing Disciplinary Variations by Journals

It is quite clear from the discussion in previous section that there are disciplinary variations in altmetric coverage of scholarly articles. An important and relevant question worth exploring here would be to find out if the source of publication (i.e. journal) has any role in higher altmetric coverage of articles. In order to explore this question, a part of data was taken out and analysed. This data comprised of top 100 journals (ranked by Web of Science article count) with the condition that they should have at least 500 articles covered in Altmetrics. These journals are then tagged with a primary discipline, based on data available either on their homepage or Wikipedia. Thus, each journal is categorized into one of the 14 broad disciplines. **Table 3** presents the data for these journals. It can be observed that MED discipline accounts for 35 out of these 100 journals followed by BIO with 24 journals and CHEM with 20 journals. These three disciplines taken together account for about 80% of the top 100 journals. In terms of coverage, MUL discipline has highest number of papers covered in all the four altmetric platforms, though it has only 6 out of 100 journals. In terms of coverage percentage, ENG discipline has the highest coverage in Twitter (86.2%) and Mendeley (85.6%) followed by GEO discipline. Disciplines like MED and BIO have somewhat lesser, but still a significant coverage of articles in Altmetrics. For example, MED has coverage percentage of 59.3% in Twitter and 65% in Mendeley for its articles in the selected sample. Similarly, BIO discipline has coverage percentage 62.2% in Twitter and 69.1% in Mendeley. However, the lesser overall covered disciplines like INF, MAT, MAR are found better covered in this sample. Thus, it is very difficult to conclusively say that publication in a particular journal gives an article a higher chance of altmetric coverage. The disciplinary variations are, however, still seen.

Table 3. Disciplinary Distribution of 100 Most Productive Journals (ranked by WoS count) across Platforms

Discipline	*#of Journals*	*# Articles in WoS*	*Twitter*		*Facebook*		*News Mention*		*Mendeley*	
			#of Articles	*Coverage %*	*#of Articles*	*Coverage %*	*#of Articles*	*Coverage %*	*#of Articles*	*Coverage %*
MED	**35**	**42,988**	25,476	59.3	7,922	18.4	4,724	11	27,955	65
BIO	24	30,596	19,027	62.2	4,380	14.3	3,734	12.2	21,127	69.1
CHE	20	46,407	17,522	37.8	2,986	6.4	1,828	3.9	21,437	46.2
PHY	18	39,474	13,941	35.3	2,775	7	1,483	3.8	17,748	45
MUL	6	46,502	**29,894**	64.3	**8,360**	18	**8,205**	**17.6**	**33,738**	72.6
ENV	4	6,306	2,363	37.5	287	4.6	284	4.5	3,000	47.6
GEO	3	2,897	2,279	78.7	271	9.4	221	7.6	2,401	82.9
MAR	3	4,149	1,576	38	202	4.9	373	9	2,125	51.2
AGR	2	1,694	1,240	73.2	328	19.4	157	9.3	1,360	80.3
SS	2	1,270	953	75	474	**37.3**	67	5.3	1,040	81.9
ENG	1	1,149	990	**86.2**	8	0.7	11	1	983	**85.6**
INF	1	793	482	60.8	67	8.4	76	9.6	524	66.1
MAT	0	0	0	0	0	0	0	0	0	0
AH	0	0	0	0	0	0	0	0	0	0

For a more detailed analysis, the journals and their data are arranged in a different ranking order. **Table 4** shows the data for top 50 journals, ranked both by absolute altmetric counts (on the left side) and by altmetric coverage percentage (on the right side). Looking at left part of the table, it is observed that PHY and CHEM have 13 journals each in the list, followed by 12 journals from MED and 11 journals in BIO. The top few journals having highest altmetric absolute count are MUL discipline. Thus, in terms of absolute counts, MUL, MED, BIO are main disciplines. However, when the journals are sorted by altmetric coverage percentage, then journals from GEO and SS disciplines are also found listed. However, out of 50 journals in the list, 19 journals are still from MED discipline followed by BIO with 16 journals. Interestingly, ranking by altmetric coverage percentage results in only one journal from PHY figuring in the list. Thus, disciplinary variations are seen in this part of analysis as well.

To analyse the impact of journal even further, another sample data was extracted. This sample comprised of 50 journals, ranked by 2016 Impact Factor of journals. **Table 5** presents Web of Science article count for these top 50 journals along with their altmetric counts and coverage percentage. Here, most of the journals belong to either Medical Science or Biological Science. However, of particular interest would be the journals, which are from other disciplines. It is interesting to note that some of these journals have better coverage levels than the typical coverage level of that discipline. For example, the journal 'Nature Materials' has altmetric coverage percentage of 78.6%. Similarly, the journal 'Annual Review of Astronomy and Astrophysics' has an altmetric coverage percentage of 93.6%. Other journals like 'Nature Nanotechnology' has altmetric coverage percentage of 90.1% and 'Reviews of Modern Physics' has altmetric coverage percentage of 88.1%. This is higher coverage percentage than the overall coverage percentage of the respective disciplines. But at the same time several other journals like ‘Progress in Materials Science’, ‘Progress in Polymer Science’, ‘Accounts of Chemical Research’ and ‘Behavioural and Brains Sciences’ have altmetric coverage percentage around or below 50%. However, most of the journals in MED, BIO etc. continue to have higher altmetric coverage percentage. Therefore, it can be observed that there is a definite impact of the discipline of an article in its altmetric coverage. Articles from some disciplines have higher altmetric coverage. There are also some exceptions to this, where some journals in disciplines having relatively low altmetric coverage percentage, have higher altmetric coverage. Therefore, the journal has also some role to play in altmetric attention potential of an article.

Table 4. Top 50 Journals Based on Altmetric Counts and Coverage Percentage

Sorted on Altmetric absolute Count			***Sorted on Altmetric Coverage Percentage***		
Journal	*Discipline*	*TP_ALT*	*Journal*	*Discipline*	*Coverage %*
PLoS ONE	MUL	15310	PLoS Pathogens	BIO	98.9
Scientific Reports	MUL	11017	Nature	MUL	94.3
Nature Communications	MUL	2849	Atmospheric Chemistry & Physics	GEO	93.6
Proceedings of the National Academy of Sciences of the United States of America	MUL	2821	Bioinformatics	MED, BIO	93.3
British Medical Journal	MED	2574	Cell Reports	BIO	93
Oncotarget	BIO, MED	2454	American Journal of Public Health	MED	93
Angewandte Chemie. International Edition	CHE	2374	Angewandte Chemie. International Edition	CHE	92.5
Applied Physics Letters	PHY	1873	NeuroImage	MED	92.2
Dalton Transactions: An International Journal of Inorganic Chemistry	CHE	1809	Dalton Transactions: An International Journal of Inorganic Chemistry	CHE	92.1
RSC Advances	CHE	1590	Journal of Clinical Oncology	MED, BIO	92
Journal of the American Chemical Society	CHE	1526	Geophysical Research Letters	GEO	90.9
Physical Review B	PHY	1489	Nature Communications	MUL	90.7
Medicine	MED	1454	Blood	MED	90.7
Journal of Biological Chemistry	BIO	1433	Current Biology	BIO	89.7
Frontiers in Microbiology	BIO	1419	New England Journal of Medicine	MED	89.6
Frontiers in Plant Science	BIO	1393	Inorganic Chemistry	CHE	89.4
Physical Review D	PHY	1373	Proceedings of the National Academy of Sciences of the United States of America	MUL	89
Physical Review Letters	PHY	1343	British Medical Journal	MED	88.9
International Journal of Molecular Sciences	PHY, CHE, BIO	1321	Journal of Alzheimer's Disease	MED	88.2
Frontiers in Psychology	MED	1286	Clinical Cancer Research	MED	87.8
Chemistry - A European Journal	CHE	1248	Nucleic Acids Research	BIO, CHE	87.2
ACS Applied Materials & Interfaces	CHE, PHY	1155	PeerJ	BIO, MED	86.4
Monthly Notices of the Royal Astronomical Society	PHY	1146	Industrial & Engineering Chemistry Research	CHE, ENG	86.3
Geophysical Research Letters	GEO,	1146	BMC Genomics	BIO	86.3
The Astrophysical Journal	PHY	1135	Journal of Neuroscience	MED	86.2
Science	MUL	1132	Neurology	MED	86.2
Inorganic Chemistry	CHE	1114	PLoS Neglected Tropical Diseases	BIO	85.8

Chemical Communications	CHE	1107	Science	MUL	85.7
International Journal of Cardiology	MED	1092	JAMA: Journal of the American Medical Association	MED	84.5
Tumor Biology	MED,BIO	1024	eLife	BIO	84.4
Industrial & Engineering Chemistry Research	CHE, ENG	992	Pediatrics	MED	84.4
PeerJ	BIO, MED	982	Journal of Immunology	MED	84.3
Physical Chemistry Chemical Physics (PCCP)	CHE	956	Antimicrobial Agents and Chemotherapy	BIO	83.1
Journal of Physical Chemistry - Part C	CHE	910	Nutrients	AGR	82.3
BMJ Open	MED	903	Psychiatry Research	SS	82.3
Science of the Total Environment	ENV	883	Journal of Affective Disorders	SS	81.7
Blood	MED	871	Frontiers in Plant Science	BIO	80.8
Sensors (14248220)	PHY	862	Frontiers in Microbiology	BIO	79.7
Nature	MUL	859	Applied & Environmental Microbiology	BIO	79.4
Cell Reports	BIO	857	Journal of Dairy Science	AGR	78.9
Astronomy and Astrophysics	PHY	839	PLoS ONE	MUL	77
Physical Review A	PHY	838	Journal of Biological Chemistry	BIO	76.4
eLife	BIO	830	Water Research	ENV	75.6
Molecules	CHE	813	International Journal of Molecular Sciences	PHY, CHE, BIO	75.3
Journal of Neuroscience	MED	810	Journal of Medicinal Chemistry	MED, CHE	74
Journal of Clinical Oncology	MED, BIO	807	Journal of Virology	MED	73.8
NeuroImage	MED	807	Surgical Endoscopy	MED	72.5
Environmental Science & Technology	ENV	806	Journal of the American Chemical Society	CHE	72.4
Physical Review E	PHY	804	Nano Letters	MAR, CHE	72.2
Biochemical & Biophysical Research Communications	BIO, PHY	783	BMC Infectious Diseases	MED	71.9

Conclusion

This paper presents a comprehensive analytical study to explore whether there are apparent disciplinary variations in altmetric coverage of articles. A large sample of data from Web of Science along with corresponding data from altmetric.com is obtained and analysed. Results obtained show interesting patterns. Medical Sciences and Biology account for more than 50% of all instances in Altmetrics. In terms of coverage, disciplines like Biology, Medical Science and Multidisciplinary Sciences have more than 60% of their articles covered in Altmetrics, whereas disciplines like Engineering, Mathematics and Material Science have less than 25% of their articles covered in Altmetrics. The coverage percentages further vary across different altmetric platforms, with Twitter and Mendeley having much higher overall coverage than Facebook and News. Disciplinary variations in coverage are also found in different altmetric platforms, with variations as large as 7.5% for Engineering discipline to 55.7% for Multidisciplinary in Twitter. Some journals are also found to have a higher altmetric coverage in comparison to the average altmetric coverage level of that discipline, which shows that the source of publication may also have some impact on altmetric coverage of article.

Table 5: Top 50 Journals (Sorted by 2016 Impact Factor (IF)) with corresponding WOS & Altmetric Values

Journal	*2016 IF*	*Discipline*	*TP_WOS*	*TP_ALT*	*Coverage %*
CA-A Cancer Journal For Clinicians	131.723	MED	41	27	65.9
New England Journal Of Medicine	59.558	MED	838	751	89.6
Nature Reviews Drug Discovery	47.12	MED, BIO	162	127	78.4
LANCET	44.002	MED	522	460	88.1
Nature Biotechnology	43.113	BIO	173	126	72.8
Nature Reviews Immunology	39.416	MED	126	100	79.4
Nature Materials	38.891	MAR	238	187	78.6
Nature Reviews Molecular Cell Biology	38.602	BIO	126	122	96.8
Nature	38.138	MUL	911	859	94.3
Annual Review of Astronomy and Astrophysics	37.846	PHY	760	711	93.6
JAMA-Journal of The American Medical Association	37.684	MED	683	577	84.5
Chemical Reviews	37.369	CHE	260	179	68.8
Nature Reviews Genetics	35.898	MED	120	113	94.2
Annual Review of Immunology	35.543	MED	23	21	91.3
Nature Nanotechnology	35.267	MAR	141	127	90.1
Science	34.661	MUL	1321	1132	85.7
Nature Reviews Cancer	34.244	MED	87	85	97.7
Chemical Society Reviews	34.09	CHE	267	163	61
Reviews Of Modern Physics	33.177	PHY	42	37	88.1
Living Reviews in Relativity	32	PHY	2	1	50
Nature Genetics	31.616	MED	214	190	88.8
Nature Photonics	31.167	PHY	126	109	86.5
Progress In Materials Science	31.083	MAR	37	15	40.5
Physiological Reviews	30.924	MED, SS	39	30	76.9
Nature Medicine	30.357	MED	166	159	95.8
Nature Reviews Neuroscience	29.298	MED	125	107	85.6
Cell	28.71	BIO	537	505	94
Nature Chemistry	27.893	CHE	162	156	96.3
Progress In Polymer Science	27.184	MAR	35	13	37.1
LANCET Oncology	26.509	MED	324	239	73.8
Energy & Environmental Science	25.427	ENV	314	159	50.6
Nature Methods	25.328	CHE	182	159	87.4
Nature Reviews Microbiology	24.727	BIO	107	103	96.3
Materials Science & Engineering R-Reports	24.652	MAR, PHY	10	4	40
Immunity	24.082	MED	189	160	84.7
Annual Review of Pathology-Mechanisms of Disease	23.758	BIO, MED	23	19	82.6
LANCET Neurology	23.468	MED	108	78	72.2
Cancer Cell	23.214	BIO, MED	199	164	82.4
Cell Stem Cell	22.387	BIO	147	138	93.9
Annual Review of Plant Biology	22.131	AGR, ENV, BIO	67	1	1.5
Accounts Of Chemical Research	22.003	CHE	273	160	58.6
Annual Review of Biochemistry	21.407	BIO	28	3	10.7
LANCET Infectious Diseases	21.372	MED	246	213	86.6
Journal Of Clinical Oncology	20.982	MED, BIO	877	807	92
Behavioral And Brain Sciences	20.415	MED	262	141	53.8
World Psychiatry	20.205	MED, SS	91	81	89
Cancer Discovery	19.783	MED	126	104	82.5
BMJ-British Medical Journal	19.697	MED	2896	2574	88.9
Nature Immunology	19.381	MED	158	138	87.3
Living Reviews in Solar Physics	19.333	PHY	4	4	100

There are however some limitations of this study, which can be addressed in future work. The most important of them is the fact that disciplinary tagging of articles is based on 'WC' field of Web of Science, which classifies an article into a discipline based on its source of publication and not the actual article contents. It would, therefore, be interesting to take some large data sample, classify that into different disciplines using some Machine Learning Classifier (that processes article contents to tag it into a discipline), and then see if the disciplinary variation patterns are similar to those observed in this work. This would also establish the usefulness of Web of Science publication-source based disciplinary classification. Another interesting thing to explore could be to look in detail at the data from some particular journals (that have higher altmetric coverage) and to identify if there are some specific characteristics that helps a journal in attaining higher altmetric coverage, than the typical altmetric coverage level of that discipline.

Acknowledgments

The authors would like to acknowledge the access provided to data of altmetric.com by Stacy Konkiel. The authors also acknowledge the enabling support provided by the Indo-German Joint Research Project titled 'Design of a Sciento-text Computational Framework for Retrieval and Contextual Recommendations of High-Quality Scholarly Articles' (Grant No. DST/INT/FRG/DAAD/P-28/2017) for this work.

References

Banshal, S. K., Singh, V. K., Kaderye, G., Muhuri, P. K., & Sánchez, B. P. (2018). An altmetric analysis of scholarly articles from India. Journal of Intelligent & Fuzzy Systems, 34(5), 3111-3118.

Bar-Ilan, J. (2014). Astrophysics publications on arXiv, Scopus and Mendeley: a case study. Scientometrics, 100(1), 217–225.

Chen, K., Tang, M., Wang, C., & Hsiang, J. (2015). Exploring alternative metrics of scholarly performance in the social sciences and humanities in Taiwan. Scientometrics, 102(1), 97–112.

Costas, R., Zahedi, Z., & Wouters, P. (2015a). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology, 66(10), 2003-2019.

Costas, R., Zahedi, Z. & Wouters, P. (2015b). The thematic orientation of publications mentioned on social media. Aslib Journal of Information Management, 67(3), pp. 260-283.

Gruzd, A., & Goertzen, M. (2013). Wired academia: Why social science scholars are using social media. In 2013 46th Hawaii international conference on system sciences (HICSS) (pp. 3332–3341). IEEE.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. Scientometrics, 101(2), 1419–1430.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. Journal of the Association for Information Science and Technology, 65(4), 656-669.

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. Scientometrics, 101(2), 1027–1042. https://doi.org/10.1007/s11192-014-1229-3

Htoo, T. H. H., & Na, J.-C. (2017). Disciplinary Differences in Altmetrics for Social Sciences. Online Information Review, 41(2), 235–251.

Mohammadi, E., & Thelwall, M. (2014). Mendeley Readership Altmetrics for the Social Sciences and Humanities : Research Evaluation and Knowledge Flows. *Journal of the Association for Information Science and Technology*, *65*(8), 1627–1638. https://doi.org/10.1002/asi

Ortega, J. L. (2015). Disciplinary differences in the use of academic social networking sites. Online Information Review, 39(4), 520–536.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: an extended analysis of citations. Scientometrics, 107(2), 723–744.

Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social web. First Monday, 15(7). http://firstmonday.org/ojs/index.php/fm/article/view/2874/257. Accessed June 2018.

Priem, J. (2014). Altmetrics. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact, 263-288.

Rupika, Uddin, A., & Singh, V. K. (2016). Measuring the university-industry-government collaboration in Indian Research Output. Current Science, 110(10), 1904.

Shema H., Bar-Ilan J., & Thelwall, M. (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. Journal of the Association for Information Science and Technology, 65(5), 1018-1027.

Sotudeh, H., Mazarei, Z., & Mirzabeigi, M. (2015). CiteULike bookmarks are correlated to citations at journal and author levels in library and information science. Scientometrics, 105(3), 2237–2248.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly Use of Social Media and Altmetrics: A Review of the Literature. Journal of the Association for Information Science and Technology, 68(9), 2037–2062.

Thelwall, M., & Kousha, K. (2015). ResearchGate: Disseminating, communicating, and measuring scholarship? Journal of the Association for Information Science and Technology, 66(5), 876-889.

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. Scientometrics, 108(1), 337-347.

Thelwall, M., & Kousha, K. (2017 a). ResearchGate versus Google Scholar: Which finds more early citations? Scientometrics, 112(2), 1125–1131.

Thelwall, M., & Kousha, K. (2017 b). ResearchGate articles: Age, discipline, audience size, and impact. *Journal of the Association for Information Science and Technology*, *68*(2), 468–479.

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. Scientometrics, 115(3), 1231–1240.

Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric. com scores to predict longer term citation counts? Journal of Informetrics, 12(1), 237-248.

Vogl, S., Scherndl, T., & Ku, A. (2018). # Psychology: a bibliometric analysis of psychological literature in the online media. Scientometrics, 115(3), 1253–1269. https://doi.org/10.1007/s11192-018-2727-5

Wang, X., Fang, Z., Li, Q., & Guo, X. (2016). The poor altmetric performance of publications authored by researchers in mainland China. Frontiers in Research Metrics and Analytics, 1, 8.

Williams, A. E., & Woodacre, M. A. (2016). The possibilities and perils of academic social networking sites. Online Information Review, 40(2), 282–294.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of “alternative metrics” in scientific publications. Scientometrics, 101(2), 1491–1513. https://doi.org/10.1007/s11192-014-1264-0

How international is internationally collaborated research? Heritage composition of Russia’s international collaboration network

Maria Karaulova[1] and Abdullah Gök[2]

[1]*maria.karaulova@manchester.ac.uk*
Manchester Institute of Innovation Research (MIoIR), University of Manchester, Oxford Road, M13 9PL, Manchester (UK)
[2]*abdullah.gok@strath.ac.uk*
Hunter Centre for Entrepreneurship, Strathclyde Business School, University of Strathclyde, 130 Rottenrow, G4 0GE, Glasgow (UK)

Abstract

International research performance indicators attain increased attention in science policy. More studies point to ongoing global bias in production, composition and assessment of research performance metrics (Rafols et al., 2012; van Leeuwen et al., 2001). This research examines international collaboration. It is an indicator routinely operationalised as co-authorship of articles between organisations by taking organisational address as a proxy of the collaborating country (Katz and Martin, 1997). We use geographical approximation of author heritage rooted in the morphology of the surname and find that in a significant minority of internationally collaborated papers, co-authors are likely to have the same origin. In other words, we observe an overestimation in the international collaboration indicator. The findings indicate that if a significant share of international collaborations of a national research system occurs with researchers previously affiliated with this system, internationally co-authored publications are therefore only ‘inter-national’ on a formal inter-organisational level. This contributes to the evidence that stresses more complex nature of scientific collaboration (Bozeman and Corley, 2004) and may have fundamental implications on the use of international collaboration indicator.

Introduction

In this research, we highlight and analyse the heritage bias in indicators for internationally collaborated research. International research performance indicators are seen to reflect relative competitiveness of a country in producing leading research (in terms of cited papers) and its commercialisation (in terms of assigned patents). An increasing number of studies point to ongoing global bias in production, composition and assessment of research metrics (Rafols et al., 2012; van Leeuwen et al., 2001). When funding, assessment and human resource decisions are made on the basis of indicators, we need to understand what exactly they do and do not tell us.

In science and innovation policy, especially when the issues of development are concerned, countries that have extensive links with other nations are appraised positively by international bodies and in national science policy. Internationalisation is a policy goal in itself, it is seen as important in advancing scientific research (Science Europe, 2014). International collaboration can also be a means to other policy goals. For instance, international collaboration in the European Research Area has been regarded, alongside with mobility, as a sign of increasing integration, and is also designed to level out research capacities of the Western and the Eastern European countries (EC, 2012).

International collaboration as an indicator is not a neutral metric and contains multiple assumptions about what is assumed to be a ‘scientific collaboration’. A landmark paper by Katz and Martin (1997) alongside with later research by Bozeman and Corley (2004) bring forth these assumptions. Most notably, these authors highlight how not ever co-authored publication is a result of a collaboration and not every collaboration results in co-authored publications (see further Youtie and Bozeman, 2014). As these studies expose rich underlying social processes that are not fully captured in instances of co-publication, co-authoring of research articles, books and conference papers has come to be regarded as the dominant way in which scientific collaboration is operationalised and measured bibliometrically.

International scientific collaborations identified this way have been growing at explosive rates and is widely reported (Glänzel, 2001). The assumptions behind what the growth of this metric means have been linked mainly to research inrernationalisation, knowledge exchange and other benefits for the collaborating authors, organisations and countries, because the diversity increases creativity and allows to find better solutions for complex problems (Melkers and Kiopa, 2010). International collaboration is recognised as a capacity-building factor of domestic research indicating the increase in research quality (Bornmann et al., 2015). The criticisms of the growth of internationally collaborated research focus mostly around its uneven distribution, uneven contributions by different authors, and uneven benefits that the authors in central and in the peripheral regions gain from it (Schubert and Sooryamoorthy, 2009). A further example is that patterns of international cooperation in nanotechnology are still centred on the developed countries, which are key nodes in international networks (Shapira and Wang, 2010).

In this paper, we further problematise the assumptions between the supposed gains from international collaboration and the way it is measured. An accepted view of a *bibliometric definition* of international collaboration is an instance when a research output is co-authored by two or more authors affiliated with organisations located in two or more countries. This definition of international collaboration is used in all publications we reviewed for this research. Single-authored outputs with multiple affiliations are not usually regarded as instances of international collaboration. The assumption behind this measurement is, as noted above is to say that when a scientist located in Spain co-authors an article with a scientist located in Germany, it is in effect a Spanish and a German scientists collaborating with each other. Such collaboration entails positive effects related to creativity and diversity (Gkypali et al., 2017), and improves the cohesion between the German and the Spanish research systems. The bias behind these assumptions is the objective of our research.

Research Objectives

We explore the extent and the potential implications of same-heritage international scientific collaboration, taking Russian international collaboration network as a case study. Within the broader goal to unpack the types of interactions and exchanges that occur in international collaborations, we investigate the extent and the role of the bias behind the bibliometric definition's assumption that when authors located in two countries collaborate, these authors are nationals of those respective countries (see Figure 1). This issue becomes increasingly important in the case of when the developing countries aim to connect to the global scientific knowledge flows.

First, we expect to uncover the heritage bias in international collaboration, and we expect it to be substantial. That is to say, that a significant share of internationally collaborated publications between Russia and the rest of the world means co-publication between Russian scientists in Russia and Russian scientists who reside in other countries.

Second, we will investigate the structure of this bias. As shared heritage points to cognitive proximity between two scientists and makes collaboration easier to accomplish, we expect that the extent of same-heritage collaboration will vary depending on the collaborating country, discipline, it will change with time, it will vary depending on which domestic organisation is collaborating, both in terms of the type of the organisation and its geographical location, and depending on the funding received. We also expect to see different same-heritage collaboration dynamics in the centres of excellence and among the ‘star’ scientists.

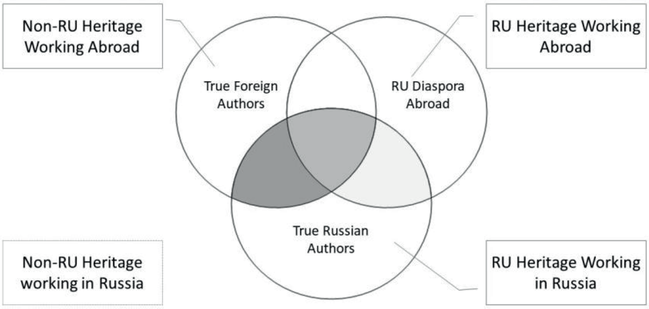

Non-RU Heritage Working Abroad
RU Heritage Working Abroad
True Foreign Authors
RU Diaspora Abroad
True Russian Authors
Non-RU Heritage working in Russia
RU Heritage Working in Russia

Figure 1 International Collaboration Bias Structure

By unpacking the inherent international collaboration bias, we, therefore, question the assumed relationship between co-authors in established international collaboration metrics. Ultimately, international collaboration indicators may point to reproduction and reinforcement of relationships between global centres and peripheries, and to knowledge channelling, rather to knowledge exchange.

Background

Heritage links memory, language and places with the construction of identity, values and communities (Smith, 2006). Heritage can be associated with the concept of human capital and links knowledge, skills accumulated in the duration of a person’s life with their behaviour. Studying heritage refers more to accounting for intrinsic and specific features of cultural capital within social groups, which make these groups distinctive in how they might approach problems and look for solutions.

In science studies, heritage has a twofold operationalisation. First, a researcher’s heritage is located in the organisations where she was first socialised into scientific profession. Through the process of socialisation, doctoral researchers learn about the rules, norms and expectations associated with doing scientific research (Knorr-Cetina, 1999). These norms are localised in organisations and countries. Second, heritage refers to the broad shared culture of social groups, which, most importantly, includes language. Second generation migrant researchers and entrepreneurs maintain their parents’ country of origin heritage and are able to act as mediators in the global innovation system (Khadria, 1999). Such broad understanding of heritage may also have an effect on the dynamics of international collaboration.

In the case of Russia, the country has developed a peculiar research and innovation system during the Soviet period, and even over 20 years after the USSR broke down, the research system is heavily path-dependent (Karaulova et al., 2016). At the same time, throughout the 1990s, Russia experienced high rates of ‘brain drain’, especially of the best recognised scientists in physics and mathematics (Graham and Dezhina, 2008). As of the mid-2000s onward, the Russian state set out the course towards internationalisation of domestic research, with the purpose to link the country with the global scientific community and benefit from knowledge exchange.

The set of related studies that analyse similar dynamics have so far focused mostly on the role of Chinese overseas diaspora in gatekeeping and/or mediating international collaboration (Freeman and Huang, 2015; Jin et al., 2007). This research has suggested some avenues of analysis, but has not discussed the implications of what the structural differences in international collaboration dynamics mean for how we should interpret science indicators.

Methodology

In our previously published research, we developed a method to identify the national heritage of authors based on the morphology of their surnames (Karaulova et al., 2019). By employing this approach, we infer Russian heritage from author surnames in internationally co-authored publications. Surname data has been used in bibliometric analyses to determine contribution of recognisable ethnic groups to the development of particular discipline (Kissin, 2011), to determine effects of inter-ethnic collaboration on quality of publications (Freeman and Huang, 2014), or to highlight the contribution of ethnic and gender minorities (Lewison, 2001). Taking Russia as a case study of this research has another benefit: in a country that was internationally isolated for the large part of the 20th century, geographical approximation of 'Russian' surnames is consistent with the actual population, i.e. most Russians still live in Russia (Revazov et al., 1986).

We take advantage of the fact that Russian surnames have persistent morphological regularities, and the majority of them can be identified from a small set of specific suffixes. Combined with first name data, this procedure has very high rates of recall and precision. By using the author name data, we can distinguish the heritage of a researcher from their work address and therefore analyse the extent to which Russian scientists in Russia collaborate with Russian scientists abroad, and what such international collaborations might mean.

We analyse publications indexed in the Web of Science that have at least one author with an address in Russia. The dataset includes 709,360 publications, the date range is 1995-2015. All disciplines, languages and document types are included in the data. Within the dataset 82.2% of publications have a co-author and 34.1% of those collaborated papers have a co-author with an address outside of Russia.

Initial Findings

Using the surname-based lexicological method, each author in the dataset is marked either as Russian heritage or non-Russian heritage. After applying the two-step Russian heritage identification procedure, we classified 95,7% of the records with the address in Russia as "Russian heritage authors", which broadly corresponds with our previous estimates that the Russian science system does not employ many non-Russian researchers (Karaulova et al., 2019).

The findings from a small pilot study of a random sample of records published in 2015 showcase that the method outlined above can be used to address our research objectives. In the random sample, we found a significant bias in internationally collaborating research: only 18% of the internationally co-authored publications were co-authored by authors in Russia and by non-Russian authors abroad, whilst the vast majority of these papers had at least one Russian heritage author based in Russia, one Russian heritage author based abroad and one non-Russian heritage author based abroad, which possibly suggests a mediation function.

Countries that have extensive international collaboration networks and are the 'core' of science globalisation, such as the USA (Wagner and Leydesdorff, 2005), have lower rates of overseas diaspora involvement in the structure of collaboration networks with Russia. While the share of research papers collaborated with the participation of Russian heritage authors reached 40% for major international partners of Russia, the results are more telling for minor partners. Countries that have relatively strong science base, but do not have traditionally close links with Russia, such as Portugal, Belgium or Australia, demonstrate very high level of overseas diaspora involvement in the share of publications co-authored with Russian scientists.

Discussion and Conclusions

If a significant share of international collaborations of a national research system occurs with researchers previously affiliated with this system, internationally co-authored publications are therefore only 'inter-national' on a formal inter-organisational level, but in fact occur between co-authors that share academic upbringing and culture. This finding contributes to the evidence that stresses more complex nature of scientific collaboration (Bozeman and Corley, 2004) and may have fundamental implications on the use of international collaboration indicator and on science policy decisions.

We found that in a significant minority of internationally collaborated papers, co-authors are likely to have the same origin. In other words, we observe an overestimation in the international collaboration indicator. This inherent bias in the established international collaboration indicator may overestimate the impact of international collaboration on periphery countries in comparison with its impact on advanced core countries. This paper makes a call for revision and further detalisation of the indicator that is sensitive to unequal science development dynamics.

When bibliometric tools are used to measure international collaboration and cooperation, invariably, assumptions are made about the social reality of these tools. Globally, the findings of this study are valid for national science policy of countries that rely on international collaboration networks to foster the development of domestic science and technology through knowledge transfer and spillovers.

References

Bornmann, L., Wagner, C., Leydesdorff, L., 2015. BRICS countries and scientific excellence: A bibliometric analysis of most frequently cited papers. Journal of the Association for Information Science and Technology 66, 1507–1513.

Bozeman, B., Corley, E., 2004. Scientists' collaboration strategies: Implications for scientific and technical human capital. Res Policy 33, 599–616. https://doi.org/10.1016/j.respol.2004.01.008

EC, 2012. Enhancing and focusing EU international cooperation in research and innovation: A strategic approach. Brussels.

Freeman, R.B., Huang, W., 2014. Collaborating With People Like Me: Ethnic co-authorship within the US (Working Paper No. 19905). National Bureau of Economic Research.

Freeman, R.B., Huang, W., 2015. China's "Great Leap Forward" in Science and Engineering, in: Geuna, A. (Ed.), Global Mobility of Research Scientists: The Economics of Who Goes Where and Why. Academic Press, pp. 155–177.

Gkypali, A., Filiou, D., Tsekouras, K., 2017. R&D collaborations: Is diversity enhancing innovation performance? Technological Forecasting and Social Change 118, 143–152. https://doi.org/10.1016/j.techfore.2017.02.015

Glänzel, W., 2001. National characteristics in international scientific co-authorship relations. Scientometrics 51, 69–115. https://doi.org/10.1023/A:1010512628145

Graham, L.R., Dezhina, I., 2008. Science in the New Russia: Crisis, Aid, Reform. Indiana University Press.

Jin, B., Rousseau, R., Suttmeier, R.P., Cao, C., 2007. The role of ethnic ties in international collaboration: The overseas chinese phenomenon. Int Soc Scientometrics & Informetrics-Issi, Leuven.

Karaulova, M., Gök, A., Shackleton, O., Shapira, P., 2016. Science system path-dependencies and their influences: nanotechnology research in Russia. Scientometrics 107, 645–670. https://doi.org/10.1007/s11192-016-1916-3

Karaulova, M., Gök, A., Shapira, P., 2019. Identifying author heritage using surname data: An application for russian surnames. Journal of the Association for Information Science and Technology 0. https://doi.org/10.1002/asi.24104

Katz, J.S., Martin, B.R., 1997. What is research collaboration? Res. Policy 26, 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1

Khadria, B., 1999. The migration of knowledge workers: second-generation effects of India’s brain drain. Sage Publications Pvt. Ltd.

Kissin, I., 2011. A surname-based bibliometric indicator: publications in biomedical journal. Scientometrics 89, 273–280. https://doi.org/10.1007/s11192-011-0437-3

Knorr-Cetina, K.D., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Boston, MA.

Lewison, G., 2001. The quantity and quality of female researchers: A bibliometric study of Iceland. Scientometrics 52, 29–43. https://doi.org/10.1023/A:1012794810883

Melkers, J., Kiopa, A., 2010. The Social Capital of Global Ties in Science: The Added Value of International Collaboration. Review of Policy Research 27, 389–414. https://doi.org/10.1111/j.1541-1338.2010.00448.x

Rafols, I., Leydesdorff, L., O’Hare, A., Nightingale, P., Stirling, A., 2012. How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. Res. Policy 41, 1262–1282. https://doi.org/10.1016/j.respol.2012.03.015

Revazov, A., Paradeeva, G., Rusakova, G., 1986. Suitability of Russian surnames as a “quasi-genetic” marker. Genetika 22, 699–704.

Schubert, T., Sooryamoorthy, R., 2009. Can the centre–periphery model explain patterns of international scientific collaboration among threshold and industrialised countries? The case of South Africa and Germany. Scientometrics 83, 181–203. https://doi.org/10.1007/s11192-009-0074-2

Science Europe, 2014. The Importance of International Collaboration for Fostering Frontier Research.

Shapira, P., Wang, J., 2010. Follow the money. Nature 468, 627–628. https://doi.org/10.1038/468627a

Smith, L., 2006. Uses of Heritage. Routledge.

van Leeuwen, T.N., Moed, H.F., Tijssen, R.J.W., Visser, M.S., van Raan, A.F.J., 2001. Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. Scientometrics 51, 335–346. https://doi.org/10.1023/A:1010549719484

Wagner, C.S., Leydesdorff, L., 2005. Network structure, self-organization, and the growth of international collaboration in science. Res. Policy 34, 1608–1618. https://doi.org/10.1016/j.respol.2005.08.002

Youtie, J., Bozeman, B., 2014. Social dynamics of research collaboration: norms, practices, and ethical issues in determining co-authorship rights. Scientometrics 101, 953–962. https://doi.org/10.1007/s11192-014-1391-7

Gender, age, and broader impact: A study of persons, not just authors

Lin Zhang[1], Huiying Du[2], Ying Huang[3], Wolfgang Glänzel[4], Gunnar Sivertsen[5]

[1]linzhang1117@whu.edu.cn
School of Information Management, Wuhan University, Wuhan (China)
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, Leuven (Belgium)

[2]dhy9596@126.com
North China University of Water Resources and Electric Power, Zhengzhou (China)

[3]huangying_work@126.com
Department of Public Administration, Hunan University, Changsha (China)
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, Leuven (Belgium)

[4]wolfgang.glanzel@kuleuven.be
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, Leuven (Belgium)
Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

[5]gunnar.sivertsen@nifu.no
Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo (Norway)

Abstract

This research-in-progress investigates whether there are differences in terms of traditional citation impact and broader impact (Mendeley, Usage and Social Media) for researchers with different social variables such as age, gender and academic position. To do this, 30,003 journal articles in *Web of Science* from 2011-2017 with Norwegian *first authors* that can be identified as *persons* in two other Norwegian national databases, are selected as our data sample. Among a total of 14,204 identifiable researchers as the first authors, 7,767 (55%) are men and 6,437 (45%) women. Percentile-based indicators of four different impacts are compared among different researcher cohorts. The preliminary results show that although there is a consistent male dominance in traditional citation impact, the dominance is less present in the younger cohorts where females are also better represented. Female researchers score relatively higher on broader impact indicators than on traditional citation impact indicators, sometimes clearly exceeding their male colleagues. The differences already revealed will be further studied by including academic position as another social variable, to better understand the distinct impact characteristics according to age and gender, and to provide new insights for funding and evaluation policy.

Introduction

In an earlier study (Zhang & Sivertsen, 2017), we found that while the average publication productivity of researchers increases with seniority, average citation impact does not. Citation impact was the highest for post-docs and in general for young researchers in their 30's in our study of 17,750 researchers in Norway. Men were on the average more productive than women, but the difference in citation impact was smaller. Previous studies (e.g., van den Besselaar & Sandström. 2016; Larivière & Costas (2016); Barrios et al., 2013; Aksnes et al., 2011) had found that productivity and citation impact are correlated, which might imply that funding should primarily follow productivity (Sandström & van den Besselaar (2016) or that individuals are best assessed by qualitative peer review to avoid strengthening the accumulative advantages of senior researchers (Larivière & Costas, 2016). We instead concluded that the opportunities to be productive within WoS varies with social variables, and that lower productivity among women and among researchers in early careers needs not inhibit high citation impact.

We reached this conclusion by using different methods. While the previous studies were based on author name disambiguation within data from WoS, we were able to study persons, not just authors. We matched WoS data with two other data sources at a national level where the author names and addresses in the publications could be linked to real persons and institutions. From these other data sources, we also knew the age, gender and academic position of the researchers.

In this new study, we extend the perspective by looking at the *broader impact* of the publications of the same kind of sample of researchers. We extend our datasets to: 1) include all Norwegian researchers when they appear as first authors in journal articles in Web of Science (WoS, 2011-2017); 2) add the altmetric data for each article under study from *PlumX Metrics* created by Plum Analytics. As research-in-progress, we will present here the results from using the social variables *age* and *gender*.

We define the term *broader impact* very narrowly to reflect the limitations of our data sources (see also the data and methods section below):

- The publications we study need to be scientific and indexed in WoS. We measure their scientific impact on the basis of their citations within WoS.
- The publications need to be matched to *PlumX Metrics* data with their DOI (Digital Object Identifier).
- Selected statistics from *PlumX Metrics* are aggregated into three major broader impact indicators named *Mendeley, Usage* and *SocialMedia* (explained in the methods section below).

The term *broader impact* is thereby used for possible traces of impact that the scientific publications might have beyond becoming references in new scientific publications, as measured by citations within WoS. Our three broader impact indicators, Mendeley, Usage and SocialMedia, are meant to reflect different types of influence beyond scientific impact. However, although the three indicators represent different kinds of influences from usefulness in research via academic readership to broader attention and discussion in social media, it is still important to acknowledge their limitations regarding what is usually understood as the *broader* or *societal* impact of research.

Broader impact is usually defined much more broadly, e.g. by the National Science Foundation of the USA as ‘the potential to benefit society and contribute to the achievement of specific, desired societal outcomes’.[1] With another word and a similar definition, the Research Excellence Framework of the UK defines *societal impact* as ‘an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia.”[2] In our view, our three broader impact indicators, and in general so-called *altmetrics* (Wilsdon et al., 2017), can only in a very limited way support such broad definitions of broader or societal impact with relevant data. The reason is that these indicators are based on the influence of *scientific publications*. Altmetrics essentially measures readership, or the direct influence, of these publications (Haustein, 2014). However, the interaction between research and society is seldom mediated directly through scientific publications, and not even through the written words of scientists (Gibbons et al., 1994). Such interactions need to be studied with the use of data sources reaching beyond written communication (Bornmann, 2013). However, indicators of online views, downloads, readership, and social media attention can supplement traditional bibliometrics for the understanding and evaluation of research activities (Bornmann, 2014).

[1] https://www.nsf.gov/od/oia/publications/Broader_Impacts.pdf

[2] https://re.ukri.org/research/ref-impact/

Data and methods

Our study depends on the matching of four data sources at the individual article and author/researcher level:

1. *National Citation Report for Norway (NCR, 1981-2017)*, a data set provided by Clarivate Analytics with a representation of all articles in Web of Science with minimum one address in Norway, and their accumulated citation counts.
2. *The Norwegian Science Index (NSI)*, a subset of the Current Research Information System in Norway (Cristin), with complete coverage since 2010 of all peer-reviewed scientific and scholarly publication outputs from Norwegian institutions, including books, edited volumes, and conference series (Sivertsen & Larsen, 2012). This database is used to match data sources 1 and 3, and also to match data sources 1 and 4 with DOIs not registered in Web of Science.
3. *The Norwegian Research Personnel Register (NRPR)*, which has been updated since 1977 for all researchers at public research institutions in Norway, including age, gender, educational background, affiliations and positions.
4. *The PlumX Metrics*, which provides insights into the ways people interact with individual pieces of research output in the online environment. It categorizes metrics into 5 separate categories: Citations, Usage, Captures, Mentions, and Social Media.

As a first step, we selected 30,003 journal articles in *NCR* from 2011-2017 by using four criteria: Firstly, they were covered by NCR. Secondly, at least one of the authors was affiliated with one of Norway's four largest universities (Oslo, Bergen, Trondheim, Tromsø). The time and resources to perform research are equal among these four institutions. Thirdly, the publications could be linked to *PlumX Metrics* through their DOI number in NCR in order to obtain their broader impact information. The broader impact data for each publication was retrieved from *PlumX Metrics* by the end of August 2018. The fourth criterium was to include only publications with Norwegian *first authors* that can be identified as *persons* in NSI and NRPR. We selected only first author publications for the same reason as in Thelwall (2018a; 2018b): First authors most often perform most of the research. Most of the articles in our data are multi-authored with a representation of both men and women. To study gender differences in impact more distinctly, we chose to include only those publications where the persons we study are first authors.

The four selection criteria resulted in a total of 14,204 identifiable persons as the first authors. Among these, 7,767 (55%) are men and 6,437 (45%) and women. The persons are placed in five-year age cohorts according to their age in the publication year. The first and last cohorts are extended to include a few outliers. Table 1 shows how they are distributed by major area of research, gender and age cohorts. The classification of persons in major areas of research was done according to the journal-based classification of publications in NSI, which consists of four major areas and 82 subfields. Persons were placed in only one major area each. If in doubt, we also consulted their organizational affiliations (e.g. a department of sociology would place the researcher in the category Social Sciences).

Table 1. Gender and age cohort distribution in four major area of research

	Health Sciences		Natural Sciences and Engineering		Social Sciences		Humanities	
Age cohort	Women	Men	Women	Men	Women	Men	Women	Men
21-30	662	445	573	1082	33	36	16	9
31-35	830	582	580	1045	65	60	15	19
36-40	775	632	302	675	61	69	30	27
41-45	649	472	187	373	64	57	23	24
46-50	392	306	91	202	37	52	16	22
51-55	377	280	45	157	40	41	12	15
56-60	249	222	35	137	28	52	13	18
61-65	120	154	7	92	19	33	8	9
66+	61	189	5	132	13	35	4	12
Totals	4115	3282	1825	3895	360	435	137	155
Gender balance	56 %	44 %	32 %	68 %	45 %	55 %	47 %	53 %

After testing the *PlumX Metrics* data with regard to reliability and validity for meaningful statistical analysis (several indicators have low frequencies), we selected three different metrics by aggregating specific indicators from *PlumX Metrics*:

1. The 'Mendeley' indicator is based on the number of readers a paper has had in Mendeley.
2. The 'Usage' indicator represents the frequency of the users' abstract view or full-text view. The two frequencies are summed up.
3. The 'SocialMedia' indicator represents the number of times a publication is referred to in Twitter and Facebook. The two frequencies are summed up.

We could compare these three *broader impact* indicators with a fourth indicator of *scientific impact* that could be readily retrieved from the NCR database: Web of Science citations, normalized by WoS subject category and year, and expressed as percentiles after ranking the articles from the most cited to the least cited. The percentile indicator in NCR compares to all articles in the world in the same subfield and year. We adopted the same *percentile* method for the indicators based on PlumX data. Here, we could only measure the percentiles within our data, a subset of Norwegian publications covered by our matched data sources. We normalized by the year and major area of research of the *publication,* using the NSI classification with four major areas and 82 subfields mentioned above. Within each major area of research, we selected and tagged the top 10% papers according to impact in each of the three categories 'Mendeley', 'Usage' and 'SocialMedia'.

For all four indicators, we measure a *cohort's share* in the *top 10 percent high impact publications* (according to the four indicators) divided by the *same cohort's general share in all publications*. Values above, below or equal to 1 will therefore show whether the cohort in focus is over, below or equal to the impact performance that can be expected in general among its peers. With the term 'cohorts', we refer to any selected groups of researchers by the options that are shown in table 1. Since the persons in our data are placed in one group only, the groups may also be aggregated without duplicating.

Results

The aim of this research-in-progress paper is to present the most immediately interesting results that inspires our further investigations. As seen in Figure 1, the impact of female researchers is relatively higher as we move from traditional citation impact to impact in Mendeley. This is as expected from the results in Thelwall (2018a). One of the possible reasons may be that female researchers are more oriented towards the educational relevance of their research, which might attract more student audiences (Thelwall, 2018b).

As we look beyond traditional citation impact and Mendeley in our study, we observe, to our surprise, that there is an even stronger tendency among female researchers to have relatively higher impact on usage (online views) and mentions of their papers in social media. The difference by gender is particularly large on the usage indicator. Our further investigation will seek explanations for this by going deeper into field-specific patterns. We will also include the academic position of the researchers as a variable.

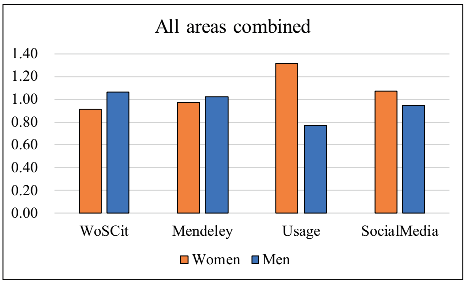


Figure 1. Gender and impact*

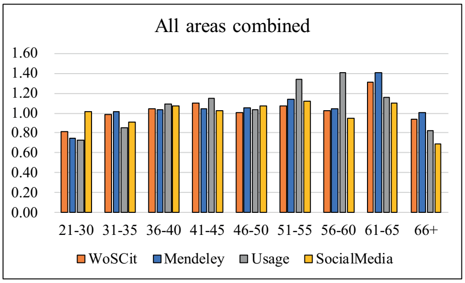


Figure 2. Age and impact

* All areas combined. The vertical axis indicates the shares in top 10% high impact publications compared to (divided by) the general shares in all publications. Shares have been calculated for all according cohorts for each impact indicator. The same applies for Figures 2-6.

Our aggregated results regarding the age of the researchers are presented in Figure 2. Here, the general patterns seem to imply that impact in research by three of the indicators, citation impact, Mendeley impact, and usage impact, comes with age. Impact in social media may be high also among younger researchers. We will investigate further why the four impact indicators have very deviating results in some age cohorts (researchers younger than 31 or older than 50).

Figure 3 shows the most impressive result of our preliminary investigations. Why are the publications from female first authors consistently more attracted by online views (full text or abstracts) than publications by male first authors? And why are the older female researchers, and not the older male researchers, so clearly more outstanding than any other cohorts by this indicator? Again, we need to get down to research fields, academic positions, and possible purposes and uses of research, to explain these impressive findings.

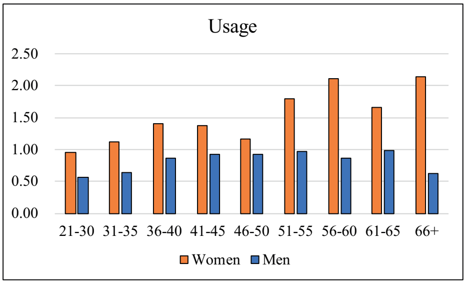


Figure 3. Age, gender and usage impact

Preliminary conclusion and further research

Although there is a consistent male dominance in traditional citation impact, the dominance is less present in the younger cohorts where females are also better represented. Female researchers score relatively higher on broader impact indicators than on traditional citation impact indicators, exceeding their male colleagues most clearly by usage impact. The preliminary results already provide evidences that it is worthwhile looking at the performance of young female researchers before all the funding is prioritized towards older male professors. In our further research, we will study each major area of research and also go down to subfields to find possible explanations for these impressive findings. We will look carefully at the publications with the highest impact by all four indicators. We know already that these publications are not the same across the four indicators. We will characterize what type of research and research purposes that have the highest impact by different measures. We will also study the two social variables, gender and age, more closely, and we will add academic position as a third social variable.

Acknowledgements

This work is supported by the National Natural Science Foundation of China: Grants 71573085, and by the Research Council of Norway: Grant 256223 FORINNPOL.

References

Aksnes, D.W., Rørstad, K., Piro, F., Sivertsen, G. (2011). Are female researchers less cited? A large‐scale study of Norwegian scientists, Journal of the Association for Information Science and Technology, 62(4): 628-636.

Barrios, M., Villarroya, A., Borrego, Á (2013). Scientific production in psychology: a gender analysis[J]. Scientometrics, 95(1): 15-23.

Bornmann, L. (2013.) What is societal impact of research and how can it be assessed? A literature survey. Journal of the American Association for Information Science and Technology, 64(2): 217-233.

Bornmann, L. (2014.) Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. Journal of Informetrics, 8(4), 895-903.

Gibbons et al. (1994) The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies, London: Sage.

Haustein, S. (2014). Readership metrics. In B. Cronin & C. R. Sugimoto (Eds.), Beyond bibliometrics: harnessing multi-dimensional indicators of performance (pp. 327-344). Cambridge, MA, USA: MIT Press.

Larivière, V. & Costas, R. (2016). How many is too many? On the relationship between research productivity and impact. PLoS ONE 11(9): e0162709. doi:10.1371/journal.pone.0162709.

Sandström, U. & van den Besselaar, P. (2016). Quantity and/or quality? The importance of publishing many papers. PLoS ONE 11(11): e0166149. https://doi.org/10.1371/journal.pone.0166149.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. Scientometrics, 91(2), 567-575.

Thelwall, M. (2018a). Do females create higher impact research? Scopus citations and Mendeley readers for articles from five countries. Journal of Informetrics, 12(4), 1031-1041.

Thelwall, M. (2018b). Does Female-authored Research have More Educational Impact than Male-authored Research? Evidence from Mendeley. Journal of Altmetrics, 1(1). doi:10.29024/joa.2

van den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: a longitudinal case study. Scientometrics, 106(1), 143-162.

Zhang, L. Sivertsen, G. (2017). Productivity versus citation impact: A study of persons, not just authors. 16th International Conference on Scientometrics & Informetrics (issi2017.org), Conference Proceedings: 970-975.

Assessing the Impact of a Highly-Cited Paper

Paul Alkemade

p.f.a.alkemade@tudelft.nl
Applied Physics, Lorentzweg 1, 2628 CJ, Delft University of Technology, Delft, The Netherlands

Abstract

The use of citation indicators to gauge impact in science has increased significantly in the last decades. Yet, in the normal sense of the word, the impact of a scientific paper is still largely elusive. This work presents the results of full textual analysis of all citations of a single highly-cited paper. For this purpose, a reference impact score *RIS* is defined, which is derived from the context and content of the citations. First results suggest that the *RIS* reflects the impact of the cited paper on the research in its citing articles. Low-*RIS* citations in particular are often perfunctory and redundant, whereas high-*RIS* citations reveal regularities that might be related to impact.

Introduction

A cornerstone of science is peer review. Journal editors ask reviewers to estimate among others the expected impact of a submitted manuscript within or beyond its field of research. But reviewers rarely assess the *actual* impact of papers. This activity is usually handled many years later by bibliometricians, who can apply various statistical methods for the assessment of impact. The most known method is citation counting. The interest nowadays in the role of citations in science, scientific achievements, and science policies is large and to some degree controversial, see *e.g.* the San Francisco Declaration (2012). Some bibliometricians hold the normative view that citation counts correlate well with the impact of scientific papers, authors, and journals. Others, in contrast, adhere to the social-constructivist view that there is a plethora of motives for citing or not citing and argue that also external factors influence citing. Read Tahamtan (2018) for a recent discussion of these opposing views.

Some assume that the impact of a paper can be best derived from textual analysis of the context and content of citations. In the 1970s, Murugesan (1978) proposed a method to categorise citations in five opposing dichotomies, such as 'conceptual versus operational' and 'confirmative versus negational'. Alternative schemes have been explored as well, usually to limited data sets because textual citation analysis is laborious and cannot be automated easily. For an overview, read Ding (2014) or Bornmann (2012). An example of a rare *normative* study on citation context and content is by Maricic (1998), who analysed citations to a few hundred papers from one institute and ranked the citations according to location, frequency, and importance.

In this work I present the results of a textual analysis of 145 citations in 102 articles to a single paper by Abbas (2014) in a leading journal in Nanoscience and -technology, the field of my expertise. The citing articles[1] in the period of January 2014 to December 2018 are collected via Web of Science and Google Scholar. In particular, I introduce an appraisal or scoring scheme for the impact of the cited paper on each citing article, based on citation characteristics that have been identified in the literature as likely being related to the impact of publications. This score, called the *reference impact score* (*RIS*), quantifies and combines these characteristics into a single number. Although the presented analysis is based upon a subjective interpretation by one person, I will hypothesize that, after harmonization and validation, the *RIS* can reveal elements of the impact of publications on the advancement of science that are otherwise barely discernible.

[1] Here, the cited work is called 'a paper' and the citing work 'an article'. 'A reference' is an item in the reference list of an article.

The reference impact score

The *RIS* method assigns a value, a reference impact score *RIS*, to one reference in a citing article. Table 1 gives an example of a *RIS* scoring scheme for four characteristics of a citation, related to its context and content. The choices for scores are personal, but they have I hope a more general usefulness, at least in my own discipline. First of all, the *RIS* depends on the location of the citation in the article. Citations in the introduction section are part of a sketch of the relevant research field. Few criteria exist for citing in introductions and the number of citations in introductions is usually high (Boyack 2018). Hence, a rather low score of 2 points is given to citations in the introduction. I assume that the conclusion section is the climax of an article. The intermediate sections describe the actions, observations, and interpretations that provide evidence for the conclusions. Therefore, the score gradually rises to 7 points for citations in the conclusion section. The exception are citations that describe the motive or aim of the article, most often at the end of the introduction, but logically linked to the conclusion. Their scoring is 5 points. Review articles have a completely different structure and, therefore, a general score of 3 is giving, independent of the citation's location.

Table 1. Example of a scoring scheme for citations. The score depends on the context (categories I, III and IV) and function (category II) of the citations in the narrative of a citing article. The reference impact score *RIS* is the highest citation score of a citing article plus bonuses for additional citations. Review articles in category I get score 3. Category II has double weight *w*.

Category	*Label*	*Meaning*	*Score*[2]
I.	i	introduction section: background sketch	2 / 3
'Location'	q	introduction section: aim or motive	5 / 3
(*w*=1)	m	methods section	3 / 3
	r	results section	4 / 3
	rd	results & discussion section	5 / 3
	d	discussion section	6 / 3
	c	conclusion & summary section	7 / 3
II.	o	using a graphical object	1
'Function'	i	introducing the reader	2
(*w*=2)	m	referring to the method	3
	s	stating an opinion	4
	r	presenting a result of the paper	5
	d	discussing a topic in the paper	6
III.	1	sentence with a single reference	6
'Grouping'	1-2	sentence with one group of multiple references	4
(*w*=1)	2-1	sentence with multiple groups of single references	4
	2-2	sentence with multiple groups of multiple references	2
IV.	few	< 41 (<120 for reviews)	3
'# of refs.'	average	41...50 (121...200 for reviews)	2
(*w*=1)	many	>50 (>200 for reviews)	1
'Reference impact score'		*RIS* = highest citation score	var.
		+ bonus for 2nd highest-scoring citation (if present)	5
		+ bonus for 3rd highest-scoring citation (if present)	3
		+ bonuses for all other citations (if present)	2

Each citation has a function (category II) in the narrative of the article, sometimes important sometimes not. Using a graphical unit from one of the cited paper's figures is rarely important, hence 1 point. Introducing a reader to the research field is, as argued above, useful but most citations are not very important. Also the score for function rises gradually, up to 6 points for a citation that discusses a particular topic. Assigning scores for function is rather difficult,

[2] The first score is for experimental and theoretical articles, the second for reviews.

because it depends both on the interpretation of the content of the citation and on the value system held in the specific discipline. Scoring of the grouping of the citation between other citations in the sentence (category III) is simpler. A citation like '*One can use the method by Young et al. {11}*' has more importance (6 points) than one with joining citations like *'One can use the Young method {11-14}'* (4 points). Also multiple groups of citations, like *'One can use the method by Young {11} or the slow-electron method {15}'*, receive 4 points. If the citation is grouped with other citations and there are multiple citation groups, *e.g. 'One can use the Young method {11-14} or the slow-electron method {15-17}'*, the score for reference {11} is 2 points. The importance of a cited paper is relatively low if the reference list is long. Therefore, the total number of references is another category (IV). Three different scores are possible here. In this example, the bounds are such that each class comprises one third of the citing articles.

The scores of categories I-IV are combined with weights into a total citation score. A paper that is cited multiple times in an article is obviously more important than one that is cited only once. One could add the individual scores, but repeating a statement that was already true in *e.g.* the introduction, does not make it more true. Therefore, the encompassing reference impact score *RIS* of a reference in an article is defined as the score of its highest scoring citation plus bonuses for possible additional citations. For instance, if the first citation is characterized by *'i m 2-1 few'*, the score is 2+2x3+4+3=15. If the second citation is *'i d 2-2 few'*, the score is 2+2x6+2 +3=19. Hence, the reference impact score *RIS* becomes 19+5=24.

Results

Table 2a presents as example six citations in four articles. In general, a citation is a phrase, a full sentence, or a few sentences. Interpretation of the content of the sentence(s) determines which parts belong to the citation; they are printed in black or red in Table 2a. Fragments that are not part of the citation are grey. The citation is red if it negates an observation, interpretation, or conclusion in the cited paper. The analysis for these six citations is in Table 2b. Columns 3–12 give the corresponding dichotomous Murugesan and Moravcsik (MM) classification. Column 13 lists the location of the citation, the meaning of the labels are in Table 1. Column 14 is the interpreted function of the citation. Column 15 denotes the grouping of the citation

Table 2a. Selection of citations of Abbas (2014) by (1) Mondal (2014); (2) Fox (2015); (3) Rivera (2016); and (4a-c) Stanford (2017). Black and red fragments are citations, grey not. The red citations are in variance with Abbas (2014).

No.	*Citation*
1	Thus, in order to make it a material of same caliber as silicon, engineering of the band gap of graphene becomes essential [7-11,12].
2	For example, graphene nanoribbons with dimensions below 10 nm have been fabricated by this method [32,33,34]. This ribbon width may be useful in order to exploit the properties of quantum confinement. However, the nanoribbons must also have well-defined edge orientations and good crystallinity. Whether this can be achieved by He^+ milling has not yet been investigated due to a lack of understanding and control of the beam-sample interaction.
3	In the high-confinement regime, the radiation into plasmons is dominant and can be harvested as far-field light through suitable outcoupling techniques such as gratings and nanoantennas [39, 44,45,46].
4a	Abbas et al. have used He patterning to create arrays of 5nm GNRs [133] as shown in Fig. 22(a). However, Raman spectra shown in Fig. 22(b) of the GNRs indicate a significant amount defects in the arrays from the patterning process that results in a rise of the D/G peak ratio, since the D peak is forbidden in defect-free graphene. This could ultimately limit the application of He-milled GNRs for high performance FET applications.
4b	*(Copied figure)* [133].
4c	*(Table element of graphene processing methods)* [133].

Table 2b. Citation classification and scoring. Column 2: E=experimental, T=theoretical, R=review. Columns 3–12 refer to the MM classification of Murugesan (1978); columns 13–16 refer to the *RIS* scheme of Table 1; column 17 is the topic of the citation (a=application, f=fabrication, p=physics). Column 18 is the citation score, derived from columns 13 to 16, and column 19 is the reference impact score *RIS*. Note that not all boxes are filled.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
	article type	*conceptual*	*operational*	*organic*	*perfunctory*	*evolutionary*	*juxtapositional*	*confirmative*	*negational*	*indispensable*	*redundant*	*location*	*function*	*grouping*	*# of refs.*	*topic*	*citation score*	***RIS***
1	E	1			1	1		1			1	i	d	1-2	60	a	11	**11**
2	E		1		1	1			1		1	q	d	1-2	48	p	23	**23**
3	T		1		1			1			1	d	i	1-2	50	a	16	**16**
4a	R								1			-	d	1	160	f	24	**32**
4b								1				-	r	1	160	f	20	
4c								1				-	m	2-2	160	f	14	

and column 16 the total number of references in the citing article. Column 17 gives the topic of the citation (there is no scoring for the topic); 'f' means the fabrication method in the cited paper, an 'a' the application, and a 'p' means a physical phenomenon. The numbers in the final column 19 are the reference impact scores *RIS*, based on columns 13–16 according to the scoring and weighting scheme of Table 1. Here, the article by Stanford (2017) has three citations to Abbas (2014). For this article, *RIS*=(4+2x6+6+2)+5+3=32.

Table 3 presents the results of the dichotomous MM classification, together with the original results by Murugesan (1978). Because reviews have a different structure and function, only the experimental and theoretical articles are considered here. Dichotomous classification is often difficult, because there is rarely a sharp boundary between the class pairs. A slightly different formulation could have shifted a citation into the opposing class. Nevertheless, I deliberately did not use 'not indexed', because the boundary between classifiable and non-classifiable is even vaguer. Theoretical articles are evolutionary with regard to experimental articles –one follows from the other and vice versa– and at the same time juxtapositional –because a theoretical approach is an alternative to an experimental one. Therefore, this dichotomy has not been indexed for theoretical articles. Review articles, for which only the confirmative-negational dichotomy is used, are not included in Table 3.

Table 3. Citation characteristics for the MM classification. First row: 3500 citations in physics journals around 1975 by Murugesan (1978). Second row: 99 citations to Abbas (2014) in this work. *r/a* = mean number of references per article. Numbers in column 3 and later are percentages.

year	*r/a*	*conc.*	*oper.*	*n.i.*	*orga.*	*perf.*	*n.i.*	*evol.*	*juxt.*	*n.i.*	*conf.*	*nega.*	*n.i*	*indi.*	*redu.*	*n.i.*	
1975	16	27	64	9	39	60	1	46	51	3	89	9	2	<66	34	?	%
2014	46	10	90	0	18	82	0	32	41	27	86	14	0	7	93	0	%

An unmistakable difference between the 1975 and 2014 analyses is the tripling of the number of references per article (*r/a*). Boyack (2018) found almost a doubling between 2000 and 2015, up to 39 references per article. Also the percentages of operational, perfunctory (=of short duration), and redundant citations are now much higher. This rise might be related to the increase in references per article.

Table 4. Summary of the *RIS* analysis of 101 experimental, theoretical, and review articles, that cite Abbas (2014). a) gives citations counts for categories I to III. b) gives topics ('res'=rest); and c) gives citing article counts per year (2014-18) for the ~50% lowest and highest *RIS* articles.

a)

Location		**i**	**q**	**m**	**r**	**rd**	**d**	**c**
Exp.		48	3	2	3	4	10	2
Theor.		16		2	1	2	6	
Review								
Function	**o**	**i**	**q**	**m**	**r**	**s**	**d**	**c**
Exp.		49		1	2	2	18	
Theor.		23			1		3	
Review	3	13		3	9	13	5	
Grouping	**1**	**1-2**	**2-1**	**2-2**	**-**			
Exp.	10	28	8	26				
Theor.	4	9	4	10				
Review	23	8	2	10	3			

b)

Topic	**f**	**a**	**f+a**	**p**	**res**	***sum***
Exp.	43	17		12		72
Theor.	20	7				27
Review	32	7	1	3	3	46

c)

Counts	**'14**	**'15**	**'16**	**'17**	**'18**	***sum***
E, low RIS	4	8	3	5	8	28
high RIS	3	6	7	7	2	25
T, low RIS	3	3	2	3	1	12
high RIS	2	3	2	3	3	13
R, low RIS	0	2	3	3	4	12
high RIS	0	2	6	3	1	12
low RIS	7	13	8	11	13	52
high RIS	5	11	15	13	6	50
All	12	24	23	24	19	102

Table 4 summarizes the statistics of the *RIS* analysis. There are 72 citations in 53 experimental articles. In all theoretical articles but one, there are only single citations, whereas review articles have on average almost two citations per reference. Most citations appear in the introduction and the second most in the discussion section. This characteristic is stronger for experimental than for theoretical citations. Similarly, the most abundant function of a citation is to introduce the reader, then to discuss an issue. The introductory function is very strong for theoretical citations. The function of review citations exhibits a much wider spectrum. Grouping of citations is common in experimental and theoretical articles, but not in review articles.

In Figure 1 the *RIS* of the articles is shown as function of time delay between the publication dates of the cited paper and its citing articles. Different symbols and colours indicate different citation characteristics.

Discussion and conclusion

Although the number of analysed cited papers in this work is absolutely low –unity– the number of citing articles is rather high, ~100. This moderately high number allows for statistical analysis of differences in citing behaviour between articles. I have assigned a reference impact score (*RIS*) to the context and content of the citations. My hypothesis is, that authors of articles with a higher *RIS* are more influenced by the cited paper than others. I found indications that this is true for same-type (namely experimental) articles. Ignoring self-citations, the first high-*RIS* articles appeared with a delay of 11 months, see Figure 1a and Table 4c. After four years, the experimental high-*RIS* articles started to dwindle. Citing review articles appeared with a similar delay, probably due to a long production process. Figure 1a shows that most organic (+), negative (purple rim), and indispensable (+) citing articles have a high *RIS*. Figure 1b gives a mixed picture: organic citing theoretical articles have a low *RIS*, whereas negative citing reviews have a high *RIS*. These trends suggest that the *RIS* is indeed a meaningful indicator for the impact of cited papers on the research activities of citing authors. Besides, the impact on the majority of the citing authors seems low: perfunctory and redundant citations dominate, impairing the validity of citation counting as a measure of quality.

This analysis of all citations to one highly-cited paper by one interpreter has limited general validity. Hopefully, the attribution of scores to the various characteristics of citations (Table 2) can be harmonised within a discipline of science, thus with sufficient consensus on the relative importance of the characteristics. Furthermore, the categorisation of citations (Table 3) must be

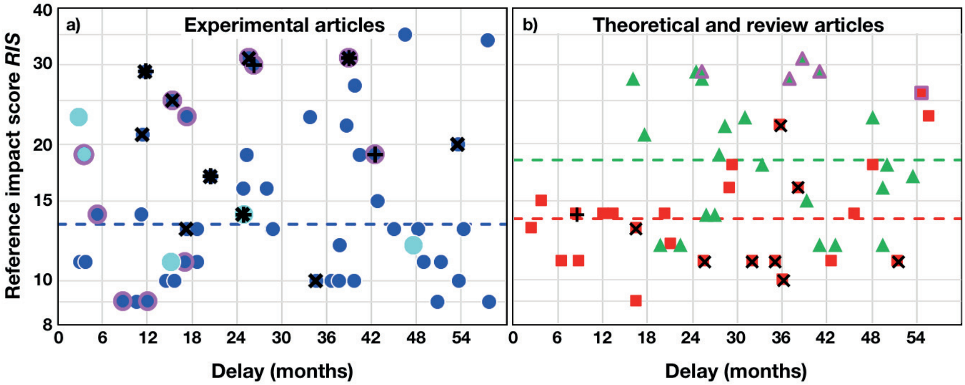


Figure 1. Reference impact score *RIS* versus time delay between the publication of cited and citing works. a) experimental articles (light-blue: self-references). b) theoretical (red) and review (green) articles. Purple-rim: negational; +: indispensable; x: organic. Dashed lines are medians, separating the low and high *RIS* articles. The vertical scale is plotted logarithmically.

validated such that categorisations by different specialists will be sufficiently correlated. Peers play a crucial role in the publication of scientific papers and with their expertise they could also appraise citations and uncover impact. Often, however, they have limited interest in 'old' papers, their expertise concerns a small field of science, and –not unimportantly– they have limited time. Computational linguistics, *e.g.* along the lines of Taskin (2018), is a very different approach to textual citation appraisal. Maybe, a clever combination of peer appraisal and computational linguistics can reveal impact in the normal sense of the word.

References

Abbas, A.N. *et al.* (2014). Patterning, characterization, and chemical sensing applications of graphene nanoribbon arrays down to 5 nm using helium ion beam lithography. *ACS Nano*, 8, 1538-1546.

Bornmann, L. *et al.* (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics* 6, 11–18.

Boyack, K.W., van Eck, N.J., Colavizza, G. & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, 12, 59–73.

Ding, Y. *et al.* (2014). Content-Based Citation Analysis: The Next Generation of Citation Analysis. *Journal of the Association for Information Science and Technology*, 65, 1820–1833.

Fox, D.S. *et al.* (2015). Nanopatterning and electrical tuning of MoS_2 layers with a subnanometer helium ion beam. *Nano Letters,* 15, 5307-5313.

Maricic S. *et al.* (1998). Citation Context versus the Frequency Counts of Citation Histories. *Journal of the Association for Information Science and Technology*, 49, 530–540.

Mondal, T. *et al.* (2014). Stress generation and tailoring of electronic properties of expanded graphite by click chemistry. *ACS Applied Materials & Interfaces*, 6, 7244-7253.

Murugesan, P. & Moravcsik, M.J. (1978). Variation of the nature of citation measures with journals and scientific specialities. *Journal of the American Society for Information Science*, 17, 141-147.

Rivera, N. *et al.* (2016). Shrinking light to allow forbidden transitions on the atomic scale. *Science,* 353, 263-269.

San Francisco Declaration (2012). Retrieved May 29, 2019 from: https://sfdora.org/read/

Stanford, M. *et al.* (2017). Review Article: Advanced nanoscale patterning and material synthesis with gas field helium and neon ion beams. *Journal of Vacuum Science & Technology B*, 33, 030802.

Taskin, Z. & Al, U. (2018). A content-based citation analysis study based on text categorization. *Scientometrics*, 114, 335–357.

Tahamtan, I. & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, 12, 203–216.

Paragraph-based intra- and inter- document similarity using neural vector paragraph embeddings

Bart Thijs[1]

[1] *bart.thijs@kuleuven.be*
KU Leuven, ECOOM, FEB, Naamsestraat 61, 3000 Leuven (Belgium)
Univ Grenoble Alpes, LIG, SIGMA, 38000 Grenoble (France)

Abstract

Science mapping using document networks is based on the assumption that scientific papers are indivisible units with unique links to neighbour documents. Research on proximity in co-citation analysis and the study of lexical properties of sections and citation contexts indicate that this assumption is questionable. Moreover, the meaning of words and co-words depends on the context in which they appear. This study proposes the use of a neural network architecture for word and paragraph embeddings (Doc2Vec) for the measurement of similarity among those smaller units of analysis. It is shown that paragraphs in the 'Introduction' and the 'Discussion' section are more similar to the abstract, that the similarity among paragraphs is related to -but not linearly- the distance between the paragraphs. The 'Methodology' section is least similar to the other sections. Abstracts of citing-cited documents are more similar than random pairs and the context in which a reference appears is most similar to the abstract of the cited document. This novel approach with higher granularity can be used for bibliometric aided retrieval and to assist in measuring interdisciplinarity through the application of network-based centrality measures.

Introduction

Document networks with weighted edges based on similarities using either citation links (Small, 1994), lexical similarity (Wang & Koopman, 2017) or combinations (eg. Ahlgren & Colliander, 2009; Thijs & Glänzel, 2018) and unweighted approaches using direct citations links (eg. Boyack, 2017) have been used for science mapping exercises. In these studies, documents were assumed to be indivisible units, each with unique links to neighbouring nodes holding a single value indicating the strength of the similarity in the case of a weighted variant. Then, these networks were subject of community detection or clustering approaches which resulted in hard clustering assigning documents to single groups or clusters of papers. However, it becomes more and more clear that the basic underlying assumption in these models is questionable. A document is often not to be reduced to such a single point or entry in the knowledge space.

A recent paper (Thijs & Glänzel, 2018) using full texts from the journal Scientometrics identified papers that shifted easily from one cluster to another after slight changes in the weighting parameter in the combined citation-lexical approach. A set of papers on institutional performance was split into two groups with the first focusing on university ranking and name disambiguation problems associated with this topic and the latter one related to institutional performance in social sciences. The first group was merged into the topic of '*Research Assessment*' while the latter one becomes part of the set of papers labelled as '*Field and Regional Studies*'. Both these clusters were not specifically labelled as dealing with institutional research. It is impossible to indicate whether one or the other grouping was better. Both had similar quality scores.

A similar observation was made by Boyack (2017) when he compared local cluster solutions in the field of Astronomy with the topics identified in his global science map. Several typical Astronomy topics from the global map did contain publications that were not in the initial data set due to their lack of compliance with the retrieval strategy. Other papers from the initial set were to be found in topics that were clearly not primarily on Astronomy. These papers had a large portion of their links to non-Astro papers. Working on the same data set, I identified

papers studying the effect of absence of gravity on the growth of plants connecting both agriculture with astronomy (see Kiss et al., 2014 as an example).
This leads to the proposition that a more fine-grained approach should be applied in science mapping where the unit is no longer the document but at lower levels like sections, paragraphs or even sentences. This has already been alluded to by several studies using an enhanced co-citation analysis which incorporates also the proximity of the cited references. Several studies indicate the co-cited publications are more similar if the distance of their in-text citation is smaller (see Gipp & Beel, 2009). Complementary to these co-citation-based findings, others have reported different lexical properties of the subsequent sections in scientific papers. Bertin et al. (2016) reported different use of verbs and rhetorical structures surrounding references across different sections.

Other approaches that move away from the assumption of a document as an indivisible unit are probabilistic topic modelling techniques like LDA (Blei, 2012) where documents are linked to different topics with a weight relative to the probability that the topic is relevant to the document. With LDA, one has to set the number of topics prior to the analysis. Gal et al (2017) applied this technique to a set of publications from cardio-vascular research with an initial set of 200 topics where after expert validation only 166 remained relevant. Three issues remain unsolved when using topic modelling approaches. First, there is the initial decision on number of topics, next, the document remains the unit of analysis and the probability that a document is related to a particular topic is attributed to the document as a whole and finally, the use of bag-of-word approach in the learning phase neglects the differences in meaning a word can have depending on its context. Leydesdorff and Hellssten (2006) demonstrated how words and co-words retrieve their meaning from their presence in sentences and broader context.

This study proposes a new approach that moves the granularity towards smaller units of text namely the paragraphs in the different sections across a full paper and applies an analytical technique that tries to capture the meaning of words and phrases not only from its position relative to other words in its neighbourhood but also from the overall subject or topic covered by the paragraph. Vector word embeddings using neural networks architecture like GloVe (Pennington, Socher & Manning, 2014) and Word2Vec (Mikolov et al. 2013) are able to map words to a low dimensional space, with high performance. These word embeddings have however a single representation in the vector space neglecting the different meaning a word can have across different contexts. This is solved by adding an additional paragraph or document layer to the learned model which holds this context information (Quoc & Mikolov, 2014).

This study will use the Doc2Vec implementation in GenSim (Rehurek & Sojka, 2010) for the calculation of the word and paragraph embeddings and for the calculation of the similarities between low level units or fragments of text extracted from all PlosOne publications up to december 2018. As such, the current research is the first to apply these techniques at such a large scale with the following objectives. First, I'll try to measure the intra document similarity between the different paragraphs and between the paragraphs and the abstract. It is assumed that paragraphs close to each other have higher similarity, that paragraphs in the 'Introduction' and the 'Discussion'-section have higher similarity with the abstract and each other and that paragraphs in the 'Methodology' are least similar to all other paragraphs Next, the research focusses on documents linked through a citation. It is assumed that the context surrounding the reference is most similar to the content of the cited paper. In a last section of the study, the location of the reference in the citing document is mapped with the different paragraphs in the cited document in order to retrieve the cited information relevant to the citing document. The results from this study can have applications like bibliometric aided information retrieval or can assist in the identification of interdisciplinary research.

Data & Methodology

Data

This study uses publications from PlosOne downloaded from PubMed Central ftp://ftp.ncbi.nlm.nih.gov/pub/pmc. The downloaded set contains publications indexed until December 7th, 2018 and it holds 204,846 documents from 2006 onwards. The papers are provided in XML-format following the '*Journal Article Tagging Suite*' (JATS) standard. This schema divides the information in three main elements: <front>, <body> and <back> with an underlying structure of elements and attributes and complies with ANSI standard Z39.96-2012 (ANSI, 2012). This format, provided as an XML-schema, is then converted by the *Java Architecture for XML Binding* library (JAXB) into generated Java source code. This generated Java library serves then as a unmarshalling toolbox which can convert any XML-document compliant with the JATS-schema into a set of Java Objects (POJOs). This toolbox incorporates parts of the *CorpusHandling* library developed by *CyCorp* and available under Apache license (version 2.0) from GitHub (see https://github.com/cycorp/CorpusHandling/). Each XML-document is unmarshalled into a Java object and parsed in order to extract:

- Bibliographic information like title, article number, publication year
- Sections and paragraphs holding the actual text fragments of the paper
- In-text references identified by the <Xref>-tag
- References at the end of the paper

It is assumed that papers published in PlosOne adhere to the IMRaD structure (Introduction, Methodology, Results and Discussion) or a variation where the Methodology section is at the end of the paper (IRDaM) following the description of the distribution of sections across PlosOne publications (Bertin, et al. 2013). The title heading each section is used to classify the text fragment to one of the following classes:

I. Introduction; Background
II. Data; Material; Methodology; Design
III. Results
IV. Discussion; Summary; Conclusion
V. Other sections

Paragraphs are identified by XML-element tags <*sec*> and <*p*>, extracted and given a sequence number. Sentences are extracted and numbered within each paragraph. Figure 1 presents a paragraph from the first paper published in PlsOne (Harris et al, 2006).

> Early investigations focused on the role of neurons in subcortical stations and primary somatosensory cortex (SI) in coding low frequency "flutter" vibrations (below 50 Hz) **[1]–[3]**, while more recent work has emphasized the role of cortical areas "downstream" from SI, such as the second somatosensory cortex (SII) and regions of frontal cortex **[4]**, **[5]**. Which of these different areas, and which features of the neural activity within these areas, are essential components in forming the percept of a vibration? A series of psychophysical experiments with humans provided evidence that neural processes in SI contribute to frequency discriminations. In a task designed to resemble that performed by monkeys in the aforementioned neurophysiological studies, subjects compared two sequential vibrations and reported which had the higher frequency.

Figure 1. Text fragment taken from Harris et al (2006)
In-text references are marked in bold and underlined.

Next, in-text references (<*xref*>) are linked with the complete reference at the end of the paper and available identifiers like PMID, PMCID or DOI of cited papers are retrieved. This enables the linking of individual paragraphs to the cited paper. The position of the in-text reference with respect to extracted sentences in the paragraph is recorded. The in-text references in Figure 1 are marked in bold and underlined. The first <xref>-element is linked to the first three entries in the reference list at the end of the paper as it indicates the range between reference 1 and 3. The next two elements refer to the fourth and fifth entry. Each element in the text is replaced by the corresponding PMCID, PMID or DOI depending on the available data in the reference list.

Methodology

After extraction of the data from the XML-file, a set of processing steps is applied in order to obtain the vector word and paragraph embeddings. A pre-processing procedure as described in Thijs et al (2017) and Glänzel & Thijs (2017) based on the Stanford Natural Language Processing library (Chen & Manning, 2014) and the Lucene text search engine is used for the extraction of sentences, application of Part of Speech tagging, stemming, removal of stop-words and selection of noun phrases. Document identifiers like PMCID of the cited references are processed as noun phrases and retained at the original position within each sentence. Table 1 presents the results after pre-processing of the text fragment in figure 1. A list of all cited documents is added at the end of each paragraph as an additional 'sentence'. The choice to include the cited references in the final paragraph embeddings is not without consequences. It adds a bibliographic-coupling-like component to the embeddings.

Table 1. Parsed content of paragraph in Harris et al (2006) per section, paragraph and sentence.

Section	Paragraph	Sentence	Parsed Content
I	5	0	earli investig role neuron subcort station primari somatosensori cortex si low frequenc flutter vibrat 50 hz pmc2118947 pmc4959494 pmc4977839 recent work role cortic area si second somatosensori cortex sii region frontal cortex pmc12368806 pmc10884334
I	5	1	differ area featur neural activ area essenti compon percept vibrat
I	5	2	seri psychophys experi human neural process si frequenc discrimin
I	5	3	task monkey aforement neurophysiolog subject two sequenti vibrat higher frequenc
…	…	…	…
I	5	15	pmc2118947 pmc4959494 pmc4977839 pmc12368806 pmc10884334 …

The mathematical representation of the text fragments or paragraphs is based on vector representations built by a neural network architecture in an unsupervised machine learning algorithm. The applied methodology was first developed for distributed word embeddings at Google by Mikolov et al (2013) as a more complex substitute for simple vector-based representations like N-gram models. These embeddings are used to predict a word given the surrounding words in its context. The context is a sliding window with a fixed word length. The context is also applied to the identifiers of the cited references. Quoc & Mikolov (2014) extended the model for the inclusion of document or paragraph embeddings to outperform traditional bag-of-word approaches. Just like the original word embeddings model, the paragraph is represented by a vector in the same space as the words. It complements each fixed word length context used for the prediction of the words in the paragraph. The vector representation is unique for each paragraph and it is not shared among paragraphs and can be thought of as the representation of the topic the paragraph is dealing with.

The neural network used for training this model is a single layered architecture with a fixed dimensionality. In contrast to the LDA approach, these dimensions are not linked to topics and no external validation of the validity of the dimensions is required.

The Python implementation included in the Gensim library (Rehurek & Sojka, 2010) is used in this study. The algorithm is named '*Doc2Vec*' and takes the paragraph as a list of words as input with an additional tag identifying paragraph. This tuple is called a '*TaggedDocument*' in the library. The abstracts are tagged by the PMCID and paragraphs with tags containing the PMCID, section classification and sequence number of the paragraph. A cosine is calculated between the vector embeddings to measure the similarity between the text fragments.

The first set of analyses focusses on the intra-document similarity between paragraphs and abstract and among paragraphs. Figure 1 provides a schematic overview of the different analytical steps in this study. The intra-document similarity is indicated at the left-hand site. Within paper A, the abstract is compared with each paragraph and each paragraph with all subsequent paragraphs within the same section and across sections. The sequence number of the paragraphs are used to indicate the distance between the paragraphs in the text.

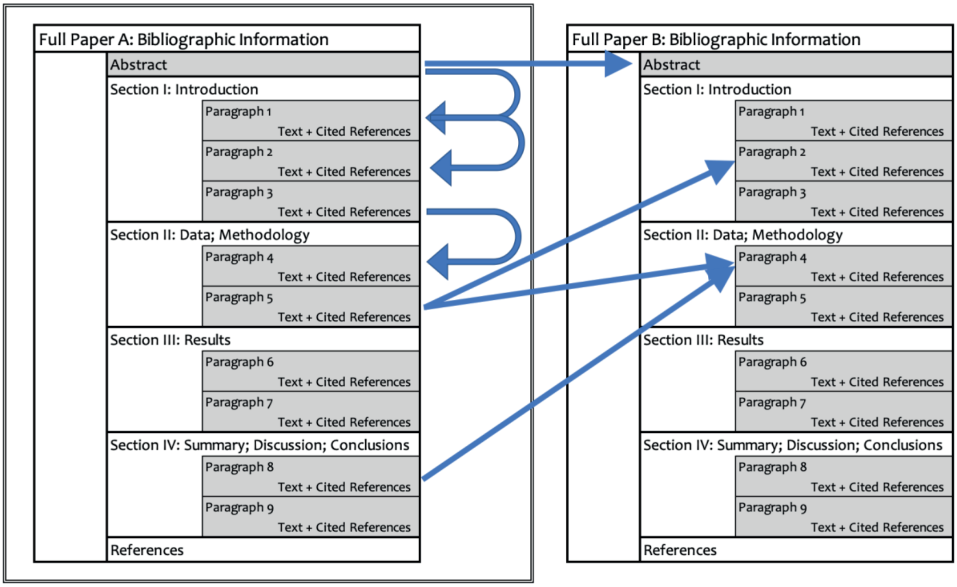


Figure 2. Schematic overview of the different comparisons.

The second set of analyses focusses on the similarity between citing and cited pairs of documents. In fig 2. there is a citation from paper A to paper B. The similarity of both abstracts in a citing-cited document pair is compared to the similarity in a randomly selected document pair. As the paragraph holding the in-text reference to the cited paper can be located, a next analysis compares the similarity between the citing paragraph and cited abstract and paragraphs across all sections within the citing-cited document pair.

Results

Descriptive Statistics

204,846 PlosOne publications have been downloaded and processed. 99% of these are recorded as '*research article*' in the Web of Science database and the remaining 1% as '*review*'. Table 2 provides the distribution of papers over publication years and the average number of paragraphs, together with the share of documents with the IMRaD sections in any order. Almost all documents have an introduction and discussion section.

Table 2. Descriptive statistics for downloaded PlosOne papers per year. Sections are classified based on the header of the section.

Publication Year	Publications	Average number of paragraphs	Introduction (I)	Methodology (II)	Results (III)	Discussion (IV)
2006	137	29.32	100.0%	99.3%	84.7%	100.0%
2007	1230	30.97	100.0%	98.5%	86.8%	99.4%
2008	2820	31.14	96.3%	95.7%	86.6%	96.0%
2009	4537	31.90	97.0%	96.3%	87.8%	96.8%
2010	6925	32.17	97.5%	96.9%	88.6%	97.3%
2011	14043	32.17	98.2%	97.3%	89.7%	97.9%
2012	24102	32.28	97.3%	96.0%	89.0%	97.0%
2013	32973	31.68	95.6%	93.4%	86.0%	95.3%
2014	30467	32.66	98.6%	96.1%	87.8%	98.4%
2015	28126	33.60	99.8%	96.8%	87.5%	99.6%
2016	22092	34.10	99.8%	96.7%	88.0%	99.6%
2017	20499	34.32	99.5%	96.2%	87.3%	99.2%
2018	16895	34.01	94.9%	91.2%	83.0%	94.6%
Total	**204846**	**32.92**	**97.9%**	**95.5%**	**87.3%**	**97.7%**

The publications contain on average 32.92 paragraphs and 3.86 different sections. This is below the values reported by Bertin et al (2013). This probably due to differences in parsing and extraction of the XML-elements. Subsections indicated by <sec>-elements as a child from another <sec>-element are not considered as separate sections and obtain their classification from their parent element. Section and paragraph elements without text as value were not considered as separate paragraphs.

The Neural Network Model

The final *Doc2Vec*-model is trained on 6.95 million text fragments from abstracts and paragraph texts. The neural network contains 400 nodes and training is done over ten iterations. Before training, a vocabulary was created with 3.90 million unique words. The total number of words included was 440 million. A sliding window of 7 words was used to establish the context for each word. It took about 15 hours to train this model on an average server requiring not more than 27Gb of RAM.

Intra-document similarity

First, the analysis focuses on intra-document similarity. Figure 3. shows the distribution of the cosine similarity of the abstract with distinct paragraphs across the four identified sections. The '*Introduction*' is most similar to the abstract, while the '*Methodology*' section is least similar. The inclusions of in-text references in the final paragraph embeddings and the absence of references in abstracts can act as a damping factor for the similarity between abstract and actual text fragments. However, the higher amount of references in the introduction and discussion (Bertin et al 2013) does not prevent the higher similarity between these sections and the abstract.

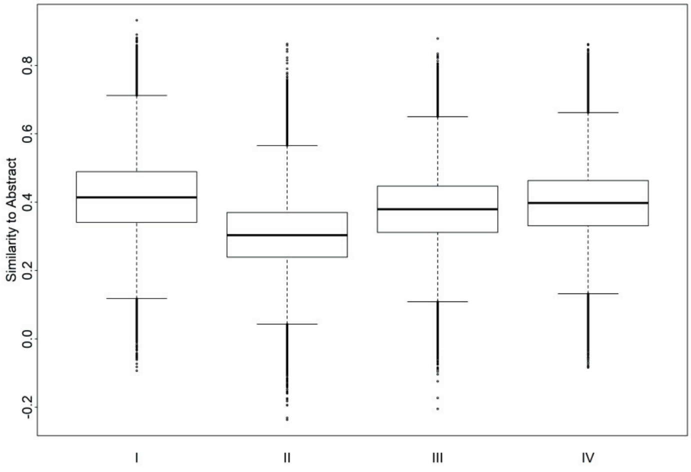


Figure 3. Distribution of similarity between Abstract and different sections in PlosOne papers.

Next, the similarity is calculated between paragraphs within sections and across section in each paper. The average similarities are presented in table 3. The intra-section similarity ranges from 0.32 for the '*Introduction*' to 0.29 for the '*Methodology*'. Looking at similarities across section, it can be observed that '*Introduction*' and '*Discussion*' are more similar, and '*Methodology*'-paragraphs are least similar to paragraphs in other sections.

Table 3. Average similarity of paragraphs across sections

	I	II	III	IV
I: Introduction	**0.32**	0.26	0.27	0.31
II: Methodology	0.26	**0.29**	0.27	0.26
III: Result	0.27	0.27	**0.31**	0.28
IV: Discussion	0.31	0.26	0.28	**0.30**

It is worthwhile to complement this analysis by adding the distance between paragraphs to the analysis. Figure 4 plots the average similarity between two paragraphs against the distance between them in the text. As each paragraph gets a sequence number in the processing phase it is easy to calculate the distance between them. The plot distinguishes between two groups, namely the distance between paragraphs inside one section opposed to the distance across sections. The solid line indicates the within section similarity and starts with the highest value. It rapidly declines with an increasing distance. The similarity between paragraphs across sections stars much lower and takes an increase and slow decline afterwards.

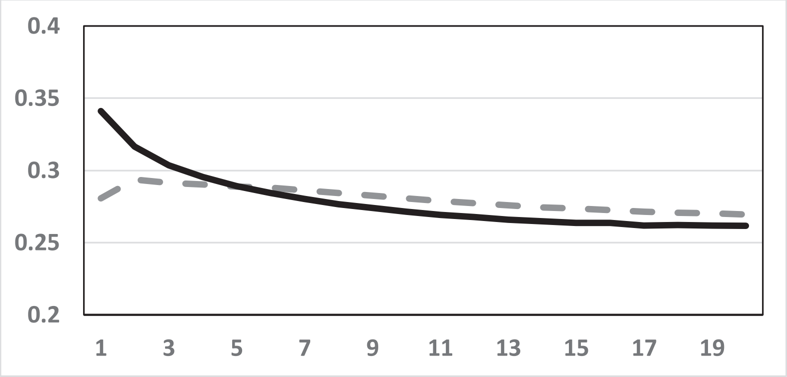


Figure 4. Average similarity between paragraphs related to the distance within the document. (Solid line: within one section, Dashed: across sections)

The overall image in figure 4 can easily be explained by the low similarity between the '*Methodology*' and '*Results*' sections with the two other sections. The main structure of PlosOne papers is either IMRaD or IRDaM with the '*Methodology*' or '*Result*' section in between '*Introduction*' and '*Discussion*' creating higher distance between these sections with higher similarity. Remarkable is the crossing of the two lines near a distance of 5 between paragraph. From then on, paragraphs from different sections are more similar than within sections. Probably, topics or themes already raised in a previous section are retaken in the light of the obtained results or applied methodology.

Between document similarity

The analyses in this next section will all focus on similarity across documents.

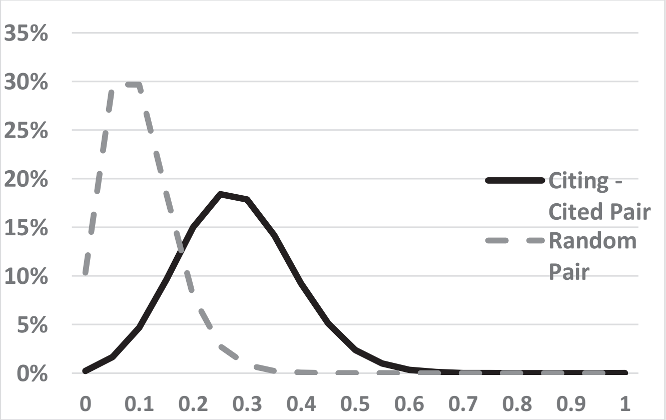


Figure 5. Distribution of similarity as measured through vector document embeddings of abstracts of document pairs (solid line: Citing-Cited document pair, dashed line: random pairs).

In order to have a baseline or reference point, the similarity between abstracts in citing-cited document pairs is gauged against the similarity between two randomly selected abstracts. The distribution of both sets of similarities have been plotted in figure 5.

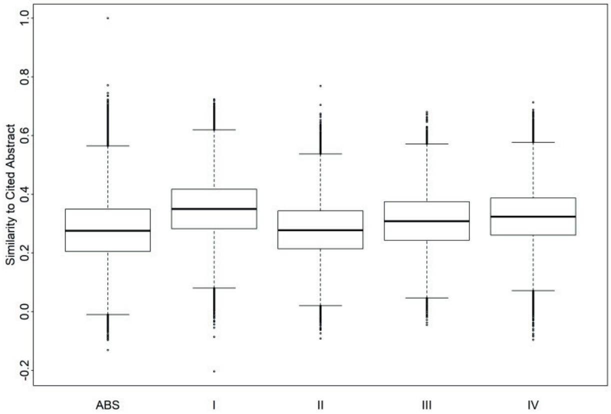


Figure 6. Distribution of similarity between abstract or citing section and abstract of cited document.

The similarity between random selected abstracts is just below 0.09, while the average for the citing-cited pair of documents is 0.28. The distribution of similarities of abstracts of citing and cited pairs of documents is also in figure 6. Here it is contrasted by the distribution of similarity of the paragraph in which the reference appears and the abstract of the cited document, grouped by citing section. Once more, paragraphs in the '*Introduction*' show the highest similarity with the cited abstract. The median in the second box is highest while the first box (abstract to abstract) has the lowest median. This shows clearly that the information in these individual paragraphs bear different content or information than the abstract.

For the last analysis, the similarity is calculated between the citing paragraph and all paragraphs in the cited document. A citation does not contain -it exceptionally does- a reference to the exact location of the relevant concept or topic in the cited document. This last analysis selects only those PlosOne papers cited at least 5 times by other PlosOne papers. Figure 7. plots the average similarity between citing paragraphs and cited paragraphs across different sections. A plot for each section at the citing side is given. Each plot contains a box per section in the cited document. Paragraphs from the '*Introduction*' and '*Discussion*' section are most similar with the '*Introduction*' in the cited document with 'Discussion' ranked second. This pattern changes when looking at the citing paragraphs in the '*Methodology*'-section. Here cited '*Methodology*' and '*Introduction*' score equally.

Discussion and Conclusion

The results obtained in this study support the statement that a more fine-grained approach using paragraphs is applicable in science mapping and that it will provide additional insights in the topic structure underlying scientific papers. As earlier observed (see Bertin et al. 2013), each section in a publication serves different purposes with distinct reference distribution. Here it is shown that there is also a textual difference between sections but also within sections. The further paragraphs are separated from each other in a section the less similar they are. The use of vector word and paragraph embeddings can be useful for several applications in quantitative science studies. In the following section, applications of intra- and inter document similarity are presented.

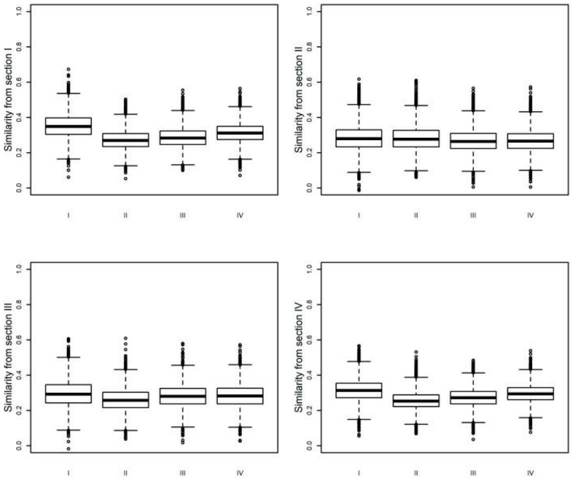


Figure 7. Distribution of similarity between citing section and different sections in cited document.

Applications and limitations

The use of intra-document similarity between paragraphs can extend the study of interdisciplinarity. Currently two main approaches are applied for the study of interdisciplinarity of scientific publications namely the use of subject classifications of cited references (Leydesdorff & Rafols, 2011) and the disciplinary profile of the researchers involved (Abramo et al., 2012). Using the lexical information embedded in the distinct paragraphs and sections combined with the similarity to cited documents can provide a novel third approach. Network-based statistics like node distance, centrality and modularity are appropriate measures for central concepts in the study of interdisciplinarity like disparity, balance and variety (Wang et al 2015).Another application of this fine-grained approach is in information retrieval. Context based word embeddings provide enhancements at both *needle* and *haystack* side. Key words in search strategies can be complemented by their specific context which defines their meaning and the same model is used to characterize the paragraphs at the haystack side. Moreover, other applications of word embeddings (see eg. Mikolov et al., 2013) have shown that mathematical operations on vectors like subtractions are possible and retain their topical characterization. This allows the creation of search strategies starting from a set of keywords without the need to list all possible alternatives or variations but also to provide a set of keywords or papers that are irrelevant to the search and should be excluded.

The novel approach presented in this study does not come without limitations. At first, there is the need for open access to the full paper in the required JATS-format. The procedure could be rewritten to be applicable on HMTL data or even on parsed PDF. However, the main advantage of JATS is that it specially targeted towards scientific journal articles and parsing is less prone to errors. The scoring of additional documents not in the original dataset is possible through the neural network algorithm. The obtained model can even be trained to incorporate these additional documents, but the procedure initially starts with the creation of a word vocabulary which cannot be updated. This puts a burden on the extensibility of the model as the proposed approach also takes the publication identifiers of the cited documents. These identifiers are merged into vocabulary as if they were words. The set of cited documents in the additional data set will thus be limited to the original set of cited documents. Other approaches for word embeddings like LSH or random projects suffer also from this limitation. The model reduces the document space from extreme sparse with hyper dimensionality into a dense matrix with limited predefined number of dimensions. Using such a dense matrix for the creation of document networks results in a near complete network where a similarity can be calculated for nearly any given pair of documents. It is very hard to use these near complete weighted networks as a basis for clustering techniques or community detection. Only the application of thresholds on the similarity can solve this issue which comes with computational constraints as the similarity of each pair of documents has to be calculated prior to the application of the threshold. Hashing algorithms like LSH can be used to solve this issue.The creation of the neural network model involves tuning several parameters like number of underlying nodes or dimensions, learning rate, learning iterations, minimum threshold for rare words, down-sampling rate for frequent words, sliding window length for the word context. With the last option, frequent words are removed with a probability relative to the inverse of their frequency which results in actual larger windows. The Word2Vec also provides two different learning approaches. Each of these parameters can have an influence on the final obtained model. More research is required to study the effect of this hyperparameter tuning on the final validity of the model and resulting vector embeddings.

Conclusion

Vector word and paragraph embeddings provide a novel approach for the calculation of within and between document similarities. The technique is used to create neural network based mathematical representations of text fragments of smaller size like paragraphs. Within such a vector space, the cosine of the angle between the vectors can be used to indicate the similarity between the underlying text fragments. The Word2Vec and Doc2Vec implementations provide an easy to use library for the creation of the word embeddings and similarity calculations. The application of the technique shows that the paragraphs in the '*Introduction*' and '*Discussion*' section are most similar to the abstract but that the '*Methodology*' has a much lower similarity with abstract. Combined with lower number of references in this section, the paragraphs are less presented in document-based approaches using abstracts and citations for the creation of document networks. When looking at citing-cited pairs of documents, the paragraph containing the actual reference to the cited paper shows a higher similarity with the abstract of the cited paper. This is especially the case with paragraphs from the introduction. The novel approach can have several applications in quantitative science studies like the study of interdisciplinarity or bibliometric aided information retrieval, but the technique suffers still from limitations which can damper the validity of the obtained results.

References

Abramo, G., D'Angelo, C. A., Costa, F. D. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. *Journal of the Association for Information Science & Technology*, 63(11), 2206–2222.

Ahlgren, P., Collinader, C., (2009), Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, 3 (1), 49-63.

Bertin M., Atanassova I., Larivière V., Gingras, Y. ,(2013) The distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. In: *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics*. Vienna, Austria, 591-603.

Bertin M., Atanassova I., Sugimoto CR., Larivière V., (2016). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics* 109:1417-1434.

Blei, David (April 2012). "Probabilistic Topic Models". *Communications of the ACM*. 55 (4): 77–84.

Boyack, K.W., (2017). Investigating the effect of global data on topic detection. *Scientometrics*. 111 (2), 999-1015.

Chen, D., Mannig, C.D., (2014). A Fast and Accurate Dependency Parser using Neural Networks. In: *Proceedings of EMNLP 2014*. Doha, Qatar.

Gal, D., Thijs, B., Sipido, K., Glänzel, W., (2017) Topic modelling based network maps in cardiovascular research. In: *Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics*. Wuhan, China, 591-603.

Gipp, B, Beel, J. (2007) Citation Proximity Analysis (CPA) - A New Approach for Identifying Related Work Based on Co-Citation Analysis. In: *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics*. Rio de Janeiro, Brazil, 571-575.

Glänzel, W., Thijs, B., (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset. Scientometrics, 111 (2), 1071-1087.

Harris, J.A., Arabzadeh, E., Fairhall, A.L., Benito, C., Diamond, M.E. (2006). Factors affecting frequency discrimination of vibrotactile stimuli: implications for cortical encoding. *PlosOne*, 1(1), e100.

Kiss, J.Z., Aanes, G., Schiefloe, M., Coelho, L.H.F., Millar, K.D.L., Edelmann, R.E., (2014). Changes in operational procedures to improve spaceflight experiments in plant biology in the European Modular Cultivation System. *Advances in Space Research*, 53 (5), 818-827.

Leydesdorff, L., Hellsten, I., (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies,' 'Frankenfoods,' and 'stem cells'. *Scientometrics*, 67 (2), 231-258.

Leydesdorff, L., Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1), 87–100.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

Pennington, J., Socher, R., Manning, C.D., (2014). GloVe: Global Vectors for Word Representation. (available at: https://nlp.stanford.edu/pubs/glove.pdf)

Quoc, L. & Mikolov, T., (2014), Distributed Representations of Sentences and Documents. In: *Proceedings of the 31th International Conference on Machine Learning, ICML*. Beijing, China, 1188-1196.

Rehurek, R., Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proc. LREC Workshop on New Challenges for NLP Frameworks*

Small, H., (1994). A SCI-map case-study – building a map of AIDS research. *Scientometrics*, 30 (1), 229-241.

Taşkin Z and Al U. (2018.) A content-based citation analysis study based on text categorization. *Scientometrics* 114(1):335-337

Thijs, B., Glänzel, W., Meyer, M.S. (2017) Improved lexical similarities for hybrid clustering through the use of noun phrases extraction. MSI Working Paper Series. University of Leuven, Leuven, Belgium

Thijs, B. & Glänzel, W. (2018), The contribution of the lexical component in hybrid clustering, the case of four decades of "Scientometrics". *Scientometrics*, 115(1), 21–33.

Wang, J., Thijs, B. & Glänzel, W. (2015). Interdisciplinarity and Impact: Distinct Effects of Variety, Balance and Disparity. *Plos One*, 10(5): e0127298

Wang, S., Koopman, R., (2017). Clustering articles based on semantic similarity. *Scientometrics*, 111 (2) 1017-1031.

Performance Model's development: A Novel Approach encompassing Ontology-Based Data Access and Visual Analytics

Marco Angelini[1], Cinzia Daraio[1], Maurizio Lenzerini[1], Francesco Leotta[1], Giuseppe Santucci[1]

[1]*{angelini, daraio, lenzerini, leotta, santucci}@diag.uniroma1.it*
DIAG Department, Sapienza University of Rome, Via Ariosto, 25 00185, Rome (Italy)

Abstract

The quantitative evaluation of research is currently carried out by means of indicators calculated on data extracted and integrated by analysts who elaborate them by creating illustrative tables and plots of results.

In this paper we propose a new approach which is able to move forward, from indicators' development to performance model's development. It combines the advantages of the Ontology-based data Access (OBDA) integration with the flexibility and robustness of a Visual Analytics (VA) environment. A detailed description of such an approach is presented in the paper.

Introduction: An advanced models' development approach

In recent decades, the rapid changes taking place in the production, communication and evaluation of research have been signs of an ongoing transformation. It has been stated that "we are living a sort of Middle-Age guided by the information and communication technologies (ICT) revolution, or the so-called *forth revolution* as described by Floridi (2014) which emphasizes the importance of information" (Daraio, 2019, p. 636). Largely, the current Middle-Age of research evaluation might be understood as the transition from a traditional evaluation model, based on bibliometric indicators of publications and citations to a modern evaluation, characterized by a multiplicity of distinct, complementary dimensions. This step is guided by the development and increasing availability of data and statistical and computerized techniques for their treatment, including among others the recent advancements in artificial intelligence and machine learning. Daraio and Glänzel (2016) show that that the complexity of research systems requires a continuous information exchange.

These changes produce different effects (see further details and references in Daraio, 2019, Table 24.2, p. 644) i) on the *demand side* (those that ask for research assessment) including an increase of institutional and internal assessments, ii) on the *supply side* (those that offer research assessment) including proliferation of rankings, development of Altmetrics, open access repositories, new assessment tools and desktop bibliometrics), iii) on *scholars* (the increase of "publish or perish" pressure, impact on the incentives, behaviour and misconduct, and increasing critics against traditional bibliometric indicators), iv) on the assessment process (increasing the complexity of the research assessment) and on the indicators' development.

Daraio (2017a) showed that the formulation of models of metrics (in this paper we will use metrics and indicators as synonyms) is necessary to assess the meaning, validity and robustness of metrics. It was observed that developing models is important for *learning* about the explicit consequences of assumptions, test the assumptions, highlight relevant relations; and for *improving*, document/verify the assumptions, systematize the problem and the evaluation/choice done, explicit the dependence of the choice to the scenario. Moreover, there are several *drawbacks* in modelling, which have to be taken into account. The main pitfalls relate to the targets that are not quantifiable; the complexity, uncertainty and changeability of the environment in which the system works, to the limits in the decision context, and, last but not least, to the intrinsic complexity of calculation of the objective of the analysis.

In this paper we depart from the traditional approach to indicators' development, based on the selection of a specific set of indicators, collection of the relevant data, cleaning of the gathered data, computation of the indicators and illustration of them in a plot or table. According to this traditional approach if you want to add a new data source or you want a different indicator you have to restart the process from the scratch.

We support an alternative approach based on an OBDA system for R&I data integration and access. An Ontology-Based Data Access (OBDA) system is an information management system constituted by three components: an ontology, a set of data sources, and the mapping between the two. An *ontology* in Description Logic (DL) is a *knowledge base*. It is a couple (pair) O=<TBox,ABox>, where TBox is the Terminological Box that represents the *intensional* level of the knowledge or the *conceptual* model of the portion of the reality of interest expressed in a formal way; and ABox is the Assertion Box that represents the *extensional* level of the knowledge or the *concrete* model of the portion of the reality expressed by means of assertions (instances). An ontology populated by instances and completed by rules of inference is defined as *knowledge base* (see e.g. Calvanese et al. 1998). The *data sources* are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The *mappings* are precise specifications of the correspondence between the data contained in the data sources and the elements of the ontology. The main purpose of an OBDA system is to allow information users to query the data using the elements in the ontology as predicates.

The OBDA system, implemented with *Sapientia*, represents the ontology of multidimensional research assessment (Daraio, Lenzerini et al. 2015) and permits the extraction of relevant data coming from heterogeneous sources - maintained independently, and reasoning about the Performance Indicators (PI) of interest.

Daraio, Lenzerini et al. (2016a) showed the advantages of an OBDA system for R&I integration and Daraio, Lenzerini et al. (2016b) showed that an OBDA approach allows for an unambiguous specification of indicators according to its four main dimensions: ontological, logical, functional and qualitative. See also Lenzerini and Daraio (2019) where a detailed illustration of the usefulness of an OBDA approach for reasoning over the ontology about indicators of performance is reported. Even the simplest indicator of performance, such as number of publications has different conceptual aspects that the ontological commitment of the domain offers to the analyst (for additional details the reader is referred to Fig. 15.9 and 15.10 of Lenzerini and Daraio, 2019, pag. 368 and pag. 369).

The main contribution of this paper is making a step further, on our previous researches and to propose a new approach for the multidimensional assessment of research and its impact based on the combination of OBDA and Visual Analytics. This novel approach allows for the development and evaluation of performance models instead of the traditional indicators' building system.

Combining OBDA and Visual Analytics

The traditional way to define indicators relies on an *informal definition* of the indicator as the relationship between variables selected among a set of data collected and integrated "ad hoc", specific for the user needs (*silos based* data integration approach). This means that when a new indicator has to be calculated, the process of data integration has to restart since the beginning because the dataset created "ad hoc" for an indicator is not reusable for another one.

The contribution of an OBDM approach to overcome this traditional indicator development approach is twofold. First of all, it permits the *free* exploration of the *knowledge base* (or information platform) created to identify and specify new indicators, not planned or defined in advance by the users. This feature would be particularly useful to face two recent trends in user requirements, namely *granularity* and *cross-referencing* (see Daraio and Bonaccorsi, 2017 for a discussion on university-based indicators). Secondly, it allows us to specify a given indicator in a more precise way as described in Lenzerini and Daraio (2019).

In this paper we develop further this approach combining it with the main strengths of Visual Analytics. Visual Analytics (Cook & Thomas, 2005, Keim et al., 2008) is "the science of analytic reasoning facilitated by interactive visual interfaces"; through the connection of the analytical calculation with visualization and interaction by the human user, this interdisciplinary approach enhances the exploratory analysis of data, allowing to represent multidimensional data in a simple way through innovative visual metaphors. Further it allows navigation in the data space, in order to obtain an overview of the eventually tunable to the required level of detail, the ability to apply complex analysis workflows that aim at explainability, the ability to obtain summary reports of the findings discovered during the analysis phase. See Figure 1 for an overview.

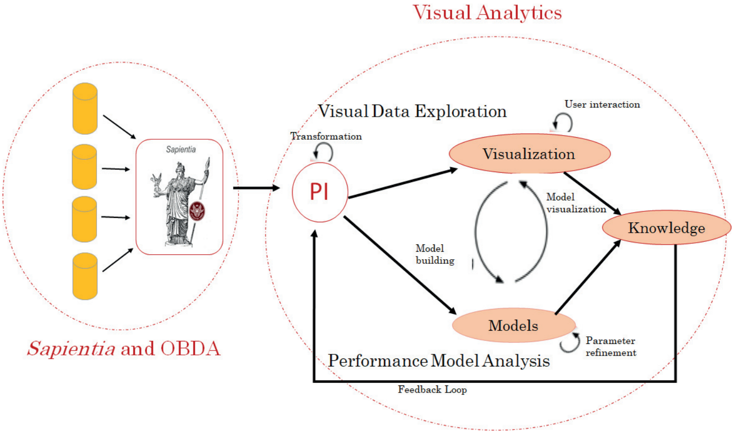


Figure 1. An illustration of our approach that combines *Sapientia*, OBDA and Visual Analytics. PI states for Performance Indicator.

The Visual Analytics approach developed in this paper allows us to move from Performance indicators development to Performance model development, by exploring and exploiting the modelling and the data features within the flexibility of a Visual Analytics environment.

This allows a multi-stakeholder viewpoint on the model of PI, the assessment of the sensitivity and robustness of the PI model in a multidimensional framework.

In the next section we outline the main features of *Sapientia* (the Ontology of Multidimensional Research Assessment). After that we present our Visual Analytics environment for the performance model's development together with an illustration of its potentialities. The final section concludes the paper.

OBDA at work through *Sapientia: The Ontology of multidimensional research assessment*

Sapientia, the Ontology of Multidimensional Research Assessment (Daraio et al. 2015, 2016a, 2016b), models all the activities relevant for the evaluation of research and for assessing its impact (see Figure 2 for an outline of its modules). For impact, in a broad sense, we mean any effect, change or benefit, to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia.

The *Sapientia* ontology has been developed using the Graphol visual language (http://www.dis.uniroma1.it/~graphol/, Lembo et al. 2016), that can be easily translated into standard ontology languages like Owl.

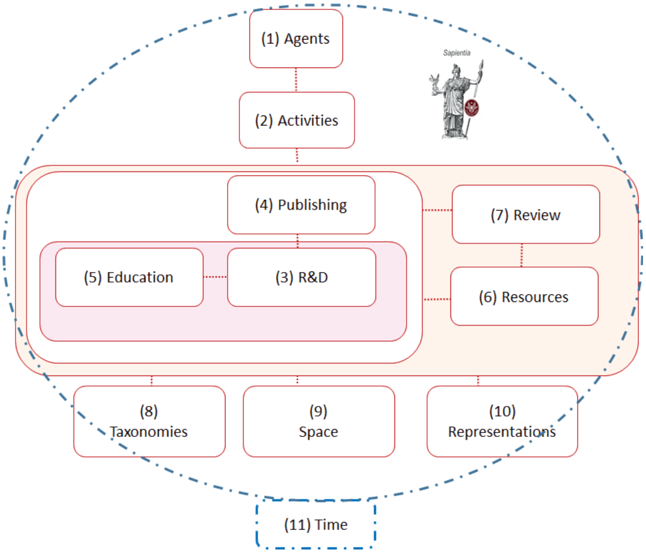


Figure 2. Modules of *Sapientia* 3.0. 1. Agents: describes all human actors and institutions involved in the education, research and innovation process. 2. Activities: describes the activities and projects the agents of the previous module are involved in. 3. R&D: describes the different products (e.g., publications, patents) that are produced in the knowledge production process. 4. Publishing: describes how knowledge products are published and made available to the public. 5. Education: introduces concept related to universities and courses. 6. Resources: describes all the ways and institution can be funded. 7. Review: describes the process entities related to the publishing activity. 8. Taxonomy: describes the elements that allows to define taxonomies applied to the different modules. 9 and 11: Space and Time: allow to describe respectively geographical entities and time instants and ranges. 10. Representation: allows to describe the fact that single instances of other modules can be represented in different ways by the different sources used in *Sapientia.*

Sapientia acquires information from multiple sources, whose content can be overlapping. The same entity modelled in the *Sapientia* ontology can be represented in more than one data source, and even one data source could present (due to internal inconsistencies or design choices) the same entity multiple times in different forms.

Hence, we have the need to identify duplicated items and integrate the information obtained for each entity from any of the available sources.

In particular, at the ontology level we have created the concept of Representation. Entities modelled in the ontology of which we have different views from different data sources may have their own representation, which specializes the general Representation concept. This makes it possible to keep track in the ontology, through the mappings, not only of the modelled entities, but also of the way in which the information relative to the entities has been gathered from the data sources.

Data acquisition from the external sources makes use of the web service standards (REST, SOAP) when available.

For less frequently updated sources and sources that do not implement an API, data acquisition leverages in some cases the open source edition of Pentaho Data Integration (http://community.pentaho.com/projects/data-integration/).

Imported data are saved in a relational database (MySql). Each source is modeled independently so that its peculiar structure can be fully exploited.

Sapientia extract information, among others, from the following datasets:

- *Scopus*. A very large abstract and citation database of peer-reviewed literature, containing information about scientific journals, books and conference proceedings. Scopus provides information about authors' affiliations as well. The available REST interface allows to retrieve: document information, document citations data, percentiles data end journal percentiles data.
- *ETER*. The ETER (European Tertiary Education Register) consortium acquired extensive information pertinent to tertiary educational institution of many European countries. Data have been acquired by the consortium for the years 2011-2016 and are publicly available (https://eter-project.com).
- *DBLP*. A service that provides open bibliographic information on major computer science journals and proceedings. Data is available through massive XML files.
- The *InCites* (https://incites.thomsonreuters.com) dataset contains research indicators organized on a geographical base. Data can be downloaded in the form of CSV files that are then imported using an ad-hoc procedure.
- *Geonames* (http://www.geonames.org/) is a dataset that contains information about geographical areas at any level. The dataset can be freely download, and has been employed to match geographical entities from the different data sources.
- *Web of Science* database is going to be included as well.

The data manipulation layer of the *Sapientia*, which allows to populate the ontology from the data sources, is composed of an indexing module, an entity-resolution module and a normalization module.

In general terms, the *indexing module* creates and maintains up to date the indices that are used by the *entity resolution module* to implement the blocking functionalities that allow to keep the time complexity of the entity-resolution algorithms under control. This module has the dual purpose of easing the definition of the mappings toward the *Sapientia* Ontology, and creating the basis for a common interface of the entity-resolution algorithms. Indices inside the *Sapientia* application are implemented using the Hibernate search (http://hibernate.org/search/) library and the Lucene indexer and searcher (http://lucene.apache.org/).

Entity resolution is the task of connecting matching entities between different data sources. As this kind of process is exponential in complexity with the number of data sources and entities per data sources, it is split in two phases:

- *Blocking*, which allows very quickly, by employing indices to create groups of potential matching entities
- *Entity matching*, which finds matching entities inside clusters identified by blocking.

After matching entities have been recognized by entity resolution, the *normalization step* is employed in order to provide a uniform representation for the information contained in different and heterogeneous data sources. These uniform representations are called mappable entities. These mappable entities are mapped to ontology entities through an operation called mapping. *Sapientia* uses the Mastro Ontology-Based Data Access (OBDA) management system (http://www.dis.uniroma1.it/~mastro/?q=node/1). The Sapientia ontology, however, is defined over a richer language than the one supported by Mastro. Hence, we used the OWL2DL tool in order to obtain a simplified version of the Sapientia ontology that conforms to the DL-light language supported by Mastro.

The definition of the mappings in Mastro is XML based. There are three types of ontology predicate mappings: concept, role and attribute.

As suggested by the names, the concept predicate mapping refers to entities, the role predicate mapping puts entities in relation, populating a role, while the attribute mapping relates an entity with a constant, which is the value of its attribute.

Some examples of extraction and mapping of relevant data

In order to show the potential of the proposed approach, we will show how indicators can be extracted from the ontology and grouped according to a specific level of analysis. In the illustration identified as European denoted Nomenclature of Territorial Units for Statistics (NUTS) code. The modules of the ontology interested in this query are:

- The *Agents module*, which contains the concept of University as a specialization of the concept of Organization. An Organization has an Organization State, which represents the evolution of the Organization in time, and that refers to the Residence.
- The *Space module* that contains the concept of Residence as a specialization of a Position. A Position has an Entrance, which is localized in an Address inside a City. The City is a Territory, and European Cities are European Territories that can be aggregated by NUTS codes.
- The *Taxonomy module* where an Organization is contained in a Taxonomic Unit. Each Taxonomic Unit has a State that has indicators as attributes.

For a specific university denoted by its Eter ID, we can for example compute the cardinality of academic staff with the following SPARQL query :

```
select ?academic_staff {
     ?org sapientia:has_place_in ?taxon_unit .
     ?org a sapientia:University .
     ?taxon_unit sapientia:has_state_of_taxonomic_unit ?state_tax .
     ?state_tax a sapientia:Present_state .
     ?state_tax sapientia:teacher_population ?academic_staff .
}
```

In order to group by a specific NUTS codes, it is possible to extend the previous query as follows:

```
select SUM(?academic_staff), ?nuts2 {
     ?org sapientia:has_place_in ?taxon_unit .
     ?org a sapientia:University .
     ?taxon_unit sapientia:has_state_of_taxonomic_unit ?state_tax .
     ?state_tax sapientia:teacher_population ?academic_staff .
     ?state_tax a sapientia:Present_state .
     ?org sapientia:has_state_of_organization ?org_state .
     ?org_state a sapientia:Present_state .
     ?org_state sapientia:has_residence ?resid .
     ?resid a sapientia:Legal_residence .
     ?resid sapientia:has_entrance ?entr .
     ?entr a sapientia:Address .
     ?entr sapientia:is_in_the_city ?city .
     ?city a sapientia:European_territory .
     ?city sapientia:is_territory_part_of ?region .
     ?region a sapientia:Small_europen_region .
     ?region sapientia: NUTS2ref ?nuts2 .
     ?region sapientia: NUTS2ref ?nuts1 .
     ?region sapientia: NUTS2ref ?nuts3
}
GROUP BY ?nuts2
```

Where the results have been grouped by NUTS2. It is possible to easily modify the query in order to group by other levels of NUTS. In a similar way, *mutatis mutandis*, it is possible to extract the data and indicators that will be used for the Performance Indicator and model development that is described in the next sections.

The Visual Analytics environment

This section describes the Visual Analytics environment and its main features. The developed solution uses Visual Analytics techniques to represent data from publications and education obtained from the OBDA approach described in the previous section, and complete the workflow. The system is implemented through Web technology. Clearly the large quantity of indicators and basic sizes for the different units of analysis, including the territorial ones, and in the different years of analysis increases exponentially the cardinality of data to be analyzed; in this respect, the display part allows to obtain a visual overview of the data in a very simple form, and the interaction capabilities allow the user to navigate in this overview and conduct detailed analysis up to the desired level. The user is also supported in the discovery of any elements of analysis of interest through a process of *data exploration* that does not require a prior analysis goal.

In addition to the data exploration capacity there is a second area explicitly aimed at analyzing the model development and performance computed on these indicators, based on the definition and exploration of performance models. The environment is instantiated on European research and education institutions as a case study. The user can, on one hand, analyze the performance of the various institutions with respect to a performance model, in order to analyze the positioning of the institutions of interest; additionally, it allows to explore different performance models and to evaluate their goodness and fitness. Further, it is also possible to evaluate the goodness of the proposed models, analyzing their variability and conducting sensitivity analysis in order to evaluate which parameters of the model (whether inputs or resources, contextual factors or outputs) contribute more to the performance of the institution with respect to the chosen model. The following subsections will provide a description of the features of Visual Analytics environment.

Data Exploration Environment

The first panel that composes the Visual Analytics environment is the data explorer environment. This environment consists of three main views depicted in Figure 3.

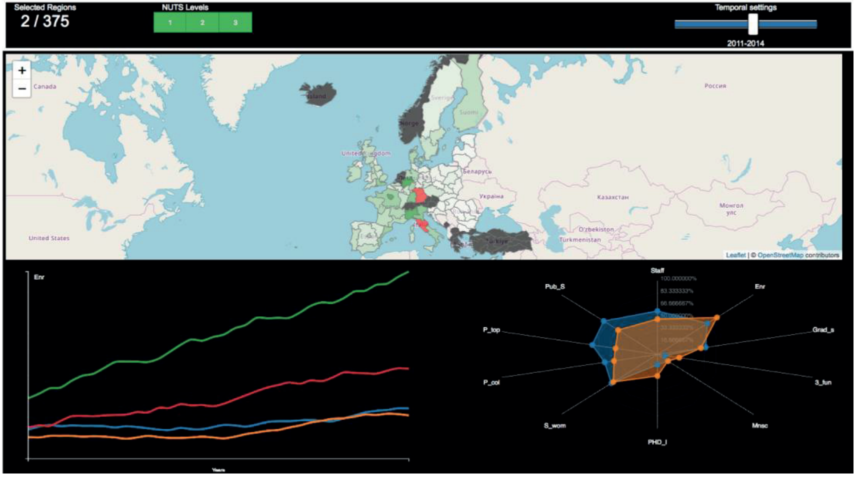


Figure 3. Data Exploration Environment

These three views are:

-*Geographic view*: which allows for geolocating of the different institutions with respect to territorial units on a geographic layer (using Leaflet.js framework, based on OpenStreetmap) is used. The map is navigable on 5 different levels of detail, where the first four follow the NUTS categorization from 0 (Nations) to 3 (Provinces) and the last one relates to single institutions. The user can at any time change the level of aggregation through a tab that shows the different available levels.

The color of each element of the map reflects an indicator (basic or derived) of, on a green scale that identifies the values (white: low value, dark green: high value). The gray color codifies the absence of data for the particular territorial unit. A slider allows the analyst to scroll through the various years and conduct a temporal analysis on the available data, looking for institutions showing a high variability through a "time-lapse".

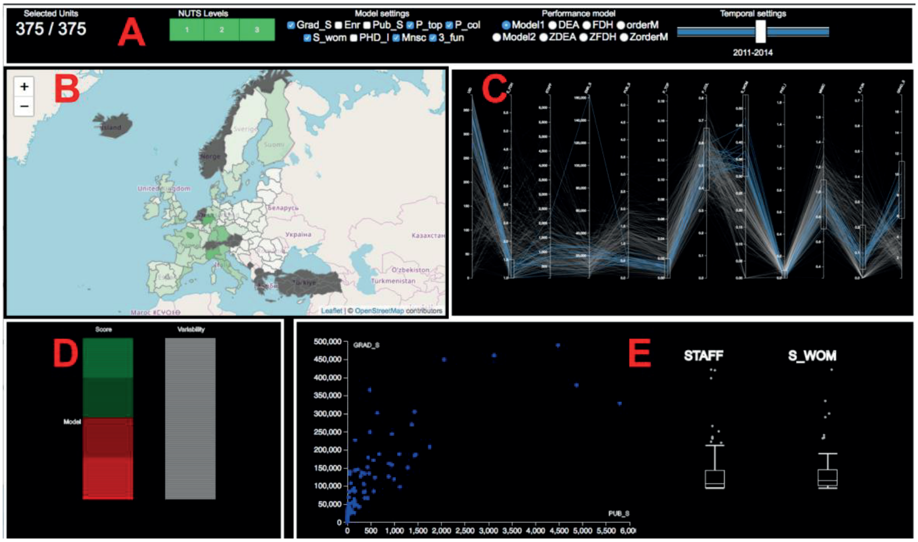
-*Radar view*: this view follows the visual paradigm of the radar diagrams (Von Mayr, 1877) , which represent the dimensions of a dataset one per axis, with the axes arranged in radial form starting from the center. The indicators are arranged one per axis and the graph presents several lines that join the points on each axis in the number of one per institution or territorial unit. When the user selects one or more territorial units, the corresponding splines are highlighted, in order to allow an easy visual comparison between the different territorial units selected on their different dimensions. It is also possible to highlight a dynamic average trend, consisting of a line that connects the different averages on the respective axes, in order to compare the performance of a territorial unit, or generally of a given unit of analysis, not only to other units but also to the aggregate behavior between the territorial units.

-*Linechart view*: This visualization allows analyzing the time course of the evaluation measures used for the units. It is possible to analyze both multiple territorial units to compare the trend of the same measure on them, and to analyze multiple measures on the same territorial unit, in order to have an overview of the progress of the unit itself, and a combination based on multiple territorial units and multiple measures. In this case the color-coding outlines all the measures belonging to each single territorial unit.

The combined use of these views, possibly guided by the definition of specific PIs, allows more powerful dynamic exploration of the model data of the territorial units compared to the classical approaches, making the user able to obtain an overview of the general trend and specific details on the individual units, subsequently allowing to refine the analysis through the visual selection of appropriate subsets of information. The approach therefore allows the exploration of s*pecific scenarios* chosen by the user in *real-time*, without precomputation, which better support the formation and validation of hypotheses and the identification of areas of interest on which to conduct further analysis or to be used for reporting activity.

Performance Model Analysis Environment

This environment is dedicated to the analysis of performances of the model used for analyzing the units. The visual environment is therefore more complex than the Data Exploration one, as shown in Figure 4.



Selected Units 375 / 375

NUTS Levels 1 2 3

Model settings ☑ Grad_S ☐ Enr ☐ Pub_S ☑ P_top ☑ P_col ☑ S_wom ☐ PHD_I ☑ Mnsc ☑ 3_fun

Performance model ● Model1 ○ DEA ○ FDH ○ orderM ○ Model2 ○ ZDEA ○ ZFDH ○ ZorderM

Temporal settings 2011-2014

Figure 4. Performance Model Analysis Environment

The environment consists of a command bar (A), a geographical view borrowed from the Data Exploration environment (B), a view based on parallel coordinates (C), a view of the rankings due to the selected performance model (D) and finally a view based on scatter-plot and box-plot that allows to conduct sensitivity analysis on the parameters of the selected model (E). The features of the individual views are described below.

-*Command bar*: this area identifies the main analysis commands that will affect the selections in all remaining views. From left to right we have:

-the counter of the territorial units active with respect to the total (the territorial units contained in the current selection)

-a tab that allows to select the aggregation level on which to conduct the analysis

- the parameters and measures of the performance model, which can be activated using the appropriate checkboxes. This command allows to re-parameterize a model (among those available) in order to conduct a different type of analysis of performance.

-The model selector, which allows you to choose between 8 families of performance models, ranging from custom model defined by the Analyst (Model 1 and Model 2) to efficiency models, Data Envelopment Analysis (DEA), Free Disposal Hull (FDH), orderM, and their conditioned variants ZDEA, ZFDH, ZorderM. An overview on these performance models can be found in Daraio (2019).

-The time selector, which allows to evaluate the result of the chosen model with respect to a temporal interval that can be controlled by means of a slider.

-*Geographical view:* this visualization follows the same operating principle illustrated for the Data Analysis Environment. In this instance, however, the color linked to each individual territorial unit is proportional to the unit's performance score with respect to selected model. In this way the user can immediately get an overview of the different performance levels given the chosen hierarchical level, model and time interval. The user can zoom in on the map in order to get more details on individual portions of the map. It is also possible to use the map as a highlighting mechanism: by clicking on one or more units, these are highlighted in red on the map and in all other coordinated views, allowing to identify a subset of data of interests starting from geographical coordinates of the unit.

-*Parallel coordinates*: this view shows all the dimensions that are part of the model (inputs, possible conditioning factors, outputs) plus the year of analysis and the ID of the units. The purpose of this visualization is to explore the relationships that exist between these quantities, in order to decide whether or not to keep them in the selected model. From the visual point of view, each of the dimensions is represented as a vertical axis, and each unit as a line that joins the values it has on each axis. Through brushing operations on individual axes, it is possible to perform multi-filter operations on several dimensions, making possible to select very complex filtering expressions while maintaining the ease of creating these filters: by dynamically define new intervals on the various dimensions, and immediately verify the cardinality and the characteristics of the resulting subset, the analyst can explore several combinations and discover relations among dimensions (see Figure 5).

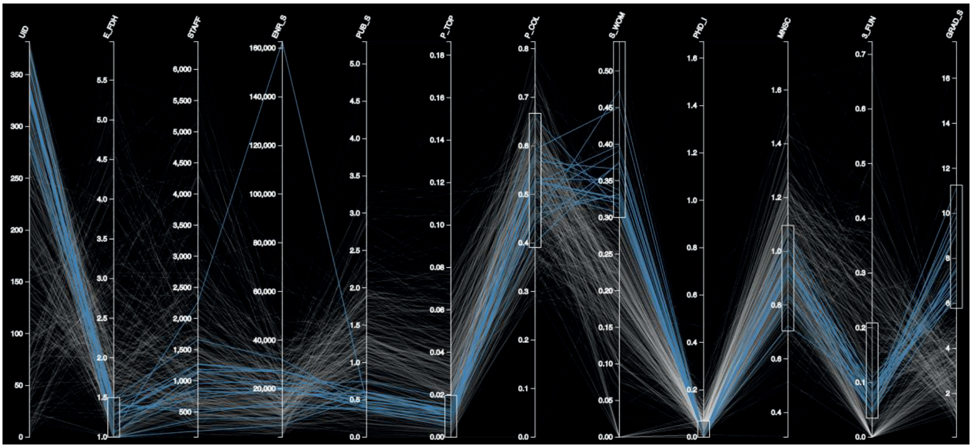


Figure 5. Example of parallel coordinates filter. Axes, from the left: UID is the institution id number, E_FDH is the FDH (in)efficiency score (equal to 1 means efficient; the higher it is, more outputs the unit could proportionally produce to become efficient) STAFF is number of academic staff in FTE (Full Time Equivalent), ENR_S is number of total enrolled students per academic staff, PUB_S is number of publications in WoS (fractional count) per academic staff, P_TOP is number of publications in top 10% of highly cited journals per academic staff, P_COL is percentage of papers done with international collaborations, S_WOM is share of women professors on total academic staff, PHD_I is PhD intensity, MNCS is Mean Normalized Citation Score (1 corresponds to the world average, >1 above (<1 below) world average), 3_FUN is share of third party funds in %, GRAD_S is total number of graduates per academic staff. The filter shows that among the most efficient units in teaching and research (i.e. E_FDH = [1 1.5]) there are those teaching oriented institutions (with the highest values of GRAD_S) in which the S_WOM is the highest ([0.30-0.50]): these are universities with almost zero PhD intensity that are able nevertheless to produce a small fraction of P_TOP publications with MNCS around the world average.

In addition, by drag and drop interaction, it is possible to exchange all the axes with each other, in order to better highlight any correlations, anti-correlations or similarity characteristics on specific subsets of data among dimensions. Any findings, as mentioned above, serve to better understands the results coming from the performance model used.

-*Rank analysis*: This view supports the tasks of exploring the performance scores of the individual units, and the sensitivity analysis on the model, in terms of estimating the contribution of each individual parameter of the model to the performance scores. The visualization is composed of two bars representing rankings, where the units are ordered according to the performance score from top (high performance score) to bottom (low performance score). Each unit is represented as a rectangle, whose color derives from the calculation of the distribution of the performance scores and from the assignment of a color to each of the 4 quartiles (the 3rd and 4th quartiles with deeper shades of green, the 1st and 2nd with deeper shades of red). An informative tooltip, activated by mouse-hover on each rectangle, allows to obtain accurate information on the performance of the unit. The second bar is initially completely gray, and is activated when individual elements (inputs, conditioning factors) of the model are selected / deselected from the command bar: in this way it is possible to evaluate the displacement in the rank of each single unit with respect to addition/deletion of a parameter of the model, and therefore be able to evaluate the stability of the model compared to the performance scores produced, and the sensitivity of the performance model in terms of contribution that any parameter produces in the ranking (see Figure 6).

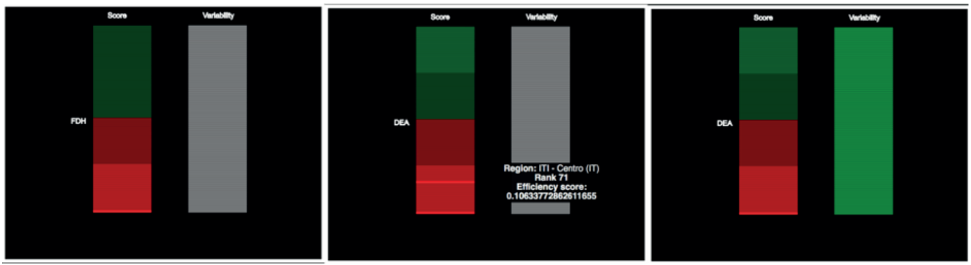


Figure 6. rank analysis obtained using a complete FDH model (left); the same chart is instantiated through a DEA model, and the tooltip reports the score for the "Italia Centro" territorial unit (center); finally, the result on the variability obtained by removing the output factor PUB_S and including P_TOP (right). As you can see, the whole bar is green, which means that the units rank remains stable with respect to this input, which could be replaced by another more significant input.

Sensitivity analysis: This view expands the sensitivity analysis capabilities, already introduced in the Rank Analysis view. The visualization uses two different visual paradigms to relate the different parameters that constitute the performance model: in the first one, a scatter plot, the relation between the conditioning factors (if present) and the outputs is reported. Input factors are instead reported as a distribution in the form of a box-plot for each input factor. The interactivity of this chart allows to select disjoint sets of values from each box-plot and inspect the propagated filter on the entire visual environment. It will be possible to analyze the relationship between the various elements of the performance model in a more precise and granular form, identifying from the distribution subsets of interest which will eventually correspond to the selection of a subset of units that respect the imposed constraints. The effect will therefore support the sensitivity analysis of the model but also support the explorative analysis of the data through filter operations based on factors of the model (see Figure 7).

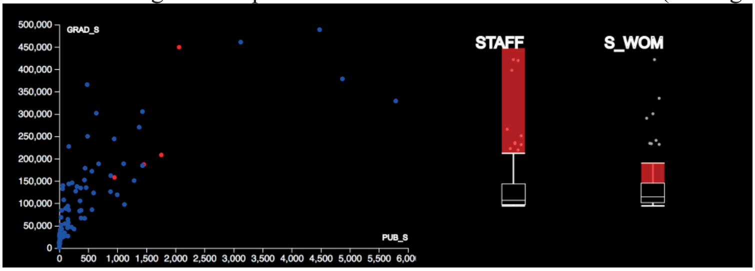


Figure 7. Example of data filtering: with respect to all the units, the selection is composed by high outliers for academic staff (STAFF) and the 4th quartile for percentage of women staff (S_WOM); the resulting points are highlighted in red in the scatter plot, and the unit can be identified by mouse-hover.

Conclusions

In this paper we leveraged on the research based on *Sapientia* and OBDA combining it with a Visual Analytics approach. The new approach proposed allows us to move from Performance indicators development to Performance model's development, by exploring and exploiting the modelling and the data features within the flexibility of a Visual Analytics environment. This allows a multi-stakeholder viewpoint on the model of PI, the assessment of the sensitivity and robustness of the PI model in a multidimensional framework as illustrated in the previous section.

Acknowledgments

The financial support of the Italian Ministry of Education and Research (through the PRIN Project N. 2015RJARX7), of Sapienza University of Rome (through the Sapienza Awards no. PH11715C8239C105 and KIMAR Project), and of Clarivate Analytics through the KOL Project, is gratefully acknowledged.

Selected References

Calvanese D., G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. JAR, 39(3): 385–429.

Calvanese D., G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati (1998). Description logic framework for information integration. In Proc. of KR, pages 2–13.

Daraio C. (2017), A framework for the assessment of Research and its Impacts, *Journal of Data and Information Science*, Vol. 2 No. 4, 2017 pp 7–42.

Daraio C. (2019), Econometric approaches to the measurement of research productivity, in *Springer Handbook of Science and Technology Indicators* edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., forthcoming.

Daraio C., Bonaccorsi A., (2017), Beyond university rankings? Generating new indicators on universities by linking data in open platforms, *Journal of the Association for Information Science and Technology*, DOI: 10.1002/asi.23679

Daraio C., Glänzel W. (2016), Grand Challenges in Data Integration. State of the Art and Future Perspectives: An Introduction, *Scientometrics*, 108 (1), 391-400.

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A., Bartolucci A. (2015). Sapientia: The Ontology of Multi-Dimensional Research Assessment, in Salah, A.A., Y. Tonta, A.A. Akdag Salah, C. Sugimoto, U. Al (Eds.), Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015, Bogaziçi University Printhouse, pp. 965-977.

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A., Bartolucci A. (2016a), Data Integration for Research and Innovation Policy: An Ontology-based Data Management Approach, *Scientometrics*, 106 (2), 857-871.

Daraio C., Lenzerini M., Leporelli C., Naggar P., Bonaccorsi A. Bartolucci A. (2016b), The advantages of an Ontology-based Data Management approach: openness, interoperability and data quality, *Scientometrics*, 108 (1), 441-455.

Lenzerini M. & Daraio C. (2019), Challenges, Approaches and Solutions in Data Integration for Research and Innovation, in *Springer Handbook of Science and Technology Indicators* edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., forthcoming.

Lembo, D., Pantaleone, D., Santarelli, V., & Savo, D. F. (2016). Easy OWL Drawing with the Graphol Visual Ontology Language. In KR (pp. 573-576).

Cook, K.A., Thomas, J.J. (2005). Illuminating the path: The research and development agenda for visual analytics. *Tech. rep.*, Pacific Northwest National Lab. (PNNL), Richland, WA (United States)

Keim, D., Andrienko, G., Fekete, J.D., Gorg, C., Kohlhammer, J., Melancon, G. (2008). *Visual Analytics: Definition, Process, and Challenges*. Springer Berlin Heidelberg, Berlin, Heidelberg, 154–175

Von Mayr, G. (1877) *Die gesetzmässigkeit im gesellschaftsleben*. Didemburg.

Bibliographic Reference List Mistakes: The Case of Turkish Librarianship

Müge Akbulut [1] and Sümeyye Akça [2]

[1] *mugeakbulut@gmail.com*
Ankara Yıldırım Beyazıt University, Dept of Information Management, Ankara (Turkey)

[2] *sumeyyeakca@ardahan.edu.tr*
Ardahan University, Dept of Information Management, Ardahan (Turkey)

Abstract

In this paper we studied bibliographic reference mistakes. Bibliographic references are an important part of scholarly publishing and are also crucial for visibility and accessibility of studies in the databases. We have searched how much of the bibliography of the works published in the Turkish Librarianship journal appeared in the Web of Science (WoS) citation database. Between 2015 and 2017, a total of 2959 studies, 2707 (91.4%) of which placed in the WoS, appear in the reference lists of 192 studies in Turkish Librarianship. 96 (38%) out of the 252 remaining non-indexed sources in WoS have been created in accordance with the original APA rules. Therefore, the reason why these sources are not indexed is either due to minor errors made by the authors or to the indexing algorithm of WoS.

Introduction

Bibliographic references are an integral part of scientific publishing in the process of creation and dissemination of information. This is also one of the areas where authors make the most mistakes. This may be due to the inattention of the authors or their lack of knowledge. Besides that, numerous bibliographic reference standards which are used for different scientific areas (Park, Mardis & Ury, 2011) also add to the mistakes. In recent years, the diversity of the resources used in research, with the widespread use of internet and electronic resources, has also made way for more mistakes in reference. Especially in classical, traditional areas (such as history, literature, archeology) these updates are not well known by the authors. Moreover, the fact that different journals in one area request different bibliographic standards is confusing the authors.

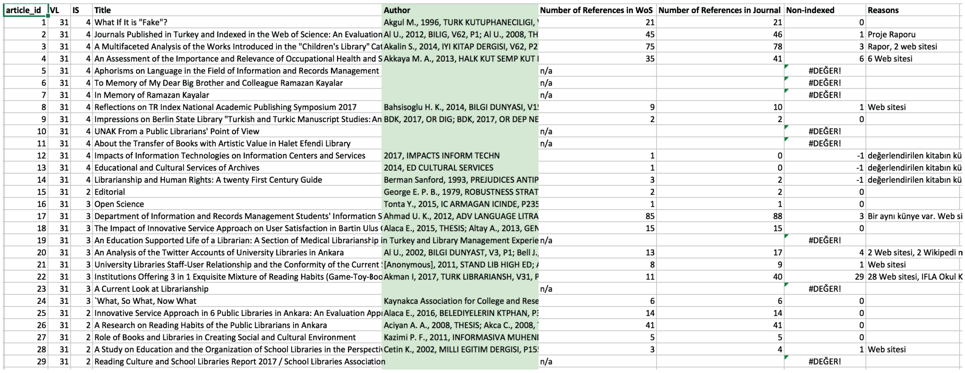
In bibliographical mistakes, it is known that there are mistakes caused by the inattention of the author as well as unethical citations (citation from secondary resources) (Bahar et al., 2012; Lok, Chan & Martinson, 2001; Oermann, Cummings & Wilmes, 2001). These mistakes may also arise from the inattention of journal editors (Oermann, Cummings & Wilmes, 2001). It does not offer an example of comprehensive bibliography preparation to many journalists. Again, the use of non-updated standards also increases the mistakes (Onwuegbuzie, Hwang, Combs, and Slate, 2012). Localized standard rules also can increase these mistakes by creating confusion.

The most important problem caused by such mistakes in given situations is the issue of appearance in citation databases of many resources with erroneous reference. When a study that has bibliographical mistakes is indexed in the citation databases, citations that do not comply with the standard or given incorrectly are not in place in the mentioned databases. For example, in a study, 19 references are shown in the database in which the study is indexed; although 20 references were used. This suggests that the missing reference is given incorrectly. It is necessary that the location and description information of the resource is given in full, so that the cited resource can be easily accessed (Moorthy, 1988).

In this study, we have searched how much of the bibliography of the works published in the Journal of Turkish Librarianship appeared in the Web of Science (WoS) citation database. We discussed the reasons why the resources not included in WoS are not indexed.

Method

The Turkish Librarianship was indexed in the ESCI (Emerging Sources Citation Index) of WoS in 2015. For this reason, we have analyzed articles indexed since 2015. In total, 192 articles covering 2015-2017 period were downloaded from WoS. With the data available, we checked from the web page of the journal and compared the relevant bibliography of articles. We were able to evaluate the bibliography of the articles written in Turkish and English, as the language of the article is included in the metadata in WoS (see Figure 1).



article_id	VL	IS	Title	Author	Number of References in WoS	Number of References in Journal	Non-indexed	Reasons
1	31	4	What If It is "Fake"?	Akgul M., 1996, TURK KUTUPHANECILIGI,	21	21	0	
2	31	4	Journals Published in Turkey and Indexed in the Web of Science: An Evaluation	Al U., 2012, BILIG, V62, P1; Al U., 2008, TH	45	46	1	Proje Raporu
3	31	4	A Multifaceted Analysis of the Works Introduced in the "Children's Library" Cat	Akalin S., 2014, IYI KITAP DERGISI, V62, P2	75	78	3	Rapor, 2 web sitesi
4	31	4	An Assessment of the Importance and Relevance of Occupational Health and S	Akkaya M. A., 2013, HALK KUT SEMP KUT	35	41	6	6 Web sitesi
5	31	4	Aphorisms on Language in the Field of Information and Records Management		n/a		#DEĞER!	
6	31	4	To Memory of My Dear Big Brother and Colleague Ramazan Kayalar		n/a		#DEĞER!	
7	31	4	In Memory of Ramazan Kayalar		n/a		#DEĞER!	
8	31	4	Reflections on TR Index National Academic Publishing Symposium 2017	Bahsisoglu H. K., 2014, BILGI DUNYASI, V1	9	10	1	Web sitesi
9	31	4	Impressions on Berlin State Library "Turkish and Turkic Manuscript Studies: An	BDK, 2017, OR DIG; BDK, 2017, OR DEP NE	2	2	0	
10	31	4	UNAK From a Public Librarians' Point of View		n/a		#DEĞER!	
11	31	4	About the Transfer of Books with Artistic Value in Halet Efendi Library		n/a		#DEĞER!	
12	31	4	Impacts of Information Technologies on Information Centers and Services	2017, IMPACTS INFORM TECHN	1	0	-1	değerlendirilen kitabın kü
13	31	4	Educational and Cultural Services of Archives	2014, ED CULTURAL SERVICES	1	0	-1	değerlendirilen kitabın kü
14	31	4	Librarianship and Human Rights: A twenty First Century Guide	Berman Sanford, 1993, PREJUDICES ANTIP	3	2	-1	değerlendirilen kitabın kü
15	31	2	Editorial	George E. P. B., 1979, ROBUSTNESS STRAT	2	2	0	
16	31	3	Open Science	Tonta Y., 2015, IC ARMAGAN ICINDE, P235	1	1	0	
17	31	3	Department of Information and Records Management Students' Information S	Ahmad U. K., 2012, ADV LANGUAGE LITRA	85	88	3	Bir aynı künye var. Web si
18	31	3	The Impact of Innovative Service Approach on User Satisfaction in Bartin Ulus (	Alaca E., 2015, THESIS; Altay A., 2013, GEN	15	15	0	
19	31	3	An Education Supported Life of a Librarian: A Section of Medical Librarianship in Turkey and Library Management Experie		n/a		#DEĞER!	
20	31	3	An Analysis of the Twitter Accounts of University Libraries in Ankara	Al U., 2002, BILGI DUNYAST, V3, P1; Bell J.,	13	17	4	2 Web sitesi, 2 Wikipedi n
21	31	3	University Libraries Staff-User Relationship and the Conformity of the Current	[Anonymous], 2011, STAND LIB HIGH ED;	8	9	1	Web sitesi
22	31	3	Institutions Offering 3 in 1 Exquisite Mixture of Reading Habits (Game-Toy-Boo	Akman I, 2017, TURK LIBRARIANSH, V31, P	11	40	29	28 Web sitesi, IFLA Okul K
23	31	3	A Current Look at Librarianship		n/a		#DEĞER!	
24	31	3	`What, So What, Now What	Kaynakca Association for College and Rese	6	6	0	
25	31	2	Innovative Service Approach in 6 Public Libraries in Ankara: An Evaluation App	Alaca E., 2016, BELEDIYELERIN KTPHAN, P	14	14	0	
26	31	2	A Research on Reading Habits of the Public Librarians in Ankara	Aciyan A. A., 2008, THESIS; Akca C., 2008,	41	41	0	
27	31	2	Role of Books and Libraries in Creating Social and Cultural Environment	Kazimi P. F., 2011, INFORMASIVA MUHEN	5	5	0	
28	31	2	A Study on Education and the Organization of School Libraries in the Perspecti	Cetin K., 2002, MILLI EGITIM DERGISI, P15	3	4	1	Web sitesi
29	31	2	Reading Culture and School Libraries Report 2017 / School Libraries Association		n/a		#DEĞER!	

Figure 1. The sample of data set

Findings

Between 2015 and 2017, 192 studies which were published in 12 issues of the Journal of Turkish Librarianship have been indexed in WoS. Eight of these studies are in English. Between 2015 and 2017, a total of 2959 studies, 2707 (91.4%) of which placed in the WoS, appear in the reference lists of 192 studies in Turkish Librarianship. 20% or more of the studies in the reference lists of 16% of 192 studies (N=30) have not been indexed. Keeping in mind that 88 of them (46%) do not have any references, more than 20% of the resources in the reference lists of 32% (N=30) of the resources that have reference have not been indexed.

As some studies have not been indexed at all, sometimes extra studies have been indexed. WoS also includes the bibliographic record of the book introduced in the book introductions into the reference list. 25 of the 192 studies in our dataset are in this way.

The indexing rate of the reference list of English sources is 98.6%. 227 of the 230 resources in the reference list of the English source have been indexed. The type of three non-indexed studies are in the website format. In Turkish sources this rate is 91.4%. In our opinion, this difference is closely related to the localization of the APA rules. Due to the syntax differences in languages, the standard structure has been corrupted and the local rules have moved away from being a machine-readable standard.

Figure 1 shows the distribution by type of 252 sources from the studies in the reference list of 192 studies, which are not indexed in WoS for a variety of reasons.

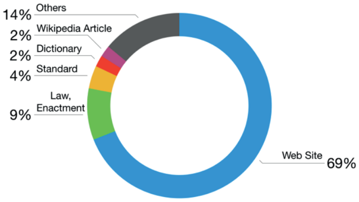


Figure 2. Distribution by type of 252 sources which are not indexed in WoS

Conclusion

- 96 (38%) out of 252 non-indexed sources in WoS have been created in accordance with the original APA rules. Therefore, the reason why these sources are not indexed is either due to minor errors made by the authors or to the indexing algorithm of WoS.

- The remaining 156 resources (62%) have not been indexed although prepared in accordance with "localized APA rules (Turkish version)".

- Disruptions related to the localized rules mostly arise from the syntax differences in the Turkish and English language rules. For example, according to APA rules, the phrase "Retrieved from" is used before the address is given. On a localized copy, "adresinden erişildi (accessed from address)" or "erişim adresi: (access address:)" phrases are used after the access address is given. Both the presence of the ":" sign and the corresponding pattern given before the access address lead to indexing problems.

- Apart from this, there are also resources not indexed by WoS even if they conform to APA style. Legal entity or organizations are indexed as title instead of author; or as [Anonymous] if the author name does not have a comma. At the entrance of the website, if there is a comma in the section up to the date, the author name is indexed as the journal name.

References

Moorthy, A. L. (1988). Towards a standard style for bibliographic references. DESIDOC Journal of Library & Information Technology, 8(4). Retrieved from http://publications.drdo.gov.in/ojs/index.php/djlit/article/view/2979).

Park, S., Mardis, L. A. & Ury, C. J. (2011). I've lost my identity - oh, there it is... in a style manual: Teaching citation styles and academic honesty. Reference Services Review, 39(1), 42–57. doi: http://dx.doi.org/10.1108/00907321111108105.

Bahar, Z., Beser, A., Elcigil, A., Karayurt, O., Vural, F., Ugur, O. & Kucukguclu, O. (2012). Accuracy of references in eight nursing journal. HealthMed, 6(6), 2066–2073.

Lok, C. K. W., Chan, M. T. V. & Martinson, I. M. (2001). Risk factors for citation errors in peer-reviewed nursing journals. Journal of Advanced Nursing, 34(2), 223–229. http://dx.doi.org/10.1046/j.1365-2648.2001.01748.x.

Oermann, M. H., Cummings, S. L. & Wilmes, N. A. (2001). Accuracy of references in four pediatric nursing journals. Journal of Pediatric Nursing, 16(4), 263–268. doi: http://dx.doi.org/10.1053/jpdn.2001.25537.

Onwuegbuzie, A., Hwang, E., Combs, J. P. & Slate, J. R. (2012). Editorial: Evidence-based guidelines for avoiding reference list errors in manuscripts submitted to Journals for review for publication: A replication case study of educational researcher. Research in the Schools, 19(2), i–xvi (Retrieved from http://search.proquest.com/ docview/1509202998).

Mapping the Life Science using Medical Subject Headings (MeSH)

Fei Shu[1], Junping Qiu[1] and Vincent Lairière[2]

[1] *fei.shu@hdu.edu.cn; casee.hdu@outlook.com; tjmxsyx@sina.com at*
Hangzhou Dianzi University, Chinese Academy of Science and Education Evaluation (CASEE), Xiasha, Hangzhou, Zhejiang (China P.R.)

[2] *vincent.lariviere@umontreal.ca at*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, Montréal, Quebec (Canada)

Abstract

Maps of scientific knowledge are generally based on citation analysis and therefore reveal how disciplines draw from each other to produce new knowledge. Although subject headings as well as their co-assignments represent the topics and their relationships within the journal article or book, they rarely have been used for mapping science. This study attempts to map the life science based on the Medical Subject Headings (MeSH) as well as their co-assignment at the paper level, which could advance the knowledge in mapping science.

Introduction

The purpose of mapping science is to visualize the scientific structure and the evolution of scientific inquiry (Börner, Theriault, & Boyack, 2015; Klavans & Boyack, 2015) by classifying science and relating the classes, which are generally derived from the analyses of scientific literature elements such as authors, journals, disciplines or other information (Klavans & Boyack, 2009). Although citation analysis is the dominant method for generating maps of science, expert judgements, subject categories, topic modelling, course descriptions, and subject headings could also be used to map the science.

Medical Subject Headings (MeSH) are controlled vocabularies for indexing journal articles and books in the life sciences, which represent all topics discussed within the journal article or book. Since a journal article or book could be assigned multiple MeSHs, the MeSH co-assignments could be used to measure the relationship between two medical topics by which the structure and evolution of life science could be mapped. The purpose of this study is to generate a map of life science using the MeSHs.

Related Works

Expert judgment was first used for mapping science. Bernal (1939) drew the first map of science representing the hierarchical structure of scientific topics by hand. Small and Griffith (1974) created the first citation-based map of science using co-citation analysis. Since then, citation analysis including direct citation (Boyack & Klavans, 2014b; Pan, Zhang, & Wang, 2013; Waltman & Eck, 2012), bibliography coupling (Boyack, 2008), co-citation (Boyack & Klavans, 2014a; Braam, 1991a, 1991b; Small, 1999) was widely used for mapping science.

Other methods in addition to the citation analysis were also used for mapping science. A map of science could be generated based on the co-occurrence of words in titles, abstracts or keywords using the co-word analysis (Ding, Chowdhury, & Foo, 2001; Leydesdroff, 1989; Peters & van Raan, 1993a, 1993b; Rip & Courtial, 1984). Balaban and Klein (2006) mapped science using undergraduate course pre-requisites at Texas A&M University. Suominen and Toivanen (2016) generated a map of science using topic modelling based on the latent patterns in texts retrieved from the Web of Science (WoS).

Subject headings was also applied to generating the map of science. Shu, Dinneen, Asadi, and Julien (2017) produced a map of science based on non-fiction books and their Library of Congress Subject Heading (LCSH) co-assignments. Leydesdorff, Comins, Sorensen, Bornmann, and Hellsten (2016) tried to compare the MeSH with cited sources using clustering and mapping. However, a map of life science based on MeSH has not been generated, which will be addressed by this study.

Method

In this study, in addition to MeSH co-assignment as discussed above, MeSH of citing/cited papers was also used to generate the map as the contrast. Although each MeSH term represents a topic discussed in the journal articles or papers, MeSH terms representing the major topics are marked in the PubMed. Each pair of the major MeSH terms between citing and cited papers also represents the relationship between two major medical topics. Leydesdorff et al. (2016) point out that the citation (citing/cited) map indicates a core structure of life science while the MeSH map shows the relevance of the life science research. Thus, two maps generated from two different approaches were compared in this study.

Data

In this study, 3,344 research papers published in four top medical journals (i.e., *The Journal of the American Medical Association, The Lancet, New England Journal of Medicine*, and *The British Medical Journal*) between 2015 and 2017 as well as their cited references were retrieved from Web of Science (WoS). A version of MEDLINE database integrated into the WoS was used as the linkage between WoS and PubMed in which a PubMed ID and MeSH terms were assigned to each journal article. As noted, not all papers are covered by both WoS and PubMed; in this study, only papers, either citing or cited, with a PubMed ID were included. Eventually, as Table 1 shows, 2,577 papers as well as their 80,782 cited references were collected under investigation; 5,119 and 16,582 MeSH terms were assigned to these citing papers and cited references respectively.

Table 1. Distribution of Papers and Cite References under Investigation in the Study

Journal	*Number of papers in WoS*	*Number of papers under investigation (citing paper)*	*Number of Cited Reference*	*Number of MeSH Terms (citing paper)*	*Number of MeSH Terms (cited reference)*
The Journal of the American Medical Association	658	516	13,889	1,749	9,589
The Lancet	963	658	25,459	2,062	8,696
New England Journal of Medicine	1,003	841	19,866	2,424	11,858
The British Medical Journal	720	562	19,704	2,926	11,812
Total	**3,344**	**2,577**	**80,782**	**5,119**	**16,582**

Note: Since one reference or MeSH term could be cited or assigned to different papers, the sum of the number of cited reference and the sum of number of MeSH term of four journals are higher than the totals in the last row.

Data Treatment

MeSH terms are organized as a 14-level tree structure, representing medical topics from broad to specific. This tree structure starts with 16 level-1 MeSH terms and 118 level-2 MeSH terms, on which the maps of life science were based. Assigned MeSH terms at level 3 or lower were re-assigned to their parent level-2 or grandparent level-1 MeSH terms. This method of reassignment to broader or more general abstraction levels, has been used in library classification mapping where its robustness has been confirmed (Shu et al., 2017).

As shown in Table 2, four datasets were finalized to produce four maps of life science: MeSH co-assignment map at level 1, MeSH co-assignment map at level 2, MeSH citation map at level 1, MeSH citation map at level 2. For each dataset, MeSH terms as well as their co-assignments or citation pairs (major MeSH terms between citing and cited papers) were imported into graph-drawing software *Gephi* to generate the visual map of life science. Each MeSH term was a node while each MeSH co-assignment or citation pair was an edge in the map. The number of assignment of each MeSH term determined the size of node while the number of co-assignment or citation pair decided the weight of each edge. Although the number of citation pairs are much higher than the number of co-assignments in these datasets, they could be normalized when producing the map through Gephi settings.

Table 2. Four Datasets for Four Maps of Life Science

Map	*Number of MeSH term*	*Number of unique MeSH co-assignment or citation pair*	*Total number of MeSH co-assignment or citation pair*
Co-assignment at level 1	16	105	104,832
Co-assignment at level 2	107	3,305	96,776
Citing/cited at level 1	16	110	818,944
Citing/cited at level 2	113	4,767	1,015,203

Results

Figure 1 shows the two maps of life science at the MeSH term level 1 containing 16 nodes/105 edges (lower) and 16 nodes/110 edges (upper) respectively. Nodes are level 1 MeSH terms while edges represent their relationship (i.e., co-assignment and citation pair respectively). Edge width is proportional to the number of co-assignment or citation pair between the two MeSH terms, and the node and label sizes are proportional to the number of assignments or citations.

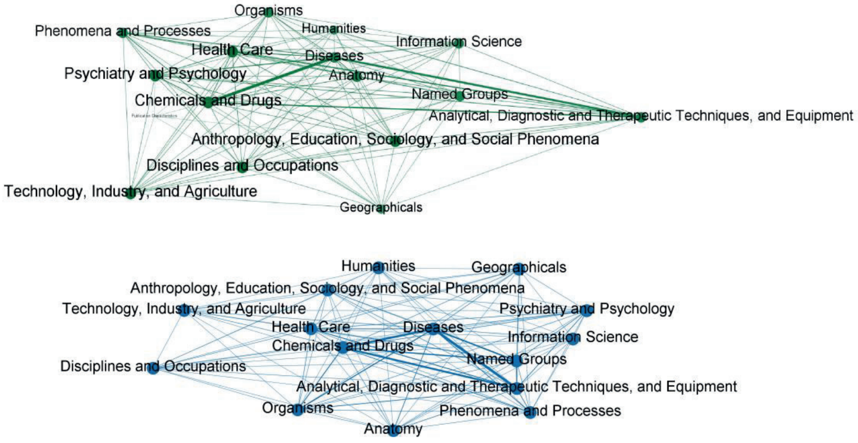

Organisms
Phenomena and Processes
Humanities
Information Science
Health Care
Diseases
Psychiatry and Psychology
Anatomy
Named Groups
Chemicals and Drugs
Analytical, Diagnostic and Therapeutic Techniques, and Equipment
Anthropology, Education, Sociology, and Social Phenomena
Disciplines and Occupations
Technology, Industry, and Agriculture
Geographicals
Humanities
Geographicals
Anthropology, Education, Sociology, and Social Phenomena
Technology, Industry, and Agriculture
Psychiatry and Psychology
Health Care
Diseases
Chemicals and Drugs
Information Science
Named Groups
Disciplines and Occupations
Analytical, Diagnostic and Therapeutic Techniques, and Equipment
Organisms
Phenomena and Processes
Anatomy

Figure 1. MeSH citing/cited map (upper) and co-assignment map (lower) at level 1.

The difference between the co-assignment map and the citation map is not significant. A strong triangle relationship among *Diseases*, *Chemicals and Drugs*, and *Analytical, Diagnostic and Therapeutic Techniques, and Equipment* was found in both maps. Indeed, as Table 3 indicates, the top 10 MeSH terms in the both maps are in the same order and similar shares.

Table 3. Top 10 MeSH terms (Level 1) in Co-assignment Map and Citing/cited Map

MeSH	*Co-assignment Map*	*Citation Map*
Diseases	23.36%	22.12%
Anatomy	16.82%	17.7%
Phenomena and Processes	14.95%	15.04%
Chemicals and Drugs	14.02%	14.16%
Analytical, Diagnostic and Therapeutic Techniques, and Equipment	6.54%	6.19%
Health Care	5.61%	5.31%
Organisms	3.74%	4.42%
Psychiatry and Psychology	3.74%	3.54%
Anthropology, Education, Sociology, and Social Phenomena	2.8%	2.65%
Technology, Industry, and Agriculture	2.8%	2.65%

The color-coded legend of level 1 MeSH terms (see right of Figure 2) were used in the level 2 maps as shown in Figure 2. Nodes are level 2 MeSH terms as the colours of nodes represent their parent MeSH terms at level 1. Some differences were found when comparing the co-assignment map (lower of Figure 2) and the citation map (upper of Figure 2). The distribution of MeSH of citing/cited papers is skewed as some large nodes and wide edges appear in the citation map, while the distribution of MeSH co-assignments is more balanced. However, comparing with Figure 1, Figure 2 is visually complex due to high connectivity between the nodes and overlapping edges.

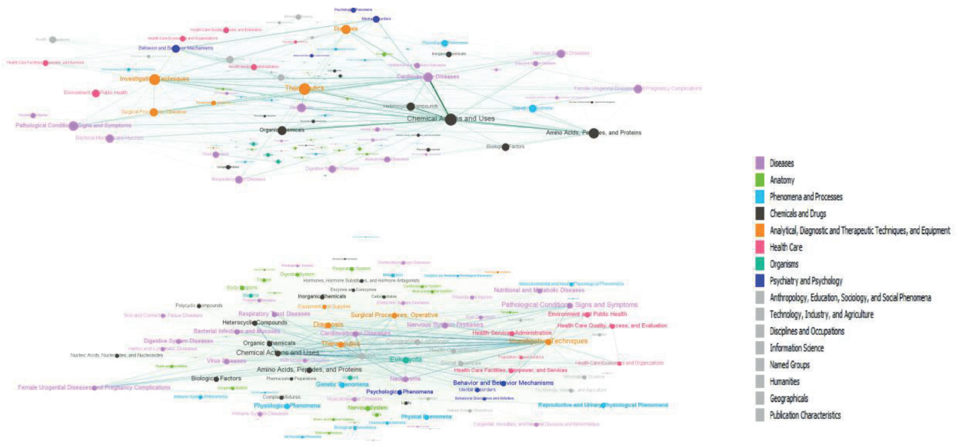

Diagnosis
Behavior and Behavior Mechanisms
Investigative Techniques
Therapeutics
Environment and Public Health
Surgical Procedures, Operative
Pathological Conditions, Signs and Symptoms
Cardiovascular Diseases
Heterocyclic Compounds
Chemical Actions and Uses
Organic Chemicals
Amino Acids, Peptides, and Proteins
Biological Factors
Female Urogenital Diseases and Pregnancy Complications
Hormones, Hormone Substitutes, and Hormone Antagonists
Inorganic Chemicals
Polycyclic Compounds
Respiratory Tract Diseases
Surgical Procedures, Operative
Nutritional and Metabolic Diseases
Bacterial Infections and Mycoses
Digestive System Diseases
Nervous System Diseases
Health Services Administration
Health Care Quality, Access, and Evaluation
Nucleic Acids, Nucleotides, and Nucleosides
Virus Diseases
Eukaryota
Neoplasms
Genetic Phenomena
Psychological Phenomena
Health Care Facilities, Manpower, and Services
Physiological Phenomena
Nervous System
Physical Phenomena
Reproductive and Urinary Physiological Phenomena
Diseases
Anatomy
Phenomena and Processes
Chemicals and Drugs
Analytical, Diagnostic and Therapeutic Techniques, and Equipment
Health Care
Organisms
Psychiatry and Psychology
Anthropology, Education, Sociology, and Social Phenomena
Technology, Industry, and Agriculture
Disciplines and Occupations
Information Science
Named Groups
Humanities
Geographicals
Publication Characteristics

Figure 2. MeSH citing/cited map (upper) and co-assignment map (lower) at level 2.

Discussion and Conclusion

In this study, four maps of life science were generated using the MeSH terms assigned to 2,577 papers published in four top medical journals between 2015 and 2017 as well as their cited reference. Few difference was found when comparing the co-assignment map with the citation map at the MeSH level 1. It indicates that the MeSH co-assignment representing the relationship among different medical topics could also be used to map the life science comparing to traditional citation-based maps generated by the citing/cited relations. The results of this study could form a foundation for future studies mapping the life science using MeSH terms.

In addition, this study found the difference in terms of the MeSH term distribution between the co-assignment map and the citation map at the MeSH level 2, which could partly be due to the different functions between subject headings and citations. Subject headings emphasize the correlation of all related topics discussed in the journal articles or books while citations measure the similarity of citing and cited documents, which has been addressed by Leydesdorff et al. (2016).

As a research-in-process paper, this study only sampled 2,577 research papers from four top medical journals, a full dataset containing all medical articles should be investigated in the future studies. In addition, different visualization methodologies, in turn stemming from choice of visualization software, may also influence the visualization of the map, which should also be addressed in the future studies.

References

Balaban, A. T., & Klein, D. J. (2006). Is chemistry "The Central Science"? How are different sciences related? Co-citations, reductionism, emergence, and posets. *Scientometrics, 69*(3), 615-637.

Bernal, J. D. (1939). The social function of science. *The Social Function of Science.*

Börner, K., Theriault, T. N., & Boyack, K. W. (2015). Mapping science introduction: Past, present and future. *Bulletin of the American Society for Information Science and Technology, 41*(2), 12-16.

Boyack, K. W. (2008). Using detailed maps of science to identify potential collaborations. *Scientometrics, 79*(1), 27-44.

Boyack, K. W., & Klavans, R. (2014a). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology, 65*(4), 670-685.

Boyack, K. W., & Klavans, R. (2014b). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics, 8*(3), 569-580.

Braam, R. R. (1991a). Mapping of Science by Combined Co-Citation and Word Analysis. I. Structural Aspects. *Journal of the American Society for Information Science, 42*(4), 233-251.

Braam, R. R. (1991b). Mapping of Science by Combined Co-Citation and Word Analysis. II: Dynamical Aspects. *Journal of the American Society for Information Science, 42*(4), 252-266.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*(6), 817-842.

Klavans, R., & Boyack, K. W. (2009). Toward a Consensus Map of Science. *Journal of the American Society for Information Science and Technology, 60*(3), 455-476.

Klavans, R., & Boyack, K. W. (2015). Exploring the relationships between a map of altruism and a map of science. *Bulletin of the American Society for Information Science and Technology, 41*(2), 30-33.

Leydesdorff, L., Comins, J. A., Sorensen, A. A., Bornmann, L., & Hellsten, I. (2016). Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level. *Scientometrics, 109*(3), 2077-2091. doi:10.1007/s11192-016-2119-7

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*(4), 209-223.

Pan, J., Zhang, X., & Wang, X. (2013). Mapping the Structure and Evolution of Science. *Journal of information processing and management, 52*(8), 255-266.

Peters, H. P. F., & van Raan, A. F. J. (1993a). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy, 22*(1), 23-45.

Peters, H. P. F., & van Raan, A. F. J. (1993b). Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. *Research Policy, 22*(1), 47-71.

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics, 6*(6), 381-400.

Shu, F., Dinneen, J. D., Asadi, B., & Julien, C.-A. (2017). Mapping science using Library of Congress Subject Headings. *Journal of Informetrics, 11*(4), 1080-1094.

Small, H. (1999). Visualizing Science by Citation Mapping. *Journal of the American Society for Information Science, 50*(9), 799-813.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science studies*, 17-40.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human‐assigned subject classification. *Journal of the Association for Information Science and Technology*.

Waltman, L., & Eck, N. J. (2012). A new methodology for constructing a publication‐level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.

How to interpret algorithmically constructed topical structures of research specialties? A case study comparing an *internal* and an *external* mapping of the topical structure of invasion biology

Matthias Held[1] and Theresa Velden[2]

[1] *held@ztg.tu-berlin.de*
Zentrum für Technik und Gesellschaft, Technische Universität Berlin, Hardenbergstr. 16-18, D-10623 Berlin (Germany)

[2] *velden@dzhw.eu*
Deutsches Zentrum für Hochschul- und Wissenschaftsforschung Berlin, Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

In our paper we seek to address a shortcoming in the scientometric literature, namely that, given the proliferation of algorithmic approaches to topic detection from bibliometric data, there is a relative lack of studies that validate and create a deeper understanding of the topical structures these algorithmic approaches generate. To take a closer look at this issue, we investigate the results of the new Leiden algorithm when applied to the direct citation network of a field-level data set. We compare this 'internal' perspective which is constructed from the citation links within a data set of 30,000 publications in invasion biology, with an 'external' perspective onto the topic structures in this research specialty, which is based on a global science map in form of the CWTS microfield classification underlying the Leiden Ranking. We present an initial comparative analysis of the results and lay out our next steps that will involve engaging with domain experts to examine how the algorithmically identified topics relate to understandings of topics and topical perspectives that operate within this research specialty.

Introduction

While algorithms and their application to extract topical structures from bibliometric data proliferate, there is a shortage of studies that validate their results and contribute to a deeper understanding of the variation in topical structures that these algorithmic approaches create. So far only a small set of studies exists that systematically investigate the validity of solutions obtained and the difference made by alternative choices (see e.g. Haunschild et al. 2018, Sjögårde 2018, Klavans & Boyack 2017, Velden et al. 2017, Šubelj et al 2016, Boyack & Klavans 2010, Klavans & Boyack 2011, Shibata et al. 2009). We are concerned that failure to invest into the systematic comparisons and careful validation of the interpretation of algorithmically extracted topical structures undermines our ability to provide robust interpretations of their results and sound guidance on the choice and appropriate use of algorithmic topic extraction or field classification approaches. This study is a contribution to what we conceive of as a larger, much needed, research program. This program should address, on the one hand, the question of the constructive nature of a topic extraction result, relying on decisions on dataset, data model, algorithm and its parameters (Gläser et al. 2017). On the other hand, it should address the validity of interpretations of results by examining the degree of agreement between a theoretical definition of the topic concept (Havemann et al. 2017) with the actual operationalization through the chosen approach, and by exploring their correspondence to perceptions of topical structures held by their creators "in-the-wild", the researchers themselves.

This paper is a case study that takes a closer look at the topical structures obtained when using the newly released Leiden algorithm for community detection (Traag et al. 2018) to produce a local map of the field of invasion biology. As a first step, we compare this 'internal' perspective that is based exclusively on relations in the direct citation network[1] of approx. 30,000 publications in invasion biology, with an 'external' perspective that is generated by projecting this data set of publications onto a global map of science. We use the global map that underlies the field classification of the Leiden Ranking[2] and consists of approx. 20 mio publications grouped into 4047 microfields. It has been produced by CWTS using the Smart Local Moving algorithm for community detection (Waltman & Van Eck 2013). Such an external perspective captures the embedding of publications in a field into the global network of scientific publications and is expected to highlight interdisciplinary connections to other areas of research (Boyack 2017).

Klavans and Boyack (2011) argue that under certain conditions[3] a global science map can be expected to produce a more 'accurate' map of a field than local maps can - where accuracy is measured by textual coherence of the clusters obtained. However, as Haunschild et al. (2018) found in a case study of the topic of 'overall water splitting', also global maps may fail to adequately capture research fields. In this study, given the sparseness of evidence so far, rather than dismiss the local map as less accurate per se, we keep an open mind. Our interest is to investigate the capability of either perspective, internal or external, to capture understandings of topics that operate within the research specialty of invasion biology[4]. In the following we present an initial bibliometric comparison of the results obtained with these two mapping approaches, and discuss our next steps that will involve engaging with field experts who do research in invasion biology to discuss interpretations of the results of these alternative algorithmic mappings of their field of research.

Data & Methods

In this study we use a data set that is based on a lexical query developed by researchers in invasion biology (Vaz et al. 2017) in order to capture publications belonging to their research specialty:

"Ecological invasion*" or "Biological invasion*" or "Invasion biology" or "Invasion ecology" or "Invasive species" or "Alien species" or "Introduced species" or "Non-native species" or "Nonnative species" or "Nonindigenous species" or "Non-indigenous species" or "Allochthonous species" or "Exotic species".

Using the lexical query above, 30,731 document IDs from the Web of Science database were retrieved on August 28, 2017 (Figure 1). For this set, we were able to retrieve the relevant metadata (titles, abstracts, source, publication year, document types, cited references) from the 2018 stable version of the Web of Science database hosted by the 'Kompetenzzentrum Bibliometrie' (KB). We decided to restrict the time window to the years 2000-2017[5], and the document types to article, letters and reviews. The further analysis was done using the giant

[1] Direct citation networks are a popular choice in bibliometrics for the production of global science maps, given their relative sparseness. Previous studies (Klavans & Boyack 2017, Velden et al. 2017a, Shibata et al. 2009) suggest its usefulness to extract taxonomic topic structures.

[2] http://www.leidenranking.com/

[3] "All things being equal", meaning if data source, data model, algorithm and so forth are the same. They also suggest that local maps will be less accurate in particular if boundary forces (links to concepts outside the field) are stronger than core forces.

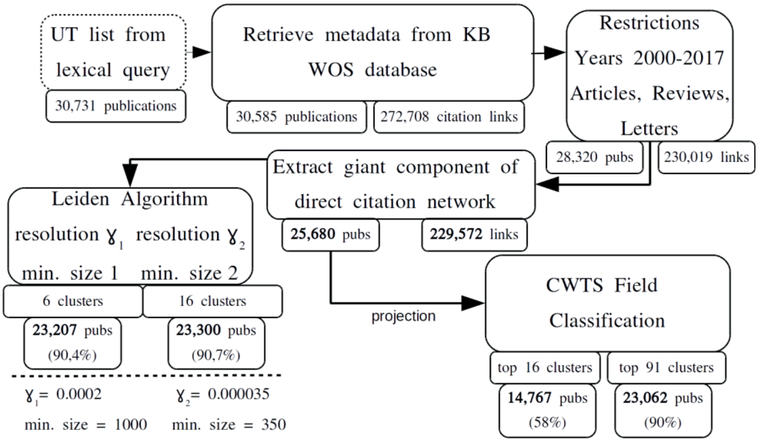
[4] Our interest in validating algorithmically generated science maps derives from theory-guided empirical work we are engaged in: M. Held is involved in the project MIMAL that explores linkages between bibliometric patterns at the micro level and the macro level; T. Velden is involved in the project 'Field specific forms of open science' that compares four fields of science and uses bibliometric maps of research specialties to support comparisons in ethnographic science studies.

[5] A specific reason for this choice was the desire to increase comparability with another bibliometric study of the field by Enders et al, (in preparation), as well as the general consideration that the field of invasion biology has experienced critical growth in the early 2000's such that the large majority of publications in the field is included in the chosen time window.

component of the direct citation network. The network of the remaining 25,680 publications and 229,572 citation links[6] served as input for the Leiden algorithm and the projection onto the CWTS microfield classification (explained below).

For clustering the direct citation network, we chose the recently released Leiden algorithm (Traag et al. 2018), a community detection algorithm which has been developed to overcome a decisive shortcoming of a widely used community detection algorithm, the Louvain algorithm (Blondel et al. 2008), namely the production of badly connected clusters. It further avoids the use of modularity as quality function due to its known shortcoming of a resolution limit and instead chooses the quality function Constant Potts Model (CPM) that has been shown to be resolution-limit free (Traag et al. 2011, 2018). For the CPM we chose two resolution values and minimum cluster sizes. Different from the methodology introduced in Waltman & Van Eck (2012), we do not merge clusters below the threshold, and instead discard them. The publications from those discarded clusters amount to less than 10% of the publications in both solutions Leiden$_6$ and Leiden$_{16}$. The algorithm was started with a random seed, run with 100 iterations with ten random starts each.

Figure 1: Schematic representation of the processing steps.



To contrast this ‘internal’ perspective of a clustering of a research specialty with an ‘external’ perspective that takes the embedding of publications in the global citation network of science into account, we project our field data set onto the CWTS microfield classification. It consist of 4047 microfields that have been extracted with the SLM algorithm on the weighted direct citation network of more than 20 million publications published in 2000-2017 and indexed by the Web of Science[7]. Of the 25,680 UTs included in the giant component, 25,627 can be found in the micro fields of the CWTS field classification. We defined as clusters in our projection cluster solution the largest intersections between our field data set and a microfield (in terms of absolute number of publications).

Finally, in order to find characteristic terms to describe the content of clusters, we extracted the noun phrases from titles and abstracts of the publications of each cluster in each cluster solution, using part of speech tagging and chunking available in the Python package ‘nltk’. Terms which had been used in the lexical query to delineate the field were excluded. To obtain a measure for how well each of the remaining terms describes the content of each cluster, we used the

[6] Following Waltman & Van Eck (2012), we produced a weighted version of the direct citation network to account for a potential variation of in citation practices within the field of invasion biology.

[7] http://www.leidenranking.com/information/fields (Accessed January 25, 2019)

differential cluster labelling by Koopman & Wang (2017), which is based on normalized mutual information (NMI). The higher the value, the more significant the term to characterize the cluster and differentiate it from the rest. For labelling the clusters in the Leiden_6 and Leiden_{16} solution, the terms of all publications from the giant component were included. In order to label the clusters in the projection solution, we only included terms and publications which occurred in the projection clusters. An additional set of labels was produced with the same approach, using journal names instead of extracted terms. The cluster labels eventually used in this paper were manually derived from those NMI score ranked lists of terms and journals by considering information provided on habitat, organisms, research problem, and the subject area of journals.

In order to visualize the relationships between the topics identified, we use topic affinity networks that evaluate the strength of citation links between clusters to determine the affinity between topics. The existence of a link between topics in the affinity network indicates a surplus of connectivity between the two topics compared to a random null model (see Velden, Yan & Lagoze 2017 for details).

Results

Topics and sizes of the 6, respectively 16 clusters in the Leiden_6 /Leiden_{16} solutions are given in the data appendix of the arXiv version of this paper[8]. An analysis of how the publications are regrouped from the six clusters of the Leiden_6 solution to the 16 clusters of the Leiden_{16} solution suggests a continuity of the overall topic structure extracted by the two clustering runs. While two topics, 'marine invasion' (Leiden_6 C3) and 'trees and pests' (Leiden_6 C5) are largely preserved, other clusters get split into smaller, refined topics. Given that the two solutions are independent and not the result of a hierarchical clustering approach, this seems noteworthy and encouraging regarding an internal consistency of results achieved with the Leiden algorithm at different levels of resolution.

Sizes and topics of some of the larger projection clusters, as well as information of the CWTS microfields that the projection clusters are embedded in are also given in the data appendix. But for a few exceptions, the projection clusters constitute only about 5% of a microfield in the CWTS classification. Adopting the terminology used by Klavans and Boyack (2011) when comparing a local and a global mapping of the field of information science, microfield m402 may be considered a 'core' microfield for invasion science (53% of its publications overlap with the invasion data set and constitute projection cluster C1 on 'invasive plants'). Three microfields may be considered 'boundary' microfields, namely m2749 (34% overlap, projection cluster C5 on 'marine aquatic invasion, ballast water, ascidians'), m1774 (17% overlap, projection cluster C2 on 'freshwater aquatic invasion, great lakes'), and m2568 (17% overlap, projection cluster C10 on 'freshwater aquatic invasion, crayfish'). All other microfields may be considered 'boundary-crossing', i.e. largely outside the field of investigation.

In Figure 2 we compare the Leiden_{16} with the projection cluster solution based on their topic affinity networks. Both solutions agree in that they include a cluster related to 'invasive plants' that consists of almost 25% of publications in the giant component. They differ in that the cluster size concentration of the projection cluster solution is lower: to cover a similar large proportion of publications from the giant component as the Leiden_{16} solution (> 90%), one has to include the 91 largest clusters of the projection cluster solution, down to a size of 37 publications. The alluvial diagram in Figure 3 shows the regrouping of publications between the Leiden_{16} solution and the 91 largest topics in the projection solution. While some topic continuity can be observed and the core of some topics clearly persists, other topics get

[8] Available online at http://arxiv.org/abs/1905.03485.

fragmented. The zoom-in in Figure 3 shows how the topics of 'marine aquatic invasion' (C2 Leiden$_{16}$) and 'ballast water' (C15 Leiden$_{16}$) split-up into numerous topics in the projection solution (these topics are located next to each other in the area of the affinity network of the projection solution circled in Figure 2). We offer two observations: first, a technical one, namely that our labelling procedure fails to extract hints of organisms or research angle when cluster sizes fall below about 100 documents (see labels in Fig. 3 for C54, C56 or C58). Second, the refinement of topics often seems driven by a focus on a specific organism: crab, algae, oyster, jellyfish, etc. Occasionally, it is driven by a specific habitat: Mediterranean, Suez canal, Antarctica.

Discussion & Future Work

The alternative internal and external mappings in this paper provide two perspectives onto the topical structure of the field of invasion biology. Common feature of those topical structures is the concentration of almost 25% of publications in a topic relating to invasive plants. This is in alignment with statements by experts that suggest that observational studies of terrestrial plants dominate the empirical literature in the field (Jeschke & Heger 2018, p. 162). Further, organisms and habitat seem to constitute important dimensions for delineating topics - likely not a surprising observation for a domain expert. Notable exceptions are the projection based topics on 'genetic diversity' (C12), and 'models, climate change' (C7) that seem rather method or research angle-driven.

The projection of the invasion biology data set onto microfields of the global map promises insights into links between topics in invasion science to neighbouring fields. However, it suffers from an enormous spread. One third of publications are represented by 'invasive plants' as a core topic and three aquatic invasion-related boundary topics. The remaining ⅔ of publications are associated with and embedded in hundreds of different microfields with each less than a 10% share. This spread of invasion biology publications across microfields suggests that a delineation and representation of the research field of invasion biology through selection of a microfield from the global map might be ill-conceived - given one trusts the lexical query used by us and by Vaz et al. (2017) to delineate the field of invasion biology. This aligns with the finding by Haunschild et al. (2018) on the field of overall water splitting. The question what these microfields are representative of and their role as a topical context for research in a research specialty such as invasion biology, deserves further investigation.

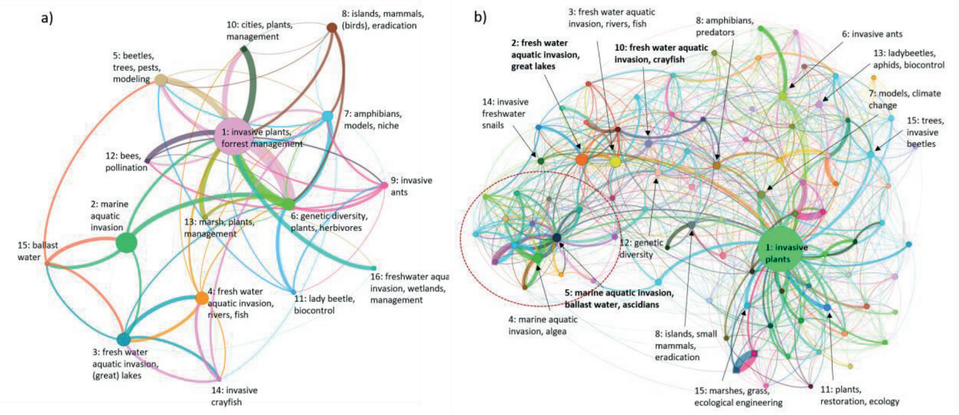
Before moving on, we plan to improve our labelling approach by implementing entity recognition of taxonomic species. This way we expect to increase the meaningfulness and precision of the content labels we extract, allowing us e.g. to contrast the content of the projection clusters with the other publications in the microfield they are embedded in.

The next major step in our study will be to explore in interviews and informal discussions with domain experts from the field of invasion biology as well as through ethnographic observations in an ongoing study, the relationship between the topic structures constructed by the chosen algorithmic approaches with the lived experience of researchers in the field of invasion biology. Specifically, we plan to pursue the following avenues:

1. How do individual research group leaders' research trails (Gläser & Laudel 2015) relate to the topical structures of the Leiden$_{16}$ and projection solutions? Do research topics that can be delineated within those trails align with or transcend the field topics we have constructed algorithmically?
2. Existing, theoretical work on empirical evidence in the field of invasion biology identifies theoretical work on key hypothesis of the field as well as empirical studies

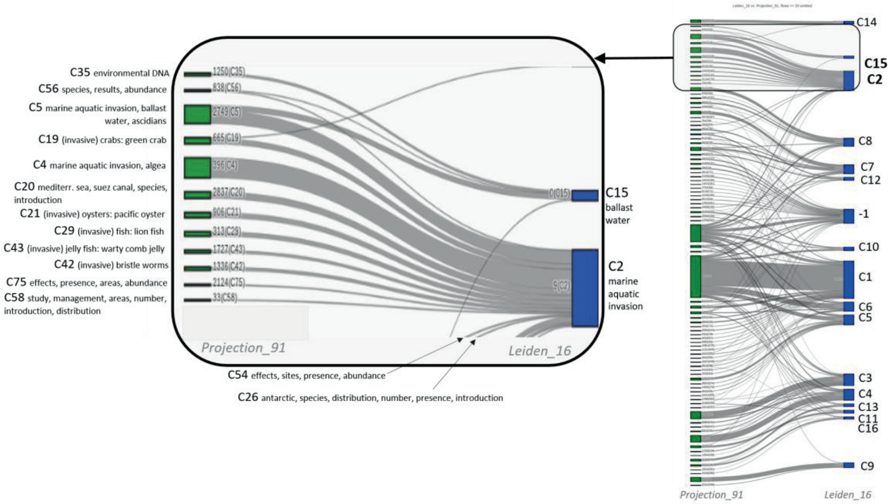
that support or challenge those hypotheses[9]. This offers the opportunity to relate the topical structures we have generated to relevant topical perspectives generated from within the field and discuss these relations with domain experts.

Figure 2: Topic affinity networks[10] of a) Leiden$_{16}$ solution and b) projection solution.



a)
8: islands, mammals, (birds), eradication
10: cities, plants, management
5: beetles, trees, pests, modeling
7: amphibians, models, niche
1: invasive plants, forrest management
12: bees, pollination
9: invasive ants
2: marine aquatic invasion
13: marsh, plants, management
6: genetic diversity, plants, herbivores
15: ballast water
16: freshwater aqua invasion, wetlands, management
4: fresh water aquatic invasion, rivers, fish
11: lady beetle, biocontrol
3: fresh water aquatic invasion, (great) lakes
14: invasive crayfish
b)
3: fresh water aquatic invasion, rivers, fish
8: amphibians, predators
2: fresh water aquatic invasion, great lakes
10: fresh water aquatic invasion, crayfish
6: invasive ants
13: ladybeetles, aphids, biocontrol
7: models, climate change
14: invasive freshwater snails
15: trees, invasive beetles
12: genetic diversity
1: invasive plants
5: marine aquatic invasion, ballast water, ascidians
4: marine aquatic invasion, algea
8: islands, small mammals, eradication
15: marshes, grass, ecological engineering
11: plants, restoration, ecology

Figure 3: Alluvial showing flow between clusters in Leiden$_{16}$ and Projection 91.



Conclusions

We report first results on a comparison of an internal and an external mapping of topical structures in the research specialty of invasion biology. Both maps exhibit some common features, like the importance of work on invasive plants, and the relevance of concepts of habitat and organism for distinguishing topics. The next step in the study will be dedicated to relating the algorithmically identified topic structures to topical concepts emerging from social and theoretical processes within the research specialty.

[9] https://hi-knowledge.org/

[10] Node size reflects number of publications (viz. gephi), links reflect disproportionately strong affinity. Link curvature indicates link direction (clockwise).

Acknowledgments This research of M. Held was supported by the German Federal Ministry of Education and Research (Grant 01PU17003).

References

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008), *J. Stat.Mech.Theory Exp*. 10008, 6.

Boyack, K. W. (2017). Investigating the effect of global data on topic detection. *Scientometrics*, 111(2), 999-1015.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *JASIST*, 61(12), 2389–2404.

Gläser, J., & Laudel, G. (2015). A bibliometric reconstruction of research trails for qualitative investigations of scientific innovations. Historical Social Research/Historische Sozialforschung, 299-330.

Gläser, J., Heinz, M., & Havemann, F. (2015). Epistemic Diversity as Distribution of Paper Dissimilarities. In Proceedings of *ISSI*.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*(2), 981-998.

Haunschild, R., Schier, H., Marx, W., & Bornmann, L. (2018). Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting. *Journal of Informetrics*, *12*(2), 436-447.

Havemann, F., Gläser, J., & Heinz, M. (2017). Memetic search for overlapping topics based on a local evaluation of link communities. *Scientometrics*, *111*(2), 1089-1118.

Jeschke, J.M.; Heger, T. (eds) (2018). Invasion biology: hypotheses and evidence. CABI, Wallingford, UK.

Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for information Science and Technology*, *62*(1), 1-18.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *JASIST*, *68*(4), 984-998.

Koopman, R., & Wang, S. (2017). Mutual information based labelling and comparing clusters. *Scientometrics*, *111*(2), 1157-1167

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *JASIST*, 60(3), 571–580.

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133-152.

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PloS one*, *11*(4), e0154404.

Traag, V. A., Van Dooren, P., & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, *84*(1), 016114.

Traag, V., Waltman, L., & van Eck, N. J. (2018). From Louvain to Leiden: guaranteeing well-connected communities. *arXiv preprint* arXiv:1810.08473.

Vaz, A. S., Kueffer, C., Kull, C. A., Richardson, D. M., Schindler, S., Muñoz-Pajares, A. J., ... & Honrado, J. P. (2017). The progress of interdisciplinarity in invasion science. *Ambio*, 46(4), 428-442.

Velden, T., Yan, S., & Lagoze, C. (2017). Mapping the cognitive structure of astrophysics by infomap clustering of the citation network and topic affinity analysis. *Scientometrics*, 111(2), 1033-1051.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017a). Comparison of topic extraction approaches and their results. *Scientometrics*, 111(2), 1169-1221.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *JASIST*, *63*(12), 2378-2392.

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 471.

Context matters: on the usage and semantics of hedging terms across sections of scientific papers

Dakota Murray[1], Vincent Larivière[2], Cassidy R. Sugimoto[1]

[1] *dakmurra@iu.edu; sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, USA

[2] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

Abstract

The electronic availability of the full-text of scholarly publications has made possible large-scale analysis of in-text citations. Many such studies have made use of signal terms—individual words or phrases that suggest some phenomenon of interest. Most studies utilizing signal terms treat them as having stable meaning; however, the usage and meaning of these terms may differ between rhetorical contexts. We conducted a preliminary analysis to investigate the extent to which the usage and semantics of *hedging* signal terms differed between the introduction, methodology, and discussion sections of scientific papers. We sampled three million sentences containing citations from a large database of full-text publications and counted the occurrences of sixteen hedging terms in each section. We found that the incidence of hedges varied across sections with distinct patterns for individual terms. We modelled semantic relationships within a section using word2vec and constructed rudimentary measures to compare the similarity of hedging terms between models trained on each section. We observed that the meaning of hedging terms differed between sections with distinct patterns for individual terms. Our results suggest that signal terms are not independent of their context which has implications for bibliometric full-text analysis.

Introduction

For more than 40 years, sociologists and information scientists have worked towards a better understanding of the function of citations (Moravcsik & Murugesan, 1975; Cronin, 1981; Small, 2004). Early analyses, while insightful, were limited by the lack of accessible data (i.e. electronic journals) and computational tools. However, the increased ubiquity of digital scholarly publications and availably of large-scale databases has made possible sophisticated textual analyses of in-text citations. Some analyses attempted to characterize the contextual attributes of in-text citations, such as their position and distribution within publications (Bertin et al., 2016a; Boyack et al., 2018). Others instead sought to develop machine learning algorithms to automatically classify *citances* (sentences containing citations) according to their function (Teufel, Siddharthan, & Tidhar, 2006), sentiment (Catalini, Lacetera, & Oettl, 2015; Jha et al., 2017), importance (Valenzuela, Ha, & Etzioni, 2015) and more.

A common input feature for citation classifiers is the presence of *signal terms* within a citance. For example, the terms "excellent" and "poor" may prove useful for the task of citation sentiment classification. Other papers have made use of specific dictionaries of signal terms to study more abstract concepts: for example, Small et al., (2017) uses a list of "discovery" terms to track scientific discoveries, whereas Chen et al., (2018) identifies terms that suggest uncertainty. A paper by Di Marco and Mercer (2004) described how a certain type of signal term—*hedges*—might be used to classify citations; however, they also noted how the semantic meaning and usage of hedging terms might differ between sections of a paper. This speaks to a wider issue with the use of signal terms in bibliometric analysis and classification: rhetorical context matters. The meaning or usage of a hedging, discovery, or uncertainty signal term may differ depending on whether the term appears in the introduction, methodology, or discussion. A term is not independent of its context (Bertin et al., 2016b).

In this paper we present a preliminary analysis of the extent to which the incidence and semantic meaning of common signal terms differ between rhetorical contexts of a manuscript. We focus on a particular type of signal term: *hedging* as defined by Hyland (1996), discussed by Di Marco and Mercer (2004), and used in Chen et al. (2018). We consider three distinct rhetorical contexts: the introduction, methodology, and discussion sections. This analysis will inform future textual analyses of scientific publications and establish a foundation for further investigations into how semantics differ between rhetorical and scientific contexts.

Data and Methods

We used data from the Elsevier ScienceDirect database, obtained and managed by the *Centre for Science and Technology Studies* at Leiden University. This data contains full-text for nearly five million English-language full-length articles, short communications, and review articles published between 1980 and 2016. More information about this dataset, including a comparison to the existing PubMed dataset, and a descriptive analysis, can be found in Boyack et al. (2018).

We first sampled 300,000 full-length English-language articles that were published after 2014. From these publications, we sampled approximately one million citances (citation sentences) listed in a section with a name that included the substring "intro"; we repeated this sampling for citances in sections with titles containing the substrings "method" and "discussion" resulting in three corpora. We tokenized sentences within each of these corpora into unigrams and bigrams and tallied the counts and proportions of each token.

We represented the semantic relationships between terms in each corpus using word2vec, a method of training vector space word embeddings from a corpus of text (Mikolov et al., 2013). Each trained word vector represents a single word; words with similar contexts will have high cosine similarities. For example, "researcher" would likely have a similar context as "scientist", and "scholar" and so their vectors would have high cosine similarities. We trained separate word2vec models for each corpus using *genism v3.6.0*. We adhered to the same training parameters as Chen et al., (2018), though we decreased the number of training vectors from 200 to 100 due to our smaller training data; we constructed selective models by including only terms that occur at least 100 times in the corpus.

To compare the meaning of a target term between word2vec models, we need a measure of similarity between models. However, due to the stochastic nature of word2vec training, direct comparison of word vectors between models is essentially meaningless, even when trained on the same data. Some researchers have developed techniques to compare models, but there is little consensus on appropriate technique (Kutuzov, 2018).

We considered two approaches to assess the similarity of a target word's semantic meaning between models. The first approach leverages *global* relationships between the target word and a large sample of words shared between the two models. For two word2vec models we defined their shared common vocabulary. We then measured the cosine similarity between the target term and each word in the shared vocabulary. We used the squared correlation between these two sets of similarities as a measure of semantic similarity. As an additional validity check we trained two word2vec models on an identical corpus of 200,000 abstracts and calculated their global semantic similarity. The smallest value of the 16 hedging terms was 0.983 with a mean of 0.992. That the correlation was high between two word2vec models trained on the same data suggests that this measure of global similarity is relatively robust.

For the second approach we examined only *local* relationships between the target term and a small list of similar words. For a target term, we selected a small number of the most similar words from the introduction word2vec model. We then compared the rankings of these words to the corresponding rankings of the target word in the methods and discussion word2vec models. We created an ad-hoc measure of change in total ranking by dividing the change in rankings between each model by the position of the word in the introduction model similarity list; this index downweighs terms appearing lower in the ranking list. Interpretation of this model's results is less straightforward than for global semantic similarity, but it provides insight into the particular relationships of individual terms.

Results and Discussion

We first assessed the extent to which the incidence of hedging phrases differed across citances by section of scientific articles. Figure 1 shows the counts of the stemmed version of each term within the introduction, methods, and discussion sections. Overall, more hedging terms were used in paper's discussion sections than in either the introduction or methods sections. This difference likely resulted from the rhetorical characteristics inherent to each section, rather than other confounding factors: the total number of citances sampled for the three sections was approximately equal. Moreover, the average number of tokens for a citance in each section differed by at most two tokens (around 28 tokens for methods, and 30 for discussion section).

Whereas hedging tended to be most common in the discussion, perhaps due to the section's typically speculative nature, we also observed differences at the level of individual terms. The most common hedges such as "should", 'may", "indicate", and "might" largely appeared in the discussion section. However, "will" was primarily used in the introduction, whereas "predict" was instead more often used in the methodology. There were also differences in the usage of words between the introduction and methodology: the usage of "should", "must', and "cannot", was roughly similar between the two sections. These findings speak to the heterogeneity of hedging terms—usage of individual terms differed by section.

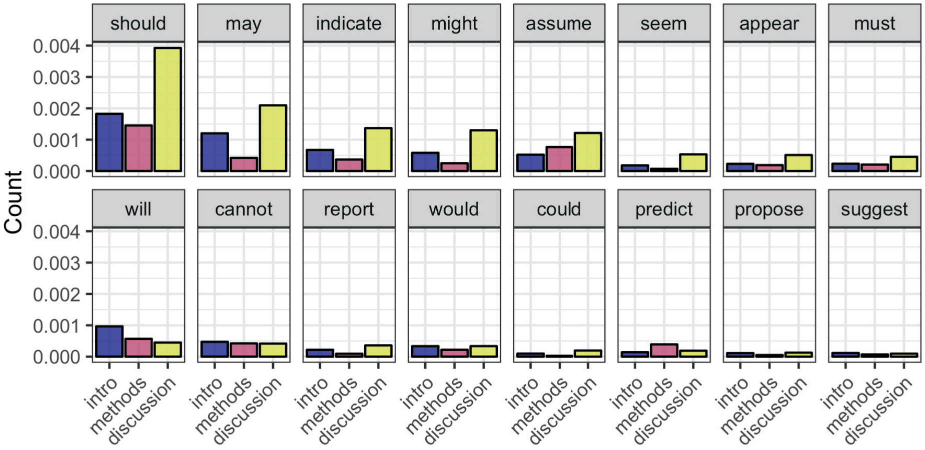


Figure 1. Counts of hedging terms appearing in the introduction, methods, and discussion corpuses. Each corpus includes approximately one million citances. Terms are arranged in order of total instances. These counts include the stemmed version of each term; for example, the count for "indicate" also includes the counts for terms such as "indicates" and "indicated".

In addition to differences in incidence, so too might the *semantic meaning* of individual hedging terms differ by section. For example, "predict" might be used in an introduction to establish a hypothesis: "we predict that *x* is related to *y*". However, in the methods section, this term may instead be used in a more technical context: "we used linear regression to predict *y* given *x*"; in this case, the meaning of the word no longer serves as a hedge in the methodology.

We modelled the semantic meaning of terms by training word2vec models for each section. We investigated the extent to which the semantic meaning of each hedging term differed between each section by calculating an ad-hoc measure of global semantic similarity for each hedging term, and between every combination of two models (Figure 2). We found that, overall, the usage of almost all hedging terms was most similar between the introduction and discussion sections and tended to be least similar between the methodology and discussion sections. Generally, the semantic meaning of hedging terms was similar between the introduction and methods and between the methods and discussion sections.

As in our analysis of the incidence of hedging we also observed heterogeneity in patterns of semantic similarity by individual terms. For example, the term “may” exemplified the general trend of high similarity between the introduction and discussion, but low similarity between these and the methodology section; however, the terms “must”, “indicate”, and “propose” ran counter to this trend. Model comparisons for the term “must” showed a similar trend as “may”, though with a much higher degree of similarity between the methodology and other sections. The semantic similarity of the term “indicate” was instead roughly similar between the introduction and methods, as well as between the introduction and discussion. The term “propose” presented a clear contrast to all other hedging terms, such that the highest semantic similarity was observed between the introduction and methodology sections. This analysis provides evidence that the semantic similarity of hedging differed between sections, and that these patterns of similarity were heterogeneous across individual hedging terms.

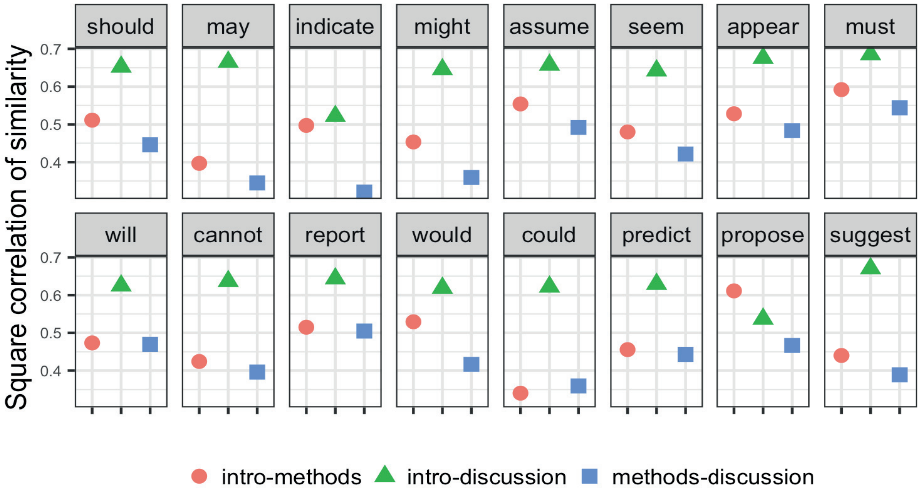


Figure 2. Global term similarity between each hedging term and for each combination of word2vec models. Global term similarity is defined as the adjusted R^2 of the line of best for the top 100,000 most common words shared between both models.

One limitation of this analysis of semantic similarity was that it relies on a measure of *global* semantic similarity, that is, the similarity between a target term and thousands of terms in the shared vocabulary between two models. This measure weighs similarities equally between a target word and all terms in the shared vocabulary. However, a small vocabulary of the most similar words may prove more useful for determining semantic similarity. For example, if we compared the semantic meaning of “dog” across two models, we would be more interested its similarity to “animal” and “pet” than its similarity to “computer”. A natural approach is to assess semantic similarity of terms at the *local level*—that is, for a select list of the top *n* most similar terms. Following this approach, we identified the top fifteen most similar terms from the introduction model and traced their change in rankings across the methodology and discussion sections. For this preliminary analysis, we only considered three hedging terms: “indicate”, “must”, and “propose”, each of which exhibited distinct patterns of similarity between the three models. The results of this analysis are shown in table 1.

From this small vocabulary, we observed a mix of both similar and divergent patterns of similarity to those observed in figure 2. Using our global measure of semantic similarity, we found that the term “must” was similar between all three models—here, we also found that “must” had the smallest total change in rankings between the introduction and other models. Moreover, the total change was smaller for the discussion section than for the methods section, potentially corresponding to the higher similarity observed between the introduction and discussion models from figure 2. However, for the term “indicate” we observed roughly similar total changes in rankings between the introduction and methods and discussion models. “Indicate”, with three terms missing in the methods model, also highlights an issue with comparisons between word2vec models—certain terms were excluded because they did not meet the minimum inclusion threshold of 100 instances. Whereas the patterns observed for “indicate” contrasted with our findings from figure 2, those for “propose” were supportive: the total change in rankings terms for the methods model was smaller than the total for the discussion model, though we note that both were especially large compared to other signal terms. One potential explanation for these large changes in rankings may be that “propose” was relatively rare in each corpus. Regardless of comparisons with figure 2, table 1 shows that, even at the local level, the semantics of hedges were heterogeneous between sections.

Table 1. Rank and change-in-rank of three selected hedging terms. For each term, this table lists the top 15 most similar words in the word2vec model trained on introduction citances, measured by cosine similarity. For each term we included the change in rank of each of the 15 similar words between the introduction word2vec model, and the methods (Met) and discussion (Dis) models. Terms not included in a model’s vocabulary are assigned a value of “-”. An ad-hoc change index—the sum of the absolute values of the change in rank, divided by the rank of the term within the introduction model–is shown at the bottom of the table.

Must	**ΔMet**	**ΔDis**	**Indicate**	**ΔMet**	**ΔDis**	**Propose**	**ΔMet**	**ΔDis**
should	0	0	reveal	-6	-1	proposes	0	-
will	0	0	demonstrate	-4	-16	we_propose	0	-10
would	-1	0	suggest	-15	2	proposed	-2	-227
might	-4	-4	show	0	-4	put_forward	-9	-172
may	-13	0	reflect	2	0	proposing	-	-
can	-7	2	imply	-	3	presented	-701	-1056
cannot	0	1	highlight	-8	-54	devised	-27	-
do_not	-6	-2	confirm	-141	-11	formulate	-43	-
could	6	2	give	-3	-29	introduced	-37	-829
does_not	-9	-3	relate	-4	-23	gave	-171	-1212
tends_to	-10	-1	give_rise	-	-20	find	-19	-13
they_do	-10	1	find	-33	-40	developed	-185	-676
tend_to	-10	-3	confer	-	-23	we_introduce	6	-
could_potentially	-	5	provide	-22	-7	introduce	-22	-115
able_to	-12	-12	correlate_with	-63	-43	employ	-12	-12
Change Index	10.7	3.8	**Change Index**	41.1	36.1	**Change Index**	169.2	579.5

Conclusion

With this preliminary analysis we found evidence that hedging was not independent of its position within a publication. We observed that hedging terms were more common in the discussion sections of papers, though patterns varied by individual term. We also found evidence that the semantic meaning of hedging terms differed between sections; overall, hedging was most common between the introduction and discussion though distinct patterns were observed for individual terms. We found evidence of this heterogeneity using two distinct approaches. These findings confirm discussions by Di Marco and Mercer (2004) and have direct implications for any study utilizing dictionaries of hedges or other signal terms. These findings also demonstrate that the meaning of a word cannot be fully understood when decontextualized from its context, and that the contextual factors should be considered during textual analysis of scientific publications. The results of this analysis are subject to several limitations. The most important avenue for future research is to further validate our measure of semantic similarity between models and develop more sophisticated techniques. Additionally, this analysis is limited by a relatively small sample of data—one million citances per section—which while large, falls short of the corpus sizes typically used when training more robust word2vec models. Expanding the data size may address some of the issues of terms missing from model's shared vocabulary. Future work will attempt to address these limitations while expanding the scope of analysis to include additional sections (e.g., conclusion, abstract), to distinguish between the semantic meaning of hedging between disciplines, and to examine other common signal terms, such as terms indicating uncertainty (Chen et al., 2018) and discovery (Small et al., 2017).

Acknowledgements

The analyses were carried out by Dakota Murray during a research visit to CWTS. We thank Elsevier for making the data available to CWTS and Nees Jan van Eck for processing. We would also like to thank Ludo Waltman and Wout Lamers from CWTS, and Kevin Boyack from SciTech Strategies, for their feedback.

References

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016a). The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology, 67(1), 164-177.

Bertin, M., Atanassova, I., Sugimoto, C. R., & Lariviere, V. (2016b). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, *109*(3), 1417–1434.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73.

Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *PNAS*, *112*(45), 13823–13826.

Chen, C., Song, M., & Heo, G. E. (2018). A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty. *Journal of Informetrics*, *12*(1), 158–180.

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation, 37*(1), 16-24.

Hyland, K. (1996). Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics*, *17*(4), 433–454.

Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Marco, C. D., & Mercer, R. E. (2004). Hedging in Scientific Articles as a Means of Classifying Citations.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 3111–3119). USA: Curran Associates Inc.

Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*(1), 86–92.

Small, H. (2004). On the shoulders of Robert Merton: Towards a normative theory of citation. *Scientometrics, 60*(1), 71-79.

Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, *11*(1), 46–62.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic Classification of Citation Function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110). Stroudsburg, PA, USA: Association for Computational Linguistics.

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. Presented at the 29th AAAI Conference on Artificial Intelligence, AAAI 2015, AI Access Foundation.

Scholars mobility and its impact on the knowledge producers' workforce of European regions

Márcia R. Ferreira[1], Juan Pablo Bascur[2] and Rodrigo Costas[3]

[1] *m.r.ferreira.goncalves@cwts.leidenuniv.nl;*
Centre for Science and Technology Studies (CWTS), Wassenaarseweg 62A, Leiden, 2333AL (The Netherlands)

[2] *j.p.bascur.cifuentes@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Wassenaarseweg 62A, Leiden, 2333AL (The Netherlands)

[3] *rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Wassenaarseweg 62A, Leiden, 2333AL (The Netherlands)
DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University (South Africa)

Abstract

Knowledge production increasingly relies on mobility. However, its role as a mechanism for knowledge recombination and dissemination remains largely unknown. Based on 1,244,080 Web of Science publications from 1,435,729 authors that we used to construct a panel dataset, we study the impact of inter-regional publishing and scientists' mobility in fostering the workforce composition of European countries during 2008-2017. Specifically, we collect information on scientists who have published in one region and then published elsewhere, and explore some determinants of regional and international mobility. Preliminary findings suggest that while talent pools of researchers are increasingly international, their movements seem to be steered by geographical structures. Future research will investigate the impact of mobility on the regional structure of scientific fields by accounting for the appearance and disappearance of research topics over time.

Background

After Europe's enlargement in 2004, a growing number of scholars had the opportunity to collaborate and live unconstrained by national borders. Some estimates indicate for instance that in 2010 there were about 1.59 million full-time researchers in the EU27 and that the number of researchers in the active population is increasing (See, for instance, More2 final report "Support for continued data collection and analysis concerning mobility patterns and career paths of researchers"). As national science systems become more globalized, they also become more dependent on the movement of researchers. This movement has been thought to be a driver of innovation and scientific breakthroughs, especially in receiving locations (Ganguli, 2015; Stephan & Levin, 2001).

However, one aspect of scientific mobility has been relatively neglected, namely, the mobility patterns of researchers across regional borders in Europe. As a result, the inter-regional structure of scientific mobility flows is still poorly understood. Yet, scholarly knowledge exchanges based on workforce mobility are vital to the transfer of tacit knowledge that cannot be transferred through formal communication channels (e.g., research articles, journals, books, e-mails) (Gertler, 2003). When mobile scholars move, they bring information and know-how, skills, and ideas that differ from natives and that are essential for knowledge recombination, interactive learning, and novelty. Based on previous theoretical and empirical work (e.g., Agrawal et al., 2006; Jaffe et al., 1993) we hypothesize that mobility is an effective mechanism for disseminating knowledge and capabilities across research institutions, fields and locations and is therefore important to our understanding of how fields evolve over time. Previous studies already point in that direction. Moser et al. (2014) showed that German-Jewish scientists fleeing from Nazi Germany into the US played a central role in the emergence of new chemistry sub-fields in which incumbents also participated. Ganguli (2015) showed that, after the 1991 breakdown of the Soviet Union, Russian scientists who migrated to the US were much more cited by US scientists than those who did not migrate. In a large-scale study of the Web of Science database, Sugimoto et al. (2017) found that scholars who were mobile (i.e., more than one affiliation to different countries) have approximately 40% more citations than non-mobile scholars.

Three exploratory questions will be asked at this preliminary stage: (1) How is scientific mobility distributed across European regions? (2) To what extent are international and regional mobility related? (3) Which regional mobility network patterns can be seen across countries? To answer these questions, we use Web of Science publications to construct a longitudinal data set for 264 regions from 32 European countries. In this research-in-progress paper, we briefly discuss the methodology, describe the data and discuss the preliminary findings and our plans for future research.

The data

Both publication data and regional data are used to compute regional mobility. Specifically, we reconstruct career trajectories of 1,435,729 distinct authors based on 1,244,080 publications (articles and reviews) published between 2008 and 2017. These data were collected from the CWTS in-house version of the Web of Science provided by Clarivate Analytics. Authors' profiles were isolated using the author-disambiguation algorithm developed by Caron and van Eck (2014). Since there is no consistent method for tracking scientific mobility, and categories of highly skilled workers are often too ambiguous to identify specific groups of scientists, we use author affiliations available in scientific publications. This is used to determine who moved between European regions, where and in what year. For each author we assign their affiliation addresses to the NUTS ('Nomenclature of Territorial Units for Statistics') classification structure, provided by Eurostat[1]. European countries differ in terms of size, population, number of researchers, availability of funding, and number of universities, therefore, we carry out the analysis at the regional level. We focus on regions with an average of at least 50 publications per year resulting in 264 NUTS2 European regions from 32 countries – EU-28 plus Norway, Iceland, Switzerland, and Turkey.

Following on Robinson-Garcia et al. (2019) we define mobile scientists as those who have affiliations in at least two different regions either simultaneously (dual affiliation) or over the period of analysis (i.e., published in more than one region between 2008 and 2017). Accordingly, the dataset consists of 1,435,729 authors out of which about 14% (203,925) have published in more than one region. The mobility of scholars across regions is quantified by observed changes (movements) in the reported affiliation and corresponding region as stated by the scholars themselves in publication documents.

Preliminary evidence

How is scientific mobility distributed across European regions?

We here present preliminary evidence on regional mobility. A total 44% of scholars' movements involves NUTS2 regions belonging to the same country (regional mobility), while 55% of remaining movements involves international mobility. With respect to the distribution of movements of scholars across the European geography, 61 regions – out of 264 – concentrate 67% of scientists' movements. The countries with the highest numbers of mobile researchers are depicted in Figure 1. It follows that mobility flows of scholars tend to be clustered regionally within Europe. The regions in which we find heavy movement, such as Île-de-France, were also those where mobile researchers tended to locate more generally. Mobile researchers are noticeably absent from Eastern European regions, as these locations were also less likely to have lower scientific activity and lower levels of research funding.

[1] https://ec.europa.eu/eurostat/web/nuts/background

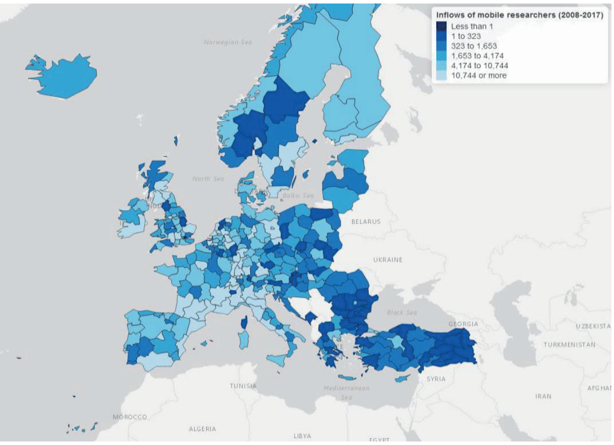


Figure 1. Total mobility inflows of scientists are shown for NUTS2 regions. In the map, lighter colours represent larger inflows.

To what extent are international and regional mobility related?

Table 1 shows the shares of internationally mobile and regionally mobile researchers in Europe's top ten countries ranked by the share of mobile researchers in each country. As can be seen, the shares of international mobility and regional mobility are significantly different. All countries have higher shares of internationally mobile researchers than regionally mobile researchers.

Table 1. Share of internationally and regionally mobile researchers for the top 10 countries with the highest share of mobile researchers (2008-2017).

Country	*%* *mobility*	*%* *international mobility*	*%* *regional mobility*	*#* *regions*
1. Switzerland	31%	25%	6%	7
2. Netherlands	28%	19%	10%	12
3. Belgium	27%	19%	8%	11
4. Germany	26%	15%	11%	38
5. Italy	26%	16%	9%	21
6. Sweden	25%	20%	5%	8
7. Austria	25%	21%	4%	9
8. United Kingdom	24%	15%	10%	40
9. France	23%	14%	8%	27
10. Spain	23%	15%	7%	18

The top countries in the table are also high-income countries separated by relatively short distances, making it easier for researchers to internationalize in those countries. Switzerland has the highest share of mobile researchers (31%), with 25% of mobile scholars having affiliations in at least two countries (internationally mobile) and 6% in at least two regions of Switzerland (regionally mobile). France (23%) and Spain (23%) have the lowest shares of mobile researchers. Switzerland has the highest share of internationally mobile scholars (25%) compared to other countries, whereas Germany (15%), the United Kingdom (15%), Spain (15%), and France (14%) have the lowest share. This can be explained in part by the number and size of these countries' regions (See Figure 2, right panel). France, for example, has more regions (n=27) than Switzerland (n=7) that are larger in spatial scale and scientific workforce. Such conditions may offer researchers a more diverse set of options and point to the role of geography in facilitating and/or constraining scientific mobility.

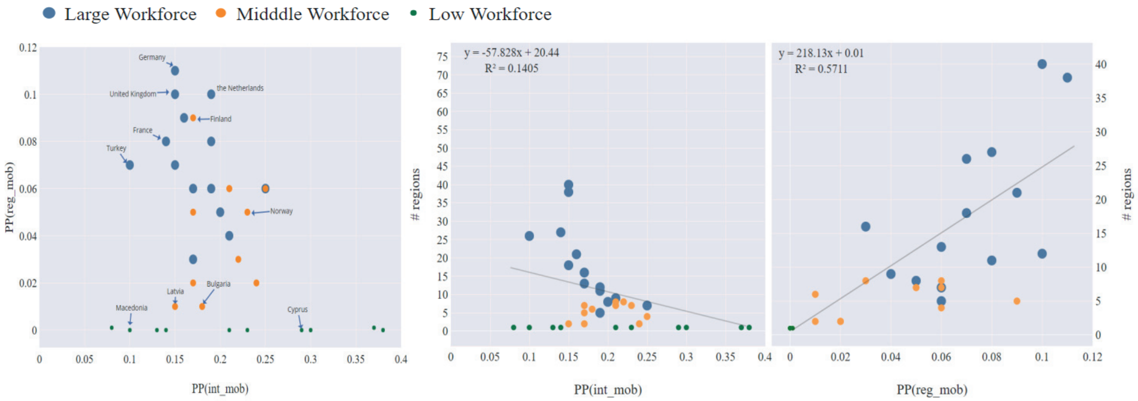


Figure 2. Scatterplots of proportion of regional mobility and international mobility by country (2008-2017). Categories are based on country's workforce size (i.e. overall scholars identified in WoS).

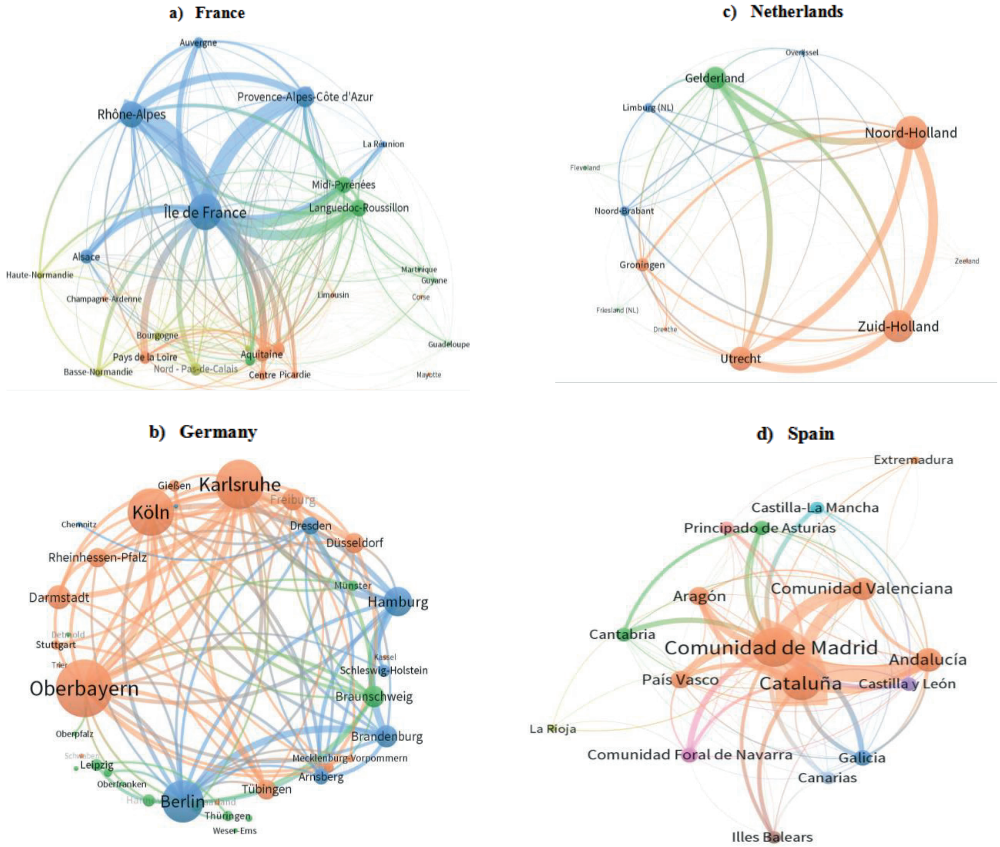


Figure 3. Undirected network of regional mobility flows for Germany, Spain, France and the Netherlands (2008-2017). The network visualizations were created using the VOSviewer software (van Eck & Waltman, 2010).

The left panel of Figure 2 suggests that there is no clear relationship between the proportion of scholars who are regionally mobile and scholars who are internationally mobile. The size of countries' workforce (as measured by the total number of scholars publishing in the WoS and identified in the countries) also does not seem to be related to the prevalence of any kind of mobility. As expected, smaller countries with few or no regions (e.g. Cyprus) also do not have any regional mobility. In fact, the two panels on the right side indicate this obvious but noteworthy pattern: larger countries with relatively high number of regions tend to have higher proportions of regional mobility (right panel). The middle panel of Figure 2 shows that larger countries with multiple regions tend to have slightly lower shares of international mobile scholars.

Which regional mobility network patterns can be seen across countries?
Figure 3 shows the regional mobility networks of four European countries: France, Netherlands, Germany, and Spain. The nodes represent the different regions in the countries, the size of the nodes is determined by the total number of mobile scholars affiliated with the region, and edges are established by the number of common scholars that have been affiliated to each pair of regions. Although a thorough network analysis is not performed here, the graphs point to two main patterns. Firstly, a concentrated pattern (e.g., France and Spain), in which a few regions (e.g., Ile de France in France and Madrid and Cataluña in Spain) conform strong nodes, having multiple mobility linkages with the other regions in the country. Secondly, a more distributed pattern (e.g. Germany and the Netherlands), in which there are more diverse connected regional nodes (e.g. Oberbayern, Berlin, Köln, Karlsruhe and Hamburg in Germany; and Noord-Holland, Zuid-Holland, Utrecht and Gelderland in the Netherlands), without clearly dominant nodes. Some proximity patterns can also be seen in some of the networks such as the stronger linkages among the northern Spain regions (e.g., Cantabria and Asturias) or among regions in the Randstad area of the Netherlands (e.g., Zuid-Holland, North-Holland and Utrecht). The fact that different networks show very different mobility patterns suggests that national scientific systems in Europe can be organized differently.

Conclusion and future research

In this paper, we have presented a first attempt to study mobility patterns among European regions using bibliometric methods. As a proof of concept, it can be concluded that regional mobility is also traceable by bibliometric means. Our preliminary results suggest that mobility patterns between regions differ from mobility patterns between countries (Robinson-Garcia et al., 2019), since they have different incidence across countries. Results indicate that, there is an unequal distribution of regionally mobile scholars across European countries. Moreover, national regional mobility does not seem to have a strong relationship with the level of international mobility of the country. Another conclusion is that, there are two important factors in the consideration of regional mobility: the existence of a regional geographical structure in the country (otherwise regional mobility is not possible), and the existence of a sizeable workforce in the country.

Overall, our results point that regional mobility is a social phenomenon that deserves attention by itself. Thus, these initial observations motivate us to examine the consequences of mobility for the distribution of scientific portfolios (i.e., thematic delineation of regions). As the inflow of researchers into regions is also likely to generate a positive effect on the scientific innovation in receiving locations, we will test whether knowledge generated by scholars in their origin location affects the scientific portfolios of their new location over the years.

We ask, therefore, two interrelated questions: What is the effect of the influx of mobile scientists on the emergence of scientific fields of receiving regions? How stable is the network of inter-regional mobility flows? We will use this approach to identify co-location effects driven by inter-regional publishing. Our expectation is that the international and regional movement of scholars can explain differences in scientific portfolios in European regions. This will include studying the inflow and outflow dynamics of regions in attracting (or sending) scholars; as well as controlling for other demographic aspects such as age and gender of these regionally mobile scholars.

Finally, similar to the identification of internationally mobile researchers by Robinson-Garcia et al. (2019), this study is dependent on the number of publications to identify mobility patterns, the coverage of the database used to extract the publication data, and the completeness of the author-affiliation information. To minimize the influence of these biases, we are currently expanding the methodology to include other document types such as conference publications, book reviews, and letters, and comparing mobility statistics derived from publication data from the Web of Science with highly skilled migration statistics.

Acknowledgements

We thank Alfredo Yegros and Wout Lamers for their help in assigning affiliation addresses to the NUTS2 regional classification. This research is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP) and by the European Commission DG EAC, U-Multirank project.

References

Agrawal, A., Cockburn, I., McHale, J., (2006). Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. Journal of Economic Geography, 6(5), 571–591.

Almeida, P., Kogut, B. (1999). Localization of knowledge and the mobility of engineers in regional networks. Manage. Sci. 45 (7), 905–917. Management Science, 45(7), 905-917.

Caron, E., van Eck, N.J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In Proceedings of the 19th International Conference on Science and Technology Indicators (pp. 79-86).

Jaffe, A., Trajtenberg, M. Henderson, R. (1993). Geographic localization of knowledge flows as evidenced by patent citations. Quarterly Journal of Economics, CVIII, 577-598.

Ganguli, I. (2015). Immigration and ideas: What did Russian scientists “bring” to the US? Unpublished manuscript, Stockholm School of Economics.

Gertler, M. (2003). Tacit knowledge and the economic geography of context or the undefinable tacitness of being (there). Journal of Economic Geography 3(1), 75–99.

Moser, P., Voena, A., Waldinger, F. (2014). German Jewish émigrés and US invention. The American Economic Review, 104(10), 3222-3255.

Robinson-Garcia, N., Sugimoto, C.R., Murray, D., Yegros-Yegros, A., Larivière, V., Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. Journal of Informetrics, 13, 50-63.

Stephan, P. E., Levin, S. G. (2001) Exceptional contributions to US science by the foreign-born and foreign-educated. Population Research and Policy Review, 20, 59-79.

Sugimoto, C. R., Robinson-Garcia, N., Murray, S.D., Yegros-Yegros, A., Costas, R., Lariviere V. (2017). Scientists have most impact when they’re free to move. Nature 550, 29–31.

van Eck, N. J., Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–53.

Does Monetary Support Increase Citation Impact of Scholarly Papers?

Yaşar Tonta[1] and Müge Akbulut[2]

[1]*yasartonta@gmail.com*
Hacettepe University, Faculty of Letters, Department of Information Management, Ankara (Turkey)

[2]*mugeakbulut@gmail.com*
Ankara Yıldırım Beyazıt University, Faculty of Humanities and Social Sciences, Department of Information Management, Ankara (Turkey)

Abstract

One of the main indicators of scientific development of a given country is the number of papers published in high impact scholarly journals. Many countries introduced performance-based research funding systems (PRFSs) to create a more competitive environment where prolific researchers get rewarded with subsidies to increase both the quantity and quality of papers. Yet, subsidies do not always function as a leverage to improve the citation impact of scholarly papers. This paper investigates the effect of the publication support system of Turkey (TR) on the citation impact of papers authored by Turkish researchers. Based on a stratified probabilistic sample of 4,521 TR-addressed papers, it compares the number of citations to determine if supported papers were cited more often than those of not supported ones, and if they were published in journals with relatively higher citation impact in terms of journal impact factors, article influence scores and quartiles. Both supported and not supported papers received comparable number of citations per paper, and were published in journals with similar citation impact values. Findings suggest that subsidies do not seem to be an effective incentive to improve the quality of scholarly papers. Such support programs should therefore be reconsidered.

Introduction

The number of refereed papers that appears in scientific journals along with citations thereto is considered to be the main indicators of scientific productivity and quality of a given researcher, a research organization or a country. Many countries introduced what is called performance-based research funding systems (PRFSs) to streamline the scientific production process and improve the research performance (Jonkers & Zacharewicz, 2016).

PRFSs aim to assess the performances of researchers in a given time period. Some countries provide monetary incentives directly to the researchers in the form "piece rates" or "cash-for-publication" schemes (Heywood, Wei, & Ye, 2011) while others prefer to reward researchers' organizations by allocating funds to them (De Boer et al., 2015). Both "ex ante" and "ex post" assessments are being used for this purpose. Compared with peer review (which requires labor-intensive evaluation processes prior to funding allocation), it is relatively easier, and less costly, to carry out ex post quantitative assessments on the basis of bibliometric measures.

Notwithstanding the type of assessment carried out, research organizations or countries tend to eagerly incentivize their researchers because they in turn expect return on investment (RoI), usually as an increase in the number of papers published by their researchers as well as the citation impact of their papers. However, such monetary incentives do not necessarily produce the intended outcomes, as the existence of PRFSs does not correlate well with the research productivity or quality (Auranen & Nieminen, 2010). The effect of PRFSs on the increase in the quantity of publications is "temporary and fades away after a few years" while the average effect on the quality of publications is "nil" (Checchi, Malgarini, & Sarlo, 2019, pp. 45, 59).

This paper aims to study the effect of the publication support system of the Turkish Scientific and Technological Research Council (TÜBİTAK) on the increase of the citation impact of papers published in scientific journals. The support system is based on the concept of "cash-for-publication" and has been in place since 1993. The authors of papers get rewarded on the basis of Journal Impact Factors (JIFs) and, more recently, Article Influence Scores (AISs) of journals in which their papers are published. The higher the JIF values, the more money the authors get paid, with a cap of 7,500 Turkish Lira (circa 3,000USD in 2015). We compare all papers published between 2006 and 2015 and indexed in Web of Science (WoS) with at least one author whose address is based in Turkey (henceforth "TR-addressed papers") with those supported by TÜBİTAK to see if supported papers received more citations and if they were published in higher quality journals in terms of JIFs, AISs and quartiles (Q1 through Q4). The rest of the paper is organized as follows: The next section briefly reviews relevant studies. The Data Sources and Method section describes the data and sampling technique used to select the TR-addressed papers along with the matching algorithms written to identify the supported ones. The Findings and Discussion section presents detailed findings. The paper ends with Concluding Remarks.

Literature Review

Performance-based research funding systems (PRFS) came into being in 1980s. The Research Excellence Framework (REF) of UK is one of the oldest PRFSs based on peer review and has been in use since 1986 (De Boer et al., 2015, p. 113). Yet, bibliometric measures such as JIF have become the dominant method used for research evaluation purposes within the last two decades and they are readily available through Journal Citation Reports (JCR) published annually by Clarivate Analytics.

Several countries have developed primarily JIF-based PRFS where JIF values are used (sometimes in combination with peer review) to determine the amount of monetary support per paper. The PRFS use around the world has been reviewed in a number of studies (e.g., Auranen & Nieminen, 2010, De Boer et al., 2015; European Commission, 2010; Geuna & Martin, 2003; Hicks, 2012; Pajić, 2014). The practices tend to vary from country to country. Some reward researchers directly through what is called "cash-for-publication" schemes (e.g., China and Turkey) while others support the affiliated research units or universities (e.g., UK and South Africa) (Heywood, Wei, & Ye, 2011; Tonta, 2017a; De Boer et al., 2015; Harley, Huysamen, Hlungwani, & Douglas, 2016; Lee & Simon, 2018).

PRFSs tend to produce unintended consequences or "side effects" (Geuna & Martin, 2003), cause researchers to develop "opportunistic behavior" (Abramo, D'Angelo & Di Costa, 2018), and eventually become "perverse incentives" (Tomaselli, 2018). Muller (2017) studied the subsidies from the viewpoint of "rent seeking" theory in economics and explored the impact of distorted incentives on academia, academics and society at large. According to rent seeking theory, academics "compete for artificially contrived transfers" in various forms (e.g., grant funding, monetary incentives for publications and citations). These transfers are usually redirected by public institutions from social surplus to rent seeking academics on the basis of bibliometric measures that are thought to measure academic success better and "provide greater reassurance of quality" (p. 59). Such measures are therefore increasingly supplanting (rather than supporting) peer review used to judge the quality of scholarly output, and universities are "creating institutional rules and practices that actively incentivize rent-seeking behavior" (Muller, 2017, p. 61). In South Africa, for instance, the amount of monetary support per paper (which may be as high as 10,000USD per a single-authored paper) is the same, regardless of where the paper has been published as long as the outlet is "accredited" by the Department of Higher Education and Training. Thus, one can submit their work to lower quality journals with relatively lower standards of peer review in order to collect subsidies quickly and more often. This may, in turn, have created a "powerful perverse incentive" and encouraged at least some researchers to "game" with the system and produce "fraudulent –or ethically questionable– publications" (Muller, 2017, pp. 63-64). Muller (2017) underlines the dilemma of such incentives as follows:

> Under the rent seeking conceptualization of such systems, appeals to individual or institutional integrity are not likely to be successful. The system directly creates

> incentives for the activities cautioned against, undermining cultures of ethical practice, and therefore only measures that carry suitable material punishment are likely to counteract these undesirable effects. (p. 64)

The side effects of PRFSs are not limited only with researchers publishing in lower quality outlets or "seeking out 'easier' publication types" (Sīle & Vanderstraeten, 2019, p. 86). Subsidies tend to discourage types of research that require more time to carry out using novel experiments prolonging the publication process, thereby giving way to papers with little or no societal impact whatsoever (Geuna & Martin, 2003; Tonta, 2017b, pp. 27-30). Moreover, some researchers simply prefer to publish in "predatory journals" and set up what is called "citation circles" to benefit more from PRFS (Good, Vermeulen, Tiefenthaler, & Arnold, 2015; Teodorescu & Andrei, 2014, pp. 228-229). South African researchers published as much as five times more papers in predatory journals than those in the United States or Brazil did (Hedding, 2019). While the number of South African publications has doubled (as pointed out earlier) after the introduction of the subsidy program, the ones published in predatory journals increased 140 times during the same period (2004-2010) (Mouton & Valentine, 2017).

Turkey, too, has a bad reputation and ranks third (after India and Nigeria) in the world among 146 countries in terms of number of papers published in predatory journals (Demir, 2018, p. 1303). Beall's (now defunct) list of predatory journals includes 41 such journals originating from Turkey, second highest after India (Akça & Akbulut, 2018, p. 264). Although no study has so far been carried out on TR-addressed papers published in predatory journals, it is likely that several such papers have been subsidized in the past under the support program. For example, more than 80% of the subsidies for papers in anthropology in 2015 went to a single predatory journal in this field in which Turkish researchers have published a total of 127 papers, most of which had had nothing to do with anthropology (Tonta, 2017b, p. 80).

Despite the side effects and undesired outcomes of PRFSs, there appears to be a commonly held belief in research funding and research performing institutions that subsidies would increase the number of papers and their citation impact. Researchers motivated by such subsidies would produce more papers with higher quality. However, the relationship between subsidies and the increase in productivity and quality is not clear-cut (Auranen & Nieminen, 2010). While there appears to be some evidence that subsidies increase the number of papers to some extent, this is not reciprocated with a similar increase in the quality of papers in terms of their citation impact (Butler, 2003, 2004; Good et al., 2015; Osuna, Cruz-Castro, & Sanz-Menéndez, 2011). For instance, the number of South African publications has almost doubled in seven years (2004-2010) after the implementation of the subsidy system. Yet, their citation impact (number of citations per paper) has decreased steadily (Pillay, 2013, p. 2). A small-scale study carried out at the University of Cape Town after the implementation of PRFS showed that the number of output is negatively correlated with both the number of citation counts of papers and their field-weighted citation impact. Although the variance explained was relatively modest, findings indicate to some extent that greater subsidy seems to be "associated with lower citation impact," which may, in part, be due to the fact that the PRFS currently in use "does not factor in research quality and impact" (Harley et al., 2016). Similarly, the number of TR-addressed papers listed in citation indexes has increased 19-fold between 1993 (when TÜBİTAK's support program began) and 2015 (from 1,500 papers to more than 28,000) (Tonta, 2017b, p. 32). Turkey has jumped from 37th place in 1993 to 18th in 2008 in the world ranked by the number of indexed publications. Yet, findings of an interrupted time series analysis based on 390,000 TR-addressed publications listed in the WoS database between 1976 and 2015 (of which 157,000 or 40% were subsidized between 1997 and 2015) showed that the support program seems to have had no impact on the increase in the quantity of TR-addressed publications (Tonta, 2017a). Moreover, the citation impact of TR-addressed papers has constantly decreased throughout the years and is well below (40%) that of the world average of all papers (Çetinsaya, 2014, p. 127; Kamalski et al., 2017, p. 4).

It appears that PRFSs do not help much in terms of improving the quality of research and force researchers to choose between "cash or quality" (Hedding, 2019). We test this conjecture with reference to Turkey and see if the subsidy system currently in use in Turkey has improved the citation impact of TR-addressed papers by comparing the number of citations per paper, JIFs, AISs and quartiles of journals in which both supported and not supported papers were published. We present below the data sources and method used to analyze the data.

Data Sources and Method

We used the well-known bibliometric measures of number of citations per paper, JIF, AIS and JCR quartiles to compare the citation impact of supported and not supported TR-addressed papers. JIF is defined as the "average" citation impact of papers published in a given journal within a given time period. AIS takes into account five years' worth of citation data for a given journal and weights citations on the basis of JIFs. If citations come from high impact journals, they are weighted more heavily. AIS is similar to Google's PageRank algorithm in that it uses the whole JCR citation network to calculate the AIS for a given journal. Unlike JIS, AIS indicates if each article in a journal has above- or below-average influence, 1.000 being the average of all journals included in JCR's citation network (Article, 2019). AIS is more stable and can therefore be used in interdisciplinary comparisons where journals have varying publication and citation patterns, although both metrics are highly correlated (r = .9) (Arendt, 2010). JIF is used to categorize journals under at least one subject category, and journals under each subject category are divided into four quartiles based on their JIF values (the first 25% of the journals with highest JIFs constitutes Q1, the second 25% Q2, etc.).

In order to identify all TR-addressed papers published between 2006 and 2015 and indexed in Web of Science (WoS), we used the following advanced search query (December 2, 2017):

> **AD=**(Turkey OR Turquie OR Türkei OR Türkiye OR Turquia)
> **Timespan:** 2006-2015. **Indexes:** SCI-EXPANDED, SSCI, A&HCI. **PubType**: Article

We found a total of 225,923 TR-addressed papers and downloaded them. We obtained the payment information for 100,919 TR-addressed papers whose authors sought financial support from TÜBİTAK through its Support Program of International Publications (UBYT). Altogether some 44% of all papers were supported (range: 59% in 2007; 28% in 2015).

We then stratified all TR-addressed papers by year and scrambled them within each year (in case they had an inherent order by author or journal name, for example) so that certain records would not appear disproportionally in the sample. We wrote a macro to select every 12th and 75th records (these numbers were randomly chosen) out of the stratified list of all TR-addressed records (225,923). Sample size being 2% of the population, we obtained a total of 4,521 records in the sample using the stratified probability sampling technique. The sample size for each year ranged between 1.86% and 2.05%, average being 1.99% (which is quite close to 2%).

Next, we wrote a second macro to match up journal data from JCR and InCites with respective years to identify bibliometric characteristics of journals (e.g., JIF, AIS, Times Cited, Quartiles and so on) in which TR-addressed papers appeared along with the number of citations that each paper received, if any (February 2, 2018). Data were then added to all the records (seven journals did not match due to inconsistencies in journal names, which were added manually).

Finally, we wrote a third macro to match the list of papers supported by TÜBİTAK with all papers (supported or not). (Some 64 records did not match due to inconsistencies in paper titles, which were added manually.) This enabled us to compare both paper and journal characteristics for both supported and not supported papers (e.g., number of citations for papers, and JIF, AIS and quartile for journals). The matching algorithm seems to have worked quite well. Altogether, 44% of all papers were supported by TÜBİTAK. In the sample, the percentage of supported papers was somewhat lower: 1,679 out of 4,521 (or 37%). The difference (7%) is due to inconsistencies in data (such as punctuation marks and abbreviations used in titles of papers and journals). Nevertheless, we looked at the data more closely. Papers appeared in 9,463 different journal titles. The ones supported by TÜBİTAK appeared in 2,336 different journal titles. Some 2,153 journals (or 92%) were represented in the sample, of which 986 (or %42) had published at least one supported TR-addressed paper between 2006 and 2015. We do not expect such small fluctuations to have any considerable impact on the analysis that follows.

We used MS Excel and SPSS 23 for data analysis and visualization; independent samples t-test for significance, as our sample size was relatively large (4,521) (Lumley, Diehr, Emerson, & Chen, 2002); and chi-square for test of independence. We used an alpha level of .05 for all statistical tests.

Findings and Discussion

As indicated earlier, the number of papers published between 2006 and 2015 is 225,923. Table 1 and Fig. 1 provide average and median JIFs and AISs of journals in which all TR-addressed papers appeared during this period. The median JIFs range between 0.998 (2012) and 1.379 (2015) while median AISs range between 0.321 (2012) and 0.457 (2010). Close to half the papers were published in low impact journals (i.e., JIF below 1.000), and their AIS values were less than half (around or below 0.400) of that of the world average (1.000). Although there seems to be a slight increase in recent years in median JIFs and AISs of papers, this has probably more to do with the continuing increase in JIFs over the years (Fischer & Steiger, 2018).

Table 1. Average and median JIFs and AISs of journals in which all TR-addressed papers published (2006-2015)

Year	Average JIF	Average AIS	Median JIF	Median AIS
2006	1.481	0.508	1.087	0.411
2007	1.311	0.454	1.091	0.406
2008	1.444	0.536	1.098	0.403
2009	1.409	0.481	1.072	0.399
2010	1.546	0.526	1.245	0.457
2011	1.327	0.459	1.053	0.384
2012	1.401	0.455	0.998	0.321
2013	1.626	0.504	1.231	0.351
2014	1.769	0.537	1.234	0.342
2015	1.988	0.574	1.379	0.368

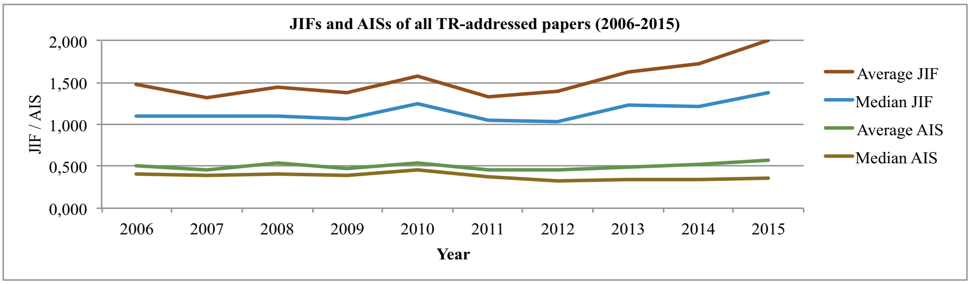


Fig. 1. Average and median JIFs and AISs of journals in which all TR-addressed papers published (2006-2015)

Findings below are based on the stratified probability sample of 4,521 papers (2006-2015). The great majority (90%) of the papers in the sample were Science papers. Social Science and Arts and Humanities papers constituted about 9% and 1% of all papers, respectively. Papers in the sample were cited a total of 55,383 times. The average number of citations per paper was 12 (*SD* = 42). Half the papers received five or fewer citations (min. = 0, max. = 2,246). Only 1% (or 45 papers) received 100 or more citations while 32% received 10 or more. Some 13% of papers were not cited at all. As expected, Science papers received the overwhelming majority of the total number of citations (over 92%) followed by Social Science papers (7%) and Arts and Humanities papers (less than 1%).[1]

Table 2 provides descriptive statistics and statistical test results for papers in the sample. As indicated earlier, 37% (1,679) of papers in the sample were supported by TÜBİTAK. Supported papers collected 43% (23,654) of all citations. On average, supported papers were cited slightly more often (M_{all} = 14.1, *SD* = 22.5) than not supported ones (M_{all} = 11.0, *SD* = 49.8). This difference was significant *t* (4,519) = -2.39, *p* < .05; however, the effect size was rather small (*r* = .04). Similarly, supported science papers were cited somewhat more frequently on average (M_{sci} = 14.5, *SD* = 22.7) than not supported ones (M_{sci} = 11.3, *SD* = 52.1). Again, although the difference was significant *t* (4,081) = -2.28, *p* < .05, the effect size was infinitesimal (*r* = .04).

Table 2. Descriptive statistics and test results

	Citation impact	Supported papers				Not supported papers				t	p	r
		N	Mdn	M	SD	N	Mdn	M	SD			
All papers	# of cit.	1679	7.0	14.1	22.5	2842	4.0	11.0	49.8	-2.39	**.02***	.04
	JIF	1624	1.2	1.6	2.3	2696	1.1	1.5	2.0	-1.07	.28	.02
	AIS	1520	.4	.5	.9	2405	.4	.5	0.9	-.39	.70	.01
Science	# of cit.	1508	7.5	14.5	22.7	2575	4.0	11.3	52.1	-2.28	**.02***	.04
	JIF	1482	1.3	1.6	2.4	2473	1.1	1.5	2.0	-1.17	.24	.02
	AIS	1388	.4	.5	.9	2218	.4	.5	.9	-.48	.63	.01
Social Science	# of cit.	171	4.0	10.6	20.1	267	3.0	8.7	13.8	-1.18	.24	.05
	JIF	142	1.2	1.3	1.0	223	1.0	1.4	1.3	-.43	.67	.02
	AIS	132	.4	.4	.3	187	.4	.5	.50	.52	.61	.03

Notes: *N*: Number of papers; *# of cit.*: Number of citations; *Mdn*: Median; *M*: Mean; *SD*: Standard Deviation; *t*: t-test; *p*: p value; *r*: effect size; *cit.*: citation; *JIF*: Journal Impact Factor; *AIS*: Article Influence Score; *: statistically significant at alpha level .05.

That during a 10-year period a supported paper in general and a supported science paper in particular received on average about three more citations than a not supported one did and that the difference was significant has probably more to do with the sample size than a true effect. This is because insubstantial differences and small effect sizes can still be significant in a relatively large sample (as is the case in this study). To put the difference into a better perspective, assuming that any given paper had on average six years to collect citations, a supported paper received about half a citation more per year than a not supported one did. This can hardly be considered a substantial difference. In fact, the statistically significant difference per paper between the numbers of citations for supported (M_{ssci} = 10.6, *SD* = 20.1) and not supported (M_{ssci} = 8.7, *SD* = 13.8) Social Science papers disappeared (*t* (436) = -1.18, *p* > .05), as Social Science papers constituted less than one tenth of all papers in the sample. Fig. 2 provides a comparative view of the number of citations per paper for all papers as well as for

[1] Note that 49 Arts and Humanities papers that received a total of 289 citations were excluded from further analysis as bibliometric characteristics of Arts and Humanities journals are not listed in JCR.

Science and Social Science papers. Boxplots show the means, medians, and first and third quartile values for both supported and not supported papers.

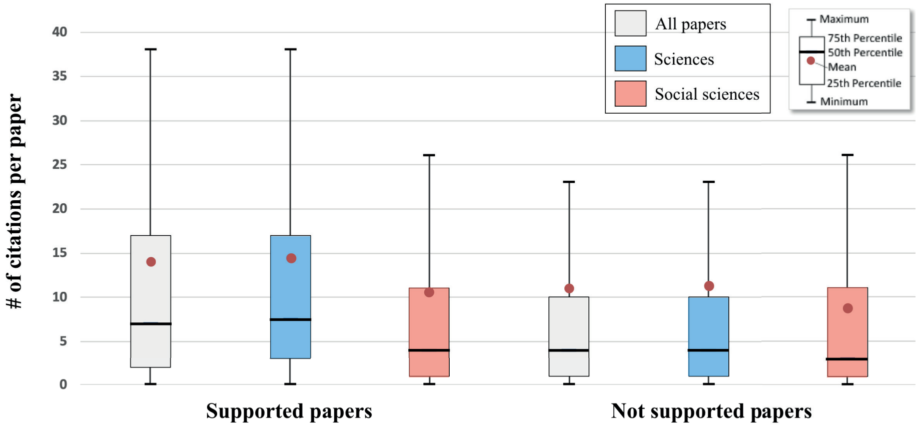


Fig. 2. Number of citations per paper for supported and not supported TR-addressed papers

Table 2 also provides data on citation impact values (e.g., JIF and AIS) of journals in which TR-addressed papers were published. On average, the JIF values of journals publishing all supported papers (M_{all} = 1.6, *SD* = 2.3), supported Science papers (M_{sci} = 1.6, *SD* = 2.4), and supported Social Science papers (M_{ssci} = 1.3, *SD* = 1.0) were quite similar to those publishing not supported ones (M_{all} = 1.5, *SD* = 2.0; M_{sci} = 1.5, *SD* = 2.0; and M_{ssci} = 1.4, *SD* = 1.3, respectively). The differences were not statistically significant in all three cases (t_{all} (4,318) = -1.07, $p > .05$; t_{sci} (3,953) = -1.17, $p > .05$; and t_{ssci} (363) = 0.41, $p > .05$, respectively). Likewise, the differences between the AIS values of supported and not supported papers were not statistically significant, either (see Table 2 for details). Fig. 3 provides the boxplots for JIF and AIS values of journals publishing both supported and not supported papers for all papers as well for Science and Social Science papers. JIF and AIS data of both supported and not supported papers were highly skewed with heavy tails, indicating that papers were mostly published in relatively mediocre or low impact journals. Average JIF and AIS values of journals obtained from the sample are consistent with those of all journals in the population (see Table 1, Fig. 1).

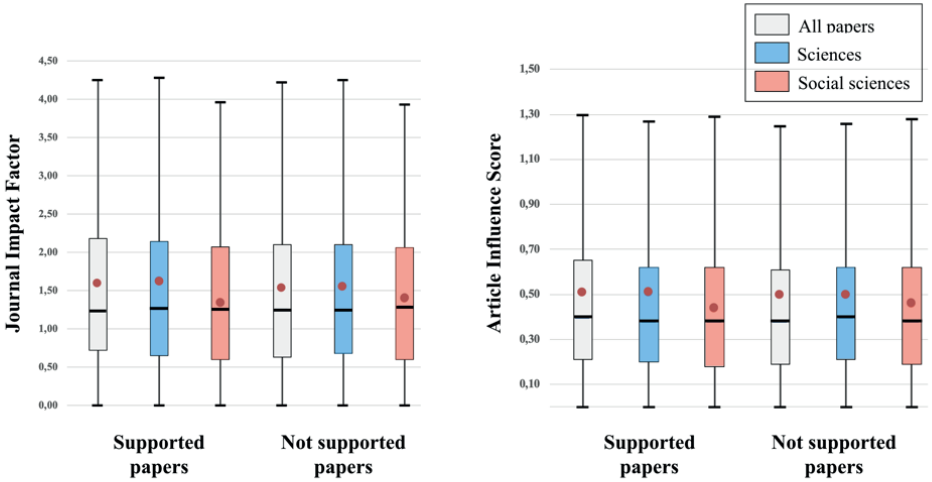


Fig. 3. Journal Impact Factors (left panel) and Articles Influence Scores (right panel) of journals publishing supported and not supported TR-addressed papers

Fig. 4 below provides the percentage distributions of JIF values of supported and not supported papers. Note that the percentages of supported and not supported papers were quite similar to each other, supporting the results of the statistical tests further. Correlation between JIF and AIS values of journals publishing TR-addressed papers was quite high (Pearson's r = .946), explaining 90% of the variance in the data[2] and confirming the findings of similar studies (e.g., Arendt, 2010). We therefore do not provide the distributions of JIF and AIS values of supported and not supported Science and Social Science papers separately, as they are similar to those in Fig. 4. Percentages of supported Science and Social Science papers in the sample were 37% and 39%, respectively. The percentage of supported Social Science papers that appeared in journals with no (zero) JIF values (17%) was much higher than that of Science papers (2%) because Social Science papers got supported more generously to increase their number (Tonta, 2017b).

We also looked at if supported papers were published in journals listed in higher JCR quartiles under their respective subject categories. It is interesting to note that more papers that appeared in journals with the lowest citation impact (Q4) were supported (28%) than those with the highest citation impact (25%) (Table 3). Almost half (48%) the supported Science papers appeared in Q1 and Q2 journals (as opposed to 36% of supported Social Science papers) (Table 4). The percentage of Science papers published in Q1 journals (26%) is almost twice that of Social Science papers (14%) (Fig. 5). The percentage of supported Social Science papers with no quartiles was the same as those with no JIFs (17%). The difference between the quartile distributions of supported and not supported Science papers was statistically significant ($X^2(4) = 39.6, p < .05, r = .01$). This may be due to the more restrictive support policy towards Science papers published in the lowest quartile of journals (Tonta, 2017b, pp. 23-24). The percentages of supported papers by quartiles suggest that the support system seems to have failed to be more selective, thereby rewarding more papers that were published in journals with lower JCR quartiles and thus lower citation impact.

[2] Not all journals in which TR-addressed papers were published had both JIF and/or AIS values listed in JCR. The correlation coefficient is based on 3,961 papers with both values. Papers that were published in journals with no JIS and/or AIS were also excluded.

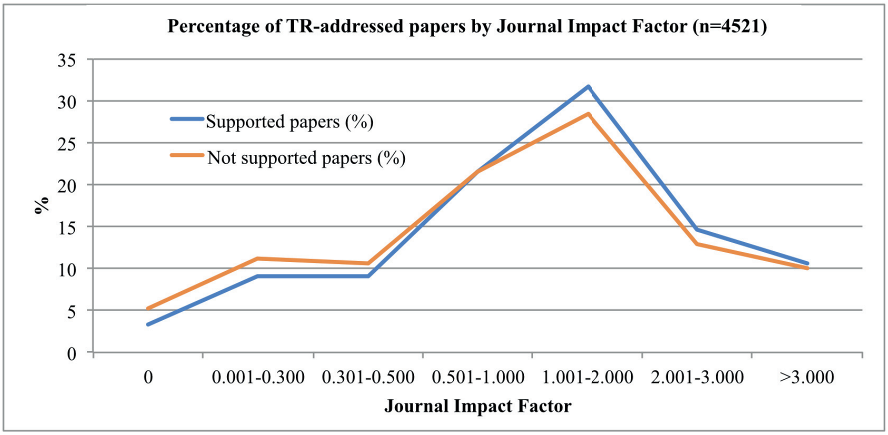


Fig. 4. Percentage of TR-addressed papers by Journal Impact Factor (n=4521)

Table 3. Distribution of TR-addressed papers by JCR Quartiles (n=4521)

	Quartiles of supported papers						Quartiles of not supported papers					
	N/A	Q1	Q2	Q3	Q4	Total	N/A	Q1	Q2	Q3	Q4	Total
N	55	415	370	364	475	1679	146	536	542	683	935	2842
%	3	25	22	22	28	100	5	19	19	24	33	100

Table 4. Distribution of papers by JCR quartiles (Social Science vs. Science (n=4521)

		Quartiles of supported papers						Quartiles of not supported papers					
Subject	Freq / %	N/A	Q1	Q2	Q3	Q4	Total	N/A	Q1	Q2	Q3	Q4	Total
Social Sciences	N	29	24	38	35	45	171	44	41	48	50	84	267
	%	17	14	22	21	26	100	17	15	18	19	32	100
Sciences	N	26	391	332	329	430	1508	102	495	494	633	851	2575
	%	2	26	22	22	29	100	4	19	19	25	33	100

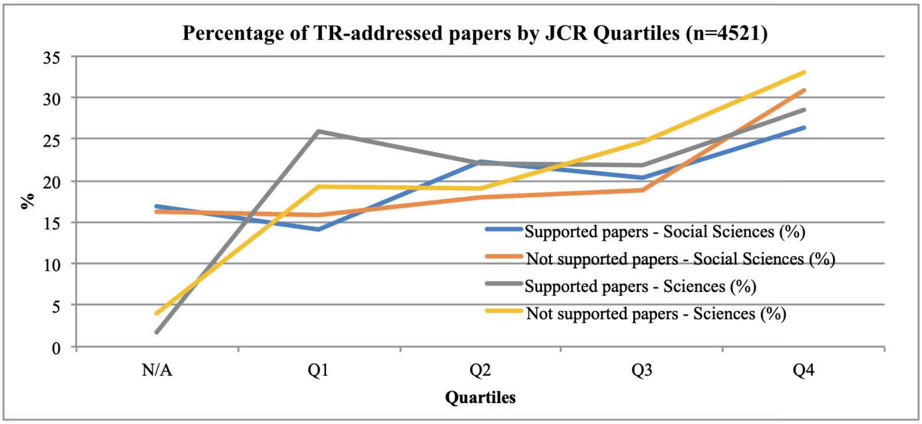


Fig. 5. Percentage of papers by JCR Quartiles (Social Science vs. Science) (n=4521)

The main findings of this study with regards to about 226,000 TR-addressed papers published between 2006 and 2015 are as follows: They were published in relatively low impact journals. More than half appeared in journals with AIS values well below the world average of AIS value for all journals (1.000) indexed in WoS. TR-addressed papers did not get cited very often, as half the papers received between zero (13%) and five citations within the study period. Supported and not supported papers collected comparable number of citations per paper. Supported and not supported papers were not significantly different from each other in terms of JIFs, AISs, and quartiles of journals in which they were published. The distributions of the supported and not supported Science and Social Science papers by citation impact values did not differ much, either.

Findings of the current study corroborate to some extent the findings of earlier studies of PRFSs (Auranen & Nieminen, 2010; Butler, 2003, 2004; Checchi, Malgarini, & Sarlo, 2019; Geuna & Martin, 2003; Good et al., 2015; Osuna, Cruz-Castro, & Sanz-Menéndez, 2011). We have not observed a negative correlation between JIFs, AISs, and quartiles of supported and not supported TR-addressed papers. Yet, the average number of citations per paper and JIF, AIS and quartile values were quite similar for both supported and not supported papers, indicating that the support system of TÜBİTAK has not increased the citation impact of TR-addressed papers, which confirms the findings of an earlier study (Tonta, 2017b).

We lack empirical data as to why the support system did not have any considerable effect. One reason might be that researchers have had to write papers for tenure and academic promotion anyway. Or, they may have found support through other sources (e.g., project budgets or other academic incentives). Yet another reason might be that the amount of support (based primarily on JIF values) was perhaps not attractive enough for some researchers, especially when we consider the fact that many TR-addressed papers were published in relatively low impact journals and the total amount per paper has to be divided by the number of co-authors (Tonta, 2017b).

Concluding Remarks

Findings indicate that both supported and not supported TR-addressed papers were somewhat similar in terms of average number of citations per paper. They have been published in journals with similar JIFs, AISs, and quartiles. Contrary to the expectations of the research funders, payments transferred to researchers through the support program do not seem to have played much role in improving the citation impact of TR-addressed papers. This suggests that subsidies based on bibliometric measures function poorly as incentives to increase the quantity and quality of the scholarly papers.

The support system seems to have rewarded the authors of papers who published in mediocre or low impact journals relatively more often. Despite comparatively lower "piece rates" paid for papers published in such outlets, many researchers sought financial support nonetheless. This might be an indication that subsidies may have encouraged some researchers to develop "opportunistic behavior" and act like "rent seekers" interested only in reaping the relatively modest monetary benefits of the support program without much consideration for the quality of their papers, a conjecture begging further research.

In conclusion, the support program of TÜBİTAK and similar academic incentives of the Turkish Higher Education Council should be reconsidered.

Acknowledgements

We thank Mr. M. Mirat Satoğlu of TÜBİTAK for providing data for supported papers, and Dr. Umut Al of Hacettepe University for reviewing an earlier version of this paper.

References

Abramo, G., D'Angelo, C.A. & Di Costa, F. (2018). When research assessment exercises leave room for opportunistic behavior by the subjects under evaluation. Retrieved July 16, 2019 from: https://arxiv.org/abs/1810.13216.

Akça, S. & Akbulut, M. (2018). Türkiye'deki yağmacı dergiler: Beall listesi üzerine bir araştırma. *Bilgi Dünyası*, 19(2): 255-274. doi: 10.15612/BD.2018.695

Arendt, J. (2010). Are article influence scores comparable across scientific fields? *Issues in Science and Technology Librarianship*, 60. Retrieved July 16, 2019 from: http://www.istl.org/10-winter/refereed2.html.

Article Influence Score. (2019). Retrieved July 16, 2019 from: http://ipscience-help.thomsonreuters.com/incitesLiveJCR/glossaryAZgroup/g4/7790-TRS.html.

Auranen, O., & Nieminen, M. (2010). University research funding and publication performance— An international comparison. *Research Policy*, 39(6): 822–834.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy*, 32(1): 143–155.

Butler, L. (2004). What happens when funding is linked to publication counts? In H.F. Moed et al., (Ed.), *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems* (pp. 389–405). Dordrecht: Kluwer.

Checchi, D., Malgarini, M., & Sarlo, S. (2019). Do performance-based research funding systems affect research production and impact? *Higher Education Quarterly*, 73: 45-69. Retrieved July 16, 2019 from: http://checchi.economia.unimi.it/pdf/91.pdf.

Çetinsaya, G. (2014). Büyüme, kalite, uluslararasılaşma: Türkiye yükseköğretimi için bir yol haritası. (2nd ed.). Ankara: Yükseköğretim Kurulu. Retrieved July 16, 2019 from: https://tinyurl.com/y6lokjt2.

De Boer, H. et al. (2015). *Performance-based funding and performance agreements in fourteen higher education systems*. (Report for the Ministry of Culture and Science. Reference: C15HdB014) Enschede: Center for Higher Education Policy Studies University of Twente. Retrieved July 16, 2019 from: http://bit.ly/2DZNVWP.

Demir, S.B. (2018). Predatory journals: Who publishes in them and why? *Journal of Informetrics*, 12(4): 1296-1311.

European Commission. (2010). Assessing Europe's University-Based Research. Retrieved July 16, 2019 from: http://bit.ly/2oNukmM.

Fischer, I. & Steiger, H-J. (2018). Dynamics of Journal Impact Factors and limits to their inflation. *Journal of Scholarly Publishing*, 50(1): 26-36

Geuna, A., & Martin, B. (2003). University research evaluation and funding: An international comparison. *Minerva*, 41(4): 277–304.

Good, B., Vermeulen, N., Tiefenthaler, B., & Arnold, E. (2015). Counting quality? The Czech performance-based research funding system. *Research Evaluation*, 24(2): 91–105.

Harley, Y.X., Huysamen, E., Hlungwani, C., & Douglas, T. (2016). Does the DHET research output subsidy model penalise high-citation publication? A case study. *South African Journal of Science*, 112(5–6): 1-3.

Hedding, D.W. (2019, January 15). Payouts push professors towards predatory journals. *Nature*, 565, 267. doi: 10.1038/d41586-019-00120-1.

Heywood, J.S., Wei, X., & Ye, G. (2011). Piece rates for professors. *Economics Letters*, 113(3): 285–287.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2): 251–61.

Jonkers, K., & Zacharewicz, T. (2016). *Research performance based funding systems: A comparative assessment*. Luxembourg: Publications Office of the European Union. Retrieved July 16, 2019 from: http://publications.jrc.ec.europa.eu/repository/bitstream/JRC101043/kj1a27837enn.pdf.

Kamalski, J., Huggett, S., Kalinaki, E., Lan, G., Lau, G., Pan, L., & Scheerooren, S. (2017). *World of research 2015: Revealing patterns and archetypes in scientific research*. Elsevier Analytic Services. Retrieved July 16, 2019 from: http://www.doc88.com/p-2032803429898.html.

Lee, A.T.K., & Simon, C.A. (2018). Publication incentives based on journal rankings disadvantage local publications. *South African Journal of Science*, 114(9/10): 1-3.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23: 151-169.

Mouton, J., & Valentine, A. (2017). The extent of South African authored articles in predatory journals. *South African Journal of Science*, 113(7/8): 1-9.

Muller, S.M. (2017). Academics as rent seekers: distorted incentives in higher education, with reference to the South African case. *International Journal of Educational Development*, 52: 58-67.

Osuna, C., Cruz-Castro, L., & Sanz-Menéndez, L. (2011). Overturning some assumptions about the effects of evaluation systems on publication performance. *Scientometrics*, 86(3): 575–592.

Pajić, D. (2014). Globalization of the social sciences in Eastern Europe: Genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*, 102(3): 2131-2150.

Pillay, T.S. (2013). Subject and discipline-specific publication trends in South African medical research, 1996-2011. *South African Journal of Science*: 109 (9/10), Article #2012-0054. Retrieved July 16, 2019 from: http://dx.doi.org/10.1590/sajs.2013/20120054.

Sīle, L. & Vanderstraeten, R. (2019). Measuring changes in publication patterns in a context of performance-based research funding systems: the case of educational research in the University of Gothenburg (2005–2014). *Scientometrics*, 118: 71-91.

Teodorescu, D. & Andrei, T. (2014). An examination of “citation circles” for social sciences journals in Eastern European countries. *Scientometrics*, 99(2): 209-231. Doi: 10.1007/s11192-013-1210-6.

Tomaselli, K.G. (2018). Perverse incentives and the political economy of South African academic journal publishing. *South African Journal of Science*, 114(11/12): 1-6.

Tonta, Y. (2017a). Does monetary support increase the number of scientific papers? An interrupted time series analysis. *Journal of Data and Information Science*, 3(1): 19-39. doi:10.2478/jdis-2018-0002

Tonta, Y. (2017b). *TÜBİTAK Türkiye Adresli Uluslararası Bilimsel Yayınları Teşvik (UBYT) Programının değerlendirilmesi*. Ankara: TÜBİTAK ULAKBİM. Retrieved July 16, 2019 from: http://ulakbim.tubitak.gov.tr/sites/images/Ulakbim/tonta_ubyt.pdf.

Telling the Early Story of Solar Energy Meteorology by Applying (Co-Citation) Reference Publication Year Spectroscopy

Thomas Scheidsteger[1] and Robin Haunschild[2]

[1] *t.scheidsteger@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70659 Stuttgart (Germany)

[2] *r.haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70659 Stuttgart (Germany)

Abstract

Studying the history of research fields by analyzing publication records and topical and/or keyword searches with a full Reference Publication Year Spectroscopy (RPYS) has been introduced as a powerful tool to identify the corresponding root publications. However, for a rather new and interdisciplinary research field like Solar Energy Meteorology (SEM), this method is not feasible to get a reasonably exhaustive publication set. Therefore we apply its variant RPYS-CO to all publications co-cited with one highly important marker paper, using the CRExplorer for plotting and inspecting the spectrogram of the number of cited references. Examining its peaks and their main contributing publications, we get a list of seminal papers, which are able to adequately tell us the story of SEM up to the 1990s. Generally, we recommend this method to gain valuable insights in (new) research fields.

Introduction

Solar Energy Meteorology (SEM) studies how solar radiation can be utilized for solar energy conversion to provide heat or electricity and how the performance of these conversion processes is affected by meteorological influences. This is largely the question of its availability in time (e.g., time of day, year) and space (e.g., geographical location, angular orientation). The quality of radiation then matters when different devices are used: Concentrating devices need *direct* radiation to work whereas non-concentrating photovoltaics, i.e., solar cells, can also utilize the *diffuse* fraction of sunlight, usually scattered in the atmosphere. So, the main fields of investigations are (i) measurements and their evaluation over different time scales and (ii) modeling of radiation and its components, depending on physical (e.g., solar constant, equation of time), geometrical (e.g., position of the sun, orientation of the converter) and meteorological (e.g., cloud coverage, aerosol concentration) parameters. Both fields also involve a lot of statistical treatment.

With the availability of geostationary weather satellites, for example the European METEOSAT, methods based on satellite data have been developed in several research groups since the 1980s. One of the authors has been a member of Oldenburg University's research group on Energy Meteorology for some time, so it came naturally to investigate this field of research with bibliometric methods in order to identify seminal and landmark papers, that lead up to the state of the art at the beginning of the satellite era in SEM.

Due to its intrinsic interdisciplinarity, potential search terms tend to have multiple meanings, which leads to answer sets from title or topical searches with low precision and/or recall. Therefore, we use the bibliometric method Reference Publications Year Spectroscopy (RPYS), introduced by Marx, Bornmann, Barth, and Leydesdorff (2014), but in a variant called RPYS-CO (for co-citation), where there is no need for an exhaustive paper set covering most of the research field. In RPYS-CO, the analyzed publication set is defined by at least one marker paper. All publications co-cited with this marker paper are included in the RPYS analysis. The RPYS has successfully been applied to identify the root publications of climate change research by Marx, Haunschild, Thor, and Bornmann (2017). In that case a set of more than 200,000 papers has been used. In a subsequent approach in the same paper, Marx, et al.

(2017) refined this large set to the greenhouse effect by keeping only cited references that are co-cited with Arrhenius (1896) and were able to retrieve the results of the RPYS on the full publication set regarding the greenhouse effect, but also lesser known works of relevance. They named this RPYS variant RPYS-CO. In another very recent study, (Haunschild & Marx, 2019) compare their own results of a RPYS on density functional theory, a very frequently applied method in computational chemistry (Haunschild, Barth, & Marx, 2016), with an RPYS-CO using *one single* seminal paper with a high citation count and find a striking similarity with the results of the analysis based on a search in controlled vocabulary.

Encouraged by these results, we set out here to investigate all publications co-cited with one highly-cited marker paper, a choice discussed with and corroborated by the long term leader of the Oldenburg group: *The interrelationship and characteristic distribution of direct, diffuse and total solar radiation* (Liu & Jordan, 1960). In order to indicate the importance of this marker paper we quote its complete abstract, emphasizing in italics all those concepts and terms that proved to be prevalent in SEM for its whole history:

"Based upon the data now available, this paper presents relationships permitting the determination on a *horizontal surface* of the instantaneous intensity of *diffuse radiation* on *clear days*, the *long term average hourly and daily sums* of diffuse radiation, and the daily sums of diffuse radiation for various categories of days of *differing degrees of cloudiness*. For these determinations, it is necessary to have, either from actual measurements or estimates, a knowledge of the *total (direct plus diffuse) radiation on a horizontal surface* – its *measurement is now regularly made* at 98 localities in the United States and Canada. For localities where only an *estimate of the long term average total radiation* is available, *relation-ships* presented in this paper can be utilized to determine the *statistical distribution of the daily total radiation at these localities*." (Liu & Jordan, 1960, abstract)

Satellite-based studies often view this paper as text book knowledge. Therefore it is affected by obliteration by incorporation (Cronin & Sugimoto, 2014) and rarely cited in this area of SEM, that gained traction in the late 1980s and early 1990s. The latter are consequently a natural end date for our study of cited references, co–cited with (Liu & Jordan, 1960). The then flourishing satellite-based publications could and should be the target of further investigations using other marker papers. On the other hand, due to the recency of the research field, we do not expect decisive contributions to SEM before 1900.

There are some studies with very different time frame, focus or methodology, e.g.: Du, Li, Brown, Peng, and Shuai (2014) analysed the solar energy literature from 1992 to 2011, but with no consideration of energy meteorology topics. A bibliometric analysis on solar power research between 1991 and 2010, again after the period of our study, has been performed by Dong, Xu, Luo, Cai, and Gao (2012) using terms as, e.g., "solar radiation" in a topical search in the WoS. Their goal was to identify research trends for the twenty-first century and not to explore historical roots. In the same vein Yang, Kleissl, Gueymard, Pedro, and Coimbra (2018) tried to identify key innovations for the future of research in "solar radiation and PV power forecasting", a field mainly emerging at the turn of the millennium. They based their work on the first 1000 hits of a keyword search in Google Scholar and applied machine learning and text mining methods to full texts in order to complement conventional topical reviews.

To the best of our knowledge, the present study is one of the first using the method RPYS-CO in order to identify seminal papers for a research field – thereby complementing qualitative knowledge of experts by a quantitative evaluation of the citation counts (i.e., the reference counts within the topic related literature). Using this method we were confident to find by this method those important contributors and their papers which tell the story of the emergence of solar energy meteorology from around 1900 up to the beginning of the 1990s. So we support the suggestion of (Haunschild & Marx, 2019) that this method can help researchers to explore their field of study - in a way complementary to a usual topical or keyword-based literature search.

Method and data set

As of 8 January 2019, Liu and Jordan (1960) had 1032 citing papers in the WoS until the end of 2018. One fourth of these papers (n=257, 25%) as well as the marker paper itself have been published in a single journal, *Solar Energy.* Their four most important WoS subject categories in the data set used in this study are *Energy Fuels* (n=673, 65%), *Green Sustainable Science Technology* (n=151, 15%), *Meteorology & Atmospheric Sciences* (n=131, 13 %), and *Thermodynamics* (n=114, 11%), thereby reflecting the multiple foci of SEM.

We downloaded the bibliographic data of the 1032 papers including 36,635 cited references (CRs) from the WoS (selecting "Save to Other File Formats" and "Other Reference Software") and imported them into the CRExplorer. (The JAVA based software can be downloaded for free from http://crexplorer.net and a comprehensive handbook explaining all functions is also available.) It provides a graphical display of the number of CRs (NCR) over the reference publication years (RPY) and a tabular presentation of the NCR of all CRs. In our case there were only single occurrences of CRs before 1900. After 1995, despite a sequence of steadily increasing peaks, no specific papers of main contribution (more than a share of 10% of the NCRs in the specific RPY) could be identified. Both facts confirm our choice of the time period to be analyzed.

Much of the processing can comprehensively and reproducibly be done by using the CRExplorer scripting language: With the script in Listing 1 we `import`ed the WoS file and got 8383 unique reference variants for the reference publication years 1900 to 1995. After that `cluster`ing and `merg`ing of equivalent CR variants was done with Levenshtein threshold 0.75 and taking volume and (starting) page number into account, thereby reducing the number of CR variants by 109. Then we `remove`d all publications with only *one* citation, in order to reduce noise. In the end, we retained 1566 CRs. The results including the NCR and other indicators were `export`ed to CSV files for further inspection and plotting of the spectrogram, which can be done by using the R package BibPlots (see: https://cran.r-project.org/web/packages/BibPlots/index.html and https://tinyurl.com/y97bb54z).

```
importFile(file:"savedrecs_Liu_1960.txt",type:"WOS",
RPY:[1900,1995,false], PY:[1962,2018,false], maxCR:0 )
info()
cluster(threshold:0.75,volume:true,page:true,DOI:false)
merge()
info()
removeCR(N_CR: [0, 1])
info()
saveFile(file:"Liu1960.rpys.cre")
exportFile(file:"Liu_1960.rpys_CR.csv",type:"CSV_CR")
exportFile(file:"Liu_1960.rpys_GRAPH.csv",type:"CSV_GRAPH")
```

Listing 1. CRExplorer script to perform RPYS on the WoS papers citing Liu and Jordan (1960)

In the spectrogram, we looked for publication years with significantly higher NCR than other years, aided by the deviation of NCR from the 5-year-median of NCR (taking into account the two preceding and the two following years). For the papers that, by applying this methodology, seemed primarily responsible for the peaks a manual merging was done, if needed.

Results

Figure 1 shows the spectrogram of the RPYS-CO for the marker paper (Liu & Jordan, 1960) in terms of the NCR and their 5-year-median deviation for the whole analyzed time period, and Table 1 lists all publications, contributing substantially to the peaks of NCR and identified as relevant.

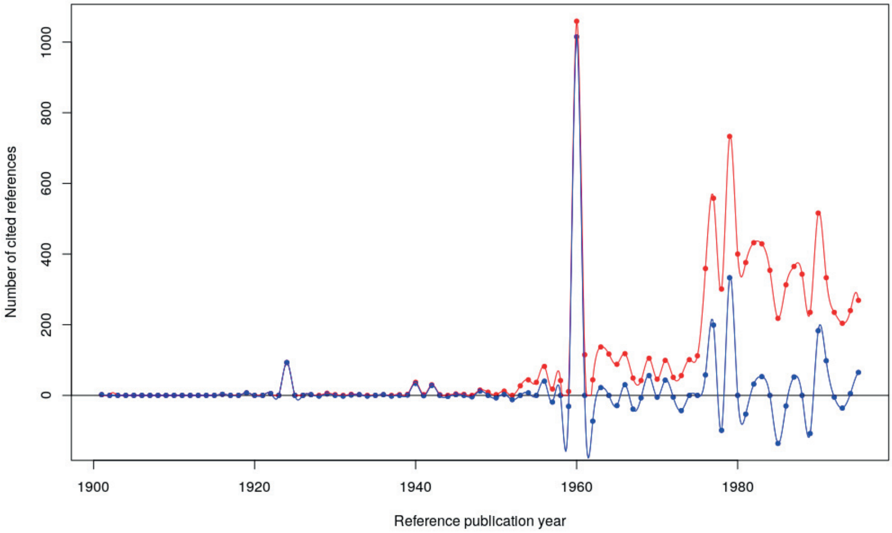


Figure 1. Overall RPYS-CO graph for Liu and Jordan (1960) with NCR (red line) and 5-year-median deviation (blue line)

Table 1. RPYS-CO for Liu and Jordan (1960): important CRs from 1900 to 1995 with number of citations NCR and indicating, if manually merged during inspection of spectrogram

#CR	*RPY*	*Cited Reference (Manually merged: **M**)*	*NCR*
CR1	1919	Kimball HH, 1919, Monthly Weather Review, V47, P769	7
CR2	1922	Angström A, 1922, Ark Mat Astron Fys,V17, P1	3
CR3	1922	Linke F, 1922, Beitr Phys Atmos, V10, P91	2
CR4	1924	Angström A, 1924, Quarterly Journal of the Royal Meteorological Society, V50, P121 ***(M)***	93
CR5	1929	Angström A, 1929, Geogr Annlr Stockhol, V11, P156 ***(M)***	6
CR6	1940	Prescott J, 1940, T Roy Soc South Aust, V64, P114	30
CR7	1942	Hottel HC, 1942, Transactions of the ASME, V64, P91 ***(M)***	23
CR8	1945	Haurwitz B, 1945, J Met, V2, P154	4
CR9	1946	Haurwitz B, 1946, J Met, V3, P123	3
CR10	1948	Haurwitz B, 1948, J Met, V5, P110	5
CR11	1953	Whillier A, 1953, Thesis, MIT Cambridge ***(M)***	17
CR12	1954	Black JN, 1954, Q J Roy Meteor Soc, V80, P231	28
CR13	1955	Hottel HC, 1955, T C Use Solar Energy, V2, P74	27
CR14	1956	Whillier A, 1956, Arch Meteorol Geophys U Bioklimatol Ser B,	38

		V7, P197 ***(M)***	
CR15	1958	Glover J, 1958, Q J Roy Meteor Soc, V84, P172	18
CR16	1960	Liu BYH, 1960, Solar Energy, V4, P1	1031
CR17	1963	Liu BYH, 1963, Solar Energy, V7, P53	71
CR18	1963	Choudhury NKD, 1963, Solar Energy, V7, P44	37
CR19	1964	Page JK, 1964, P UN C New Sources E, V4, P378 ***(M)***	87
CR20	1966	Stanhill G, 1966, Solar Energy, V10, P96	29
CR21	1966	Robinson N, 1966, Solar Radiation ***(M)***	25
CR22	1966	Kasten F, 1966, Arch Meteorol Geop B, VB14, P206	11
CR23	1969	Cooper PI, 1969, Solar Energy, V12, P333 ***(M)***	42
CR24	1969	Kondratyev KY, 1969, Radiation in the Atmosphere	23
CR25	1971	Spencer J, 1971, Search, V2, P172 ***(M)***	35
CR26	1974	Duffie JA, 1974, Solar Energy Thermal Processes ***(M)***	35
CR27	1976	Ruth DW, 1976, Solar Energy, V18, P153	68
CR28	1976	Hottel HC, 1976, Solar Energy, V18, P129	63
CR29	1976	Tuller SE, 1976, Solar Energy, V18, P259	46
CR30	1976	Hay JE, 1976, Atmosphere, V14, P278	34
CR31	1977	Orgill JF, 1977, Solar Energy, V19, P357	149
CR32	1977	Klein SA, 1977, Solar Energy, V19, P325	121
CR33	1977	Temps RC, 1977, Solar Energy, V19, P179 ***(M)***	46
CR34	1979	Collares-Pereira M, 1979, Solar Energy, V22, P155	238
CR35	1979	Klucher TM, 1979, Solar Energy, V23, P111	60
CR36	1979	Hay JE, 1979, Solar Energy, V23, P301	59
CR37	1980	Duffie JA, 1980, Solar engineering of thermal processes, 1st Ed.	63
CR38	1980	Iqbal M, 1980, Solar Energy, V24, P491	46
CR39	1981	Bendt P, 1981, Solar Energy, V27, P1	66
CR40	1982	Erbs DG, 1982, Solar Energy, V28, P293	208
CR41	1983	Iqbal M, 1983, Introduction to Solar Radiation ***(M)***	117
CR42	1987	Skartveit A, 1987, Solar Energy, V38, P271	50
CR43	1987	Perez R, 1987, Solar Energy, V39, P221	42
CR44	1988	Suehrcke H, 1988, Solar Energy, V40, P413	36
CR45	1988	Graham VA, 1988, Solar Energy, V40, P83 ***(M)***	28
CR46	1990	Reindl DT, 1990, Solar Energy, V45, P1	116
CR47	1990	Reindl DT, 1990, Solar Energy, V45, P9	54
CR48	1990	Perez R, 1990, Solar Energy, V44, P271	49
CR49	1991	Duffie JA, 1991, Solar engineering of thermal processes, 2nd Ed. ***(M)***	83

The overall RPYS-CO picture can easily be divided by the maximum NCR per RPY into two periods with regard to the reference publication years, which we are going to discuss separately: the first one from 1915 leading to, but excluding, 1960, the publication year of the marker paper, containing peaks with at most NCR=100; the second one from 1960 to 1995, with peaks between NCR=100 and NCR=900 (apart from the marker paper itself).

Time Period 1: 1915 to 1959

Figure 2 shows the RPYS-CO spectrogram for the marker paper (Liu & Jordan, 1960) for the relevant time period before it was published.

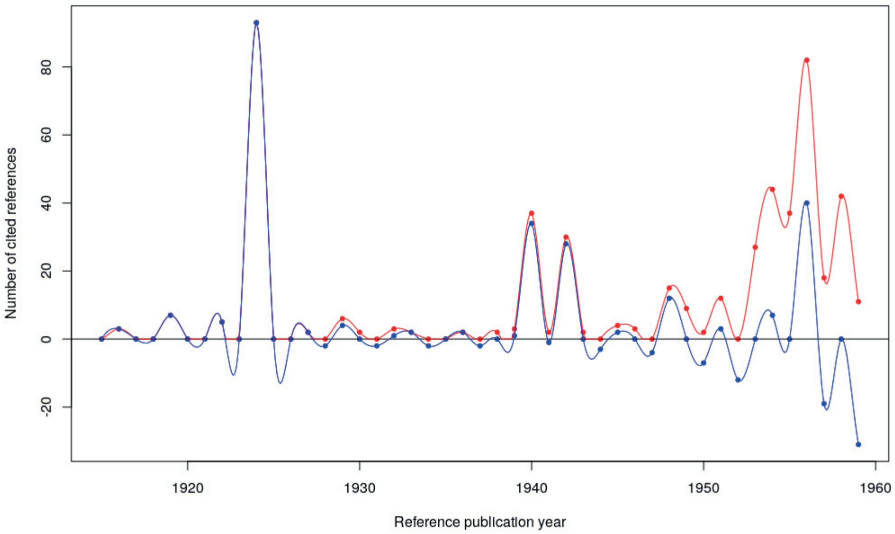


Figure 2. RPYS-CO graph for Liu and Jordan (1960) and period 1 (1915 – 1959) with NCR (red line) and 5-year-median deviation (blue line).

In this time period, we were able to identify 9 peaks with *relevant* papers for the following RPYs: 1919, 1922, 1924, 1929, 1940, 1942, 1945/46, 1948, and 1953-58. Because of the generally low NCR values in this time period, we did not want to lose reference variants of possibly relevant papers and therefore additionally looked at the CRs *before* the removal of only once referenced papers. But this did not reveal new relevant papers, instead it only confirmed the results from the reduced set.

Now we can follow the path of SEM by looking at the peak papers and drawing partially from explanations given in the citing papers. In this first period, two independent streams of research flew together: *meteorology* and *engineering*.

Meteorologists tried to develop a climatology of irradiance, emphasizing daily mean values, with no or little application to solar energy in mind. Solar irradiance varies deterministically with the sun's position on the sky dome and irregularly with changing cloudiness. The relation of the latter with sunshine has initially been measured by Kimball (CR1) and later subjected to statistical analysis by Angström (CR2, CR4, and CR5), leading to a linear relation between the duration of bright sunshine and average solar energy available on a horizontal surface at ground level, the so-called *Angström equation*. This has been generalized to the *Angström–Prescott* (CR6) *equation* by introducing a *daily clearness index*, quantifying all stochastic meteorological influences, as a measure of the atmospheric transparency (Paulescu et al., 2016). Linke (CR3) in 1922 published his *turbidity factor* for the attenuation of the sun's radiation by water vapor and aerosols in the atmosphere. Later Black, Bonython & Prescott (CR12) gave a linear regression relation between solar radiation and the duration of sunshine based on monthly values (Munkhammar & Widen, 2016) and Glover & McCulloch (CR15) improved this by including latitude effects.

Two *engineers*, Hottel & Woertz (CR7), came up with the first serious study on solar energy in 1942: the fundamental relationships given in their classic paper have since then been used for decades to model solar collectors. Hottel & Whillier (CR13) evaluated them concerning the flat-plate solar collector performance (Stanciu, Stanciu, & Paraschiv, 2016) and formulated the *Hottel-Whillier-Bliss equation* on its heat flow and available heat balance, based on considerations of the thermal usability of solar irradiation, coming from Whillier's PhD thesis (CR11) under Hottel's supervision at MIT. This latter work concerned also the relation between radiation on different time scales, showing a close interdependence of the frequency distributions of the so-called clearness index on a monthly, daily, and hourly basis (Vijayakumar, Kummert, Klein, & Beckman, 2005). Because information on sunshine duration was no longer sufficient, he later proceeded to "The determination of hourly values of total solar radiation from daily summations" (CR14) by statistical means, a subject still of great importance for SEM, where still an ever more time-resolved knowledge of solar irradiance is needed.

The *modeling* of solar irradiance components through parameterization of atmospheric phenomena is an equally important area of work in SEM. It was begun in the 1940s by Haurwitz (CR8-CR10), focusing on cloudiness, cloud density, and cloud type (Chowdhury, 1990). (We do not include a publication from the 1948 peak by Penman with NCR=8, i.e. more than CR10, because it is only concerned with evaporation by solar radiation.)

Time Period 2: 1960 to 1995

Figure 3 shows the RPYS-CO spectrogram for the marker paper (Liu & Jordan, 1960) after its publication year.

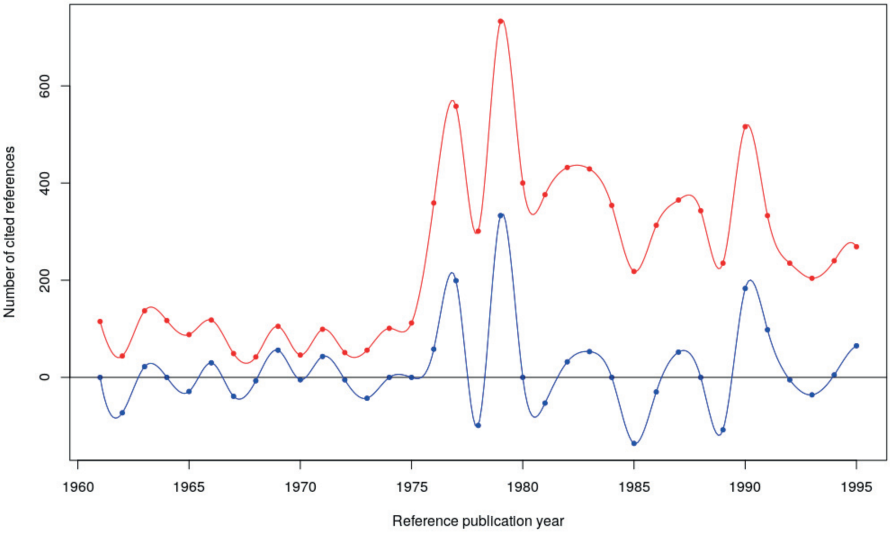


Figure 3. RPYS-CO graph for (Liu & Jordan, 1960) and period 2 (1960 – 1995) with NCR (red line) and 5-year-median deviation (blue line). For better presentation, the RPY 1960 is excluded.

Another 10 peaks could be identified from Figure 3 for the following RPYs in the second time period: 1960, 1963/64, 1966, 1969, 1971, 1976/77, 1979/80, 1981-83, 1986-88, 1990/91.

The first two peaks are mainly caused by *engineers*: After the marker paper itself (CR16) the same authors gave generalized curves to predict the "Long-Term Average Performance of Flat-Plate Solar-Energy Collectors", making use of Hottel's and Whillier's methods (CR13) and building upon the knowledge of two parameters only: i) the monthly-average daily total radiation on a horizontal surface and ii) the monthly average day-time ambient temperature (CR17). In 1961, J. K. Page presented "The estimation of monthly mean values of daily total short-wave radiation on vertical and inclined surfaces from sunshine records for latitudes 40N-40S" **(**CR19**)** on a "UN Conference on New Sources of Energy" in Rome but the conference proceeding was published in 1964. Much later, he advised advanced publicly funded projects like HELIOSAT-3 (Mueller et al., 2004).

The *meteorologist* F. Kasten (CR22) developed turbidity models as one essential ingredient for radiation model calculations and also functioned as an advisor in later solar energy projects.

Attempts to check and confirm the *diffuse-to-total radiation correlation* by Liu and Jordan (1960) against *measurements* have been undertaken for several locations in the world: Southern Israel by Stanhill (CR20), New Delhi by Choudhury (CR18), and Canada by Ruth and Chant (CR27) and Tuller (CR29).

Cooper (CR23) and Spencer (CR25) are the only representatives in Table 1 of those researchers concerned with *solar geometry*, i.e. sun-earth angle values over time – which is essential for all modeling of radiation. In this respect, our method could only capture these first works, but not later standard works like Michalsky (1988).

Robinson's "Solar Radiation" (CR21) was a meteorological standard publication, but not that much focused as Kondratyev's monograph "Radiation in the Atmosphere" (CR24), whose influence lasted until Iqbal's standard work "Introduction to Solar Radiation" (CR41) from 1983.

In the 1970s and early 80s, one focus of research literature was on *time-resolved diffuse radiation models from the scale of months down to hours*, mostly from the viewpoint of *engineering* like Duffie & Beckman's volumes "Solar Energy Thermal Processes" in 1974 (CR26) and "Solar engineering of thermal processes" in two editions in 1980 (CR37) and 1991 (CR49), but also Hottel (CR28), Orgill & Hollands (CR31), Klein (CR32), and Erbs, Klein & Duffie (CR40). Only Hay (CR30) and Iqbal (CR38) represent *geography* resp. *meteorology*. *Empirical radiation modeling*, in particular with respect to tilted surfaces (e.g., of solar panels), was done by Hay (CR36) and *meteorologists* as Temps & Coulson (CR33) and Klucher (CR35).

At the end of the 1970s, the focus switched also to *stochastic modeling*, outstandingly covered by Collares-Pereira & Rabl (CR34) with their time series analysis and production of the *first synthetic time series*, that were widely used in the community. In the same vein, Bendt presented his "frequency distribution of daily insolation values" (CR39). The time-scale was later even narrowed down to *minute data* by Suehrcke & McCormick (CR44), and Graham, Hollands & Unny (CR45) were able to *simulate daily values* of the clearness index from monthly mean values by using probability distribution functions.

In 1987, Skartveith & Olseth (CR42) presented a *diffuse fraction model*, that became essential part of later works, e.g. in HELIOSAT (Dürr & Zelenka, 2009). CR43, i.e., Perez, Seals, Ineichen, Stewart, and Menicucci (1987), also focused on the diffuse part of total irradiance and accomplished a major improvement in its error-prone computation, in order to "estimate short time step (hourly or less) irradiance on tilted planes" (Perez, et al., 1987), which has received high recognition in the community. (See *Discussion & Conclusion* for considerations to use CR43 as a second marker paper.)

Duffie & Beckman, together with their coauthor Reindl, were mainly responsible for the last high peak, taken into account in our RPYS-CO analysis, in 1990: they *evaluated statistical models for hourly radiation on the tilted surface* (CR47) and could significantly *improve on the time resolution of statistical diffuse radiation models* in CR46, whose abstract we now quote (with our emphases in italics): "The influence of climatic and geometric variables on the *hourly diffuse fraction* has been studied, based on a data set with *22,000 hourly* *measurements* from five European and North American locations. The goal is to determine if other predictor variables, in addition to the clearness index, *will significantly reduce the standard error* of *Liu- and Jordan-type correlations* (…). Stepwise regression is used to *reduce a set of 28 potential predictor variables* down to *four significant predictors: the clearness index, solar altitude, ambient temperature, and relative humidity***.”** (Reindl, Beckman, & Duffie, 1990, abstract)

We can in a sense close the circle to our marker paper after exactly 30 years by mentioning CR48, i.e., Perez, Ineichen, Seals, Michalsky, and Stewart (1990), as a successful attempt to *apply diffuse fraction models* to questions of *daylighting in buildings*, transferring irradiance to illumination, and again connecting the fields of *meteorology & radiation* to *energy & engineering*, the two-fold focus of SEM.

Discussion & Conclusion

We studied the early history of SEM by applying RPYS-CO to one highly cited and relevant marker paper (Liu & Jordan, 1960), matching the recommendation of a long-term expert, inspected RPYs with peak citation numbers in the corresponding graph and table calculated by the CRExplorer and were able to thereby identify many important papers before and after the marker paper. They give an adequate view of most of the essential contributing streams of research in SEM, as, e.g., measuring, empirical and statistical modeling of direct and diffuse radiation on the horizontal and the tilted plane on time scales from years to minutes. The topics *solar geometry, radiative transfer* calculations through the atmosphere, and *spectrally resolved* treatment of sun light can be identified as underrepresented in our RPYS-CO results. But the latter two in particular gained interest only in later years and should be studied with other marker papers.

It could also be argued that an RPYS-CO on one or few marker papers only produces a bias by favoring a limited number of research groups, maybe even enforced by the effect of self-citations. But in the given case the marker paper is obviously so well chosen as to unearth a diverse set of methodologies and approaches in the world-wide SEM community, coming from the two main domains *meteorology* and *engineering* and covering *measurement*, *modeling* and *evaluation*. (This could be different when we are going to study the satellite based methods, where European and US research groups take slightly different approaches.) Moreover, self-citations of the admittedly repeatedly occurring (co-)authors among the peak papers play no role in our study because of their sporadic appearance.

CR43 (Perez, et al., 1987) had been suggested as a second marker paper by the expert, but as it turned out, the RPYS-CO on both papers only confirms the results for Liu and Jordan (1960) alone, as Figure 4 shows. Some NCR peaks get sharpened, but there are no new ones found. Furthermore, a RPYS-CO analysis on CR43 alone would reveal less peak papers than the RPYS-CO as performed here. An even stronger confirmation results from another RPYS-CO conducted for the two top most cited papers in our list of CRs, i.e., CR34 (Collares-Pereira & Rabl, 1979) with 238 citations and CR40 (Erbs, Klein, & Duffie, 1982) with 208 citations: *all* peaks and peak papers get reproduced, except from one in 1948 (CR10)!

Moreover, none of the potential candidates in our list of CRs got nearly as many citations as Liu and Jordan (1960) in the whole Web of Science. These facts corroborate the careful choice of the marker paper and the stability of the method’s outcomes.

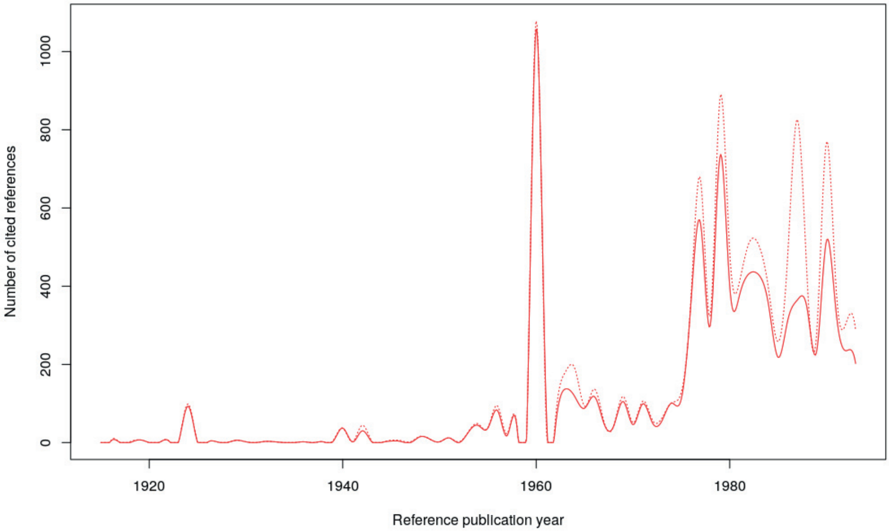


Figure 4. Comparison of the RPYS-CO for the marker paper Liu and Jordan (1960) (solid line) and the RPYS-CO for the marker paper plus the potential second marker paper Perez, et al. (1987) (dotted line).

In total, the outcome of our study meets our expectations: All relevant historical roots of SEM research were disclosed by our RPYS-CO analysis. Therefore, we recommend RPYS-CO for similar investigations by researchers to get more insight in the development of their field of work or even as a tool for studies in the history of science.

Acknowledgments

We want to thank Detlev Heinemann for very useful discussions on the "early story of Solar Energy Meteorology" and comments regarding an earlier version of this manuscript. We also thank Lutz Bornmann for helpful discussions.

References

Arrhenius, S. (1896). On the influence of carbonic acid in the air upon the temperature of the ground. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 41*(251), 237-276. doi: 10.1080/14786449608620846.

Chowdhury, B. H. (1990). Short-term prediction of solar irradiance using time-series analysis. *Energy Sources, 12*(2), 199-219. doi: 10.1080/00908319008960199.

Collares-Pereira, M., & Rabl, A. (1979). Average distribution of solar-radiation - correlations between diffuse and hemispherical and between daily and hourly insolation values. *Solar Energy, 22*(2), 155-164. doi: 10.1016/0038-092x(79)90100-2.

Cronin, B., & Sugimoto, C. R. (2014). Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact / edited by Blaise Cronin and Cassidy R. Sugimoto (pp. 131-149). Cambridge, Mass; London: MIT Press.

Dong, B., Xu, G., Luo, X., Cai, Y., & Gao, W. (2012). A bibliometric analysis of solar power research from 1991 to 2010. *Scientometrics, 93*(3), 1101-1117. doi: 10.1007/s11192-012-0730-9.

Du, H., Li, N., Brown, M. A., Peng, Y., & Shuai, Y. (2014). A bibliographic analysis of recent solar energy literatures: The expansion and evolution of a research field. *Renewable Energy, 66*, 696-706. doi: 10.1016/j.renene.2014.01.018.

Dürr, B., & Zelenka, A. (2009). Deriving surface global irradiance over the Alpine region from METEOSAT Second Generation data by supplementing the HELIOSAT method. *International Journal of Remote Sensing, 30*(22), 5821-5841. doi: 10.1080/01431160902744829.

Erbs, D. G., Klein, S. A., & Duffie, J. A. (1982). Estimation of the diffuse-radiation fraction for hourly, daily and monthly-average global radiation. *Solar Energy, 28*(4), 293-302. doi: 10.1016/0038-092x(82)90302-4.

Haunschild, R., Barth, A., & Marx, W. (2016). Evolution of DFT studies in view of a scientometric perspective. *Journal of Cheminformatics, 8*, 12. doi: 10.1186/s13321-016-0166-y.

Haunschild, R., & Marx, W. (2019). *Discovering Seminal Works with Marker Papers*. Paper presented at the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019), Cologne, Germany, April 14, 2019. http://ceur-ws.org/Vol-2345/paper3.pdf

Liu, B. Y. H., & Jordan, R. C. (1960). The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. *Solar Energy, 4*(3), 1-19. doi: 10.1016/0038-092x(60)90062-1.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the Historical Roots of Research Fields by Reference Publication Year Spectroscopy (RPYS). *Journal of the Association for Information Science and Technology, 65*(4), 751-764. doi: 10.1002/asi.23089.

Marx, W., Haunschild, R., Thor, A., & Bornmann, L. (2017). Which early works are cited most frequently in climate change research literature? A bibliometric approach based on Reference Publication Year Spectroscopy. *Scientometrics, 110*(1), 335-353. doi: 10.1007/s11192-016-2177-x.

Michalsky, J. J. (1988). The Astronomical Almanac's algorithm for approximate solar position (1950–2050). *Solar Energy, 40*(3), 227-235. doi: 10.1016/0038-092X(88)90045-X.

Mueller, R. W., Dagestad, K. F., Ineichen, P., Schroedter-Homscheidt, M., Cros, S., Dumortier, D., . . . Heinemann, D. (2004). Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. *Remote Sensing of Environment, 91*(2), 160-174. doi: 10.1016/j.rse.2004.02.009.

Munkhammar, J., & Widen, J. (2016). Correlation modeling of instantaneous solar irradiance with applications to solar engineering. *Solar Energy, 133*, 14-23. doi: 10.1016/j.solener.2016.03.052.

Paulescu, M., Stefu, N., Calinoiu, D., Paulescu, E., Pop, N., Boata, R., & Mares, O. (2016). Angstrom-Prescott equation: Physical basis, empirical models and sensitivity analysis. *Renewable & Sustainable Energy Reviews, 62*, 495-506. doi: 10.1016/j.rser.2016.04.012.

Perez, R., Ineichen, P., Seals, R., Michalsky, J., & Stewart, R. (1990). Modeling daylight availability and irradiance components from direct and global irradiance. *Solar Energy, 44*(5), 271-289. doi: 10.1016/0038-092X(90)90055-H.

Perez, R., Seals, R., Ineichen, P., Stewart, R., & Menicucci, D. (1987). A new simplified version of the Perez diffuse irradiance model for tilted surfaces. *Solar Energy, 39*(3), 221-231. doi: 10.1016/s0038-092x(87)80031-2.

Reindl, D. T., Beckman, W. A., & Duffie, J. A. (1990). Diffuse fraction correlations. *Solar Energy, 45*(1), 1-7. doi: 10.1016/0038-092x(90)90060-p.

Stanciu, D., Stanciu, C., & Paraschiv, I. (2016). Mathematical links between optimum solar collector tilts in isotropic sky for intercepting maximum solar irradiance. *Journal of Atmospheric and Solar-Terrestrial Physics, 137*, 58-65. doi: 10.1016/j.jastp.2015.11.020.

Vijayakumar, G., Kummert, M., Klein, S. A., & Beckman, W. A. (2005). Analysis of short-term solar radiation data. *Solar Energy, 79*(5), 495-504. doi: 10.1016/j.solener.2004.12.005.

Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T. C., & Coimbra, C. F. M. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. *Solar Energy, 168*, 60-101. doi: 10.1016/j.solener.2017.11.023.

Structure of Litigation Relationship Network among Dental Companies and Patent Portfolio Strategy -A Social Network Analysis

Chao-Chih Hsueh[1,3] and Mu-Hsuan Huang[2,3]

[1] *cchsueh@mail.npust.edu.tw*
Graduate Institute of Management of Innovation and Technology, National Pingtung University of Science and Technology, Pingtung (Taiwan R.O.C.)

[2] *Corresponding Author: mhhuang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taipei (Taiwan R.O.C.)

[3] Center for Research in Econometric Theory and Applications, National Taiwan University, Taipei (Taiwan R.O.C.)

Abstract

This study construct a systematic approach to explore how the advantages of the network position formed by the litigation of the dental firms with what patent portfolio they adopted based on the Actor-network Theory (ANT). The dataset is from 1995 to 2018 the 592 litigation dental patent, 640 litigation case and 240 plaintiffs and 548 defendants that formed a 550 firm matrix analyzed in the UCINET. The systematic approach include six approach: (1) calculate the social network variables-degree centrality and structure holes; (2)cluster analysis; (3)validation of cluster solution and cluster interpretation; (4)patent portfolio analysis. The cluster analysis show 6 cluster: Key player 1, Key player 2, Patent troll, Active Attacker, Bystander and Victim, to explore the position of each firm in the litigation network. The result show that key player1 and 2 use multiple patent portfolio to get competitive advantage. 3 Patent troll use only one patent or one patent portfolio to sue defendants. 23 Active Attacker firms has less patent portfolio than key players. The results will provide a broader theoretical perspective for the governance of network relations between organizations through litigation, and also provide practical advice for managers to understand the development of litigation networks and patent portfolio.

Introduction

In technology-driven industry, the IP war had become an important means for business competition. The patent right was that the country gave the right to create a legal monopoly market for the inventor. It was common for competitors to use patent litigation as a means of commercial competition. In the practice case of US patent litigation, litigation compensation may be as high as several billion dollars or quit the market due to the loss of the lawsuit. Besides, the Patent Troll had become increasingly popular in the United States in recent years and also use patent infringement lawsuits to get extra profit.

Therefore, how to develop patent portfolios strategy in response to patent litigation should be an important issue that business operators must consider in advance. For the purpose, the study took the medical equipment industry as an example to explore what the patent portfolio strategy that adopted by the manufacturers in the network position advantage formed on the entire litigation network was. In a patent litigation reconciliation funds may not be the final purpose, to file an infringement lawsuit may be a strategy to threaten competitors as a containment to stop them to enter market. From the patent strategy practices the firms would form a defensive patent pool as a countermeasure against patent litigation. It could select a number of patents with important strategic value in a specific technology field through jointly purchased or developed by different in order to react against a specific patentee’s monopoly. For example, Google spent $12.5 billion to acquiring Motorola Mobility that owned 17,000 patents, and then licensed the aggressive and powerful patents to Android smartphone makers with low prices to fight back to Apple.

Thus, in a fast and highly complex business environment, it is an important theoretical and practical basis for managers to understand what competitors will initiate a patent lawsuit, how they have sued with what kind of patent portfolio and how the defendants have reacted against the lawsuit. From the social interaction perspective, it is a key issue to analyze the relationship development of the organization in the network, to demonstrate the structure of the network and the position of firms and their competitors based on the governance mechanism and the arrangement of interests of the organization through cooperation, litigation or safeguard strategies (Jones et al., 1997).

Many studies have analyzed network structure to explore how the performance generated through economic transactions or resource exchanges among organizations (Jones, Hesterly, & Borgatti, 1997; Borgatti & Halgin, 2011; Bizzi & Langley, 2012). However, many studies have clarified structure of the network or position of actors by showing the role, ties or position of actors. This kind of study is useful to explain or understand the relationship of companies and what benefit they got through. It oversimplifies the overall development of the network, and it is difficult to meticulously describe the details of how the interactions between organizations, what governance mechanism actor adopted to react, cooperate or attack other actors and what the development of inter-organizational network relationships.

Accordingly, this study focus on both human (firms) and the non-human actor (patent portfolio), emphasizing the evolution between firms and their patent portfolio. This is the basic spirit of Actor-network Theory (ANT), which emphasizes action, evolution, and context. This study explore the development of litigation relationship network, discovers the relationship among plaintiffs and defendants and how plaintiffs used patent portfolio to obtain their network advantage or how defendants develop patent portfolios in response to patent litigation.

This paper takes the medical device industry as an example and construct a systematic approach to explore the advantages of the network position formed by the litigation of the dental firms and what patent portfolio they adopted in the entire litigation network based on the ANT theory. The results will provide a broader theoretical perspective for the governance of network relations between organizations through litigation, and also provide practical advice for managers to understand the development of litigation networks and patent portfolio.

Litigation Network and Social Network Indicators

Social network refers to a group of specific social relationships that are connected by actors. Actors and relationships are the basic elements of network analysis. Any actors are nodes in network could be an individual, an organization, a patent or a technology depends on research purpose. The connection (tie) between nodes expresses the relationship and flow between actors. In the social network analysis, many research have provide various network indicators to measure or explain various phenomena. In this study we adopted two network indicators- degree centrality and structure how to explain the structure of patent litigation network.

Centrality

Each actor has its positon in the social network structure. The structural position affects the actor's ability to control resources. When an actor is at the center of the network, it can control related resources and obtain more benefits (Ibrra, 1993). The centrality of social networks is used to measure the influence of actors in the network, and Freeman (1979) has proposed three measures, namely degree centrality, closeness centrality and betweenness centrality.

- **Degree centrality**

Degree centrality is to observe the connection relationship between a node and surrounding nodes. The more nodes a node is connected to, the more it has the informal power and influence associated with more actors in the network. Wassermann and Faust (1994) definite that the core member with more degree centrality must be the most active has the most ties to other actors in the network. Degree centrality only considers the direct link of the actor

- **Closeness centrality**

Since the degree centrality only considers the direct link of the actor, however, an actor may be connected to many nodes that may not have much connection with other ones in the entire network. Closeness centrality emphasizes the distance between actors and others, focusing on the distance between actors (Hanneman & Riddle, 2013). The connection among all points of the overall network will be measure, which can be used to determine the proximity of an actor to others. The shorter the distance from others, the higher the proximity to the center, indicating that they can obtain information faster.

- **Betweenness centrality**

Betweenness centrality treats actors as being between other peers of the network and is considered a dominant position. The more actor rely on me to connect with others, the more power I have. If two actors have more than one path to form connection, but I am not on these paths, I lose some power (Hanneman & Riddle, 2013). That is, the interaction of any two actors of the network relationship must be linked through a key point (Borgatti, Carley, & Krackhardt, 2006).

In this study, there is no broker between plaintiffs and the defendant and the distance between the nodes is not meaningful. Therefore, this study only analyzes the degree centrality. The higher the degree centrality, the more nodes are connected to the node and associated with more actors, indicating that the node is the most active in the network, and also has a strong influence.

Structure Hole

Most social structures tend to be characterized by strong or weak connections. The structure hole theory relies on a fundamental idea that the homogeneity or redundancy of information, new ideas, and behavior is generally higher within any group of people as compared to that in between two groups of people (Burt, 1993). An individual who acts as a mediator between two or more closely connected groups of people could gain important comparative advantages. Burt (1997) investigated the bankers in large financial organizations found that there is a negative correlation between network restrictions and dividends. Besides, the results also shown that managers who cross the structural hole under the same conditions will receive higher salaries (Burt, Hogarth & Michaud, 2000). Geletkanycz and Hambrick (1997) believe that when senior managers master the relationship between cross business boundaries or subcompanies are more likely to achieve good performance (Geletkanycz & Hambrick, l997).

Also, we think there has the same effect in the patent litigation network. If we compare two nodes (plantiff) has the same out-dergree that sued three defendants, node A is more likely to get complex technology information with different types of patent portfolio than node B and control the litigation network. This is so because nodes connected to B are also highly connected between each other. Other nodes has ability to sue node B and each other overlapped in the same technology fields, so connections involving node B are said to be redundant. On contrary, the position of node A makes it serve as a bridge between three different clusters that firms have different type technology portfolio. As Parchomovskt & Wagner (2005) think that patent portfolio with scale effect and diversify effect could create more value and get competitive advantage through increasing litigation potential and opportunity of cross-license. Thus, node A is likely to control non-redundant technology information from other competitors. Structural holes is used to measure non-redundant contacts and provide network benefits to node A.

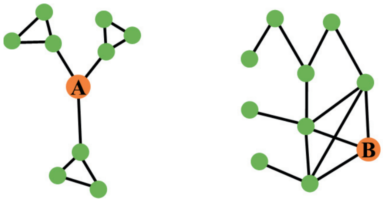


Figure 1. Two types of litigation network structure

Patent Portfolio Strategy

The concept of patent portfolio (Patent Portfolio) was first proposed by Brockhoff in 1991. Parchomovskt & Wagner (2005) think that a patent portfolio is a collection of all relevant patents for a particular technology field that can construct a patent pool to defense, increasing technology licensing opportunities, and increasing barriers to entry. Wagner (2005) proposed the concept of patent portfolio, to assume that a new patent is added to a patent portfolio, and its expected marginal benefit will be greater than the marginal cost of obtaining the patent and to achieve synergistic effect.

Parchomovskt & Wagner (2005) believes that the true value of patents is not the focus on individual patents, but the advantages of scale and technical diversity of patent portfolio of patents. Parchomovskt & Wagner (2005) think that the value created by the patent portfolio includes (1) Economic value: establishing a competitive position through patent infringement agreements and patent cross-licensing by using patent portfolio. (2) Scale effect: Scale effect is also called scale economy. The overall value of the company's patents is also the embodiment of economies of scale in the patent portfolio. (3) Monopoly effect: Monopoly effect refers to that enterprises have a strong patent portfolio strategy in a certain technology field, and form a certain “patent barrier” to protect corporate R&D activities and their freedom operation. (4) Risk minimization: Since the patent portfolio is a collection of multiple interrelated patents, it is possible to avoid costly patent litigation caused by patent infringement and to maximize the benefits under the principle of minimizing risk based on the patent portfolio.

Methodology

Data Retrieval

In this study, we collected litigation patents belong to the IPC of dental medical device introduced in table 1 from WIPS Infringement Search and Patentability Search database (https://www.wipsglobal.com/service/mai/main.wips). This study used the keywords "tooth" or "oral cavity" to search definition of IPC from the TIPO's IPC search system and manually screened the IPC definition. We deleted the 90 IPCs that were not related to dentistry from 159 IPC categories. Finally, the technology defintion of 14 IPC classes as shown in Table 1 is dental device technology.

Table 1. Major sub-class of IPC in dental medical device patent.

IPC	*Definition*
A46B 9/04	Position or arrangement of bristles in relation to surface of the brush body for toothbrushes, e.g. inclined, in rows, in groups
A61B 1/24	For the mouth, i.e. stomatoscopes, e.g. with tongue depressors; Instruments for opening or keeping open the mouth
A61B 1/247	With means for viewing areas outside the direct line of sight, e.g. dentists' mirrors
A61B 1/253	With means for preventing fogging to viewing areas outside the direct line of sight, e.g. dentists' mirrors
A61B 6/14	Apparatus for radiation diagnosis applications or adaptations for dentistry, e.g. combined with radiation therapy equipment
A61B 17/24	For use in the oral cavity, larynx, bronchial passages or nose; Tongue scrapers
A61C	Dentistry; apparatus or methods for oral or dental hygiene
A61G 15/00	Operating chairs; Dental chairs; Accessories specially adapted therefor, e.g. work stands
A61H 13/00	Gum massage
A61J 17/02	Teething rings
A61K 6/00	Preparations for dentistry
A61Q 11/00	Preparations for care of the teeth, of the oral cavity or of dentures, e.g. dentifrices or toothpastes; Mouth rinses
F03B 13/04	Adaptations of machines or engines for use in dentistry; Combinations of machines or engines with driving or driven apparatus; Power stations or aggregates
F21W 131/202	Lighting for dentistry medical use

Note: TIPO's IPC search system, from https://www.tipo.gov.tw/sp.asp?xdurl=mp/lpipcFull.asp&ctNode=7231&mp=1.

From WIPS Infringement database we gathered litigation dental patents by using 14 IPC classes. Finally, from 1995 to 2018 the 592 litigation dental patent, 640 litigation case and 240 plaintiffs and 548 defendants that formed a 550 firms matrix analysed in the UCINET.
Besides, we collected detail information of the litigation patent from US patent database in the Patentability database of WIPS that will used in the patent portfolio analysis.

Systematic Analysis procedure

This study will construct a systematic analysis procedure to recognize the position of companies in a litigation network. In the step 1, we establish a data set for social network analysis. First, we collect patent litigation data in the US dental industry and to form an adjacency matrix based on plaintiff and defendant relation. Then, we compute the value of degree centrality and structure hole of each firm in the litigation network to understand the position of each firms.
In the step 2,3 we use degree centrality and structure hole of each firms in cluster analysis. In this study we adopt the two-stage cluster analysis that we first decide the proper number of clusters based on the Ward method and classify companies into clusters based on K-means clustering according to proper number from the Ward method. And then, we use MANOVA to test the significance of clusters (positions) to validate the cluster solution and named the cluster based on the network variables.
In the step 5, we use four indicators to analysis patent portfolio of each major plaintiff. The four indicators include total patent counts, what technology types the patent portfolio has been used to

sue competitor, litigation case and defendant. The framework of this study is shown in Figure 2. Finally we discuss the conclusions and managerial implications.

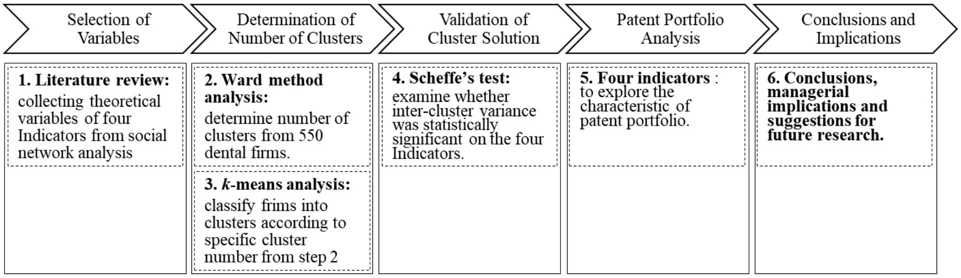

Selection of Variables
Determination of Number of Clusters
Validation of Cluster Solution
Patent Portfolio Analysis
Conclusions and Implications
1. Literature review: collecting theoretical variables of four Indicators from social network analysis
2. Ward method analysis: determine number of clusters from 550 dental firms.
3. k-means analysis: classify frims into clusters according to specific cluster number from step 2
4. Scheffe's test: examine whether inter-cluster variance was statistically significant on the four Indicators.
5. Four indicators : to explore the characteristic of patent portfolio.
6. Conclusions, managerial implications and suggestions for future research.

Figure 2. Research model and systematic analysis procedure.

- **Network Indicators**

Visualizing the relationship among companies in the patent infringement lawsuits network and generating network indicators. In this study we used two indexes to evaluate patent litigation network - degree centrality and structure hole. Degree centrality means that how many competitors a firm has sued or been sued. A firms with high degree centrality no matter what that firm have more ties to reach and it is the key player in the dental industry. The firms with high structure hole occupy the network structural positions that act as a source for larger volumes of information exchange and other transactions involving other resources.

➢ **Degree Centrality (in-degree centrality and out-degree centrality)**

Degree centrality can distinguish between the in-degrees and out-degrees of nodes in directed networks by measuring in-degree and out-degree centrality, respectively (Knoke and Burt 1983). In the directional patent networks, a tie has a start node and an end node. Degree centrality is used to measure the size of the actor's control range in the network. The in-degree centrality ($C_{D,in}$(ni)) and out-degree centrality ($C_{D,out}$(ni)) of a given node are formally defined as

$$C_{\mathrm{D},in}(n_i) = \sum_{j=1}^{k} r_{ij,in} \ ; C_{\mathrm{D},out}(n_i) = \sum_{j=1}^{k} r_{ij,out}$$

where $r_{ij,in}$ and $r_{ij,out}$ denote sued and been sued connections of node i, respectively, and node k indicates the network size. The in-degree of a node is defendant. The out-degree of a node is plaintill. Comparing in-degree and out-degree measures of a given node can reveal what positional relationship among different types of firms with what kind of patent portfolio in the patent litigation network.

➢ **Structural holes (Effective size, Constrain)**

Burt (1993) studied the structure of interpersonal networks, and analyzed what kind of network structure can bring to the network of actors more benefits or rewards. The so-called "structural hole" is a gap between non-redundant contacts. Structural holes were used to represent the competitive advantage for a node in which linkages spanning different. The concepts of redundancy is generally measured by the effective size of egocentric network of each node.

Redundancy can be calculated as average degree of ego's alters that not count the number of ties connected to ego node. So the effective size of an ego network is just the whole network scale minus the redundancy.
Calculate the degree of constraint of a node that is connected to another node:

$$C_{ij} = \left(P_{ij} + \sum_{q} P_{iq} P_{qj} \right)^{2}$$

Equation for the Burt constraint value of node i:

$$C_i = \sum_{j} C_{ij}$$

The definition of effective size is the ratio of the effective size divided by ni indicates efficiency on a scale from zero (highly redundant links and low efficiency) to one (no redundant links in the network).

Result

The Interpretation of Cluster in the US Dental Medical Device Industry.

We then use the Ward method to identify the number of clusters, which shows that when the number of groups is reduced from six to five, the agglomeration coefficient suddenly rises sharply (from 341.92 to 385.87). The agglomeration coefficient represents the square Euclidean distance between the two clusters joined. As such, small coefficients indicate that fairly homogeneous clusters are joined, whereas larger values indicate that dissimilar clusters are joined. Therefore, the most suitable number of groups is defined as six.

After deciding on the number of groups, we used k-means cluster analysis to divide the 550 firms into six clusters. Scheffe's multivariate comparison was used to identify the differences among groups, which shows that each of the six clustering variables is significant at $p<0.05$ respectively. Table 2 records the average scores and the results of the Scheffe's multivariate comparison of the four network indicators in the six groups. The interpretation of clusters is described below.

- **Cluster 1: Key player 1**

Statistically, cluster 1 (n = 1) reported the strongest emphasis on the entire network indicators among the six clusters. Key player 1 (Ivoclar Vivadent) has sued 29 defendant with out-degree is 29 and in-degree is zero. No companies sued Ivoclar Vivadent.

- **Cluster 2: Key player 2**

Cluster 2 (n=1) has second highest scores for network indicators. Key player 2 (Dentsply International) also high out-degree (the value is 22), but the firms has high in-degree (the value is 23). We found that three patent troll all sued Dentsply International

- **Cluster 3: Patent troll**

In cluster 3 (n=3), the score of network indicators is similar to cluster 4. But, in the cluster 3 the 3 firms are patent troll. The Patent Troll referred to a patent management company that acquired patents from others but did not engage in research and development, regarded the patents it owned as a weapon to initiate litigation, and used compensation or reconciliation funds as the main source of revenue. So we found that the in-degree of the 3 firms is only 0.667.

- **Cluster 4: Active attacker**

Firms in cluster 4 (n=4) display the similar scores in most variables when compared to the third clusters. The four firms are dental companies has higher out-degree (the value is 10.5) and in-degree is 1.5 mean that some firms have sued the four companies.

- **Cluster 5: Bystander**

In cluster 5 (n = 23), the 23 have average 1.29 patent litigation cases (the out-degree is 1.29) but have been sued by 6.2 litigation cases (the in-degree is 6.29). The 23 firms have less experience or ability to initial a patent lawsuit but the defendants that key player and patent troll have sued. We could say that firms in this cluster have marginal patent litigation ability. Therefore, cluster 5 is named the "Bystander" pattern.

- **Cluster 6 : Victim**

In cluster 6 (n = 517), the means of the four network indicators are all lower than the other clusters. Besides, the out-degree of 517 firms has lower than one and the in-degree is also lower than one. Therefore, cluster 6 is named the "victim" pattern.

Table 2. The cluster naming and result of MANOVA.

Position of Network	*N*	*EffiC*	*CON*	*OutD*	*InD*	*Company Name*
Key player 1	1	1.000	.070	29.000	.000	Ivoclar Vivadent
Key player 2	1	.991	.055	22.000	23.000	Dentsply International
Patent troll	3	.984	.115	12.667	0.667	511 Innovations, DE Partners Golden Rule, Randall s Asher
Active Attacker	4	.972	.156	10.500	1.500	3M Innovative Properties, 3M Company, CAO Group, Ultradent Prod
Bystander	23	.991	.293	1.290	6.290	American dental, Anatomage, Blue Sky Bio, Dentsply IH, Dentsply Sirona, Discus Dental, Hu-Friedy Mfg, Kerrrporation, Mylan Pharmaceuticals, Nobel Biocare Service, Nobel Biocare USA, OnDem and 3D Technology, Ormcorp, Procter & Gamblempany, Ranir, Shock Doctor, Sirona Dental Systems, TP Orthodontics, Technique D`Usinage Sinlab, The Ohio Willow Woodmpany, The Procter & Gamblempany, US Dental Depot, Zimmer Holdings
Victim	517	.996	.835	.890	.910	--

Visulization the Structure and Position of Firms in the Litigation Network.

We visulizated he patent-infringement lawsuits graph for 550 dental companies involving in the 640 patent litigation cases from 1995 to 2018. A node represents a company and an edge from node X to node Y represents the lawsuit relationships from X to Y.

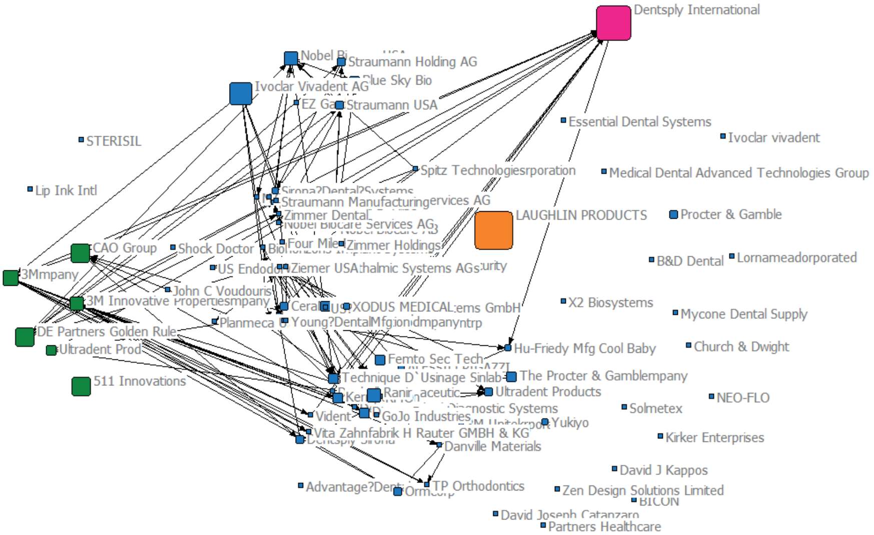


Figure 3. Position of companies in litigation related network in US dental industry.

The Patent Portfolio of major firms in the litigation network

We used four indicators (patent counts, number of patent portfolio, number of litigation cases and number of defendants) to analyze the patent portfolio of 8 major firms the litigation network. The characteristic of the patent portfolio of major firms in the litigation network shown in table 3.

Table 3. The characteristic of the patent portfolio of major firms in the litigation network.

No.	*Firms*	*No of Patent*	*No of patent portfolio*	*No of Litigation Case*	*No of Defendant*
1	Ivoclar Vivadent AG	13	5	8	27
2	Dentsply Internation	17	7	24	26
3	3M Innovative Properties	13	11	12	5
4	511 Innovations	8	1	6	23
5	DE Partners Golden Rule	1	1	12	18
6	Randall s. Asher	1	1	8	9
7	CAO	29	3	18	27
8	Ultradent Production	19	3	11	14

Conclusion and Suggestion

This study construct a systematic approach to explore how the advantages of the network position formed by the litigation of the dental firms with what patent portfolio they adopted in the entire litigation network based on the Actor-network Theory (ANT).

The cluster analysis show 6 cluster: Key player 1, Key player 2, Patent troll, Active Attacker, Bystander and Victim, to explore the position of each firm in the litigation network. Then we demonstrate the Patent Portfolio of major firms in the litigation network. The result show that key player1 and 2 use multiple patent portfolio to get competitive advantage. 3 Patent troll firms use only one patent or one patent portfolio to sue defendants. 23 Active Attacker firms has less patent portfolio than key players. The results will provide a broader theoretical perspective for the governance of network relations between organizations through litigation, and also provide practical advice for managers to understand the development of litigation networks and patent portfolio.

The IP war between technology-driven firms is likely to become more and more fierce. In the future business environment, firms should pay more attention in the management of intellectual property rights and monitor the patent litigation. Firms should upgrade their previous defensive orientation in the patent portfolio development into strategical litigation and attack competitors by using patent portfolio. Besides, it is necessary to re-examine the planning of litigation strategies in order to avoid becoming a defendant sued by patent troll.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 108L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. MOST 107-3017-F-002-004 and MOST 107-2637-H-020 -001.

References

Bizzi, L., & Langley, A. (2012). Studying processes in and around networks. *Industrial Marketing Management*, 4 (2): 224-234.

Borgatti, S. P., & Halgin, D. S. (2011). On network theory. *Organization Science*, 22 (5): 1168-1181.

Burt, R. S. (1992). Structural holes. Cambridge, MA: Harvard University Press.

Burt, R. S. (2004). "Structural holes and good ideas". *American Journal of Sociology,* (110): 349–399.

Buskens, V. & van de Rijt, A. (2008). Dynamics of Networks if Everyone Strives for Structural Holes. *The American Journal of Sociology*. 114 (2): 371–407.

Butt, R. S. (1997). The contingent value of social capital. *Administrative Science Quarterly*, 42, 339-365.

Butt, R. S., Hogarth, R. M., & Michaud, C. (2000). The social capital of French and American managers. *Organization Science*, 11, 123-147.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification, *Social Networks*, 1(3), 215-239.

Geletkanycz, M. A., & Hambrick, D. C. (1997). The external ties of top executives: implications for strategic choice and performance. *Administrative Science Quarterly*, 42, 654-681.

Granovetter, M. S. (1973). The Strength of Weak Ties. The American Journal of *Sociology,* 78 (6): 1360–1380.

Jones, C., Hesterly, W. S., & Borgatti S. P. (1997). A general theory of network governance: Exchange conditions and social mechanisms. *Academy of Management Review*, 22 (4): 911-945.

Parchomovsky, G., & Wagner, R. P. (2005). Patent Portfolios. *University of Pennsylvania Law Review*, 154, 1-77.

Wasserman, S. & Faust, K. (1994). Social Network Analysis: Methods and Applications. UK: Cambridge University Press.

Editorial practices and systematic conscious bias on Wikipedia: An initial test with articles on Traditional Chinese Medicine

Dangzhi Zhao[1] and Andreas Strotmann[2]

[1] *dzhao@ualberta.ca*
School of Library and Information Studies, University of Alberta, Edmonton (Canada)

[2] *andreas.strotmann@gmail.com*
ScienceXplore, F.-G.-Keller-Str. 10, D-01814 Bad Schandau (Germany)

Abstract

Information has a clear effect on the choices people make. As a vital source of information on the Web, critical inaccuracies or biases on Wikipedia may have major negative impact even when accidental, but even more so in the form of systematic bias through organized manipulation. We propose to study the latter in a case study of articles on Traditional Chinese Medicine (TCM), an area that has been identified as affected in the English language Wikipedia. We will classify Wikipedia contributors to TCM articles by their editing, referencing, and collaborating behaviours, applying two bibliometric methods: co-authorship analysis and author bibliographic coupling analysis. We will identify editor groups that appear as dense clusters in all three classifications, and test if they show characteristics previously identified as indicative of persistently biased editing. We will verify if these tightly knit groups of editors have worked together to systematically bias articles on TCM.

Introduction

Information clearly affects the choices people make, such as shifting their voting patterns (DellaVigna and Kaplan, 2007) or affecting their medical outcomes (Yom-Tov and Boyd, 2014). As a vital source of information on the Web, any critical inaccuracy (e.g., in medical content) or bias (e.g., on sensitive historical or political events) on Wikipedia (WP) may have major negative impact on, and potentially be detrimental to, society even when accidental, but even more so in the form of systematic bias through organized manipulation (Shiri, 2017). It is critical to investigate systematic bias on WP to see what kinds of editorial practices contribute to such bias, and if such bias and its "priming effects on audiences" fall into persistent patterns relevant to political or other agendas. "Powerful players devote massive resources to advancing their interests precisely by imposing such patterns on mediated communications" (Entman, 2007, p. 164), and Wikipedia may be no exception.

Such research will help to fill a substantial knowledge gap. It will add new empirical knowledge to debates about social practices of peer production communities, and to conversations about bias and its implications for democracy, particularly in the current context of political manipulation of public opinions on social media and "echo chamber" (or "filter bubble") effects of personalization of online services such as Google and Facebook (Introna & Nissenbaum, 2000; Shiri, 2017). It will shed light on ways to identify WP articles that are affected by systematic conscious bias and on ways to alleviate impact of such bias on the general public, thus helping enhance both traditional (e.g., libraries) and online information services (e.g., Google) in ways that contribute to democracy. After all, exposure to diverse opinions can improve civic discourse and is "one of deliberative democracy's basic tenets" (Sunstein, 2009; Yom-Tov and Boyd, 2014).

This research in progress paper is part of a research program that aims to study systematic conscious bias on WP. It attempts to map the editorial practices of editors who have contributed to articles on Traditional Chinese Medicine (TCM) on WP, and to identify editor groups who have likely committed systematic biased editing. This study could provide evidence to generate empirically-grounded hypotheses concerning various editorial practices that are more likely to be indicative of systematic conscious bias. Two bibliometric methods will be applied: co-authorship analysis and author bibliographic coupling analysis.

Background and related articles

Success and problems of Wikipedia

WP is a system unique in the history of civilization (Simonite, 2013). Both benefits and challenges of the WP system have been widely debated in academia, law, business, and other sectors of society.

WP started in 2001 with the lofty goal of compiling all human knowledge. It grew quickly into the largest encyclopaedia in the world with over 6.8 million registered and uncounted unregistered volunteer editors contributing to 2.3 million articles in the English version alone as of April 2008 (Burke & Kraut, 2008). Many WP articles were of a quality comparable with corresponding ones in Encyclopaedia Britannica as a result (Giles, 2005). WP's success also made it play a symbolic role in highlighting the potential for voluntary peer-production to generate valuable collections of information. Hansen, et al., (2009) even contend that WP "approximates features of the ideal speech situation articulated by Habermas" (Habermas, 1984). With now 5.4 million articles (WP:Statistics), WP has become a one-stop shop for information on most topics. As WP articles often intentionally show up at the top of Google search result lists and are now promoted as fact check sources on Facebook, to name just two prominent examples, WP is clearly a primary information source that people see, and in many cases even the only source that people use.

However, WP's success was to many a surprise. Among the fundamental problems of the WP system that have been criticized are unpredictable motivations of editors and an emphasis on consensus rather than authority (Denning et al., 2005). WP evolves without supervision by certified subject experts or authorities, and its largely anonymous volunteer editors are left to define, interpret and implement its policies and resolve conflicts on their own. WP editors "may be altruists, political or commercial opportunists, practical jokers, or even vandals" (Denning et al., 2005, p. 152). Bias can be introduced and maintained in an article as long as a group of editors with that bias manage to dominate the discussion and force it to a "consensus." Mechanisms used in traditional systems to ensure quality and avoid abuse of power are normally based on true identity along with social expectations, norms, and status positions, and thus cannot work for WP (Arazy et al., 2011).

WP has developed detailed policies and guidelines for contributing and for resolving conflicts, including its two fundamental principles, Neutral Point of View (NPOV) and Verifiability, which intend to ensure all important viewpoints are represented fairly and supported by trustworthy published resources. However, "ironically, Wikipedia rules are often used less to resolve disputes than as tools in waging editing struggles" (Tkacz, 2015, p. 99). A "crushing bureaucracy with an often abrasive atmosphere" formed over the years (Simonite, 2013) where editors would often rely on citing WP policies or threatening to take a matter to ARBCOM (the arbitration committee for dispute resolution) rather than on having a more substantive discussion in resolving conflicts around controversial topics (Koppelman, 2017; Martin, 2017). As a result, "conflicts among editors rarely conclude on the basis of merit but are typically ended by sheer exhaustion, the evident numerical dominance of one group, or admin intervention" (Yasseri, et al., 2012). WP in fact "represents the viewpoint of its most strident and persistent editors" (Hube, 2017, p. 717), some claim.

Admins are considered trusted custodians of WP, and have power to decide if a topic is worthy to cover, to lock articles for targeted editors, and to ban editors (Burke & Kraut, 2008). However, the process of promoting an editor to admin has been found susceptible to manipulation, and, once an admin, they are mostly free to exercise the power at will as there are few measures to hold them accountable (Picot-Clémente, et al., 2015).

Anecdotal evidence of manipulation and systematic conscious bias on WP has been reported in the literature. Organizations with strong beliefs such as "Guerilla Skeptics on Wikipedia" have been accused of persistently maintaining WP articles of their interests to promote their own viewpoints (Koppelman, 2017; McLuhan, 2013). Research has found that some editors have likely sought admin power in order to change and control discussions regarding controversial topics of their interest (Das et al., 2016). There is evidence for "covert editing efforts to shape the portrayal of individuals, organizations, and topics" (Craver, 2015; Thompson, 2016), and for "systematically biased editing, persistently maintained," to impose certain viewpoints on controversial topics such as vaccination (Lih, 2009, pp. 122-131).

Studies on bias and controversy on Wikipedia

Studies from different perspectives both about WP and using WP as data are abundant (Yasseri, et al., 2012). These studies were possible partly because nearly all edits and discussion posts are saved and available on WP. Directly related to the present research are studies on bias and controversy.

Bias has been extensively studied in the context of media as biased news can influence power distribution and hurt media's contribution to democracy (Entman, 2007). Based on the research concept "bias" defined there, studies on bias on WP, including political bias (Greenstein & Zhu, 2012), cultural bias (Callahan & Herring, 2011), and gender bias (Graells-Garrido, et al., 2015), have mostly focused on "content bias" – "favoring one side rather than providing equivalent treatment to both sides" in a conflict.

Bias on WP is often related to controversial topics. Studies on controversy on WP have largely focused on automatically detecting them, e.g., by identifying "edit wars" in the form of mutual reverts (e.g., Hube and Fetahu, 2018; Yasseri, et al., 2012). Most studies targeted articles; only a few targeted editors, as does the present study. Liu & Ram (2011) studied the impact of editor collaboration patterns on article quality. They found that clustering editors by features extracted from their editing activities is a highly promising method for characterizing editing behaviours and collaboration patterns. For example, they found Watchdogs and Cleaners, i.e., editors focusing on reverts and removing materials respectively. Das et al. (2016) proposed and validated two controversy scores in an attempt to identify "manipulating editors," i.e., editors who changed their editing behavior to focus on controversial articles in a single topic area after gaining admin status. They measured topic similarity between two WP articles by a combination of the degree to which they share references, linked terms, and WP topic categories. We use and explain some of these methods in the Methodology section.

Methodology

Overview and assumptions

We will first identify and download articles on TCM from WP. We will then examine the editorial practices of editors who have contributed to these articles. We will finally identify editor groups who have likely committed systematic biased editing.

Drawing on previous studies, we assume that editors who have likely committed systematic biased editing are editors who feature the following editorial practices.

(1) Peculiar editing behaviors. To bias an article, editors are expected to exhibit different styles of contributing to WP compared to normal editors. They might constantly monitor and act quickly to "protect" articles of their interests whereas others may contribute irregularly based on their work/life schedule (Sundin, 2011). They would rely heavily on citing WP rules and banning resisting editors through admins to avoid consensus building processes whereas

normal editors may be more interested in contributing ideas and fixing problems than learning about the nuances of WP rules (Koppelman, 2017). They would often revert edits (Martin, 2017) while normal editors tend to simply add new materials and viewpoints. They tend to spend most of their time on a small number of contested topics (Das, et al., 2016).
(2) Lasting group support. Conflicts among editors are typically ended by "sheer exhaustion, the evident numerical dominance of one group, or admin intervention" (Yasseri, et al., 2012). To bias a controversial article in the face of resistance, a number of editors sufficient to dominate the discussion must gain support from each other, and/or from admin(s) who can overrule resisting editors (Martin, 2017). Such a support structure will likely be in place in many articles to dominate an entire topic area.
(3) One-sided selection of information sources. This is a technique used to introduce bias into an article (Martin, 2017). To bias articles systematically and to meet WP's verifiability requirement at the same time, organizations or groups of editors would likely develop a core set of information sources that support their views and use them in all articles of their interest.

Data collection and analysis

To this end, we will adapt and follow the well established procedures and techniques for co-authorship analysis and author bibliographic coupling analysis. Instead of a citation database, we will use the English version of WP as data source, and will develop computer programs to collect and analyze data from WP. Instead of multivariant analysis methods such as cluster analysis or factor analysis, we will use community detection techniques to accommodate the large number of editors to be examined. We will use InfoMap, an open-source community detection algorithm recommended for identifying meaningful communities with stronger internal than external connections in networks of up to 6000 nodes (Yang, et al., 2016).
(1) Download dataset and select editors to examine
We will identify articles on TCM by starting with the articles on TCM proper or in the WP categories of TCM therapies or TCM medical plants, and then adding articles interlinked heavily with these seed articles. We will download all these articles from WP, including the entire editing and discussion history of each article. We choose the topic area TCM for this initial test because it has been reported that this area has been affected by systematic biased editing as one of the public targets of the "Guerilla Skeptics on Wikipedia" (Koppelman, 2017; McLuhan, 2013). Since biased editors are likely frequent contributors, we will exclude editors who have not contributed frequently to these articles.
(2) Characterize editing behaviour
For this, we will examine editing features on both the articles themselves and on their associated Talk pages. The former will include the features used in Liu & Ram (2011), i.e., insertion, modification, or deletion of text, link, or reference. The latter will include number of reverts, WP rule citation (R), taking (or threatening to take) a matter to ARBCOM (A), or suggesting banning of resisting editors (B). We will also include the two controversy scores introduced and tested in Das et al. (2016): a Controversy (or C-) score indicates the degree to which an editor focuses on controversial articles, and a Clustered Controversy (or CC-) score captures the degree to which an editor focuses on a particular controversial topic (in this case, TCM). We will build an editing profile for each editor, which includes these features, and, by analyzing these profiles using InfoMap, will identify communities of editors that are densely connected through common editing behaviour patterns. We will compare editing behaviours of those communities that are characterized by high R, A, B, C, and CC scores with other communities. We will determine if the articles that these featured communities have contributed to heavily belong to the same topic areas, and, if they do, will also determine if bias exists in samples of these articles through close reading.
(3) Characterize editors' co-involvement patterns

Following co-authorship analysis techniques, for each pair of editors selected above, we will compute the number of articles (N) they are both involved in, either on the articles themselves or on their associated Talk pages. We will also weigh this number by the level of their involvement in each article. For example, if editors A and B have made n and m substantial contributions respectively to an article, this article will increase their weighted co-involvement score by min(n, m), i.e., the smaller one of the two values. All the N articles will be counted this way to arrive at the total weighted co-involvement score for A and B. As a result, two K x K matrices of co-involvement scores will be produced, with K being the number of editors selected to be examined: simple and weighted. We will identify communities of editors that are densely connected through their co-involvement patterns by analyzing these matrices using InfoMap. We will then compare co-involvement patterns of those communities that overlap heavily with communities of editors identified above that are characterized by high R, A, B, C, and CC scores with other communities.

(4) Characterize editors' referencing behaviour

Following author bibliographic coupling analysis techniques, a matrix of shared reference scores will be produced for all selected editors. If editors A and B have contributed substantially to article sets S1 and S2 respectively, and n different information sources were cited in both S1 and S2, n would be the shared reference score for A and B. We will identify communities of editors that are densely connected through shared references by analyzing this matrix using InfoMap. We will then compare referencing features and preferences of those densely connected communities that overlap heavily with communities of editors identified above that are characterized by high R, A, B, C, and CC scores with other communities.

Expected major results

We expect to see heavy overlaps between communities of editors that are characterized by high R, A, B, C, and CC scores, and those that are densely connected through co-involvement and shared reference patterns. Bias is expected to exist in articles that these featured communities have contributed to heavily, and clear topical themes should emerge from these articles, indicating that these topic areas have been biased systematically.

References

Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Community-Based Collaboration in Wikipedia: The Effects of Group Composition and Task Conflict on Information Quality. *Journal of Management Information Systems,* 21(4), 71–98.

Burke, M., & Kraut, R. (2008). Mopping up: modeling Wikipedia promotion decisions. *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (pp. 27-36).

Callahan, ES., & Herring, SC. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American society for Information Science and Technology*, 62(10), 1899-1915.

Cabunducan, G., Castillo, R., & Lee, J. (2011). Voting behavior analysis in the election of wikipedia admins. Proceedings of the *2011 International Conference on Advances in Social Networks Analysis and Mining* (pp. 545 –547).

Craver, J. (2015). PR firm covertly edits the Wikipedia entries of its celebrity clients. *Wiki Strategies*. Retrieved Dec. 18, 2018 from: http://wikistrategies.net/sunshine-sachs/

Denning, P., Horning, J., Parnas, D. and Weinstein, L. (2005). Wikipedia risks. *Communications of the ACM*, 48(12): 152.

Das, S., Lavoie, A., & Magdon-Ismail, M. (2016). Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *ACM Transactions on the Web*, 10(4).

DellaVigna, S., & Kaplan, E. (2007). The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Entman, R.M. (2007). Framing Bias: Media in the Distribution of Power. *Journal of Communication*, 57, 163-173.

Giles, G. (2005). Internet Encyclopedias Go Head to Head. *Nature*, 438(7070), 900-901.

Graells-Garrido, E., Lalmas, M., & Menczer, F. (2015). First women, second sex: gender bias in Wikipedia. *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 165-174).

Greenstein, S., & Zhu, F. (2018). Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia. *The MIS Quarterly*. 42(3), 945-959.

Hansen, S., Berente, N., & Lyytinen, K. (2009). Wikipedia, Critical Social Theory, and the Possibility of Rational Discourse. *The Information Society – An International Journal*, 25(1), 38-59.

Habermas, J. (1984). *The theory of communicative action: Reason and the rationalization of society*. Boston: Beacon Press.

Hube, C. (2017). Bias in Wikipedia. *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 717-721).

Hube, C., & Fetahu, B. (2018). Detecting Biased Statements in Wikipedia. *Proceedings of the The Web Conference 2018* (pp. 1779-1786).

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, 16(3):169– 185.

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He says, she says: Conflict and coordination in Wikipedia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07* (pp. 453–462).

Koppelman, M.H. (2017). WikiTweaks: The Encyclopaedia that Anyone (Who is a Skeptic) Can Edit. *Journal of Chinese Medicine* . Feb2017, Issue 113, p35-40.

Leskovec, J., Huttenlocher, D. & Kleinberg, J. (2010). Governance in Social Media: A case study of the Wikipedia promotion process. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Retrieved Dec. 18, 2018 from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1485/1841/.

Liu, J., & Ram, S. (2011). Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality. *ACM Transactions on Management Information Systems*, 2(2), [11].

Lih, A. (2009). *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia*. London, England: Aurum.

Martin, B. (2017). Persistent Bias on Wikipedia: Methods and Responses. *Social Science Computer Review*, 36(3), 379-388.

McLuhan, R. (2013). *Guerrilla Skeptics*. Retrieved Dec. 18, 2018 from: https://monkeywah.typepad.com/paranormalia/2013/03/guerrilla-skeptics.html.

Pariser, E. (2011). *The Filter Bubble: What the Internet is hiding from you*. Penguin Press HC.

Picot-Clémente, R., Bothorel, C., & Jullien, N. (2015). Social Interactions vs Revisions, What is important for Promotion in Wikipedia? *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 888-893).

Shiri, DH. (2017). *Controversy Analysis and Detection*. Doctoral Dissertations. Retrieved November 12, 2018 from: https://scholarworks.umass.edu/dissertations_2/1084.

Simonite, T. (2013). The Decline of Wikipedia. *MIT Technology Review*. Retrieved Nov. 12, 2018 from: https://www.technologyreview.com/s/520446/the-decline-of-wikipedia/.

Suh, B., Convertino, G., Chi, E. H., & Pirolli, P. (2009). The singularity is not near: slowing growth of Wikipedia. *Proceedings of WikiSym 2009*.

Sundin, O. (2011). Janitors of knowledge: constructing knowledge in the everyday life of Wikipedia editors. *Journal of Documentation*, 67(5), 840-862.

Sunstein, C. R. (2009). *Republic.com 2.0*. Princeton University Press.

Thompson, G. (2016). Public relations interactions with Wikipedia. *Journal of Communication Management*, 20, 4–20.

Tkacz, N. (2015). *Wikipedia and the politics of openness*. Chicago, IL: University of Chicago Press.

Yang, Z., Algesheimer, R., & Tessone, C.J. (2016). A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Scientific Reports*, volume 6, Article number: 30750.

Yasseri, T. Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of Conflicts in Wikipedia. *PLOS ONE*, 7(6): e38869. https://doi.org/10.1371/journal.pone.0038869.

Yom-Tov, E., & Boyd, D.M. (2014). On the link between media coverage of anorexia and pro-anorexic practices on the web. *International Journal of Eating Disorders*, 47(2), 196–202.

Bias in Academic Recruitment: the Italian National Scientific Habilitation

Marco De Santis Puzzonia, Irene Mazzotta and Sandro Momigliano

marco.desantispuzzonia@anvur.it; irene.mazzotta@anvur.it; sandro.momigliano@anvur.it
ANVUR, Via Ippolito Nievo 35, 00153 (Italy)

Abstract

Since 2012, access to university calls for associate and full professorship in Italy is restricted to candidates who passed a national fit-for-the-role evaluation, the National Scientific Habilitation (NSH). NSH Committees evaluate candidates considering the importance, quality and originality of their publications, granted awards, received funds and membership of scientific boards. We analyse the relationship between the decisions taken by the Committees and the quality of the scientific production of candidates, by using a polynomial probit model. We also test the impact of two specific factors, gender and affiliation, on the probability of success. Relatively small effects have been detected with regard to the candidates' gender; furthermore, in Humanities and Social Sciences women are generally favoured. As for the affiliation factor, an outsider status is associated to drastically lower probabilities of success. This result may however reflect a number of factors not considered in this analysis, whose results are thus only a first step in the direction of understanding whether it exists or not an "affiliation" bias in the NSH.

Introduction

Since 2012, the Italian university recruitment of associate and full professors follows two steps: a national fit-for-the-role evaluation, the National Scientific Habilitation (NSH), and local (university) level contests, restricted to "habilitated" candidates. Habilitation is awarded by NSH Committees, one for each of 190 disciplinary fields ("Settori Scientifici", SC) identified within 14 areas. NSH Committees evaluate only eligible candidates, i.e. candidates who satisfy pre-determined requirements concerning their publication record.[1] More specifically, candidates must meet or exceed at least 2 out of 3 SC-specific thresholds. In STEM areas[2] the thresholds refer, for given time intervals, to 3 indicators: the number of articles, citations received and the h-index value. In the Humanities and Social Science areas[3], they refer to the number of articles and book chapters, articles in A-ranked journals, and books (for a more thorough description of the procedure, see Bagues et al., 2017).

At the start of the NSH "rounds", in 2012, 2016 and 2018, based on specifications by the Ministry of Education (MIUR), the Agency for the Evaluation of the University and Research System (ANVUR) calculated the thresholds for each SC, separately for candidates to Associate and Full Professor positions.[4] In 2016, these thresholds meant that, in each SC, about 2/3 of the associate professors (AP) and researchers were eligible candidates for a full professor (FP) and an AP position, respectively. Following the selection carried out by ANVUR, in 2016-18 the National Committees evaluated the eligible candidates on the basis

[1] In the first NSH round (2012-14), Committees could choose to admit to evaluation individual candidates which did not satisfy the minimal requirements, but had to explicitly motivate the decision.

[2] They include the following areas: Mathematics and Informatics; Physics; Chemistry; Earth Sciences; Biology; Medicine; Agricultural and Veterinary Sciences; Civil Engineering and Architecture; Industrial Engineering and Information Systems.

[3] They include the following areas: Classical studies, Philology, Arts and Literature; History, Philosophy and Psychology; Law; Economics and Statistics, Political and Social Sciences.

[4] The thresholds were published in the Ministerial Decree n. 602/2016 (http://attiministeriali.miur.it/anno-2016/luglio/dm-29072016.aspx).

of the following elements: importance, quality and originality of their publications according to the disciplinary field, granted awards, received funds and membership of scientific boards[5]. The aim of this paper is to investigate: a) the relationship between the decisions taken by the NSH National Committees and the quality of the scientific production of candidates; b) whether this relationship changes depending on gender and on candidates being inside or outside the Italian academic community. Underrepresentation of women in medium and high ranks of academic careers is internationally documented, and Italy is not an exception (Fig.1)[6]; stimulated by this evidence, a large literature on gender bias in academic promotions has developed (Ceci et al., 2014; Morley, 2014; European Commission, 2015; OECD, 2018). The Italian case is particularly interesting because the NSH introduced in 2012 was meant to correct allegedly widespread favouritism (likely to include a gender bias) in the previous recruitment and promotion system. While there is relatively clear evidence of gender bias in local contests (Marini and Meschitti, 2018; De Paola and Scoppa, 2015), the existence of discrimination against women in the NSH process is more controversial. In particular, De Paola, Ponzo and Scoppa (2015) do not find evidence of discrimination, while Bagues et al. (2017) document the existence of a significant gender bias. Evidence concerning possible biases in recruitment/promotion against outsiders of the academic community is relatively scanty. The main difficulty here is the fact that candidates are typically evaluated, not only on the basis of their scientific production, but also on a number of other features, some of which are difficult to gauge (Pezzoni et al., 2012; Xuhong, 2014). Typically, insiders tend to be stronger in these additional elements included in the evaluation; for example, though it is not the case of Italy, teaching experience is explicitly included. As in the following preliminary analysis we are not able to control for these additional factors, the results have to be considered only as a first step in the direction of understanding whether there exists an "affiliation" bias in NSH. Evidently, a better knowledge on the issues of gender and affiliation bias would represent an important input in the discussion for further reforms of the regulation of the Italian recruitment system.

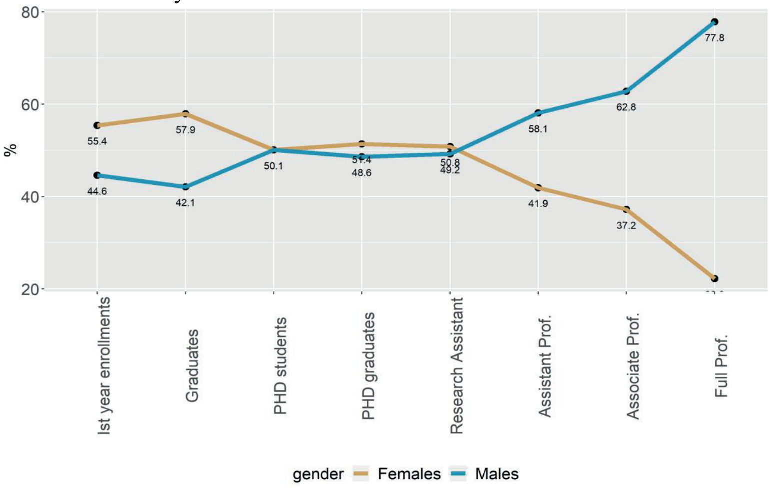


Figure 1. Percentage of females and males in the Italian academic rank in 2016
Source: ANVUR (2018); Chapter I.3.4; page 273. Translation by the authors.

[5] A detailed specification of the elements to be considered in the evaluation carried out by the National Committees is contained in the Ministerial Decree n. 76/2012.

[6] Chapter II.11.4 in ANVUR (2018) provides some analyses on gender-related issues in the Italian University system.

Data

Our initial dataset includes 28,942 eligible applications for Associate and Full professor assessed by the NSH Commissions in the period 2016-18 (2016 round).

To analyze affiliation, we consider as "insiders" Full Professors (FP), Associate Professors (AP) and Assistant Professors ("Ricercatori": formerly permanent staff, since 2010 new recruits are under fixed-term contracts) in Italian Universities. Everybody else is classified as "outsider", including professors working abroad and researchers working in Italian Universities with other positions (e.g. PhD fellows, post-doctoral fellows and fixed-term research assistants).

As a proxy of the quality of the scientific production of candidates, we use the h-index value, for STEM areas, and the number of publications in A-ranked journals, for Humanities and Social Sciences[7]. To standardize the indicators across SCs, we divide each value by its SC-specific threshold used for the initial screening of candidates (see above). The distribution of candidates across values of the quality indicator is broadly similar for males and females and for insiders and outsiders (Figures 2 and 3).

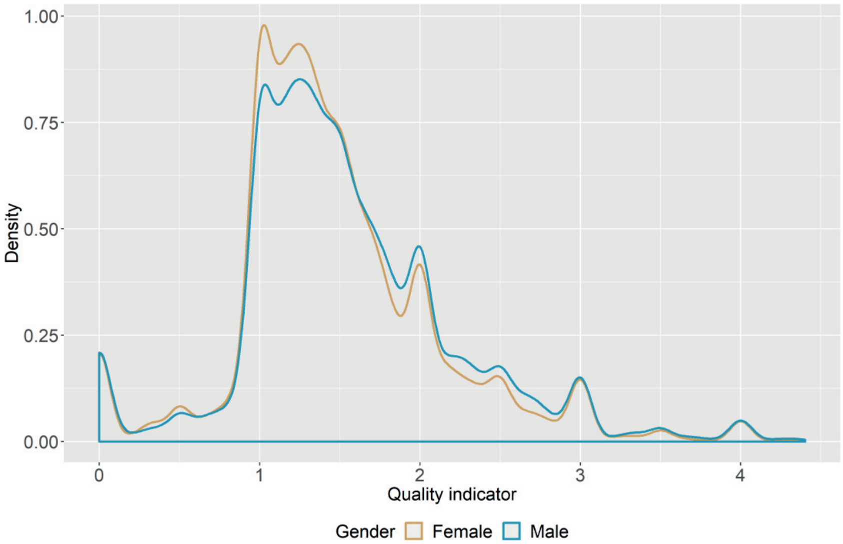


Figure 2. Density of the quality indicator distribution by gender

[7] These indicators were calculated, for each AP candidate, over 10 years and FP candidate over 15 years period.

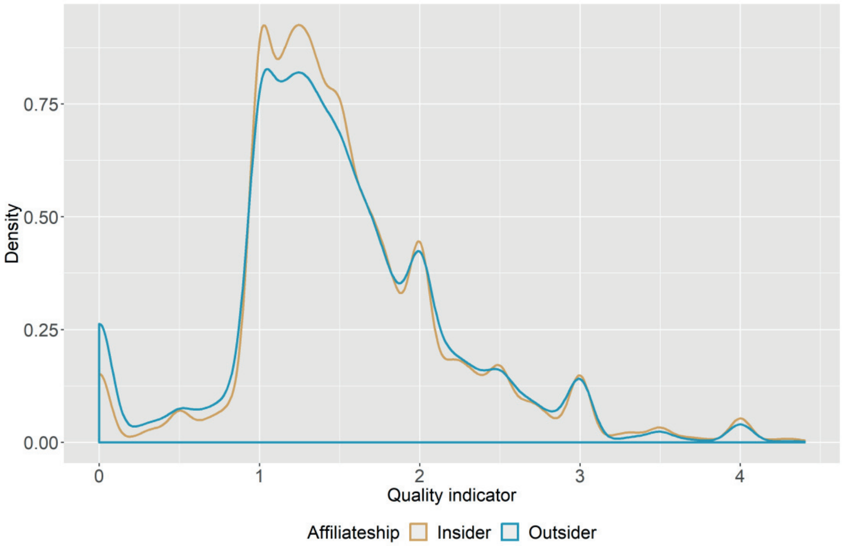


Figure 3. Density of the quality indicator distribution by affiliation

We exclude candidates with extremely high values for the indicator of quality. In particular, in STEM areas we exclude candidates with values higher than 3 times the admission threshold (about 4% of candidates). As this would imply excluding too large a number of candidates in HSS areas, there we exclude those with values above 4.4 times the admission threshold (10% of candidates). Overall, excluding outliers, we considered in our analysis 27,403 individual observations. We study the impact of gender and affiliation separately[8]. Each analysis is carried out in 4 separate cases: distinguishing between STEM and HSS areas and between the two positions for which the habilitation is sought (AP or FP; Table 1).

Table 1. Frequencies and % of success of candidates, by areas and academic rank (excluding outliers).

	HSS areas		*STEM areas*	
	Male	*Female*	*Male*	*Female*
Full Prof.	1388 (56,9%)	753 (62,2%)	5235(59,7%)	2202(61,3%)
Associate Prof.	2249 (49,1%)	1613 (59,6%)	8162 (55,5%)	5801 (53,2%)

	HSS areas		*STEM areas*	
	Outsider	*Insider*	*Outsider*	*Insider*
Full Prof.	278 (26,5%)	1863 (63,7%)	1525 (42,8%)	5912 (64,7%)
Associate Prof.	1949 (38,3%)	1913 (66,7%)	7539 (44,3%)	6424 (66,6%)

[8] The separate analysis is warranted by the very low correlation (r=0.006) which exists between the two characteristics.

The relevant data are as follows: we consider whether the candidate belongs to STEM or HSS areas, the value of its quality indicator, the gender, the position for which the habilitation is sought, the affiliation, and the outcome of the evaluation.

In estimates not provided here, we also considered as control variables, the age of applicants and a proxy for quantity of publications (the number of articles, for STEM, and the number of articles and book chapters for HSS areas). The evidence of their impact is not robust, as the significance depends on the specific setting: for example, age is never significant when analysing gender, and the quantity indicator is only significant when analysing affiliation in evaluation of AP in STEM areas. Moreover, the inclusion of these two variables does not modify in an appreciable way the results produced below (estimates with the extended model available on request).

The model

To explore the impact of gender and affiliation on the probability of success in the NSH, we estimated the same following probit model:

$$P(y=1) = F\,(\beta_0 + \beta_1\,x + \beta_2\,z + \beta_3\,z^2 + \beta_4\,z^3 + \beta_5\,z^4 + \beta_6\,xz + \beta_7\,xz^2 + \beta_8\,xz^3 + \beta_9\,xz^4) \qquad (1)$$

where y is the binary outcome variable, x is the binary variable representing either the affiliation or the gender (equal to 1 for insiders and males, respectively), z is our quality indicator, F is the cumulative distribution function of the standard normal distribution and P is the probability of success. The use of the polynomial to the 4th power for z reflects preliminary estimates, which suggested that a linear model was not appropriate, both in STEM and HSS areas. In particular, in a probit model for STEM areas (including both AP an FP candidates) coefficients were significant until the 4th power, while in HSS areas until the 2th. For homogeneity of analysis, we decided to adopt the same model (polynomial at 4th power) in both cases.[9] In the variables associated to coefficients from β_6 to β_9 we make z and x interact, allowing the latter to impact on the shape of the probability curve, not simply to shift it.

In the Appendix, together with the benchmark model detailed estimates (Tables A1 and A2; columns 1, 3, 5, 7), we present the (broadly similar) results of an alternative specification where we consider as control variables the individual areas respectively included in STEM and HSS (Tables A1 and A2; columns 2, 4, 6, 8).

Results

Gender

Figure 4 shows the estimated probability of success associated with different values of our proxy for quality of publications, distinguishing by gender. Each sub-diagram refers to one of our 4 cases, as we apply our model separately for STEM and HSS areas and for the position sought (AP and FP). In Appendix we present the full results of the probit estimations.

As for the shape of the relationship outlined in the sub-diagrams, it is clearly different in STEM areas with respect to HSS areas. In the former, it is clearly S-shaped: H-index values which are below half of the threshold are associated to a nihil probability of success; in the range 0.5-1.5 the curve is steep, and probability rises at 60%; afterwards it is mainly flat, suggesting that higher values do not have an additional impact on the NSH Commissions' assessment.

[9] Excluding the 3rd and 4th power for HSS does not modify in an appreciable way the results.

In HSS areas, even in the absence of publications in A-ranked journals,[10] there is a significant probability of success (30%); the latter increases mildly as the value of our quality indicator rises, reaching values at the end of the distribution which are close to 80% for AP and 60% for FP.

Figure 4 shows relatively small differences between males and females, more evident in the case of HSS areas, where they are generally in favour of women. In Table 2 we report the results of a χ^2-test concerning the joint significance of the coefficients ß$_1$ and from ß$_6$ to ß$_9$. They indicate that the shape of the relationship between the quality indicator and the probability of success is significantly influenced by gender only in the case of AP (top sub-diagrams in Figure 4).[11] As for the actual impact on estimated probability, in the case of AP it is on average 8% for HSS in favour of women and 2% for STEM, in favour of men (Table 3).

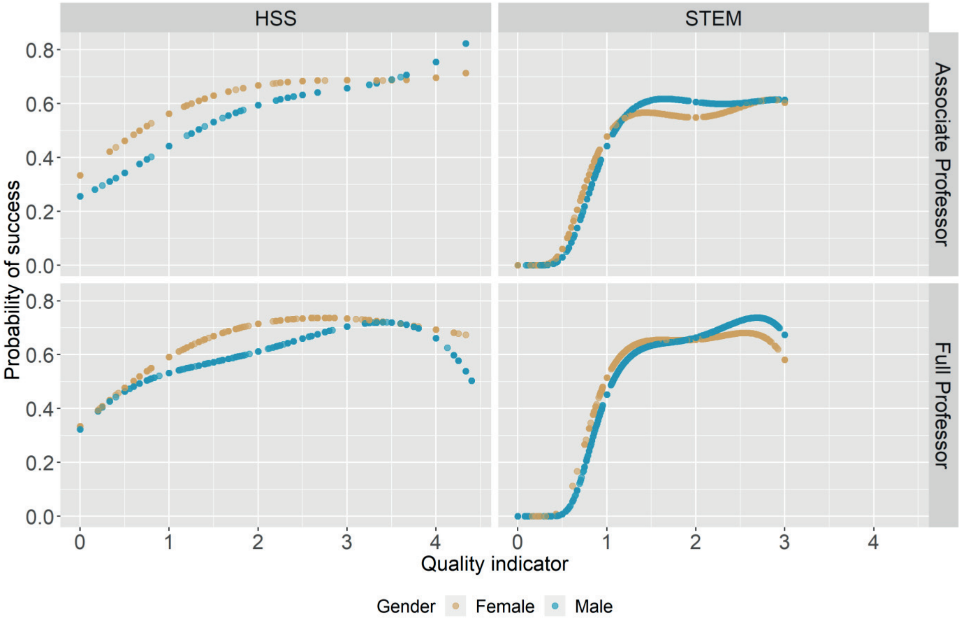


Figure 4. Estimated probability of success by scientific field and gender

Table 2. The χ^2 statistics for ß1, ß$_6$ - ß$_9$ coefficients

		HSS areas				*STEM areas*		
		full prof.		*associate prof.*		*full prof.*		*associate prof.*
Gender		0,0711		0,0000		0,0851		0,0003

[10] In this case, evidently, the candidate had met or exceeded the thresholds for the other 2 requirements (articles and book chapters, books).

[11] The result of the test for FP in HSS may also reflect the relatively small number of observations.

Table 3. Differences in estimated probability: males-females

Accademic rank	*Male-Female Probability at the median*		*Male-Female Average Probability*	
	HSS areas	*STEM areas*	*HSS areas*	*STEM areas*
Associate professor	-0,11	0,04	-0,08	0,02
Full professor	-0,10	-0,02	-0,05	-0,02

Affiliation

Figure 5 shows the estimated probability of success associated to different values of our proxy for quality of publications, distinguishing by affiliation. In Appendix we present the full results of the probit estimations. Also in this case, the shape of the relationship outlined in the sub-diagrams is different in STEM areas with respect to HSS areas, though differences are not as clear-cut as in Figure 4. The main evidence here is the large difference, for almost any given levels of our quality indicators, in the probability associated to insiders vs. outsiders. In particular, in HSS areas, even in the absence of publication in A-ranked journals, insiders have a probability of success of 50%, while the probability for outsiders is 0 for FP and about 10% for AP. Moving to the higher levels of quality, the difference decreases but remains large in both cases. In STEM areas, low levels of H-index are instead associated to a nihil probability of success for both insiders and outsiders, but a large difference in the probability between the two groups emerges for higher values. In Table 4 we report the results of a χ^2-test concerning the joint significance of the coefficients $ß_1$ and from $ß_6$ to $ß_9$. They indicate that the shape of the relationship between the quality indicator and the probability of success is significantly influenced by affiliation in all analysed cases. As for the actual impact on estimated probability, on average it ranges from 22% for FP to 37% for AP in STEM areas (Table 5).

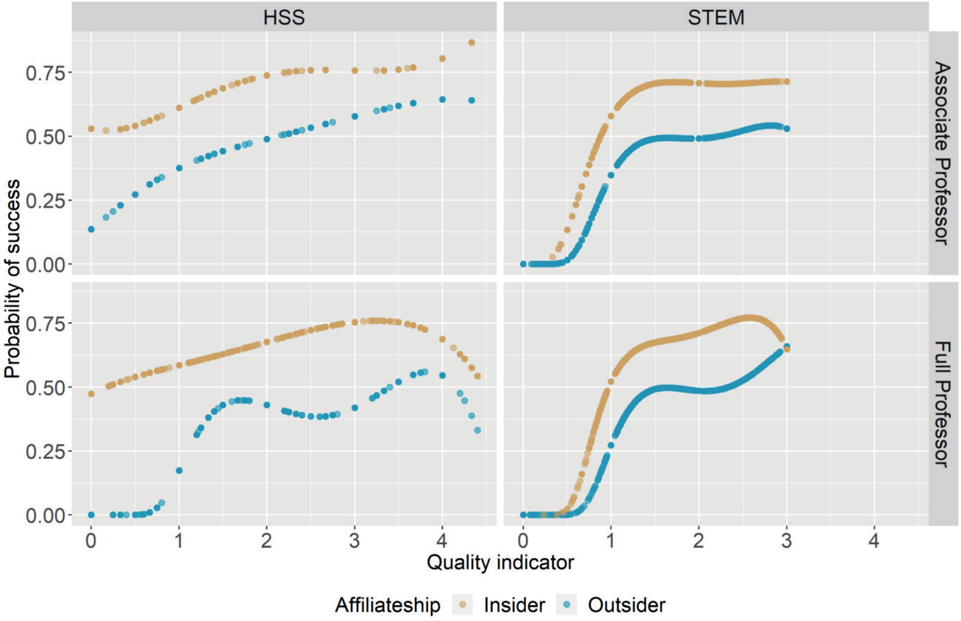


Figure 5. Estimated probability of success by scientific field and affiliation

Table 4. The χ^2 statistics for ß1, ß$_6$ - ß$_9$ coefficients

		HSS areas				*STEM areas*		
		full prof.		*associate prof.*		*full prof.*		*associate prof.*
Affiliation		0,00000		0,00000		0,00000		0,00000

Table 5. Differences in estimated probability: insiders-outsiders

Accademic rank	*Insider-Outsider Probability at the median*		*Insider-Outsider Average Probability*	
	HSS areas	*STEM areas*	*HSS areas*	*STEM areas*
Associate professor	0,24	0,21	0,28	0,37
Full professor	0,20	0,21	0,22	0,22

Conclusion and discussion

This paper contributes to the debate on gender discrimination in Italy and provides a preliminary analysis of the issue of bias against outsiders of the Italian Academia.
We focus on the decisions taken by the National Scientific Habilitation Committees – the first step in the Italian university recruitment of associate or full professors – and investigate whether the relationship between these decisions and the quality of the scientific production of candidates changes depending on gender and on candidates being inside or outside the Italian academic community.

As for the gender factor, we find relatively small differences between males and females in the relationship between the probability of success and our proxy for quality of publications. For candidates applying for a position as an Associate Professor, a $\chi 2$-test on the gender coefficients shows that these differences, albeit quantitatively small, are statistically significant, in HSS areas in favour of women, in STEM areas in favour of men. For candidates applying for a position as Full Professors, possibly due to the relatively small number of observations, the differences (in both cases in favour of women) are not significant. We interpret these data as an evidence that transparency in the NSH process limits gender discrimination to isolated cases, if any. The question remains whether this balanced assessment will eventually produce a drastic reduction of the large gender gap which exists in medium and high ranks of academic careers, or if it will be substantially undone by the second, decisive step in the recruitment/promotion process, the local contests.

As for the affiliation factor, our analysis shows that an outsider status is associated to drastically lower probabilities of success in the NSH, for all levels of our quality of publications indicator, regardless of which position or disciplinary area candidates apply for. This result, however, may reflect factors other than discrimination. As already pointed out, insiders may be stronger in some of the additional elements included in the evaluation – such as granted awards, received funds, and membership of scientific boards – which the Committees are explicitly required to take into account. The difference could also be explained by the specialization in the field of study. Insider candidates could be more able to meet the requirement of coherence with the disciplinary field they apply in. Since we could not control for these additional factors, the results of our analysis shall be considered only a first step in the direction of understanding whether it exists or not an “affiliation” bias in the NSH.

Acknowledgments

We are grateful to the Ministry of Education and CINECA, for providing the dataset of the applications to the National Scientific Habilitation, and to Alessandro Palma for helpful comments.

References

ANVUR (2018). Report on the status of the Italian academic system and research 2018. From: http://www.anvur.it/wp-content/uploads/2019/01/ANVUR-Completo-con-Link.pdf.

Bagues M., Sylos Labini M., and Zynovieva N. (2017). Does the gender composition of scientific committees matter?. *American Economic Review* 107(4) , 1207-1238.

Ceci, S.J., Stephen, J., Ginther, D.K., Kahn, S., Williams, W.M. (2014). Women in academic science: a changing landscape. *Psychol. Sci. Public Interest* 15 (3), 75–141.

De Paola, M., Ponzo, M. and Scoppa, V. (2015). Gender Differences in Attitudes towards Competition: Evidence from the Italian Scientific Qualification. *IZA Discussion Paper* 8859.

De Paola, M., and Scoppa, V. (2015). Gender Discrimination and Evaluators’ Gender: Evidence from Italian Academia. *Economica* 82 (325), 162–88.

European Commission (2015). She Figures 2015. From: https://ec.europa.eu/research/swafs/pdf/pub gender equality/she figures 2015-final.pdf

Marini, G., and Meschitti, V. (2018). The trench warfare of gender discrimination: evidence from academic promotions to full professor in Italy. *Scientometrics*, 115(2), 989-1006.

Morley, L. (2014). Lost leaders: Women in the global academy. *Higher Education Research & Development*, 33(1), 114–128.

OECD (2018). Education at a Glance 2018. From: http://www.oecd-ilibrary.org/education/education-at-a-glance-2018_eag-2018-en.

Pezzoni M., Sterzi V. Lissoni F. (2012). Career progress in centralized academic systems: Social capital and institutions in France and Italy. *Research Policy,* 41(4), 704-719.

Xuhong Su (2014). Academic scientists’ affiliation with university research centers: Selection dynamics. *Research Policy*, 43(2), 382-390.

APPENDIX

Table A1. Probit models with gender as covariate: estimates

VARIABLES	STEM areas				HSS areas			
	Associate Professors		Full Professors		Associate Professors		Full Professors	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
gender_male	-0.299	-0.518	-1.328	-0.559	-0.226**	-0.249**	-0.034	-0.043
	(1.387)	(1.360)	(3.247)	(3.182)	(0.099)	(0.100)	(0.184)	(0.185)
qual	12.962***	12.247***	16.413**	16.556**	0.747**	0.729**	0.818	0.681
	(2.576)	(2.523)	(6.821)	(6.740)	(0.342)	(0.344)	(0.566)	(0.569)
qual_2	-10.043***	-9.462***	-12.732**	-12.772**	-0.153	-0.058	-0.156	-0.004
	(2.373)	(2.332)	(5.971)	(5.912)	(0.421)	(0.423)	(0.637)	(0.641)
qual_3	3.332***	3.144***	4.316*	4.330*	-0.011	-0.057	-0.007	-0.058
	(0.926)	(0.913)	(2.229)	(2.212)	(0.172)	(0.173)	(0.251)	(0.252)
qual_4	-0.399***	-0.379***	-0.538*	-0.542*	0.004	0.010	0.002	0.008
	(0.129)	(0.128)	(0.300)	(0.298)	(0.022)	(0.022)	(0.032)	(0.032)
gender_qual	-0.675	-0.247	2.463	0.494	-0.305	-0.261	0.229	0.401
	(3.512)	(3.451)	(7.892)	(7.749)	(0.452)	(0.456)	(0.685)	(0.689)
gender_qual_2	1.567	1.185	-1.841	-0.051	0.325	0.294	-0.584	-0.749
	(3.195)	(3.148)	(6.915)	(6.807)	(0.558)	(0.561)	(0.763)	(0.767)
gender_qual_3	-0.807	-0.654	0.624	-0.064	-0.109	-0.099	0.276	0.330
	(1.234)	(1.218)	(2.585)	(2.551)	(0.228)	(0.229)	(0.297)	(0.299)
gender_qual_4	0.124	0.102	-0.075	0.020	0.013	0.012	-0.036	-0.042
	(0.170)	(0.169)	(0.348)	(0.344)	(0.029)	(0.029)	(0.037)	(0.037)
_cons	-5.907	-5.772	-7.423	-8.083	-0.429	-0.128	-0.427	-0.602
	(1.004)	(0.982)	(2.808)	(2.770)	(0.077)	(0.105)	(0.151)	(0.184)
Observations	13,963	13,963	7,437	7,437	3,862	3,862	2,141	2,141
Area FEs:	NO	SI	NO	SI	NO	SI	NO	SI
LR chi2(9)	538.58		307.49		345.80		121.75	
Prob > chi2	0.0000		0.0000		0.0000		0.0000	
LR chi2(18)		947.16		611.58				
Prob > chi2		0.0000		0.0000				
LR chi2(14)						396.40		146.45
Prob > chi2						0.0000		0.0000
chi2(5)	23.16	21.40	9.67	9.88	31.42	32.09	10.15	7.96
Prob > chi2	0.0003	0.0007	0.0851	0.0787	0.0000	0.0000	0.0711	0.1582

Standard errors in parentheses
**** p<0.01, ** p<0.05, * p<0.1*

Table A2. Probit models with affiliation as covariate: estimates

VARIABLES	STEM areas				HSS areas			
	Associate Professors		Full Professors		Associate Professors		Full Professors	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
affil_outsider	-2.531*	-2.753*	-1.496	-2.053	-1.173***	-1.276***	-8.968***	-9.695***
	(1.462)	(1.445)	(3.705)	(3.611)	(0.104)	(0.106)	(3.444)	(3.589)
qual	9.053***	8.217***	17.914***	15.640***	-0.216	-0.291	0.426	0.339
	(2.706)	(2.707)	(3.977)	(3.932)	(0.326)	(0.330)	(0.352)	(0.354)
qual_2	-6.172**	-5.507**	-14.281***	-12.108***	0.667*	0.827**	-0.241	-0.144
	(2.459)	(2.461)	(3.493)	(3.463)	(0.399)	(0.403)	(0.379)	(0.382)
qual_3	1.826*	1.608*	5.022***	4.186***	-0.278*	-0.345**	0.117	0.084
	(0.948)	(0.949)	(1.309)	(1.301)	(0.163)	(0.164)	(0.144)	(0.145)
qual_4	-0.198	-0.173	-0.648***	-0.536***	0.034	0.042**	-0.019	-0.015
	(0.131)	(0.131)	(0.176)	(0.176)	(0.021)	(0.021)	(0.018)	(0.018)
affil_qual	4.671	5.104	0.173	1.597	1.478***	1.590***	14.874**	16.437**
	(3.679)	(3.642)	(8.919)	(8.711)	(0.461)	(0.465)	(6.715)	(6.962)
affil_qual_2	-3.983	-4.340	1.691	0.274	-1.303**	-1.384**	-9.180**	-10.268**
	(3.326)	(3.299)	(7.745)	(7.585)	(0.564)	(0.568)	(4.549)	(4.694)
affil_qual_3	1.438	1.579	-1.275	-0.694	0.447*	0.471**	2.327*	2.635**
	(1.277)	(1.268)	(2.872)	(2.821)	(0.230)	(0.231)	(1.277)	(1.312)
affil_qual_4	-0.185	-0.206	0.250	0.167	-0.051*	-0.053*	-0.207	-0.237*
	(0.175)	(0.175)	(0.384)	(0.378)	(0.029)	(0.030)	(0.126)	(0.129)
_cons	-4.309	-4.226	-7.954	-7.733	0.076	0.435	-0.067	-0.238
	(1.070)	(1.070)	(1.633)	1.611	(0.075)	(0.106)	(0.100)	(0.145)
Observations	13,963	13,963	7,437	7,437	3,862	3,862	2,141	2,141
Area FEs:	NO	SI	NO	SI	NO	SI	NO	SI
LR chi2(9)	1188.16		518.26		619.61		272.57	
Prob > chi2	0.0000		0.0000		0.0000		0.0000	
LR chi2(18)		1592.60		849.64				
Prob > chi2		0.0000		0.0000				
LR chi2(14)						703.80		298.95
Prob > chi2						0.0000		0.0000
chi2(5)	663.70	657.33	217.40	243.38	296.83	327.23	61.61	57.01
Prob > chi2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Standard errors in parentheses
**** p<0.01, ** p<0.05, * p<0.1*

Augmenting a Research Information System with automatically acquired category and keyword information

Sven Blanck, Andreas Niekler and Marc Kaulisch

{marc.kaulisch,sven.blanck}@zv.uni-leipzig.de, aniekler@informatik.uni-leipzig.de
Leipzig University, Ritterstraße 26, 04109 Leipzig (Germany)

Abstract

With an ever-increasing amount of data, it is essential for many systems that documents can be retrieved efficiently. The process of information retrieval can be supported by metadata enrichment of the documents. The aim of this work is to make scientific publications and project descriptions, consisting of titles, abstracts and bibliographical references, easier to find. Therefore, we investigate text analytical methods such as keyword extraction algorithms (TFIDF, Log-Likelihood, RAKE, TAKE and KECNW) and classification approaches using a SVM with ensembles of classifier chains (Web of Science and GEPRIS categories as taxonomies) and compare their quality. We present an altered an optimized keyword extraction algorithm and a supervised subject and keyword classification approach which are, to our knowledge so far, one of the first automatic applications of this kind in informetrics and scientific information retrieval.

The most promising methods are employed and the extracted information is attached to the documents as metadata. These support a search query, using pseudo relevance feedback, to obtain further relevant search results and can also be used to derive profiles for authors, faculties, etc. The concepts developed here will serve as a basis for the Leipzig University Research Information System.

Introduction

In this paper, we present an evaluation of methods used in Natural Language Processing and Text Mining in order to generate unified keywords, classify projects and publications and improve findability of research information. Our aim is to develop a (semi-) automatic annotation approach for research information, which provides a standardized and unified way to augment this information with additional metadata such as keywords and classifications.

The self-developed Leipzig University Research Information System (leuris) targets metadata about projects, publications, prices/awards, international research partnerships, academic events, patents, transfer activities and researcher education. In each information entity, the text data such as titles, abstracts or full-texts are collected. Leipzig University aims to use this text data threefold. First, this text data is analysed in order to profile research at Leipzig University internally. Second, text data analysis improves the findability of research information for external and internal purposes. Third, the utilization of text data enables content-based classification in order to ease further internal analysis and to display annotated research information users of the information system.

Keywords or subject areas are important metadata of research projects, research proposals, conference contributions or journal articles. However, each publisher, funder or conference organizer might use a different subject category system and often leaves the keyword selection decision to the authors. Within a Current Research Information System (CRIS), this leads to a heterogeneous and large set of keywords and subjects in large collections of documents. This ultimately results in ambiguous keywords which make it hard for a CRIS to use such information for performant and effective searches. However, these manually assigned keywords can be used as metadata to optimize a CRIS – not only for search tasks but also for classifying research activities – in addition to automatically extracted keywords and classes.

In the literature there are essentially two different ways to automatically generate keywords from a given text (Beliga, Meštrović, & Martinčić-Ipšić, 2015). The first type is keyword extraction. This weights the terms that appear directly in the given text according to their importance and sets the most important ones as keywords as they declare the documents main contents. The second approach is keyword assignment. We assess the problem of keyword

assignment with the automatic classification of given category schemes as it might be difficult or impossible to create, maintain and extend a controlled uniform technical vocabulary for multiple disciplines (Singhal, Kasturi, & Srivastava, 2013). Therefore, we decided to evaluate individual methods for keyword extraction and keyword assignment. We assess the problem of keyword assignment with the automatic classification of given category schemes.
The novelty of our contribution is the alteration of known approaches to fit scientific publications or scientific texts. In detail, we changed term weighting schemes and score aggregation functions of the underlying algorithms and evaluated their results in order to fit the methods to the here described text sources. Additionally, we present a supervised subject and keyword classification approach which is, to our knowledge so far, one of the first applications of this kind in informetrics and scientific information retrieval.

Related Work

In this paper we present a Text Mining-based approach to extract full-text-based meta-information. This information will later be used in information retrieval tasks within leuris. Information retrieval and informetrics are symbiotic fields (Wolfram, 2015). In scientometrics/informetrics fields text analysis and information retrieval is among other things used for topic clustering and finding possible specialities (Sjögårde & Ahlgren, 2019).
For some time now, statistical methods of keyword extraction have offered the best results with regard to precision, recall and F1 score. Approaches such as TF-IDF (Sparck Jones, 1972) and Log-Likelihood (LL) (Rayson, Berridge, & Francis, 2004) use a reference corpus to rank individual terms. Especially the TF-IDF measure is popular as a baseline to compare and evaluate new algorithms. Up to this point, it is the workhorse in many applications such as the Solr/Lucene full-text index.
Another group of keyword extraction approaches is graph-based. As stated by Beliga et al. Graph-based approaches represent a good alternative to statistical approaches due to their independence from linguistic knowledge, domain and language (Beliga et al., 2015). Part of this work is aimed at evaluating the results of statistical methods with graph-based methods such as RAKE (Rose, Engel, Cramer, & Cowley, 2010), TAKE (Pay, 2016) and KECNW (keyword extraction using collective node weight) (Biswas, Bordoloi, & Shreya, 2017), which is a further development on TextRank (Mihalcea & Tarau, 2004) and NE-Rank (Bellaachia & Al-Dhelaan, 2012). While statistical and graph-based methods arrange the individual terms according to their relevance, RAKE and TAKE also employ methods for multi-term keyword candidate recognition using linguistic knowledge (Hulth, 2003). We apply the strategies, which create multi-term keyword candidates to the other approaches (KECNW, TF-IDF, LL) as well.
Our second approach, a supervised classification of scientific categories, has originally been described in (Brück, Eger, & Mehler, 2016; Uslu, Mehler, Niekler, & Baumartz, 2018; Waltinger, Mehler, Lösch, & Horstmann, 2009). In this work, the authors describe a classification mechanism using the word contents of short text snippets in order to assign a DDC category to each of the short texts. The authors utilize a Support Vector Machine (SVM) to classify the text snippets. Uslu et al. (2018) extend this approach by combining a SVM and a Neural Network Classifier with only a slight performance gain and we therefore stick to the original approach and will use only SVM classification for ease of understanding. However, a SVM is normally only responsible for assigning one label to a text. In our case, we need to assign multiple labels to a single document that will act as categories in later applications. One problem transformation that has proven to be practical is One-vs-Rest classification also referred to as binary relevance (BR) (Aly, 2005). Label relations amongst each other can be integrated into the learning process of a SVM with the help of classifier chains (CC) (Read, Pfahringer, Holmes, & Frank, 2009). A further improvement are ensembles of classifier chains

(ECC), which minimize the risk of a poorly chosen classification order through multiple iterations of CC and majority voting (Read, Pfahringer, Holmes, & Frank, 2011).

In our work, we use the extracted metadata consisting of keywords and categories to expand search queries and hence improve the performance of retrieval systems. As Xu & Croft (1996) confirmed in their experiments, an expansion with local feedback based on the search results of the original query is more suitable than global techniques that examine word relationships in a corpus. Therefore, the application examples in this paper use an expansion with local feedback called Pseudo Relevance Feedback (PRF), as this technique currently represents the state of the art (Ariannezhad, Montazeralghaem, Zamani, & Shakery, 2017; Keikha, Ensan, & Bagheri, 2017).

Datasets

In this paper, we are investigating methods to extract keywords and classify research information into suitable classifications on the basis of scientific abstracts. In order to do so, we need to utilize different datasets for training and testing. We used a suited dataset to evaluate the keyword extraction methods. For this purpose, we acquired a dataset from the Web of Science, which contains examples of the category *Computer Science, Interdisciplinary Applications* in order to design and evaluate the keyword extraction algorithms. Based on the assumption that there is a good chance to get meaningful keywords in the field of computer science the decision fell on this specific category. Each example in this dataset contains the title, abstract, references and a set of keywords chosen by the authors. Since the presented methods can only find keywords that are directly in the title or abstract, all author keywords that are not directly contained in the title or abstract are ignored for further evaluation. In order to avoid filtering out different inflections the titles, abstracts, bibliographical references and authors keywords where lemmatized. We use only publications with more than three keywords and out of the 1056 initial examples, 603 remain for our evaluation purposes.

We chose datasets from the Web of Science and the German Project Information System[1] to create and evaluate subject classification models. For the evaluation of the project abstract classification in German, the GEPRIS taxonomy[2] was used, which consists of four granularity levels. We had no access to a dataset containing the fourth level. Hence, only the first three levels where used in the category classification experiments. The GEPRIS dataset consists of the title, the project description and the corresponding category from the GEPRIS taxonomy. The datasets divide into the three levels as follows:

- First Level: 46547 examples, 4 categories
- Second Level: 46547 examples, 14 categories
- Third Level: 28712 examples, 48 categories

We used the Web of Science taxonomy[3] for the evaluation of publication abstract classification in English. To assign proper categories to the acquired dataset a publication automatically gets a category assigned depending on the journal in which it was originally published. Additionally, the CWTS Leiden main fields[4] provide broader topics and can easily be mapped from the Web of Science categories. This works for all but three of the Web of Science categories, which have no mapping yet. Each of the examples contained in the dataset consists of the title, the abstract, the bibliography and the corresponding category from either

[1] GEPRIS, http://gepris.dfg.de/gepris/OCTOPUS?language=en

[2] GEPRIS taxonomy, http://www.dfg.de/dfg_profil/gremien/fachkollegien/faecher/index.jsp

[3] Web of Science taxonomy, https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html

[4] Leiden main fields, http://www.leidenranking.com/information/fields

the Web of Science or the CWTS Leiden taxonomy.
We divided the dataset as follows:

- Web of Science categories: 44866 examples, 254 categories
- CWTS Leiden main fields: 42110 examples, 5 categories

As a last step we applied our evaluated methods to different datasets based on publication and project abstracts derived from the Leipzig University Research Report[5] and inserted the results into a SOLR index.

Keyword and Phrase Detection

In general, all keyword extraction algorithms presented in the related work use a score calculation for the individual terms (unigrams) to determine their relevancy within the document. RAKE and TAKE additionally use methods to extract multi-term keyword candidates from the text and rank them by using the sum of the individual term scores. We assume that the presented keyword extraction methods can be divided into three sections:

1. Keyword candidate extraction
2. Score calculation of the individual terms
3. Ranking of the keyword candidates

Keyword candidate extraction

The approaches from RAKE and TAKE were used for the candidate extraction. In short, RAKE uses a stopword list to divide a text into word sequences and uses them as candidates. The original TAKE approach uses Part-of-Speech Tags to extract noun chunks with eventually co-located adjectives from the text. This approach is based on Hulth's studies which show that linguistic knowledge with noun chunks achieves better results in a keyword extraction process than the simple use of n-grams (Hulth, 2003). Our program extracts the candidates from the title and the abstract, which are combined into one single text using a full stop as sentence separator. The bibliographical references will be used later to calculate the individual term scores. SpaCy (Honnibal & Montani, 2017) is used as a POS tagger and lemmatizer for TAKE and we used a stopword list as a separation feature for RAKE.

Score calculation of individual terms

To calculate the score we tested several methods. These include the statistical approaches TF-IDF and the Log-Likelihood-Ratio. We also tested the graph-based methods KECNW and RAKE respectively TAKE, which are both based on the same principle. In addition, the evaluation includes a combination of TF-IDF and KECNW, which will be discussed later.
In TF-IDF, the individual words are arranged in descending order according to their relevance and the n-best are used as keywords. Note, the IDF value is obtained from the reference corpus. The Log-Likelihood-Ratio uses two corpora to calculate the keywords. The first corpus is the corpus to be examined and the second serves as the reference corpus. Basically, the method carries out a statistical significance test if a selected keyword differs from its statistical expectation. The Log-Likelihood-Method was developed based on the Chi-Squared-Test. The Chi-Squared-Test is based on a normal distribution for statistical text analysis. As an alternative for short texts Dunning suggests the Log-Likelihood-Ratio (Dunning, 1993). This ratio is not based on the assumption that statistical textual analysis is a normal distribution and assumes a binomial and/or multinomial distribution. We simply calculate the RAKE / TAKE (RaTa) based score by the degree of a word divided by its occurrence frequency within a document. The degree is the number of co-occurrences with other words within a specified window in a single document. In the evaluation, a window size of 2 was chosen. KECNW is another algorithm for

[5] Leipzig University Research Report, https://fob.uni-leipzig.de/#/welcome

the automatic extraction of keywords, which was developed especially for Twitter data and based on the ideas of PageRank and TextRank. KECNW is divided into four phases: Textual pre-processing, building a graph structure, node weight assignment and keyword extraction. Each token in the texts will be become a node in the resulting graph structure. An additional filtering is applied by deleting all words which do not exceed the Average Occurrence Frequency (AOF). The edges are formed at words that directly adjoin each other in a sequence. The weighting of an edge is determined by the co-occurrence frequency of a term *i* and a term *j*, as well as their individual frequencies. In the next step, the nodes are weighted. The weighting is composed of five different components: Distance to the central node ($D_{C(i)}$), selectivity centrality ($SC(i)$), importance of neighbouring nodes ($Neigh_{Imp}(i)$), position of a node ($F(i), L(i)$) and term frequency ($TF(i)$). These values are normalized for further calculation. For details on the formulas please refer to the original paper (Biswas et al., 2017). We alter the formulas later but leave out the calculation details for reasons of compactness. Since KECNW is designed for the analysis of Twitter texts (tweets, very short), it is necessary to alter the algorithm for scientific texts. In detail, we exchanged the pre-processing and the additional weighting of the first and last word in the text. We simply, in the case of scientific texts, weighted the first and last words from the title and abstract ($F(i), L(i)$). An additional weighting of the first and last words in the literature references was omitted, since these have only a supporting function and should not influence the result excessively. In addition, we experimented with several parameters. The AOF which must be used according to the original KECNW algorithm is ignored because the abstracts are short texts and possibly important words, which occur only once, are omitted. As our experiments confirmed, omitting the AOF did indeed improve the results for our short documents.

The TF-IDF and the Log-Likelihood-ratio computation needed a comparison corpus. We choose a collection of 1 million sentences of the English Wikipedia of the year 2016 made available by a project within Leipzig University (Uwe Quasthoff & Eckart, 2015). These sentences are lemmatized and brought into a bag-of-word representation.

In addition to the normal KECNW approach, the following modifications were tested:

- *KECNW-Trigram* does not only use the direct neighbours, but also the neighbours of the neighbours as edges.
- *KECNW-Sentence* uses all words that occur in the same sentence as edges.
- *KECNW_multiply_idf* additionally uses the IDF for the node weighting. The modified node weighting function looks as follows:

$$Node\_weight(i) = (D_{C(i)} + SC(i) + Neigh_{Imp}(i) + F(i) + L(i) + TF(i)) \cdot \text{ IDF(i)}$$

- *KECNW_use_tfidf* also uses the IDF from the TF-IDF weighting scheme. The modified node weighting function then looks as follows:

$$Node\_weight(i) = D_{C(i)} + SC(i) + Neigh_{Imp}(i) + F(i) + L(i) + \left(TF(i) \cdot \text{IDF(i)}\right)$$

All these modifications slightly improved the results as our evaluation demonstrates. The best modification proves to be *KECNW_use_tfidf.* However, the quality gain is only in the range 0.05 around the F1 score results of the TF-IDF keyword extraction approach.

Ranking of the keyword candidates

The original approach of candidate scoring uses the sum of all scores of the individual terms belonging to a keyword candidate. Note, multiple words can form a longer sequence for a keyword candidate. However, this approach favours much longer candidates because the sum gets larger with more words in a candidate sequence. This is undesirable and alternatively, we used the average of the individual terms instead of the sum. As confirmed by our evaluation, this method achieves better results and is therefore used in our implementations.

Evaluation

The evaluation measures for these approaches are precision, recall and the F1 score. The keywords are tested with the specified keyword dataset from the Web of Science.

We evaluated scores for increasing numbers *n* of most relevant keywords detected by the algorithms. The results are shown in Table 1, Table 2 and Table 3. Since most authors assign multi-term keywords, each automatically extracted keyword that is present in the authors keywords only partially is counted as a hit. It is noticeable that the TF-IDF, Log-Likelihood and KECNW are very close to each other. Only RAKE/TAKE (RaTa) is far behind in comparison.

To test the average scores for multi-word keywords we extracted the candidates with the TAKE methodology (RAKE candidates were found to be not optimal). The evaluation is presented in Table 4, Table 5 and Table 6 and demonstrates that the quality of the scoring approaches is similar to the single word extraction evaluation, but a candidate only counts as a hit if the extracted keyword is completely matching an author's keyword. We used the variant *KECNW_use_idf* for testing the KECNW algorithm, since it turned out it slightly improves the results.

Table 1: Precision of the original individual terms scoring algorithms with partial hits

Method	*n = 1*	*n = 2*	*n = 3*	*n = 4*	*n = 5*	*n = 6*	*n = 7*	*n = 8*	*n = 9*	*n = 10*
TF-IDF	**0.62**	0.56	**0.518**	**0.47**	**0.43**	**0.396**	**0.366**	**0.343**	**0.321**	**0.302**
LL	**0.62**	**0.567**	**0.518**	0.469	**0.43**	0.394	0.365	0.341	0.32	0.3
RaTa	0.011	0.01	0.011	0.012	0.011	0.011	0.013	0.014	0.015	0.017
KECNW	0.605	0.558	0.51	0.464	0.421	0.391	0.36	0.335	0.312	0.295

Table 2: Recall of the original individual terms scoring algorithms with partial hits

Method	*n = 1*	*n = 2*	*n = 3*	*n = 4*	*n = 5*	*n = 6*	*n = 7*	*n = 8*	*n = 9*	*n = 10*
TF-IDF	**0.123**	0.222	**0.308**	**0.372**	**0.426**	**0.47**	**0.507**	**0.543**	**0.571**	**0.598**
LL	**0.123**	**0.224**	0.307	**0.372**	0.425	0.468	0.505	0.539	0.57	0.594
RaTa	0.002	0.004	0.007	0.009	0.011	0.014	0.018	0.022	0.027	0.033
KECNW	0.12	0.221	0.303	0.368	0.417	0.464	0.498	0.531	0.556	0.584

Table 3: F1 score of the original terms scoring algorithms with partial hits

Method	*n = 1*	*n = 2*	*n = 3*	*n = 4*	*n = 5*	*n = 6*	*n = 7*	*n = 8*	*n = 9*	*n = 10*
TF-IDF	**0.205**	0.318	**0.386**	**0.415**	**0.428**	0.43	**0.425**	**0.421**	**0.411**	**0.402**
LL	**0.205**	**0.322**	**0.386**	**0.415**	0.427	**0.428**	0.424	0.418	0.41	0.399
RaTa	0.004	0.006	0.008	0.01	0.011	0.012	0.015	0.017	0.02	0.022
KECNW	0.2	0.316	0.38	0.41	0.419	0.425	0.418	0.411	0.4	0.392

Table 4: Precision of the keywords using TAKE candidates by means of average and exact matching

Method	*n = 1*	*n = 2*	*n = 3*	*n = 4*	*n = 5*	*n = 6*	*n = 7*	*n = 8*	*n = 9*	*n = 10*
TF-IDF	0.222	0.198	**0.169**	0.149	0.138	**0.130**	**0.123**	0.115	**0.111**	**0.105**
LL	0.221	0.194	0.168	0.149	0.138	0.128	0.120	0.115	0.110	0.105
RaTa	0.054	0.060	0.067	0.067	0.067	0.065	0.064	0.064	0.063	0.063
KECNW	**0.233**	**0.199**	**0.169**	**0.152**	**0.140**	**0.130**	**0.123**	**0.116**	0.110	**0.105**

Table 5: Recall of the keywords using TAKE candidates by means of average and exact matching

Method	*n = 1*	*n = 2*	*n = 3*	*n = 4*	*n = 5*	*n = 6*	*n = 7*	*n = 8*	*n = 9*	*n = 10*
TF-IDF	0.044	0.078	**0.100**	0.118	0.137	**0.154**	**0.170**	0.183	**0.197**	**0.208**
LL	0.044	0.077	**0.100**	0.118	0.137	0.152	0.167	0.182	0.195	0.207
RaTa	0.011	0.024	0.040	0.053	0.066	0.078	0.089	0.101	0.112	0.124
KECNW	**0.046**	**0.079**	**0.100**	**0.120**	**0.139**	**0.154**	**0.170**	**0.184**	0.196	0.207

Table 6: F1 score of the keywords using TAKE candidates by means of average and exact matching

Method	*n = 1*	*n = 2*	*n = 3*	*n = 4*	*n = 5*	*n = 6*	*n = 7*	*n = 8*	*n = 9*	*n = 10*
TF-IDF	0.073	0.112	**0.126**	0.132	0.138	**0.141**	0.142	0.141	**0.142**	**0.140**
LL	0.073	0.110	0.125	0.132	0.137	0.139	0.140	0.141	0.140	0.139
RaTa	0.018	0.034	0.050	0.059	0.066	0.071	0.075	0.078	0.081	0.083
KECNW	**0.077**	**0.113**	**0.126**	**0.134**	**0.140**	**0.141**	**0.143**	**0.143**	0.141	0.139

Subject area classification

To create and evaluate the classifiers, each dataset for the Web of Science and GEPRIS classification is divided into a training (90%) and a test set (10 %). All publication and project abstracts are pre-processed. This includes lemmatization for English texts and stemming for German ones because a simple lemmatization for German texts was not available in our processing software. We extracted both, unigrams and bigrams, in order to keep a textual context in the feature set for the classifier. We used LIBLINEAR as SVM implementation (Fan, Chang, Hsieh, Wang, & Lin, 2008). Since the examples can have multiple labels, the problem was transformed to binary relevance classification. Thus, the SVM creates a separate classifier for each label, which only specifies for this label whether it applies to a given document or not. Furthermore, we extended the classification scheme to classifier chains which use the results of the previous classifications for each classifier to take label relationships into account (Read et al., 2011). Classifier Chains (CC) are an extension of the binary relevance method and integrate label relationships into the calculation of classifiers. Both use one model per label for classification. In CC, this is used to form a chain of models. The basic structure of the input vector remains the same, but starting with the second model in the chain, the previous labels are appended to the feature vector. Ensemble of Classifier Chains (ECC) solve the problem of order. To ensure that results do not depend on the order of the classifiers, a constant number of CC classifiers with randomly selected sequences are calculated. Our ensembles of classifier chains use a total of 11 iterations with randomly selected orders of classifiers. In addition, we tested whether it makes sense to change the threshold value for setting the labels. The original method calculates a threshold value by using the mean value of the estimates (either 0 or 1) for each classifier of the same label (Method A). For example, when looking at 11 iterations representing 6 predictions indicating 1 and the remaining 5 predictions indicating 0, the threshold should be less than 6/11. Another proposed approach uses the probability of a classifier that the label should be set (between 0 and 1) instead of the exact estimate (Method B). The latter methodology can considerably improve the results as shown in Table 7 and therefore a threshold value based on probabilities was used for all further evaluations.

Table 7: Comparison of using the traditional and altered threshold calculation scheme while employing ECC.

		0/1 Loss	*Hamming Loss*	*Accuracy*	*F1 Score*
Method A	GEPRIS 1	0.096	0.045	0.911	0.913
	GEPRIS 2	0.279	0.035	0.754	0.766
	GEPRIS 3	0.423	0.02	0.591	0.597
Method B	GEPRIS 1	0.097	0.044	0.912	0.915
	GEPRIS 2	0.233	0.027	0.813	0.828
	GEPRIS 3	0.539	0.033	0.619	0.686

In addition, the influence of using over- and undersampling were examined in the case of the Web of Science categories. Oversampling replicates the examples of the minority class until the minority and the majority class have the same size. Undersampling deletes random examples of the majority class until the minority and majority class have the same size. Our evaluation, as shown in Table 8, revealed that oversampling is more suitable for our purposes than undersampling with different tested ratios. Note, the quality differences between Table 7 and Table 8 are due to the different datasets. We simply demonstrate the impact of threshold selection and over- and undersampling. We present the final scores in the next section.

Table 8: Comparison of using Oversampling and Undersampling with different ratios (Minority:Majority) while employing ECC. Note, the Web of Science dataset was used for this evaluation.

	0/1 Loss	*Hamming Loss*	*Accuracy*	*F1 Score*
Undersampling 1:1	0.788	0.007	0.269	0.292
Undersampling 1:2	0.786	0.007	0.271	0.289
Oversampling 1:1	0.79	0.008	0.367	0.431

Evaluation

We applied evaluation measures as presented by Read et al (2011) for multilabel classification and the corresponding results are shown in Table 9. The experiments with ECC instead of single classifiers significantly increased the quality depending on the number of different labels. The higher the number of different labels, the more the quality improved. Therefore, all classifiers for the evaluation on Table 9 used ECC. In some examples of the test dataset it may happen that the all probabilities for assigning a label are below the required threshold. In such cases the label with the highest probability is selected for these examples. As expected, the classification quality decreases if more different labels are classified. Nevertheless, the results are sufficiently good and suitable to be used. We also observed misclassified categories, which are nonetheless often semantically similar to the correct labels.

Table 9: Evaluation scores for classification with SVM based on two different datasets and different amounts of categories. The number of categories is noted in brackets.

	GEPRIS			*Web of Science*	
	First Level (4)	*Second Level (14)*	*Third Level (48)*	*Normal (254)*	*CWTS (5)*
0 / 1 Loss	0.097	0.233	0.539	0.7	0.216
Hamming Loss	0.044	0.027	0.033	0.008	0.055
Accuracy	0.912	0.813	0.619	0.473	0.871
F1 Score	0.915	0.828	0.686	0.534	0.9

Application in Search Index

We applied our approach for keyword and category extraction on a sample of 4362 unlabelled and unprocessed publications from the Research Report of Leipzig University as described in the Datasets. Of these, we inserted the publication title, abstract, publication year and the first mentioned author into a Solr full-text index. Additionally, we automatically annotated the Web of Science categories, research areas, CWTS Leiden categories and the top 10 keywords per document using our proposed methods. The categories were extracted using ECC and the keywords were determined by using KECNW with the TAKE candidates to represent multi-term keywords. The categories were extracted using ECC and the keywords were determined by using KECNW with the TAKE candidates to represent multi-term keywords.

We illustrate the application in a search index with English examples. The usage examples of the generated metadata are divided into two areas:

1. Expansion of conventional search queries using pseudo relevance feedback (Xu & Croft, 1996)
2. Facetted search queries on metadata for analysis purposes

A conventional search query refers to a search within the title and abstracts of the texts with one or more search terms defined by the user. In the full-text index, we defined the individual components such as title and abstract as searchable fields in tokenized form. The query is built with a logical disjunction. Since we assume that a search result has a higher relevance if the search term appears in the title and not only in the abstract, we weighted the search differently. Boosting is used to append a multiplier to single search fields or search terms. We use this mechanism in order to calculate the relevance score.[6] The search query is expanded by using the results of the conventional search query. We acquire the keywords and Web of Science categories from each of the search results and reintegrate them into the search query.[7] For example, we find 33 documents when searching for the term *cardiology* within titles and abstracts. From these the keywords and the categories are extracted, which are used to expand the conventional search with logical disjunctions. In this example, a keyword or a category must appear more than twice in the result set to be embedded into the expanded search query. In addition, we define a boosting factor to each field. It is assumed that search terms, which are included in the title or the abstract of the documents, give more relevance to them. The keywords are also more important for the search than the categories and therefore have a higher boosting factor. The boosting factors are intuitively set as following:

- Title ^4
- Abstract ^2
- Tokenized Keywords ^1
- Untokenized Keywords ^1.5
- Categories ^0.5

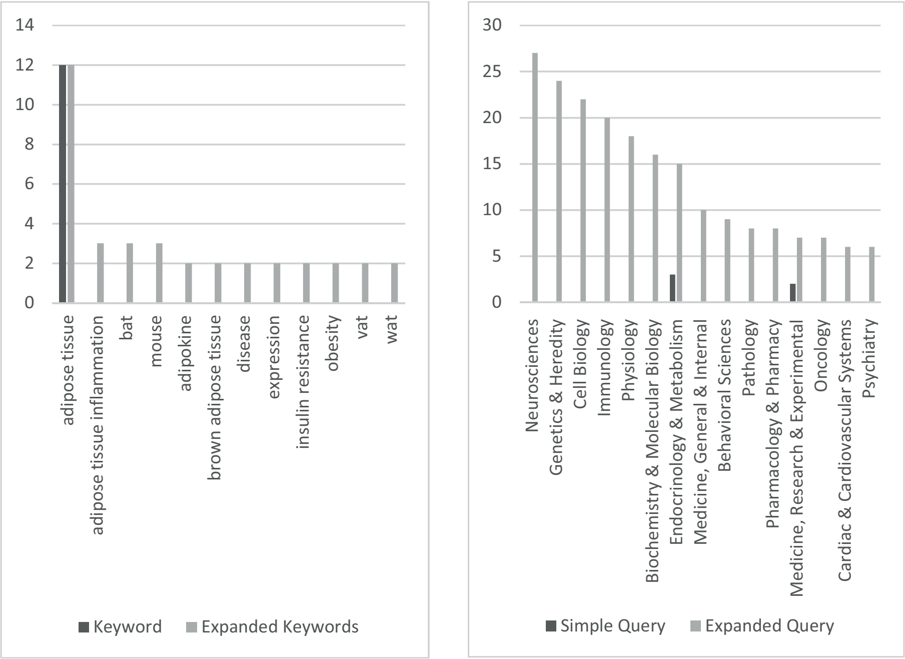
Figure 1 shows how the search query can be expanded directly on a keyword search in the title and abstract. The simple search finds only two categories occurring in documents which contain the query terms. But the documents containing the term *adipose tissue* also deliver 11 further keywords, which co-occur at least in 2 documents from the search result. These keywords expand the search query and we obtain a larger result set. Additional categories (which occur six times or more) can be found that are relevant to the enriched keywords. In this way we are

[6] https://lucene.apache.org/solr/guide/6_6/the-dismax-query-parser.html

[7] Added keywords (>2): patient, cardiology, tavr, catheter ablation, dcb, ehra, esc area, esc member country, geographical esc region, lesion revascularization, non-european esc country, predictor, pulmonary hypertension, revascularization

Added categories (> 2): Cardiac & Cardiovascular Systems, Health Care Sciences & Services, Medicine, General & Internal, Peripheral Vascular Disease, Surgery able to obtain a profile which allows us to investigate the research activities of Leipzig University relevant to the topic reflected by the original search term.

Figure 1: Histograms of collected keyword and category information well suited for profiling of academic careers, faculties or research projects.



Discussion

In this work, we presented and evaluated several known algorithms for keyword extraction and text classification in scientific publications and project descriptions. In an example, we augmented the documents in a full-text index with the results of these algorithms to support the search process and to enable analysis applications based on the document repository, such as profiling authors or faculties. The evaluation demonstrates that graph-based keyword extraction approaches such as KECNW can compete with TF-IDF based algorithms with respect to precision, recall and F1 Score. However, the computing time, which is considerably longer than that of TF-IDF based approaches, has been noticed negatively. This can possibly be improved but was not pursued in this work. Furthermore, the candidates of TAKE gave better results than those of RAKE. We also found out that the score calculation of the candidates with average values achieved better results than with summation, since the average prevents the preference of longer keywords. Nevertheless, the individual term scoring of RAKE / TAKE was not able to keep up with the results of the other presented methods.

Keyword extraction methods that only set a single term as a keyword are not meaningful enough since authors frequently assign multi-term keywords. The candidates extracted by TAKE have greater accordance with the author keywords than those extracted by RAKE. However, many keywords that are not directly contained in the text are missing since the authors derive them from the text with the help of their own knowledge. However, since authors also assign keywords that do not appear directly in the text, a fixed vocabulary is required that can be assigned to texts. One such approach is Wikify (Mihalcea & Csomai, 2007). Wikify uses the Wikipedia headings as a fixed vocabulary and assigns them to the texts. As an additional

improvement, one could create a stopword list given a specific context such as publications and project descriptions. This can prevent unimportant terms from becoming too important, such as the word *project* in project descriptions, which usually describes all these texts.
To classify the texts, the Web of Science categories with the derivable CWTS Leiden categories where used as a taxonomy for English publications. For the German project descriptions, our experiments used the first three levels of the GEPRIS classifications as taxonomies. We applied a SVM in order to classify the documents. Since the texts of these domains cannot necessarily be assigned to a single label, we transferred the problem to binary relevance and created a single classifier for each label. Due to the resulting class imbalance, it is inevitable to use oversampling as the evaluation demonstrated. Furthermore, we implemented ensembles of classifier chains in order to include label hierarchies. This is particularly profitable with a great number of different labels. The classifications results are satisfactory. Only the Web of Science categories appear to be problematic with regard to their low F1 score. Due to the large amount of different labels, which may overlap thematically or could also be superordinate fields, it is difficult to determine an exact result. With the test data, it often happens that labels are awarded that do not match the correct label but are very close to it and one would think as an assessor that the label was assigned correctly. Another way to improve the classification is to create larger training datasets. Especially in the Web of Science dataset with 254 classes it is difficult to create a good classifier with only ~200 examples per label. Finally, it would be possible to use PU Learning (Sansone, De Natale, & Zhou, 2018). PU means *positive unlabeled*. Although there are labels for all examples in the training and test datasets, it may be useful to consider the false class as unlabeled, because some examples from the false class may overlap with the labels of the positive class. For example, there are publications in the Web of Science that are only assigned to the class *Cell Biology*. However, there is also the class *Biology*, which is theoretically a superordinate field of cell biology. Since this is a class, which can be described with the same features as *Cell Biology*, it does not make any sense to train such a document as negative class example. PU Learning would consider this aspect.
Finally, we demonstrated the application of the extracted metadata in the context of search indexes and profiling. A search query expansion creates results that are additionally relevant for a search query. Furthermore, we can create categorical summaries of documents thanks to the metadata determined. The extracted metadata enables researchers to use keywords for finding other researchers working on the same topic. This aspect could promote cooperation between researchers who might not be aware of mutual research interests.

References

Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, *19*, 1–9.

Ariannezhad, M., Montazeralghaem, A., Zamani, H., & Shakery, A. (2017). Iterative Estimation of Document Relevance Score for Pseudo-Relevance Feedback. In J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, & J. Tait (Eds.), *Advances in Information Retrieval* (pp. 676–683). Springer International Publishing.

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, *39*(1), 1–20.

Bellaachia, A., & Al-Dhelaan, M. (2012). Ne-rank: A novel graph-based keyphrase extraction in twitter. *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, 372–379. IEEE Computer Society.

Biswas, S., Bordoloi, M., & Shreya, J. (2017). A Graph Based Keyword Extraction Model using Collective Node Weight. *Expert Systems with Applications*, *97*. https://doi.org/10.1016/j.eswa.2017.12.025

Brück, T. vor der, Eger, S., & Mehler, A. (2016). Complex Decomposition of the Negative Distance kernel. *CoRR*, *abs/1601.00925*. Retrieved from http://arxiv.org/abs/1601.00925

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*, 1871–1874.

Honnibal, M., & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To Appear*.

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 216–223. https://doi.org/10.3115/1119355.1119383

Keikha, A., Ensan, F., & Bagheri, E. (2017). Query expansion using pseudo relevance feedback on wikipedia. *Journal of Intelligent Information Systems*, 1–24.

Mihalcea, R., & Csomai, A. (2007). Wikify!: Linking Documents to Encyclopedic Knowledge. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 233–242. https://doi.org/10.1145/1321440.1321475

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Pay, T. (2016). Totally automated keyword extraction. *Big Data (Big Data), 2016 IEEE International Conference On*, 3859–3863. IEEE.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. *7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*, 926–936.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier Chains for Multi-label Classification. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 254–269). Berlin, Heidelberg: Springer Berlin Heidelberg.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, *85*(3), 333. https://doi.org/10.1007/s10994-011-5256-5

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1–20.

Sansone, E., De Natale, F. G., & Zhou, Z.-H. (2018). Efficient training for positive unlabeled learning. *IEEE Transactions on Pattern Analysis Annotationd Machine Intelligence*.

Singhal, A., Kasturi, R., & Srivastava, J. (2013). Automating document annotation using open source knowledge. *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences On*, *1*, 199–204. IEEE.

Sjögårde, P., & Ahlgren, P. (2019). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *ArXiv:1901.05273 [Cs]*. Retrieved from http://arxiv.org/abs/1901.05273

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Uslu, T., Mehler, A., Niekler, A., & Baumartz, D. (2018). *Towards a DDC-based Topic Network Model of Wikipedia*.

Uwe Quasthoff, D. G., & Eckart, T. (2015). *Building Large Resources for Text Mining: The Leipzig Corpora Collection*. Springer.

Waltinger, U., Mehler, A., Lösch, M., & Horstmann, W. (2009). Hierarchical Classification of OAI Metadata Using the DDC Taxonomy. *NLP4DL/AT4DL*, 29–40.

Wolfram, D. (2015). The symbiotic relationship between information retrieval and informetrics. *Scientometrics*, *102*(3), 2201–2214. https://doi.org/10.1007/s11192-014-1479-0

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 4–11. ACM.

The social sciences and their publishers: Publication, reception and changing meaning of German monographs

Christoph Thiedig

thiedig@dzhw.eu
German Centre for Higher Education Research and Science Studies, Schützenstr. 6a, D-10117 Berlin (Germany)

Abstract

Books are an important part of scholarly communication in the social sciences and humanities (SSH). A number of developments affecting both the SSH and their publishers seem to have shifted the relative importance and prestige of the various publication types in favour of the English journal article and to the detriment of books. While these developments have been studied quite extensively in other countries, empirical analyses of German SSH publishing are still rare. This paper attempts to narrow this gap by investigating publication and citation patterns of German social science monographs since the 1980s against the backdrop of changes in SSH communication and publishing. Using a comprehensive bibliographic dataset for the social sciences in Germany and corresponding citation data, it addresses the question of the relevance of scholarly monographs today.

Introduction

Publication practices in the Social Sciences and Humanities (SSH) are characterized by a pluralism of publication types. Monographs, edited volumes and articles in peer-reviewed and not peer-reviewed journals are common ways of publication. In addition, various other types of publication are produced as well, among them reports, reviews, *Festschriften*, or articles for the lay public. A lot of this communication takes place on a national level (Hicks, 2013; Nederhof, 2006). National language publishing is, thus, much more common in the SSH than in other scientific disciplines, where communication predominantly takes place in English peer-reviewed journals with international scope.
In the social sciences, books and articles are of similar importance (Hicks, 2004). However, they exhibit different characteristics e.g. with regard to the acquisition process (Powell, 1985), author characteristics (Puuska, 2010), content and methodology (Münch, 2011; Swygart-Hobaugh, 2004), and the work's reception (Hargens, 1991; Sullivan, 1994; see also Clemens et al., 1995). These characteristics attest to different publication and reception contexts which influence the creation and dissemination of knowledge in the SSH. In the following chapters, I briefly outline three developments that have seemingly shaped SSH communication considerably – the increasing use of bibliometric indicators in research evaluation practices, changes in library acquisition behaviour connected to the 'serial crisis', and corresponding developments in the field of academic publishing. I argue that these developments have shifted the relative importance and prestige of the various publication types in favour of the English journal article and to the detriment of books. I will then describe the data and methods used to empirically assess the state of monograph publication and reception in German SSH and present first analyses. A final chapter discusses the results and provides a conclusion.

Evaluation of research output

The first development concerns the use of bibliometric indicators which now routinely inform research evaluation measures (Hicks, 2012; Moed, Glänzel & Schmoch, 2004; Rijcke et al., 2016). Criticism of such evaluation practices concerns e.g. the selective coverage of SSH publications in the relevant databases deemed unrepresentative of actual publication practices (Gläser, 2006; Hicks, 2004; Nederhof, 2006) and the appearance of objectivity suggested by broadly available bibliometric indices (Weingart, 2005), leading to rash use of indicators in some disciplines despite heterogeneous citation practices and motives, small sample sizes and an insufficiently developed theory of citation (Bornmann & Daniel, 2008; Moed, 2005). Such

reactive evaluation measures might create incentive structures in which the increase of the entities measured becomes an end in itself (Bornmann, 2011; Espeland & Sauder, 2007).
Bibliometric indicator use in evaluation and allocation procedures has raised the importance of the English journal article. Butler (2003) observes an increase in SCI-indexed article publications after the introduction of a publication-based component in Australian funding allocation. For Norwegian universities, Kyvik (2003) shows a substantial gain in importance of English journal articles between 1980 and 2000, especially in the social sciences. Recently, Kulczycki et al. (2017) find an increase in the number of journal articles and English language publications in Poland between 2009 and 2014, which they attribute to changes in science policy incentive structures. For Germany, Münch (2011) warns of unjustified prominence of the English journal article and the devaluation of books and other publication types.

The 'serial crisis'

Changes in scholarly communication also affected the academic libraries and publishers. Since the 1980s, increases in journal subscription costs (Kyrillidou, 2006; Thompson, 2005; European Commission, 2006) exacerbated by consolidation in the field have put strains on libraries' stagnating budgets. The self-reinforcing cycle of shrinking subscription numbers and price increases for journal bundles commonly known as the 'serial crisis' has furthered this development. Digital infrastructure has additionally bound increasing parts of libraries' budgets. As a result, monograph acquisition stagnated or decreased (Thompson, 2005; for Germany, see Kopp, 2000; Kirchgäßner, 2008).

Developments in the publishing field

In the field of academic publishers, especially in STEM publishing, mergers and acquisitions have led to oligopolistic structures (Larivière et al., 2015). Another central development is the observable decline of scholarly monograph publishing since the 1970s (Thompson, 2005). Cost reduction, price increases, a larger title output or the exploration of new (mass) markets have been some of the coping strategies employed by publishers (ibid.).
Compared to Anglo-Saxon developments, the German publishing landscape is still shaped by a heterogeneous arrangement of small and mid-size publishers, especially in SSH publishing (von Lucius, 2005). Since the 1990s, an increase in managerial-led publishing houses and the concentration of resources as a result of mergers and acquisitions can be observed (ibid.).
For sociology publishing in Germany, Volkmann et al. (2014) observe the emergence of a fast-growing market leader able to significantly advance digital dissemination of scholarly literature via its online platform and thus shape researcher's reception practices, bearing the risk of rendering smaller publishers increasingly invisible. Library budget restrains are felt here as well, leading to lower print-runs for monographs and edited volumes. Tendencies of internationalization, on the other hand, are not very pronounced (Volkmann, 2014).

Changes in scholarly monograph publishing

The developments sketched above suggest that monograph publication becomes increasingly less attractive for both SSH researchers and publishers alike. While they provide only a partial view, they indicate significant changes in scholarly communication bearing further study. An important strand of research on this matter concerns changes in scholarly communication patterns. Despite efforts to increase coverage in databases such as Scopus or the Web of Science Core Collection (Glänzel, Thijs, & Chi, 2016), bibliometric analyses comprising books and book chapters are still dependent on national bibliographic databases.
Empirical analyses of (national) scholarly book publishing seem to accrue in recent years: Engels et al. (2018) use data from five comprehensive coverage national publication databases in order to analyze shares of monographs and book chapters among peer reviewed SSH

publications between 2004 and 2015. They observe stable shares of monographs for most of the countries studied, but a seeming decline in the share of social science monographs in Finland, Norway, and Poland. Similarly, Kulczycki et al. (2018) analyze all SSH publications from eight European countries for the period of 2011-2014. They find differing publication patterns across fields and countries, with stable publication patterns for some, but significant changes in other countries (notably Poland and the Czech Republic) relating especially to the share of monographs, which they attribute to changes in science policy and evaluation methodology respectively (for monograph publishing in Poland, see also Kulczycki, 2018). In all countries, a growth of English language publications could be observed.

Empirical analyses of German language SSH publication practices are rare, since Germany currently does not have a national SSH bibliographic database (Sīle et al., 2018). This paper attempts to narrow this gap by investigating publication and citation patterns of German language social science monographs since the 1980s. Based on the developments sketched above, I argue that the relative importance and prestige of the various publication types has shifted in favour of the English journal article and to the detriment of books. Thus, I broadly expect German-language SSH monographs to be published less frequently over time, in smaller circulation, and to be cited less frequently over time.

Data and methods

The analyses presented here are based on the Social Science Literature Information System (SOLIS) bibliographic dataset provided by GESIS – Leibniz Institute for the Social Sciences. An information service on German Social Science literature, SOLIS collected bibliographic data on articles, edited volumes and their chapters, monographs, and grey literature from sociology, political science and other social science disciplines. As the SOLIS service has been terminated, GESIS now provides the database as a dump.[i] Containing more than 478,000 entries, it covers the ‘core’ of German language sociology literature (Bärisch et al., 2008) and thus represents a unique dataset for the analyses of German language social science publication practices over time.

In order to account for changes in the reception of scholarly monographs, SOLIS entries were matched with Scopus citation data via DOI (89,908 references matched to 901 DOIs). The references’ publication types (“Book”, “Chapter/contribution”, journal article, and Other) were approximated based on the structure of the Scopus data. For “book” references, the language was determined as well. Despite its Anglo-Saxon bias (mostly English-language SOLIS publications had matching DOIs), this sample allows for comprehensive citation analyses on the basis of a curated bibliographic dataset.

In addition, reference patterns of articles in four prominent German-language sociology journals[ii] are studied. Using Web of Science citation data, 317 ‘articles’ containing 16,165 references were analyzed, covering all available data of the 1985, 1995, 2005, and 2015 volumes. Publication type and language were assigned to all references via manual coding.

Results

Fig. 1 presents the shares of German-language SOLIS publications over time. Since chapters in edited volumes are insufficiently captured in SOLIS, they have been excluded from the analyses. Neither journal articles nor monographs exhibit particularly strong trends.[iii] Overall, the share of monographs in the dataset slightly increases; a decline of scholarly monographs as measured by publication output can thus not be observed here.[iv]

The reference patterns of SOLIS publications in Fig. 2 show an increase in the citation of journal articles (and a reduction of citations to “books”) in the period of 2000-2004, with similar reference shares in the following years. This trend can be observed for both journal articles and monographs (not differentiated here). While these results suggest relatively stable citation practices, Fig. 3 indicates an overall increase in the median age of cited “books” and journal articles respectively. This trend is especially pronounced for monographs.

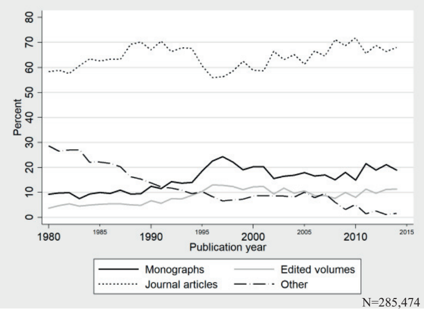
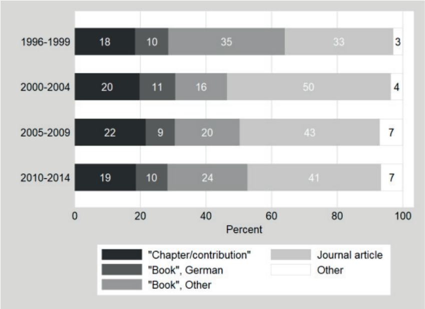


Fig. 1: Share of publications, 1980-2014



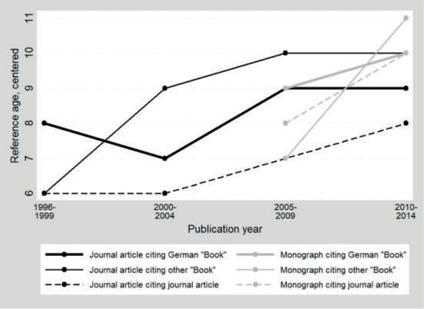
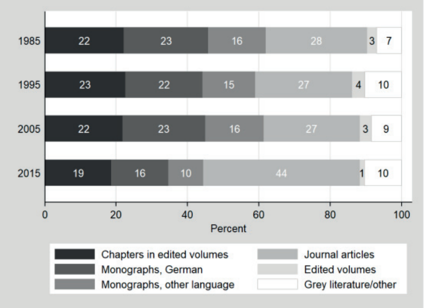


Fig. 2. Share of references by publication type   Fig. 3. Median reference age by publication type

The reference patterns of the four sociology journals in Fig. 4 show similar findings. While the share of German-language monographs is expectedly higher than in the SOLIS sample, the most recent volume shows a distinct change in citation pattern closely resembling its SOLIS counterpart: journal articles are cited more often than before, and monographs less. Analyses of the references’ median age (Fig. 5) show that, while there is an overall increase in median citation age across all publication types, both German and other books ‚age‘ the most.



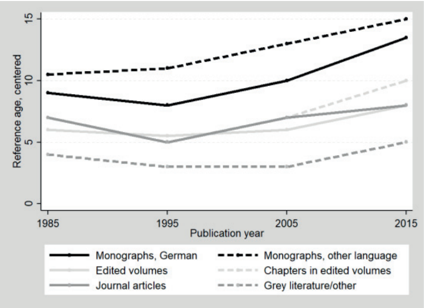


Fig. 4. Share of references by publication type   Fig. 5. Median reference age by publication type

Discussion and conclusion

The analyses of SSH communication patterns has received increasing attention in recent years and has highlighted the importance of national bibliographic databases as comprehensive data sources for such endeavors. Due to the lack of a national database, analyses of German SSH publication practices remain rare. This paper attempts to narrow this gap by investigating publication and citation patterns of German language monographs using a comprehensive bibliographic dataset for the social sciences. Regarding the question of the relevance of social science monographs today, the results are mixed. Publication output might e.g. not necessarily be an indicator of continuous relevance, since publishing more titles is a common strategy to balance smaller circulation. In addition, differences in editorial processes, both between book and journal publication as well as over time, might account for some discrepancies between reference age. Further analyses, both quantitively and qualitatively, will be needed in order to qualify these findings more thoroughly. The present study intends to provide an impetus for this as well as the further assessment and interpretation of (inter)national SSH publishing.

References

Bärisch, S., Hermes, B., Jakowatz, S., Krause, J., Riege, U., Stahl, M., et al. (2008). *Pilotstudie Forschungsrating Soziologie: Vorbereitung, Durchführung, Ergebnisse der Erhebung soziologischer Publikationen. GESIS-Arbeitsbericht: Vol. 5*.

Bornmann, L. (2011). Mimicry in science? *Scientometrics, 86*(1), 173–177.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure?: A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy, 32*(1), 143–155.

Clemens, E. S., Powell, W. W., McIlwaine, K., & Okamoto, D. (1995). Careers in Print: Books, Journals, and Scholarly Reputations. *American Journal of Sociology, 101*(2), 433–494.

Engels T., Starčič A. I., Kulczycki E., Pölönen J., & Sivertsen G. (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? In Centre for Science and Technology Studies (CWTS) (Ed.), *STI 2018 Conference Proceedings. Proceedings of the 23rd International Conference on Science and Technology Indicators* (pp. 774–780).

Espeland, W. N., & Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds. *American Journal of Sociology, 113*(1), 1–40.

European Commission (2006). *Study on the economic and technical evolution of the scientific publication markets in Europe*. Final Report.

Glänzel, W., Thijs, B., & Chi, P.-S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: The book citation index. *Scientometrics, 109*(3), 2165–2179.

Gläser, J. (2006). Die Fallstricke der Bibliometrie. *Soziologie, 35*(1), 42–51.

Hargens, L. L. (1991). Impressions and Misimpressions About Sociology Journals. *Contemporary Sociology, 20*(3), 343–349.

Hicks, D. (2004). The Four Literatures of Social Science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S & T systems* (pp. 473–496). Dordrecht: Kluwer Academic Publishers.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*, 251–261.

Hicks, D. (2013). One size doesn't fit all: on the co-evolution of national evaluation systems and social science publishing. *Confero: Essays on Education Philosophy and Politics, 1*(1), 67–90.

Kirchgäßner, A. (2008). Zeitschriftenkonsortien: Angebotsausweitung auf Kosten der Flexibilität. In E. Pipp (Ed.), *Informationskonzept für die Zukunft. ODOK '07* (pp. 137–146). Graz: Neugebauer.

Kopp, H. (2000). Die Zeitschriftenkrise als Krise der Monographiebeschaffung. *Bibliotheksdienst, 34*(11), 1822–1827.

Kulczycki, E. (2018). The diversity of monographs: Changing landscape of book evaluation in Poland. *Aslib Journal of Information Management, 70*(6), 608–622.

Kulczycki, E., Engels, T. C., & Nowotniak, R. (2017). Publication patterns in the social sciences and humanities in Flanders and Poland. In *Proceedings of ISSI 2017 Wuhan: 16th International Society of Scientometrics and Informetrics Conference, Wuhan, China, 16–29 October 2017* (pp. 95–104).

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., et al. (2018). Publication patterns in the social sciences and humanities: Evidence from eight European countries. *Scientometrics, 116*(1), 463–486.

Kyrillidou, M. (2006). The Impact of Electronic Publishing on Tracking Research Library Investment in Serials. *ARL Newsletter: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC*. (249), 6–7.

Kyvik, S. (2003). Changing trends in publishing behaviour among university faculty, 1980-2000. *Scientometrics, 58*(1), 35–48, from http://dx.doi.org/10.1023/A%3A1025475423482.

Larivière, V., Haustein, S., Mongeon, P., & Glänzel, W. (2015). The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE, 10*(6), e0127502.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation. Information Science and Knowledge Management: Vol. 9.* Dordrecht: Springer.

Moed, H. F., Glänzel, W., & Schmoch, U. (Eds.) (2004). *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S & T systems.* Dordrecht: Kluwer Academic Publishers.

Münch, R. (2011). *Akademischer Kapitalismus: Zur politischen Ökonomie der Hochschulreform* (Orig.- Ausg., 1. Aufl.). *Edition Suhrkamp: Vol. 2633.* Berlin: suhrkamp.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review. *Scientometrics, 66*(1), 81–100.

Powell, W. W. (1985). *Getting into print: The decision-making process in scholarly publishing.* Chicago: University of Chicago Press.

Puuska, H.-M. (2010). Effects of scholar's gender and professional position on publishing productivity in different publication types. Analysis of a Finnish university. *Scientometrics, 82*(2), 419–437.

Rijcke, S. de, Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation, 25*(2), 161–169.

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E., Arefiev, P., et al. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation, 27*(4), 310–322.

Sullivan, T. A. (1994). Genre in Sociology: The Case for the Monograph. In R. J. Simon & J. J. Fyfe (Eds.), *Editors as gatekeepers. Getting published in the social sciences* (pp. 159–175). Lanham, Md: Rowman & Littlefield.

Swygart-Hobaugh, A. J. (2004). A citation analysis of the quantitative/qualitative methods debate's reflection in sociology research: Implications for library collection development. *Library Collections, Acquisitions, and Technical Services, 28*(2), 180–195.

Thompson, J. B. (2005). *Books in the digital age: The transformation of academic and higher education publishing in Britain and the United States.* Cambridge, U.K, Malden, Mass: Polity Press.

Volkmann, U. (2014, October 07). *Soziologieverlage unter multiplem Veränderungsdruck*. Presentation held at the 37. Congress of the German Sociological Association (DGS). Trier.

Volkmann, U., Schimank, U., & Rost, M. (2014). Two Worlds of Academic Publishing: Chemistry and German Sociology in Comparison. *Minerva, 52*(2), 187–212.

von Lucius, W. D. (2005). Strukturwandel im wissenschaftlichen Verlag. *Soziale Systeme. Zeitschrift für soziologische Theorie, 11*(1), 32–51.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics, 62*(1), 117–131.

[i] https://git.gesis.org/open-data/solis-sofis, accessed 30th of May 2019.

[ii] *Berliner Journal für Soziologie* (BJS), *Kölner Zeitschrift für Soziologie und Sozialpsychologie* (KZfSS), *Soziale Welt* (SW), and *Zeitschrift für Soziologie* (ZfS).

[iii] The period of 1994-1998 constitutes an exception which will need to be assessed further.

[iv] The steady decline of the other/grey literature (containing mostly reports and dissertations) might indicate adjustments of the SOLIS data collection processes rather than publication behavior adjustment.

Sorting out Guidelines for the Good Evaluation of Research Practices

Cinzia Daraio[1] and Alessio Vaccari[2]

[1] daraio@diag.uniroma1.it, [2] alessio.vaccari@uniroma1.it

Department of Computer, Control and Management Engineering "Antonio Ruberti" (DIAG), Sapienza University of Rome, Rome (Italy)

Abstract

In this paper, we propose the adoption of moral philosophy and in particular normative ethics, to clarify the concept of "good" evaluation of "research practices". Our perspective is based on the idea that research is a form of social practice according to MacIntyre (1985)'s conceptualization. From MacIntyre's notion, we elaborate three typologies of researcher: the *leader*, the *good* researcher and the *honest* researcher. Reflecting on what is a "good" research practice and on what is the role of researchers in it provides insight into some aspects of both the self-assessment process and how this promotes individual improvement. Moreover, this kind of reflection helps us to describe the functions (missions) of the research practices. A "good" evaluation should take into account all the building constituents of a "good" research practice and should be able to discriminate between good and bad research practices, while enforcing the functions of good research practices. These reflections may be the starting point for a *paradigm shift* in the evaluation of research practices which replaces an evaluation centred on *products* with an evaluation focused on the *functions* of these practices.

Introduction and main objectives

In this paper, we propose to use some of the notions employed by contemporary normative ethics to develop a framework for the "good" evaluation of research practices. We define *ethics* as the sphere of our reflection, language, emotions and behaviour that concerns "good" life, where "good" indicates what favours human flourishing in the various social practices in which human life is expressed. Following the extensive literature on the subject, we define *normative ethics* as that part of moral philosophy that formulates and justifies principles of conduct and concepts that are conceptually connected to the moral good. Basic ethical principles and concepts govern our self in two ways. They help us (i) to make the right sorts of decisions (*practicality requirement*), and (ii) to form a correct evaluation of other's behaviour (*evaluative requirement*).

Usually, normative ethics (see e.g. Furner, 2014) is not considered by evaluative bibliometrics and research evaluations. The consequence of this deficiency is that although there is a proliferation of increasingly sophisticated quantitative methods to evaluate research (see for example the case of university rankings), there is still a lack of clarity on how to understand and operationalize the notion of "good" evaluation of research practice. One of the reasons for this is the lack of a framework for the assessment of research and its impacts (Daraio, 2017). We believe, however, that the concept of "good" is a crucial standard against which to evaluate research practices enabling us to assess them in light of broad human interests.

The characterization of the notion of "good evaluation" of research practices requires the description of "good research practices". This is because a good evaluation takes into account the constitutive elements of a good research practice. Our proposal is to start from a general notion of a "good social practice". From this notion, we specify the notion of "good research practice" and from the latter we specify that of "good evaluation" of research practice. This involves different moves that can be schematically indicated in three points.

- First, clarifying the notion of "social practice" explaining what it means to comply with its rules, and which elements of our *psychology* can account for its emergence. As it will become clear later on, to comply with social practices requires agents to develop specific traits of character which unable agents to grasp, produce and further the "internal goods" of the practices they join. These traits identify those who excel in following the practice. They are *exemplary* figures that the other participants in the practice want to emulate.

- Second, examining how the practice affects the life of those who inhabit it.
- Finally, setting the standards in the light of which assessing the overall effects of practices on society as a whole.

To undertake these tasks, we propose to use different resources offered by philosophical reflection on morals. In particular, we use tools borrowed from utilitarianism, virtue ethics and McIntyre (1985)'s characterization of social practice. Following this track, the paper aims to fill a significant methodological gap in the field of the evaluation of research. We argue that the evaluation of research activities, including research projects and programs, together with their outcomes, should not be limited to assess the products or quantitative aspects of the production and dissemination of recorded information, but should also take into account the *psychology* of the actors involved in this process (authors, readers, etc.), including their motivations.

Taking character qualities as essential inputs within the process of generating research outputs is not a completely new idea. Robert Merton (1973) famously illustrated the link between traits such courage, self-confidence, resilience, taste and the recognition from one's peers along with the capacity to produce excellent quality research. The way we use psychology, however, diverge in at least two ways from Merton's.

First, he uses individual qualities to explain and justify differences in capacity to acquire outstanding achievements between future Nobel laureates and average researchers. We instead use virtues to understand the difference between the activity of researchers, whose motivations cannot be described independently of the *intrinsic* (non-instrumental) desire to acquire the "internal goods" of the practice - e.g. the peculiar pleasure of undertaking new line of research, excellence in analytical skills, a particular taste for problem raising, etc. -, and those who participate in the research practice mainly out of desire to acquire goods *external* to the practice, e.g. power and wealth.

Second, unlike Merton, who merely mentions the excellences in character, we want to present a characterization of these virtuous psychological traits that highlights their constitutive role in producing a "good" research practice. We argue that a "good" practice is characterized, among other things, by the fact that its participants have an *intrinsic* (non-instrumental) interest in seeking the "internal goods" of the practice together with the capacity to grasp and appreciate them. In line with McIntyre's approach we shall argue that the possibility of achieving these "goods" depends on whether participants in the practices have, cultivate and teach others certain virtuous character traits. In the following, we will provide a detailed description of what "internal" and "external" goods of the practice are.

Our conceptual framework

Our perspective is based on a highly plausible hypothesis: the idea that a good scientific/academic research is a form of social practice *a la* MacIntyre. Following MacIntyre's formulation of a social practice (MacIntyre 1981 first ed.; 1985: 187; Vaccari 2012) we define a *good practice* as "any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partially definitive of, that form of activity, with the result that human powers to achieve excellence, and human conceptions of the ends and goods involved, are systematically extended" (MacIntyre, 1981 first ed.; 1985: 187).

That scientific/academic research can be described according to the McIntyre model is strongly justified by the well-known definition of research practices offered in the Frascati Manual. According to this document (OECD, 2015, p. 44) Research and experimental development (R&D) "comprises creative and systematic work undertaken in order to increase the stock of knowledge - including knowledge of humankind, culture and society- and to devise new applications of available knowledge". For an activity to be an R&D activity, it must satisfy five core criteria. The activity must be: 1. Novel, 2. Creative, 3. Uncertain, 4. Systematic, 5. Transferable and/or reproducible.

On the basis of this definition, we characterize a *good research practice* as any coherent and complex form of socially established cooperative human activity through which its participants, through the exercise of a set of refined human psychological qualities or virtues (called "human powers" or virtues by MacIntyre), contribute to the advancement of the body of knowledge that is constitutive of that practice. The term "good" is used to identify a refined and reason-mediated use of typically human abilities (e.g. empathy, imagination, courage, self-reliance, the ability to bind oneself to rules perceived as authoritative, etc.) that yields a meaningful and fulfilling life.

From MacIntyre's notion of research practice we elaborate three typology of researchers: the *leader*, the *good researcher* and the *honest researcher*.

The *good researcher* is a typical participant of a good research practice defined so far – she/he participates in the practice learning and developing the virtues of the practice. The *good* researcher employs typically human qualities to respond in the best way she/he can to the problems that are typical of that practice enabling her/him to creatively advance a particular stock of knowledge. From the good researcher we distinguish the *leader researcher* and the *honest* researcher. The *leader researcher* is one who achieves an outstanding level in the development of creative and social virtues enabling her/him to produce excellent outputs and to be a motivating leader in research group. Finally, the *honest researcher* is the one who does not produce outputs that are contrary to good research practices. More precisely, the honest researcher typically exemplifies the researcher who has completed her/his PhD and is at the beginning of her carrier. Within research institutions, this figure mainly carries out her/his activity in the service of more experienced researchers. Within the university, she/he carries out her teaching activity mainly as a tutor not having yet her/his institutional course during one of the terms. The figure of the honest researcher generally progresses towards that of a good researcher and, in some cases, becomes a leader. However, this may not be the case. In such a situation, an honest researcher is one who, despite having a permanent job as a lecturer or researcher for many years, continues to carry out the tasks she/he was carrying out at the beginning of her career. She/he, however, grasps the research practice in which is involved, with its "internal goods", and fulfils his/her role of being at the service of the practice.

A *good evaluation* of research practices is then an evaluation that is able to take into account the different elements that characterize a good practice, that is, both its outcomes (which can be classified in *internal* and *external goods*, following MacIntyre) as well as the virtues of these three types of researchers.

We characterize *internal goods* as both the outcomes of research and the subjective experiences related to participation in the practice of research, which does not necessarily translate into outputs. We call *external goods* the positive and measurable effects of research results or outcomes on society as a whole.

We argue that good evaluation must therefore be able to distinguish good practice from bad practice. The first is that in which researchers participate in the practice because they are motivated by both internal and external goods. A bad practice, on the other hand, is one in which participants are in no way motivated by the intrinsic desire to achieve internal goods but only act out of the desire of goods external to the practice.

To carry out a good evaluation it is not sufficient to follow abstract and impartial rules, but it is also necessary to have developed *certain virtues* that enable the evaluator both to apply those rules when they conflict with their partial interests and to interpret them in such a way as to make them applicable to the specific case.

We propose to identify the most significant virtues of the *good* evaluator with justice, empathy and practical wisdom. We will analyse them in details in the following.

***Internal* "goods" and *external* "goods" of a research practice *à la* MacIntyre and Nussbaum's theory of abilities**

Let us now examine in detail the nature of external and internal goods as well as how this distinction affects the plurality of standards that constitute good evaluation practice.

Since its products are both internal and external to scientific practice, having an impact outside the research community that potentially affects the well-being of society as a whole, it is advisable to use different styles of evaluation to assess each of them. Therefore, in addition to MacIntyre's concepts of virtue ethics, our framework will also use notions from Nussbaum (2006)'s theory of abilities, and from utilitarianism (as discussed in the next section).

We believe that nowadays research practices require hybrid forms of combination between internal and external goods. Different factors can explain this transformation, including the changes in the way in which science is produced and interacts with society (Scott, 2003).

The model of the virtues can be useful to identify those dispositions that unable researchers to grasp and respond to, in a good enough way (Swanton 2003:1), the internal and external goods of the research practice. As we will show in a moment, these goods include objects, and ways of socially interact with persons and educate them.

To better characterize the notion of social practice that we are using to describe scientific research, it is useful to articulate further the distinction put forward by MacIntyre between "internal" and "external goods".

"Internal goods" to a practice are high quality *outcomes* of the practice that (a) can only be specified in terms of some specific practice, as for example the way of conducting an empirical experiment; the practice of university teaching through lessons, seminars, individual tutoring activities; the practice of interpretation and problematization of the text of classical authors in the humanities; etc. and (b) can only be identified and recognized by the experience of participating in the practice in question. Those who lack the relevant experience are incompetent thereby as judges of internal goods (MacIntyre 1985: 189). Internal goods are reachable by those participants in the practice who practice it as an end in itself and not merely as a means to get something else, e.g. money, power, prestige. According to MacIntyre, these goods include three kinds of *outcomes*. They are

- *the high quality in performance* (e.g. ability to question a text; ability to ask relevant questions during an experiment; ability to motivate one's own research group or students in class, etc.);

- *the high quality of the outcome itself* (e.g. articles, books, research projects, discoveries, etc.);

- *the great value that comes from living a certain kind of life* – the fact that occupying a certain professional role in a research practice contributes to the unity and value of the researcher's life.

The last point needs more articulation. The idea is that those who participate in a practice by acquiring its internal goods are likely to consider it as *something that makes their lives meaningful*. They will tend to describe their lives as those of the participants in a certain practice and this will give a unitary character to the different parts of their biography.

Unlike internal goods, external ones are only "externally and contingently attached" to the practice by the accidents of social circumstance and typically includes prestige, status and money. There are always alternative ways for achieving such goods, and their achievement is never to be had "only" by engaging in some particular kind of practice (MacIntyre, 1985: 201). Moreover, external goods, when achieved, they are always some individual's property – i.e. the more someone has of them, the less there is for other people. They are characteristically objects of competition in which there must be losers as well as winners. On the contrary, internal goods include the outcome of competition to excel, but also positive externalities. This means that their achievement is a good for the whole community who participate in the practice (e.g. Bowlby's attachment theory has transformed the way of seeing the relationship between mother and child by reducing trauma in hospitalized young children; Moore's naturalistic fallacy argument has helped expose many fallacious arguments in philosophical reflection).

The evaluation of the particular practice covered by this paper requires that both internal and external goods are taken into account. On the one hand, it is necessary to assess whether the practice of the academic/scientific research under examination is actually a good practice. In doing so, account must be taken of the excellence of its outputs, the way in which they are achieved (in accordance with the rules that constitute the practice), and the impact that following the practice has on researchers' life plans. On the other hand, we need to establish what consequences following the practice has on the values protected by the democratic constitutions in which the practices have taken hold. That is, we must assess whether the practices produce outputs that are in conflict with interests such as freedom, equality, health, respect for the environment, human dignity, and sociability.

In the light of this twofold requirement, we believe it may be helpful to interpret the two types of goods in the light of the capability approach developed by Nussbaum (Nussbaum, 2006). Specifically we holds the view that internal goods of the research practice are:

1. Use the senses, imagination and rationality in a typically human way, informed by adequate education. Be able to use imagination and thought in connection with our experience

and produce works that are the result of our autonomous and reflective choices (reinterpretation of Nussbaum's point 4: 95)

2. To be able to pursue the objectives of research without ulterior purposes but as intrinsic ends. Be able to have fun and play with activities related to the practice. Moreover, be able to acquire and use specific mental capacities connected with the exercise of the practice such as the ability to apply the rules of the practice to completely new and unexpected contexts, ability to grasp the saliences of the situation required to act in accordance to the practice, etc. (reinterpretation of Nussbaum's point 9: 95).

3. To be able to have attachments to people involved in the practice and to the outcome of research; to experience gratitude towards teachers and masters and justified anger towards those who betray our trust and violate our intellectual property. Be placed in conditions where one's potential and development is not hindered by fear and anxiety (reinterpretation of Nussbaum's point 5: 95).

Following the same approach, we argue that external goods are not only money, power or the reputation of the research institution and its capacity to attract investment, but also the impact research practice has on what Nussbaum has called the "human capacities necessary to live life worthy of human dignity". These capacities, should include:

1. *Life, Bodily Health.* Being able to live to the end of a human life of normal length – i.e. not dying prematurely, or before one's life is so reduced as to be not worth living – and being able to have good health, including reproductive health; to be adequately nourished (Nussbaum's point 1 and 2);

2. *Affiliation.* Being able to live with and toward others, to recognize and show concern for other human beings, to engage in various forms of social interaction; to be able to imagine the situation of another. Protecting this capability means protecting institutions that constitute and nourish such forms of affiliation, and also protecting the freedom of assembly and political speech. Having the social bases of self-respect and non humiliation; being able to be treated as a dignified being whose worth is equal to that of others. This entails provisions of non discrimination on the basis of race, sex, sexual orientation, ethnicity, caste, religion, national origin (Nussbaum's point 7: 96)

3. *Other Species.* Being able to live with concern for and in relation to animals, plants, and the world of nature (Nussbaum's point 8: 96)

4. *Control over One's Environment.* (a) *Political*. Being able to participate effectively in political choices that govern one's life; having the right of political participation, protections of free speech and association. (b) *Material.* Being able to hold property (both land and movable goods), and having property rights on an equal basis with others; having the right to seek employment on an equal basis with others; having the freedom from unwarranted search and seizure. In work, being able to work as a human being, exercising practical reason and entering into meaningful relationships of mutual recognition with other workers (Nussbaum's point 10: 96)

In order to take account of these goods, the evaluation of research practice must also be able to assess the ability of researchers to obtain them. To this end, the virtues of the participants in the practice should also be taken into account. Following once again the MacIntyre setting, we define virtue as "an acquired human quality the possession and exercise of which tends to enable us to achieve those goods which are internal to practices and the lack of which effectively prevents us from achieving any such goods (Macintyre, 1987: 191)".

A potential issue arising from taking this approach concerns the relationship between moral virtues and the virtues that are relative to those who practice scientific research. This is one aspect of the more general issue which concerns the possible tension between the traits of character that make us good as human beings and those that make us efficient as occupying a particular social role. For example, it may be argued that the ability to take a certain detachment from suffering may be a necessary trait in a physician who allows him to make crucial decisions by looking only at facts objectively without letting himself be clouded by emotions. This same trait, however, is not desirable within family relationships where the ability to participate in the emotional life of loved ones is a fundamental part of relational life. Likewise, although a professor's loyalty to his pupils can have the useful function of creating a close-knit group that works efficiently and does the good of research. This same trait could lead the teacher to misbehaviour when, in assigning a public job, he prefers one of his students to another clearly more competent one. These are, of course, simplifications, and one could argue that the more detailed the example becomes, the more so-called conflicts are mitigated. Mitigated as it may be, however, it could be argued that some dose of conflict between the virtues of participants in social practices and moral virtues exists. And if this is true, what is the point of arguing that philosophical ethics can help us define the virtues of the academic researcher?

Our thesis is that moral virtues can be interpreted in ways that allow a typically non-conflictual relationship with the role-specific virtues. According to our proposal, which follows the general lines of Swanton (2007)'s analysis, the relationship between virtues and role-specific ones runs in two directions. On the one hand, role-specific virtues allow moral virtues to be given content, which otherwise would be too vague and generic to offer a practical guide to action and to assessing the conduct of others. This is to say for example that the virtue of courage acquires its content only when it is grounded on the paradigmatic cases of courage that human beings find themselves living in their concrete social interactions as parents, friends, members of a community, etc. On the other hand, what virtue requires in different social circumstances is delimited by the general meaning of virtue. To return to the example of courage, what is required of a brave friend is partly defined and circumscribed by the fact that courage should not be confused with recklessness and disregard for danger.

Our proposal will make use of the following concepts. Following the analysis of Churchland (1998) and Swanton (2007) we define moral virtues as prototype virtues. These have a "high degree of generality, in which contexts such as the role of the agent, his relationship with others, social conventions, and the particular narratives of his/her life have been abstracted away" (Swanton 2007, p. 211).

The transition between the prototypes or moral virtues and the corresponding role specific virtue involves two stages:

1. The *thin* account gives the specification of the field of a virtue (its domain of concern) and states that the virtue is being well disposed in relation to that field. For example, *loyalty* is being well disposed with respect to 'sticking to' relevant individuals or institutions. *Trust* is being well disposed with respect to believing, supporting, and so forth relevant individuals.
2. The *thick* account is given by so-called 'mother's knees rules and basic accounts of relevant emotional and motivational dispositions (Hursthouse, 1999). These rules are characteristically unsophisticated and vague. These provides saliences and paths to assist the development of appropriate emotional and cognitive paths on the world. These different rules articulate the content of the prototype virtues so as to differentiate it according to the different social roles that the subject finds himself occupying.

The claim that prototype virtues are vague is central to the idea that *roles demands* do not characteristically conflict with those of being good as a human being. For example, insofar as honesty is a prototype virtue, it does mean something like *telling the truth* and not lying here and now. On the other hand, in the role-specific virtue of honesty for academics, the substantive question is *what counts as excellence* in the field of quoting, divulgating and disseminating information. *Only when more specific requirements are determined by role-differentiation do we know what it would be to act honestly as a human being.* However, given that honesty is a prototype virtue, an agent with that virtue will have emotional and cognitive dispositions which make his/her not ready to lie or distort the truth.

Given that prototype virtue are vague and need to be specified by role-differentiation, *how do they provide constraints to role virtue*? Our idea is that they provide anchors for moral reflections in role contexts, alerting us to possibilities of excess. Such anchors are traits of characters whose emotional and cognitive features are deeply rooted though early training.

Indeed, those treats of character can be introduced through the narrative of exemplar stories of leaders, may be included in training for young professors and so on.

Why we introduced this distinction? Because it may be useful to understand misconduct in the scientific practice.

In the conclusion of a recent report on Fostering Integrity in Research (2017, p.208-209) by the US National Academies, it is stated: “The committee reaffirms the central recommendation from Responsible Science [a previous report of 2002] that formally places the primary responsibility for acting to define and strengthen basic principles and practices for the responsible conduct of research on individual scientists and research institutions. At the same time, the committee based its recommendations on its understanding that the integrity of research depends on creating and maintaining a system and environment of research in which institutional arrangements, practices, policies, and incentive structures support responsible conduct. Fostering research integrity is an obligation shared not only by individual researchers but also by leaders and those involved with all organizations sponsoring, conducting, or disseminating research, including corporate and government research organizations.” Hence, the primary responsibility is on individual scientists. Fraud and misconduct have for several years been identified as a relevant problem of the scientific community. For a review, see Fanelli (2009). More recently, Fang et al. (2012) found that the main reason for retractions relies on misconduct.

Our conceptual framework allows us to distinguish (discriminate) good research practices from bad research practices.

The virtues a good research practice must have

We believe that a provisional list of these virtues should include:

a *Justice*: this virtue consists in the disposition, required above all by the professors and more generally by the evaluators of the performances and outputs of others, to treat others “in respect of merit and of the desert according to uniform and impersonal standard" (MacIntyre, cit., 191-192).

b *Resilience*: together with pride, this ability is indispensable to move forward in the search. It allows us to leave behind failures (paper rejected, unfunded projects, etc.) and to focus on future projects (Hormann, 2018).

c *Empathy*: In line with extensive literature, by this term we mean the human ability to feel the emotions and feelings of other people through a vicarious feeling that is similar to that of the person with whom we sympathize. We do not believe, however, that empathy in itself is a virtuous capacity in research practices. Since empathy is an instrument for reading the other's mind, it can also be used to manipulate others researchers in malicious ways. Empathy must be cultivated in such a way that it is rooted in the benevolent tendencies of human beings (Batson 2017: 2). In this way, empathy can allow the creation of a climate of *trust* between those who work within research institutions. Mutual trust is in fact an indispensable component in these practices given the fundamental fact of the asymmetry of power that characterizes those interactions (Baier, 1991).

d *Pride*: it is evaluative attitudes towards ourselves (Ardal 1966; Cohon 2008; Taylor 2015). Unlike other emotions, which simply motivate us to pursue or avoid objects, this traits of character fix our attention on persons, casting a positive or negative light on them. If I am proud of my child's success at school, my pride does not fix my attention on the 'merits of my child,' and still less on 'me in the role of father,' but on the whole of myself. As Cohon has rightly said, "when I feel pride, I am proud of something in particular [its cause] . . . But the attitude of pride is a pleasure or satisfaction not in that particular accomplishment or possession, but in myself in my entirety" (Cohon 2008: 166). We believe that the pride associated with one's own achievements in research and the consequent approval of one's peers or superiors is a fundamental spring that drives researchers to perform at best in their area of research (Tangney, 1999).

e *Humility*: the ability to accept the authority of the standards related to the rules that define the practice. I have to recognize that other participants know rules and know how to apply them better than I do. I have to be willing to learn from these people and accept their criticism (MacIntyre 1985: 193).

f *Patience:* or the ability to curb one's own immediate emotions, which could drive us to quickly complete a research in order to obtain as soon as possible the gratification of a positive result. To be able to wait and to be guided by a cautious scepticism that prompts us to control accurately the different steps of our investigation.

g *Practical wisdom*: it is a kind of super-virtue essential to make each virtue effective. It enables the virtuous agent to acknowledge and respond properly to the items in the field of the research practice, choosing the appropriate means for their own ends. (McDowell 1979).

Our thesis is that these virtues are those traits that permit to acquire the internal goods of research practices. We also argue that the link between virtue and internal goods is not instrumental but conceptual: internal goods are not understandable or achievable except through the exercise of the virtues mentioned above. The situation is different for external goods. Even if the possession and exercise of the virtues by researchers can allow them or the institution in which they work to obtain them, this also depends to a considerable extent on other factors. In particular, by the institution's relations with other companies and organisations and by its ability to communicate and sell its results externally (Scott 2003).

Corroboration of our perspective

The framework proposed in this paper may be corroborated by considering recent comprehensive surveys on personal values (Sagiv et al. 2017) and other studies that attempted to describe what is research and research performance.

Sagiv et al. (2017) propose a comprehensive review of the numerous existing studies on personal values, integrating different streams of research in psychology, sociology, management and political science. In their study, Sagiv et al. (2017) state: "*Individuals act in ways that allow them to express their important values and attain the goals underlying them.* Thus, understanding personal values means understanding human behaviour (Sagiv et al. 2017). In contrast to the numerous studies investigating the consequences of values, much less is known about the origin of values." In particular, Table 1 of Sagiv et al. (2017) provides a series of value definitions that may be attributed to the list of virtues we proposed in the previous section.

The framework we proposed is also in agreement with recent research on the main characterization and dimensions of the research activities developed in Åkerlind G.S. (2008) and Bazeley (2010).

Conclusions and further research

The conceptual framework developed in this paper allows us to identify (define) the *good* evaluation as the evaluation that is able to discriminate between good and bad research practices. Having characterized the research practice through "internal" and "external" goods, offers us the possibility to deepen our understanding about what is a good research practice and what is the role played by researchers in it. Reflecting on what is a "good" research practice, on what is the role of researchers in it, according to the typologies of researchers we propose (leader, good and honest researcher), may be extremely useful for many reasons.

Firstly, it offers a *self-assessment tool* for researchers, to understand the functions of their research activities, their motivations and where they are in their research practice. This is an important step towards the *improvement* of research practices and individuals involved in it.

Secondly, it helps institutions to collect and *describe* the main functions of the research practices (highlighting their special features) developed by its researchers, and their motivations, to include them in their *strategic plan*. This is a further important step for the *development* and *improvement* of the organizations involved.

To conclude, a "good" evaluation should take into account all the building constituents of a "good" research practice and should be able to discriminate between good and bad research practices, while *enforcing* the functions of good research practices.

These reflections, although at their infant stage, may be the starting point for a *paradigm shift* in the evaluation of research practices. From an evaluation focused on *products* towards an evaluation focused on the *functions* of research practices. This new way of evaluate might also contribute to improvement of the research practices itself, stimulating new innovative solutions thanks to the self-assessment of the research community, providing clearer views of the strategy, missions and functions of the groups involved in the research practices.

Acknowledgments

The financial support of the Italian Ministry of Education and Research (through the PRIN Project N. 2015RJARX7) and of Sapienza University of Rome (through the Sapienza Awards no. PH11715C8239C105) is gratefully acknowledged.

References

Åkerlind, G.S. (2008). An academic perspective on research and being a researcher: An integration of the literature, Stud. High. Educ. 33(1), 17–31.

Ardal, P. S. (1966). *Passion and Value in Hume's Treatise*, Edinburgh, Edinburgh University Press.

Baier, A. (1991). *A Progress of Sentiments: Reflections on Hume's Treatise*, Harvard, Harvard University Press.

Batson, C. D. (2017). The Empathy-Altruism Hypothesis: What and So What? In Emma M. Seppälä, Emiliana Simon-Thomas, Stephanie L. Brown, Monica C. Worline, C. Daryl Cameron, and James R. Doty (eds.), *The Oxford Handbook of Compassion Science*, Oxford, Oxford University Press.

Bazeley, P. (2010). Conceptualising research performance, Stud. High. Educ. 35(8), 889–900.

Churchland, P. M. (1998). Toward a cognitive neurobiology of the moral virtues. *Topoi* 17 (2):83–96.

Cohon, R. (2008). *Hume's Morality: Feeling and Fabrication*, Oxford – New York, Oxford University Press.

Daraio C. (2017), A framework for the assessment of Research and its Impacts, *Journal of Data and Information Science*, Vol. 2 No. 4, 2017 pp 7–42.

Davidson D. (1980). *Essays on Actions and Events*, Oxford, Oxford University Press.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PloS One, 4(5), e5738.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, 109(42), 17028-17033.

Furner, J. (2014). The ethics of evaluative bibliometrics, in Cronin, B., & Sugimoto, C. R. (Eds.). *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, Boston, MIT Press. 85-107.

Hormann S. (2018). Exploring Resilience: in the Face of Trauma. *Humanistic Management Journal* 3 (1): 91-104.

Hursthouse, R. (1999). *On Virtue Ethics*, Oxford, Oxford University Press.

MacIntyre, A. (1981 first ed., 1985). *After Virtue*, London, Duckworth.

McDowell, J. (1979). Virtue and Reason. *The Monist* 62 (3):331-350.

Merton, R. K., (1973). *The Sociology of Science. Theoretical and Empirical Investigations*, Chicago, The University of Chicago Press.

Nussbaum, M. (2006). *Frontiers of justice: disability, nationality, species membership*, Chicago, Belknap Press.

OECD (2015), Frascati Manual 2015, Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities (OECD, Paris 2015).

Sagiv, L., Roccas, S., Cieciuch, J., & Schwartz, S. H. (2017). Personal values in human life. *Nature Human Behaviour*, 1(9), 630.

Swanton, C. (2007). Virtue ethics, role ethics, and business ethics. In Rebecca L. Walker & P. J. Ivanhoe (eds.), *Working Virtue: Virtue Ethics and Contemporary Moral Problems*, Oxford – New York, Oxford University Press.

Tangney, J. P. (1999). The self-conscious emotions: Shame, guilt, embarrassment and pride (pp. 541–568). In Tim Dalgleish & M. J. Powers (eds.), *Handbook of Cognition and Emotion*, New York, Wiley.

Vaccari, A. (2012). *Le etiche della virtù. La riflessione contemporanea a partire da Hume*, Firenze, Le Lettere.

The effects of research policies on the management of research information in HEIs: evidence from Germany

Sophie Biesenbender[1] and Christoph Thiedig[2]

[1] *biesenbender@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Schützenstraße 6a, 10117 Berlin (Germany)

[2] *thiedig@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

The effects of research policies on research information management (RIM) in higher education institutions (HEI) are poorly studied, despite their practical and scientific relevance. Both, science governance and science studies require valid and quality-assured research information (RI), that is data and information on the processes, output and impact of research (e.g. with regard to staff, projects, publications, or patents). This paper presents first results from a comparative project on the patterns, developments and effects of research information governance in four European science systems. By focussing on the German case, we analyse the processes behind the institutional implementation of national research information policies, the rationales and factors affecting institutional prioritization as well as organizational and technical consequences. Survey evidence on the implementation of the German "Research Core Dataset" (RCD) illustrates that institutional implementation of the voluntary RI standard is currently being driven by internal considerations in HEIs and the strategic prioritization of RIM and less so by external factors (such as e.g. data requests in the RCD standard).

Introduction

Information on the processes, output and impact of research (in the following referred to as 'research information' – RI) is central to a number of actors and stakeholders of the science system. Policy making and governance at all levels (e.g. at the level of governments or research institutions) require both qualitative and quantitative information and data on staff, projects, publications, patents etc. at different levels of aggregation (e.g. research institutions or academic disciplines). In addition, research information is essential for analyses of the science system, its processes and performance. Transformed into STI indicators, the availability and quality of research information crucially influences the robustness and generalizability of scientometric analyses and their results. In this context, research institutions in many science systems constitute a central source of research information. They are important data providers and complement third-party publication or project databases.

In many science systems, we currently observe public policies to – directly or indirectly – regulate and professionalize the institutional management of research information and its reporting. These range from binding national assessment exercises to voluntary reporting standards.

Despite differences in terms of policy scope, obligation or strategic importance, these policies – once implemented – will have an impact on how research institutions collect, manage and report research information. While many science policies have been analysed in terms of e.g. behavioural effects in the scientific community or their impact on scientific excellence (see Rijcke et al. (2016) for a recent review), the consequences with regard to the handling and management of research information and – by extension – indicator quality have not been systematically studied so far.

This paper is part of an ongoing project to fill this gap and to study the effects of public RI policies on the institutional management of research information in higher education institutions (HEI). It is guided by the following question: What are the effects of public research policies on processes of institutional information management in HEIs? The analysis is part of an ongoing comparative project on the patterns, developments and effects of RI governance in four European science systems (Germany, Great Britain, Italy and the Netherlands). The project combines qualitative interviews with desk research and survey material. The empirical analysis presented in this paper focuses on the so-called "Research Core Dataset" (RCD) for the German science system – a voluntary definitional and reporting standard for research information (Biesenbender & Hornbostel, 2016). It reports evidence of a recent survey carried out among publicly funded German higher education institutions to assess the institutional effects of the RCD.

This research-in-progress paper is structured as follows. In the following section, we provide a discussion of the context and state-of-the-art. The next section includes a case description, and introduces the conceptual approach as well as the methods of analysis. The subsequent section presents the empirical results, followed by a short discussion and conclusion.

Context and state-of-the-art

In most science systems, research and higher education institutions are currently faced with a multitude of external and internal information requirements: publicly funded organizations usually have to report on their research staff, projects, publications, patents etc. towards governments and agencies (e.g. statistical offices) as well as to private actors (for e.g. rankings or comparative analyses) (Biesenbender & Hornbostel, 2016). In addition, research information is of strategic relevance also for internal purposes, such as the implementation of performance-based funding instruments (see Bryant et al., 2018; Hicks, 2012; Liefner, 2003; Orr, Jaeger, & Schwarzenberger, 2007).

Next to other important sources of research information (e.g. project repositories of funding organizations, publication databases by private companies), data and information collected, processed and reported by research institutions are central sources for assessments and evaluations of institutional and the science system's development and performance. In some science systems, such as the Netherlands, they are used for national research portals. In other cases (such as Great Britain or Italy) institutionally reported data are necessary for carrying out national evaluation schemes (Biesenbender, Petersohn, & Thiedig, 2019). Last, also science studies scholars often directly or indirectly rely on these data sources (e.g. by using official statistics, institutional publication records or project registers).

As a consequence of the growing relevance of institutional research information and the challenges associated with their collection and handling, research information management (RIM) systems and practices have increasingly become strategic aspects in the organization and steering of research institutions (Bryant et al., 2018, p. 45) The requirements of external and internal data requests have an impact on institutional data collection and management processes, which in turn crucially determine the quality of reported information and, hence, the robustness of evaluation and assessment results. In the recent past, these processes have received increasing political attention (see e.g. Butler, 2008; Geuna & Piolatto, 2016; Wilsdon et al., 2015). Moreover, horizontal networking has become an important activity for practitioners and data providers (e.g. through the non-for-profit association euroCRIS or initiatives like the CERIF standard; Vancauwenbergh, 2017).

Despite these dynamics, the governance of research information (from a policy-analytical perspective) is a poorly studied issue in the literature (but see Daraio et al. (2019)). There is only anecdotic evidence on the effects of RI policies on the institutional level. We do not know under what circumstances research institutions invest in the professionalization of RIM and how such processes can be steered by public policies. In sum, the way in which science policies influence the institutional handling of research information is an issue that has received little scholarly attention in the literature so far (Biesenbender, 2019). Therefore, this contribution offers, first, a classification of RI policies with regard to their provisions for institutional implementation. Second, we look at the effects of public RI policies on institutional practices to manage and report institutional research information by focussing on a recent survey on the Research Core Dataset for the German science system.

Empirical case, conceptual frame, and methods

Science systems differ with regard to the approaches to directly or indirectly regulate research information management at the institutional level and to affect comparability of data across research institutions. In some science systems, this occurs through the establishment of national information systems or research portals or mandatory central evaluation schemes to be implemented through standard software. Others rely on voluntary definitional and reporting standards to harmonize the variety of information requests. In this paper, we employ a broad definition of RI policy. It includes any government policy that affects the institutional collection, processing and/or reporting of research information. This may refer to e.g. institutional adjustments in organizational structures, procedures (to collect, process, or report information) or software We suggest that public RI policies be broadly classified along two basic dimensions: First, their implications with regard to technical issues – Does the policy require or favour the implementation of specific software in research institutions (e.g. Current Research Information System – CRIS)? Second, financial relevance – Does the policy entail any financial implications for research institutions?

The ongoing project and this paper focus on national RI policies in four European science systems: Italy, the Netherlands, the UK and Germany.

The Italian VQR (short for *Valutazione della qualità della ricerca*) is a national evaluation instrument to assess the quality of academic 'products' of public research institutions. Depending on the field of research, the evaluation applies either bibliometric or peer-review instruments. VQR requirements have led the majority of participating research institutions to implement IRIS – a customized CRIS solution supporting both the management and reporting of institutional research information (Biesenbender, 2019).

In the Netherlands, the National Academic Research and Collaborations Information System (NARCIS) provides a central information portal covering a wide range of research information harvested from distributed repositories and institutional CRIS (Jippes, Steinhoff, & Dijk, 2010). Dutch CRIS development has been closely linked to national RI policy measures since the 1990s, most notably the Standard Evaluation Protocol (SEP) assessment exercise, which has since resulted in a very homogenous CRIS landscape (Galimberti & Mornati, 2017).

In the UK, the Research Excellence Framework (REF) assessment exercise has arguably been the main driver of institutional CRIS development. In its 2014 iteration, scientific output, 'impact' and the environment supporting research have been assessed using a peer-review process informed by quantitative indicators – prominently, bibliometric data (Traag & Waltman, 2018). The importance of REF performance for institutional funding prompted HE institutions to adapt their RI infrastructure in order to more efficiently fulfil REF reporting requirements.

Finally, the Research Core Dataset for the German science system represents the 'softest' example of RI policies under consideration. As a voluntary reporting standard without any financial implications for research institutions, it does not focus on the technical or procedural implementation in research institutions. The RCD is a definitional and reporting standard. Developed between 2013 and 2015 with broad involvement of actors and stakeholders of the science system, it provides definitions and aggregation rules for information on staff, early-career researchers, third-party funding, publications, patents, and research infrastructures. Implementation takes place on a voluntary basis at institutional level. Both, data owners (research institutions) and data-requesting organizations (e.g. funders) are asked to adapt their research information collection and dissemination processes in accordance with RCD specifications. In contrast to the UK, the federal governance structure of Germany does not make top-down RI governance feasible. As a reporting standard, the RCD concerns aggregates or lists of RI only, e.g. the sum of full-time equivalents at the reporting institution or records of publications of a given organizational unit. In order to cover a variety of reporting contexts, the RCD offers a number of parameters to be used to further differentiate the aggregate data. In addition, a (non-mandatory) reference data model is provided in order to support institutional and technical implementation (Biesenbender & Hornbostel, 2016).

Table 1. Typology of national research information policies

		Technical implementation	
		Prespecified	*Flexible*
Financial relevance	Yes	Italy (VQR)	Great Britain (REF)
	No	Netherlands (NARCIS)	Germany (RCD)

Table 1 classifies the four research information policies according to their financial and technical provisions for the four countries under study. We expect that the two dimensions of research information policies affect institutional implementation processes. Project results suggest that financial consequences accelerate institutional prioritization and implementation of the RI policy, while technical specifications regarding institutional implementation promote the harmonization of the science system's CRIS landscape (Biesenbender et al., 2019).

Notwithstanding, we know little about the mechanisms behind the institutional implementation of national RI policies, the rationales and factors affecting institutional prioritization as well as organizational and technical processes (Rebora & Turri, 2013). Yet, we assume that these aspects have an impact on the eventual quality of the data with regard to (a) definitional correctness and (b) reliability over time. Therefore, institutional implementation of RI policies is the focus of the ongoing empirical analysis. This paper focuses on the German country case by taking a look at the implementation of the Research Core Dataset. In a full population survey amongst public higher education institutions in Germany (N=190) carried out in April 2019, representatives responsible for institutional RI management were asked to specify among other things (a) the degree of and motivation for institutional implementation of the RCD and (b) how RCD implementation relates to ongoing activities with regard to the implementation and/or use of institutional CRIS.

We expect institutional implementation being driven by institutional considerations regarding the (potential) usefulness of the standard for external and internal reporting and favourable opportunity structures for reforming technical systems. We distinguish between different types of motivation for institutional RCD implementation: related to (a) external reporting and use scenarios (e.g. data requests in the RCD standard by third parties), (b) internal processes and use scenarios (e.g. changes in the processing and management of research information), and (c) political reasons.

Results

In the survey, representatives of all publicly funded HEIs in Germany (one contact per HEI) were asked to specify the current status of the RCD standard in the institutional handling of research information. Representatives from 94 out of 190 German HEIs took the full survey. Of these, 38 respondents indicated that they were planning to introduce the RCD standard.

Table 2 suggests a relationship between the planning or implementation of institutional CRIS and decisions to introduce the RCD in HEIs. Out of 61 HEIs that are currently planning or implementing CRIS, 30 also implement the RCD. Yet, only 5 out of 21 do so in institutions with institutional CRIS already in use. The majority of RCD-implementing institutions (29 out of 38) explicitly state that the two processes – RCD implementation and CRIS implementation – take place in the same context.

Table 2. RCD implementation in German HEIs

	CRIS				*Total*
	in use	*in implementation*	*in planning*	*not planned*	
RCD implementation	5	16	14	3	38
No RCD implementation	16	13	18	9	56
Total	21	29	32	12	94

Being asked to specify the reasons behind the institutional RCD introduction, the three most often indicated aspects reveal mostly internal motivations (see Table 3): 24 respondents indicate that the parallel introduction or adaption of institutional CRIS plays a central role for the decision for the RCD standard; 26 respondents refer to the expected improvement of institutional RI management and reporting; 22 respondents mention expected benefits in terms of outreach and institutional PR. In this context, external motivations (such as e.g. the use of the RCD standard in information requests by funders) seem to be less decisive.

Table 3. Reasons for RCD implementation in HEIs

		Frequency (out of 38)
Political:	Recommendations by the Science Council	21
External:	Use of RCD by funders	15
	Benchmarking possibilities	11
	Implementation of RCD at other institution(s)	7
Internal:	Expected improvement of RI management and reporting	26
	Introduction/adaptation of institutional RIS	24
	Expected improvement of institutional outreach and PR	22

Discussion and conclusion

The institutional implementation of national RI policies, the rationales and factors affecting institutional prioritization as well as organizational and technical consequences are complex and rarely studied processes. Yet, we assume that the institutional handling and management of research information are of high scientific and practical relevance.

In this paper, we presented an analysis of the institutional implementation of the Research Core Dataset – a standard specification for research information for the German science system and – thus – a particular type of RI policy. The RCD is of little immediate financial relevance for research institutions. In addition, it is flexible (unspecific and nonbinding) with regard to its technical implementation. We expected the degrees of institutional implementation and allocation of internal resources to mainly depend on the perceived strategic usefulness of the RCD for HEIs. The empirical findings highlight the current importance of considerations regarding the usefulness of the standard in HEIs for internal processes (e.g. professionalization in institutional RIM) and the role of favourable opportunity structures (such as the parallel implementation of institutional CRIS). The findings indicate that – so far – the RCD standard has hardly been established by external stakeholders (e.g. research funders requesting RI reports). This might explain the reservation regarding the RCD of those HEIs that have effective and functioning CRIS in place. Despite favourable opportunity structures for RCD implementation in many HEIs, which are currently in the

process of professionalizing institutional RIM, we expect that success and sustainability of the RCD standard will eventually depend on external factors in the mid and long term.
The results of this paper are (yet) of small generalizability beyond the type of RI policy studied (see Table 1). Further research will be needed to explore different types of research information policies and their effects on the information management procedures in research institutions.

Acknowledgments

Research for this paper is funded by the German Federal Ministry of Education and Research and the 16 federal governments (project period: 2017-2019; grant number: KDS2016).

References

Biesenbender, S. (2019). The governance and standardisation of research information in different science systems: A comparative analysis of Germany and Italy. *Higher Education Quarterly, 73*(1), 116–127.

Biesenbender, S., & Hornbostel, S. (2016). The Research Core Dataset for the German science system: Challenges, processes and principles of a contested standardization project. *Scientometrics, 106*(2), 837–847.

Biesenbender, S., Petersohn, S., & Thiedig, C. (2019). Using Current Research Information Systems (CRIS) to showcase national and institutional research (potential): research information systems in the context of Open Science. *Procedia Computer Science, 146*, 142–155.

Bryant, R., Clements, A., Castro, P. de, Cantrell, J., Dortmund, A., Fransen, J., et al. (2018). *Practices and Patterns in Research Information Management: Findings from a Global Survey.*

Butler, L. (2008). Using a balanced approach to bibliometrics: Quantitative performance measures in the Australian Research Quality Framework. *Ethics Sci Environ Polit (Ethics in Science and Environmental Politics), 8*, 83–92.

Daraio, C., Heitor, M., Meoli, M., & Paleari, S. (2019). Policy turnaround: Towards a new deal for research and higher education. Governance, evaluation and rankings in the big data era. *Higher Education Quarterly, 73*(1), 3–9.

Galimberti, P., & Mornati, S. (2017). The Italian Model of Distributed Research Information Management Systems: A Case Study. *Procedia Computer Science, 106*, 183–195.

Geuna, A., & Piolatto, M. (2016). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while). *Research Policy, 45*(1), 260–271.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*, 251–261.

Jippes, A., Steinhoff, W., & Dijk, E. (2010). NARCIS: research information services on a national scale'. In *The 5th International Conference on Open Repositories (OR2010)* (pp. 6–9).

Liefner, I. (2003). Funding, resource allocation, and performance in higher education systems. *Higher education, 46*(4), 469–489.

Orr, D., Jaeger, M., & Schwarzenberger, A. (2007). Performance-based funding as an instrument of competition in German higher education. *Journal of Higher Education Policy and Management, 29*(1), 3–23.

Rebora, G., & Turri, M. (2013). The UK and Italian research assessment exercises face to face. *Research Policy, 42*, 1657–1666.

Rijcke, S. de, Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation, 25*(2), 161–169.

Traag, V. A., & Waltman, L. (2018). Systematic analysis of agreement between metrics and peer review in the UK REF. *ArXiv e-prints (2018), arXiv:1808.03491[cs.DL]*.

Vancauwenbergh, S. (2017). Governance of Research Information and Classifications, Key Assets to Interoperability of CRIS Systems in Inter-organizational Contexts. *Procedia Computer Science, 106*, 335–342.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., et al. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.*

DataCite as a Potential Source for Open Data Indicators

Jonathan Dudek[1], Philippe Mongeon[2], and Josephine Bergmans[1]

[1] *j.dudek@cwts.leidenuniv.nl, j.e.bergmans@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333AL
(The Netherlands)

[2] *philippe.mongeon@ps.au.dk*
The Danish Centre for Studies in Research and Research Policy (CFA), Aarhus University, Bartholins Allé 7,
8000 Aarhus C (Denmark)

Abstract

Evaluating the impact of sharing research data is essential for comprehending the value of such initiatives in the context of Open Science. This study investigates indicators for both the output and the impact of datasets listed in DataCite. Based on metadata available for a single datacenter and research institute from the ocean sciences, the French IFREMER, originators and (re)users of datasets were collected at the levels of publishers, author affiliations, and authors. The results show that for the indicators considered, the metadata obtainable from DataCite is limited in consistency and completeness and does not allow facilitated comparisons of datasets. Consequently, meaningful and comprehensive insights are not easily generated at this point of time. In regard to measuring the (re)use of datasets, we suggest more sophisticated approaches to pursue in the future.

Introduction

Datasets are important scientific records. Making them accessible for broader audiences not only serves the reproduction of scientific findings but allows conducting further research as well. Finally, datasets can be considered a complimentary form of scientific output. In order to know whether such potential is exploited requires insights into how visible, findable, and traceable datasets are. Measuring the production and sharing as well as the (re)use of datasets, metadata plays a crucial role. Metadata are records created for datasets by the entities storing, collecting, or cataloguing them. Accordingly, an investigation of the metadata to be found in current data infrastructures can reveal how consistently and completely this information is provided, and how well datasets thus are comparable. Here, we focus on DataCite as a source of dataset metadata and use a bibliographic database (Scopus) to identify formal citations to these datasets in the scientific literature.

DataCite is an international non-profit consortium established in 2009 and combines the efforts of public research institutions, funding bodies and publishers towards open research data. The central value brought about by DataCite is to provide an infrastructure for data producing entities to assign persistent identifiers (DOIs, digital object identifiers) to datasets. Alongside DOIs, additional information on datasets is being attributed as metadata and retrievable from DataCite. (“Our Mission”, n.d.) As of January 2019, DataCite has listed over 16 million data records, with more than 13 million records enhanced by searchable metadata (“DataCite Statistics”, n.d.). How valuable is this metadata for a) understanding the origins of datasets, and b) creating links to other forms of scientific output? Approaching this question, we apply a case-study-like procedure, focusing on metadata for datasets from one single data originator. In doing so, we test two different kinds of indicators: *output* indicators and *impact* indicators. The former aim at obtaining an overview of the variety of contributors to datasets covered in DataCite. The latter investigate the frequency of dataset (re)use and the overlap between creators of datasets and (re)users. Following this step, we evaluate DataCite metadata based on how well those indicators reflect the insights sought for.

Data Sources

Each dataset recorded with DataCite originates from a so-called datacenter. Datacenters are not necessarily the entities exclusively dedicated to preserving data. Instead, the term subsumes data repositories as well as libraries, research centers, and publishers. For this study, we selected a datacenter from the ocean sciences, a field in which research data plays an important role. In addition, the datacenter selected should show some indication of data (re)use (i.e. references to the datasets in the scientific literature). A preliminary inquiry had shown that datasets by the *Institut Français de Recherche pour l'Exploitation de la Mer* (IFREMER) received the most citations of all ocean science datacenters. Hence, we selected it. IFREMER is a French research institute that manages oceanographic databases and designs and implements tools for the observation, experimentation and monitoring of the marine environment. It addresses societal challenges around climate change effects, marine biodiversity, pollution prevention, and seafood quality, and allows the scientific community to have access to the development, management and distribution of large research infrastructures, such as fleets, computational resources, testing facilities, and laboratories. ("The Institute", 2018)

We collected all 186 IFREMER datasets included in the CWTS version of DataCite, which dates to April 2018. As a second source, metadata for IFREMER-datasets was retrieved in manual searches from the repositories those datasets can be accessed at online, following their DOIs. This provided additional data on affiliations of authors of datasets, which are not included in metadata directly obtainable from DataCite. For a detailed discussion of metadata provided by DataCite, we refer to Robinson-Garcia et al. (2017). The IFREMER-datasets in our sample were registered with DataCite beginning in 2014; for 134 (72%) of the datasets metadata is provided in English; metadata for the remaining 52 (28%) datasets is in French.

Indicators

Measuring output

The datasets observed originate from several different entities, which varied depending on the source the datasets were extracted from, i.e. the publishing organisations. Among the points of origin, there are *publishers, authors, principal investigators, custodians, originators, resource providers*, and *affiliations*. However, not all datasets have all those entities assigned. Metadata in French returns even more terms. We focused on three points of origin: *authors*, *affiliations* (of authors), and *publisher*.

Not all datasets originate from IFREMER directly. Instead, various publishers and data repositories act as intermediaries. One of the most pronounced institutions is SEANOE, a publisher of scientific data in the field of marine sciences. ("About SEANOE", n.d.) Altogether, 103 (55%) datasets originate from this publisher (See Figure 1.)

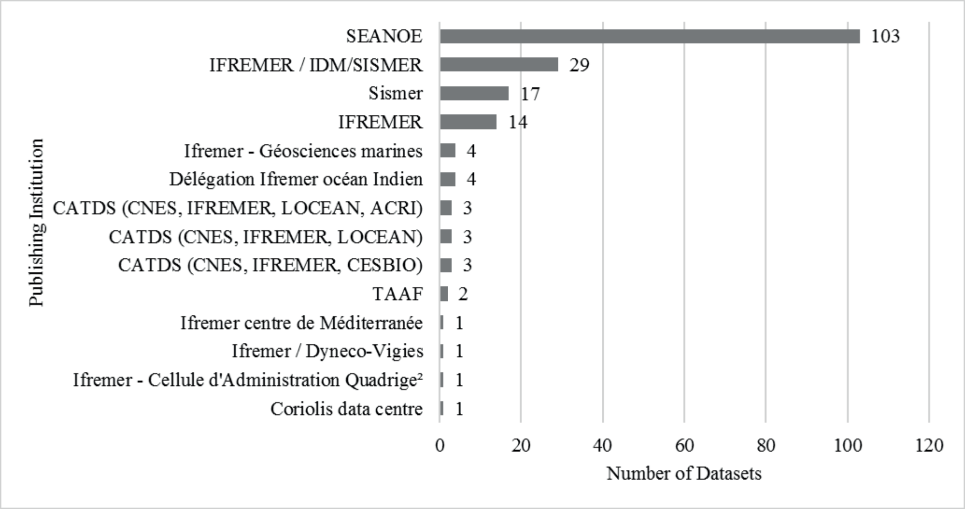


Figure 1: Number of datasets per publisher.

Authors are not necessarily affiliated with the institution serving as the publisher of a dataset. Since many datasets are results of team efforts, author teams with very mixed affiliation backgrounds can be observed. Unsurprisingly, IFREMER is the most prominent affiliation, with 133 authors affiliated to it or to a subsidiary organisation of IFREMER (see Figure 2 for the top ten affiliations of dataset authors).

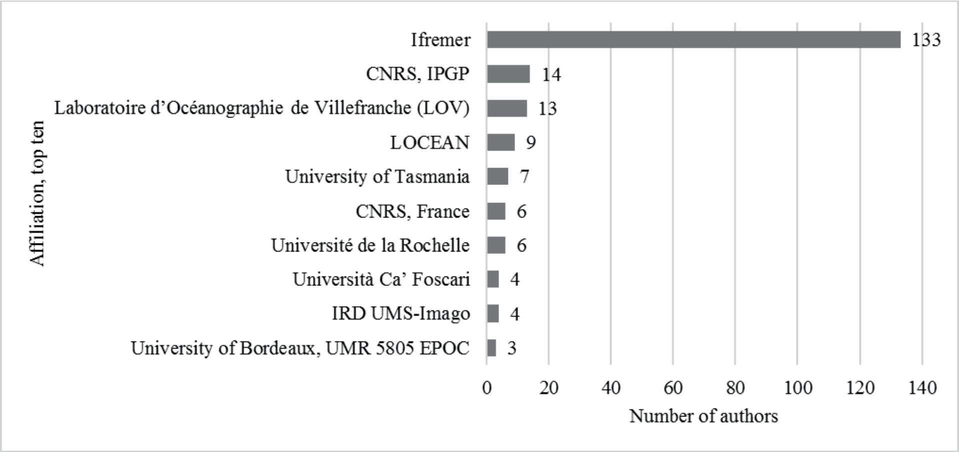


Figure 2. Top ten affiliations of authors.

The *author* field in DataCite usually contains individuals. However, there are 24 cases where organisations are listed as authors. In some of these cases, principal investigators are then provided additionally. As this is not consistently done, for the analysis of authors we focused on any entity called authors, i.e. both individuals and organisations, and did not replace institutional authors with principal investigators.

Accordingly, a total of 280 distinct authors can be identified for the datasets observed. Datasets usually are the result of several contributing investigators, with four authors per dataset on average. 71 datasets share at least one author with another dataset. At the same time, a few authors are highly prevalent, with three of them (co-)authoring more than 50 datasets.

Measuring impact

Regarding impact indicators, we sought empirical evidence of usage of IFREMER datasets by looking at the cited references of all documents indexed in the Scopus database. Overall, we identified 43 such references for a total of 12 distinct datasets. This shows that references to IFREMER datasets are quite rare. Furthermore, those few references are highly concentrated, with one single dataset out of the 12 cited datasets attracting 30 (70%) of all references. Previous work (Park, You, & Wolfram, 2018) has found that (re)used datasets are often not listed in the references, but rather mentioned in the articles' text or acknowledgements. A search for mentions of IFREMER datasets in abstracts of Scopus articles with the two keywords "dataset" and "IFREMER" returned 21 entries. The same keyword search in acknowledgements documented in the Web of Science returned 1,000 entries. This shows that there is a potential for discovering mentions of datasets in abstracts or acknowledgement texts of publications beyond formal citations in publications.

The second part of our investigation of impact aimed to provide an overview of dataset (re)users and their relationship with data producers/creators. In total, 208 different authors were found citing IFREMER datasets (our analysis is limited to formal citations), affiliated to 77 different research organizations. Figure 3 shows the top ten of those organizations.

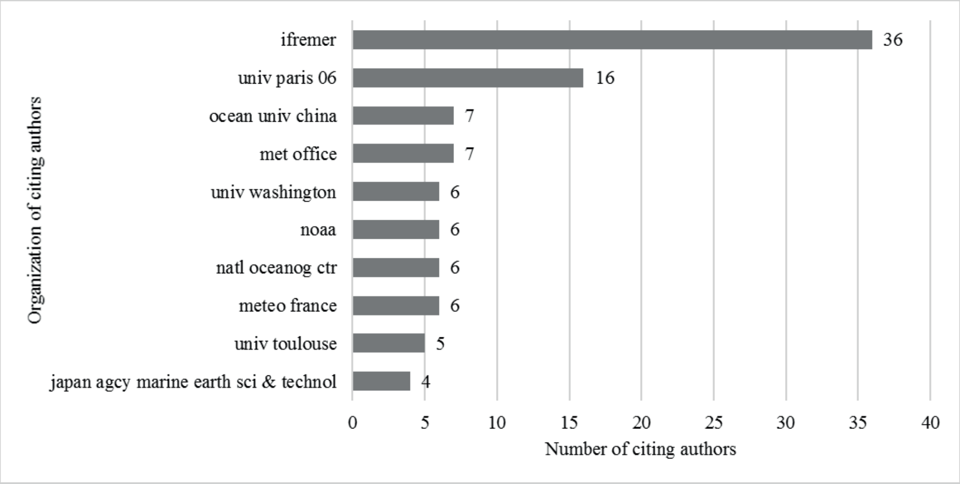


Figure 3. Top ten affiliations of authors citing datasets.

We found that, just like the data producers, the (re)use of datasets is highly concentrated: of all organizations serving as affiliations of citing authors, a small number is responsible for most of the identified instances of data (re)use. In this case, it is IFREMER leading the list, with a total of 36 affiliated authors (17% of all citing authors).

A further analysis investigated the overlap of authors of datasets and citing authors. Nine out of the twelve datasets cited share at least one author with the publication it is cited by; of the 208 unique citing authors, 31 (15%) are also authors of datasets.

From a copyright perspective, (re)use of datasets requires the permission to do so. Most datasets (67%) included in our sample are labelled with a Creative Commons (CC) license, establishing an indicator of potential (re)use. CC-licenses specify in which contexts and how intellectual work can legally be (re)used. ("About The Licenses", n.d.) For the remaining 33% of datasets, licenses are not explicitly stated; however, verbal statements on (re)use possibilities of datasets are provided in almost all cases. Figure 4 shows the share of datasets by license type and language; license types are ordered from the least restrictive (CC0) to the most restrictive (BY-NC-ND). Apparently, the extent to which datasets show a CC-license may partly depend on the language of origin, with 52% of datasets with French metadata having no license at all (compared to only 25% of datasets with English metadata).

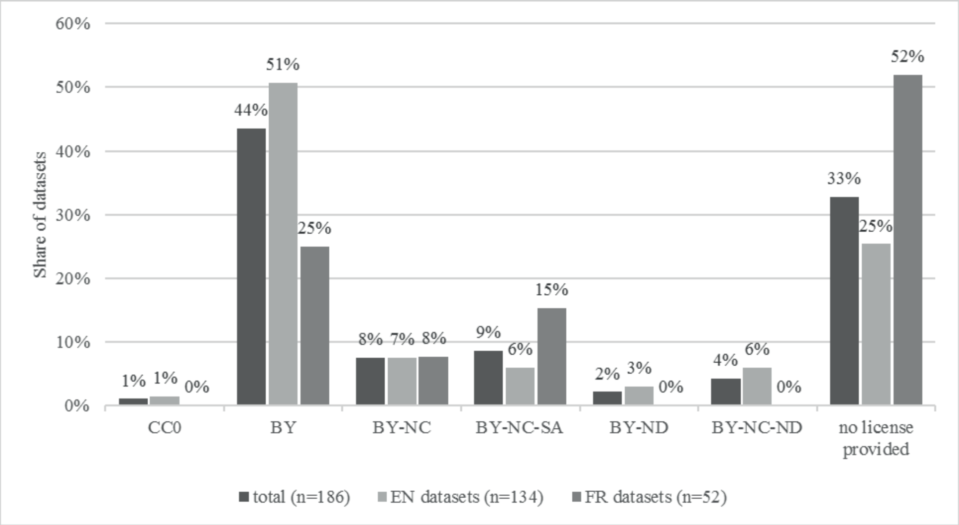


Figure 4. Shares of datasets per CC-licensing type and language of datasets.

Discussion

The study at hand reveals some of the intricacies of generating insights into the origins and the (re)uses of research data based on the metadata available from data infrastructures. Focusing on a subset of datasets originating from a selected datacenter listed in DataCite, we collected the publishers, the authors of datasets, and their affiliations. Further on, we investigated the impact of datasets by measuring counts of citations per dataset, the distributions of citing authors and their respective affiliations, and the overlap of authoring entities and citing entities. A final indicator of (potential) impact were CC-licenses assigned to datasets.

In the course of testing those indicators, the biggest challenge encountered is what we call a lack of metadata control. Herein, the necessity to extract metadata from different sources is a first hurdle: Metadata for the indicators devised is not entirely available from DataCite alone but requires querying publishers' repositories as well (next to a database like Scopus). Secondly, the metadata observed differed in how entities of origin are named and how they are listed, as well as how CC-licenses are assigned. This shows that with DataCite as a single point of access, information cannot be assembled sufficiently – even if only for the same datacenter. Instead, it appears necessary to consider metadata characteristics at the level of publishers.

In the light of the FAIR-principles of data sharing (Wilkinson et al., 2016), a dataset fulfills the requirement of *findability* by being listed in DataCite. However, in order to cover the full range of FAIR-principles for a given dataset (e.g., *reusability*), additional sources need to be included as well. For gathering and comparing datasets, this might constitute a considerable barrier: Depending on the scope of analysis and the point of entry – either starting e.g., an exploration of datasets in DataCite records, or in publisher's repositories; and either comparing datasets across publishers, or only those by a certain publisher – adaptations to different metadata can be necessary. When, as in our case, such dataset origins and usages are to be measured, this barrier becomes even more relevant.

We have shortly mentioned references to datasets beyond what can be found in reference sections (e.g., in abstracts or the acknowledgements) as a further means for estimating the (re)use of datasets. Providing respective metadata would be a worthwhile next step to pursue in addition to reporting citation metrics and serve a better understanding of (re)use, and hence, a dataset's potential for open use. However, both at the levels of DataCite, and of the publishers a consistent framework for reporting such information would need to be set in place. The urgency of this depends on the desirability of indicators of (re)use. Enabling a thorough evaluation of the opening and sharing of research data, though, does require such action.

Our investigation shows that output as well as impact indicators based on DataCite metadata alone do not represent a complete picture, necessitating caution in research evaluation. It should be noted, though, that this conclusion is limited as far as we have focused on one particular datacenter only, from one field of research only. Further research is needed into the data sharing practices of the whole of a scientific field, and then, also, regarding the comparison of different fields. With measures in place to track (re)use of datasets, broader and more general conclusions should become possible. Still, our work shows how the different sources of metadata (can) interact and currently need to be considered when evaluating the state of open data. With DataCite as a major infrastructure provider, fortunately, a central point for enhancing the visibility, comparability and traceability of research data exists. Thus, the necessary foundations for understanding better the origins and (re)uses of datasets may eventually be provided.

Acknowledgements

This study is part of the Open Science Monitor which is funded by the European Commission under grant number PP-05622-2017.

References

About SEANOE. (n.d.). Retrieved January 24, 2019, from https://www.seanoe.org/html/about.htm

About The Licenses (n.d.). Retrieved May 28, 2019, from https://creativecommons.org/licenses

DataCite Statistics. (n.d.). Retrieved January 24, 2019, from https://stats.datacite.org

Our Mission. (n.d.). Retrieved January 28, 2019, from https://www.datacite.org/mission.html

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology, 69*(11), 1346–1354. DOI: 10.1002/asi.24049

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics, 11*(3), 841-854. DOI: 10.1016/j.joi.2017.07.003

The Institute. (2018). Retrieved January 24, 2019, from https://wwz.ifremer.fr/en/The-Institute

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*. DOI: 10.1038/sdata.2016.18

Admitting uncertainty: a weighted socio-epistemic network approach to cognitive distance between authors

J. Hartstein[1]

[1] *hartstein@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

In bibliometric networks, uncertainty about the exact relevance of bibliographic information for network construction is a constant issue, which has to be addressed by modelling decisions. There is a lively debate about meaningfulness of authorship and citation as well as about fractionalization of network tie strengths to address this issue. This paper argues, that uncertainty is best expressed in probabilities, and provides a framework to conceptualize cognitive distance accordingly. Tie strengths are introduced with a foundation in basic probability theory and a distance based on multiplication of tie strengths alongside a path. Due to its high resolution, this approach is tailored for the exploration of networks between individual authors which are working in closely related research areas.

Introduction

One of the foundational challenges in network modelling and distance measurement is how to account for the uncertainty of bibliographic information in an optimal way. Uncertainty becomes even more problematic, when we zoom into a bibliometric network and the applicability of a “law of large numbers” becomes less plausible. In the following paper, we will show that there are possibilities to account for uncertainty already during the process of network construction. We will present a modelling approach for cognitive distance employing basic probability theory. We will also explain, that this is feasible with reasonable computational effort.

Our modelling approach is fitted for “zooming in”-distance measurement, i.e. distance measurement problems meeting the following criteria: distances to be measured are between (a) individual researchers, which are (b) working in the same or closely related research areas. This would mean, that we have (a) small publication portfolios, which are assumed to be (b) closely related.

The hereby presented concept is based on the ideas of author-bibliographic coupling (Zhao & Strotmann (2008)), which is often used to compute cognitive distances between individual researchers (e.g. Van den Besselaar & Sandström (2017), Wang & Sandström (2015)). The focal entities in such models are authors, their publications and the corresponding references.

We are integrating uncertainty by introducing a specific understanding of tie strength as a probability mass function. From tie strengths we construct path strengths by multiplication and summation. Finally, distance is defined as the logarithm of the additively inverse strength of the strongest path.

Socio-epistemic network construction

Cognitive distances between individuals or groups are frequently measured through distances between socio-epistemic network positions - usually positions in artifact-based social networks, where social entities are related to each other through artifactual entities. These artifact-based networks are usually obtained through secondary data analysis. Artifacts used for these network constructions include not only publications, but also knowledge bases (Broekel & Boschma

(2012)) and products (Balland et al. (2013)) and their respective attributes. For illustrative reasons, we construct such a network from authors, papers and references (see Figure 1).

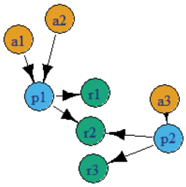


Figure 1. A bibliographic network consisting of authors ($a1, a2, a3$), papers ($p1, p2, p3$) and references ($r1, r2, r3$).

While the assignment of ties in such networks is determined through the connection between entities, the assignment of specific weights or strengths to these ties remains often unattended or at least unclear. Methodologically speaking, we should seek to identify basic requirements for the assignment of strengths to artifact-based socio-epistemic network ties.

Authorship tie strengths

The authorship network is a directed (bipartite) network with two types of nodes: authors and publications. The edges signify the authorship relationship. We give the network definition for clarity about the notation used hereinafter. For more details on network construction in general the reader is kindly asked to consult Havemann & Scharnhorst (2010), Newman (2001) or Barrat et al. (2004).

Definition. *[Unweighted authorship network.] Let $G_{ap} = (V_a, V_p, E_{ap})$ be the author-paper graph to given sets $V_a = \{a_1, \ldots, a_m\}$ of authors and $V_p = \{p_1, \ldots, p_n\}$ publications. The tuple (a_i, p_j) is in the set of edges E_{ap} if, and only if, author a_1 is in the list of authors of paper p_j.*

In the case of unweighted authorship, all authorship ties are equally strong. The upcoming question is, how to model unequally 'strong' relations between two entities.

Authorship ties in the case of certainty

Before define authorship-tie strengths in general, we will look at two minimal examples: the two-nodes graph with and without authorship edge. The first example network consists of just one author a_1 and one paper p_1, which are connected through authorship. This would mean

$$G_{ap} = (V_a, V_p, E_{ap}) \text{ with } V_a = \{a_1\}; V_p = \{p_1\}; E_{ap} = \{(a_1, p_1)\}.$$

If we assume, that the epistemic position of an author is expressed in her work, then we may deduce that the connection from author to paper is whole in this case. Also, if we assume, that the paper is an expression of the author's epistemic position (and only hers), then we may also deduce, that the connection from author to paper is 'whole' in this case. This is the 'strongest' possible relationship between author and paper which should be assigned a strength of 100%. This means, that the strength $s_{1,1}$ of the tie between a_1 and p_1 is:

$$s_{1,1} := 1.$$

We may also construct a network without edges as a minimal degenerate authorship graph:

$$G_{ap} = (V_a, V_p, E_{ap}) \text{ with } V_a = \{a_1\}; V_p = \{p_1\}; E_{ap} = \emptyset.$$

The non-authorship connection is the 'weakest' possible relationship between author and paper. Such non-ties should be assigned a strength of 0%, which would mean:

$$s_{1,1} := 0.$$

(Please note, that a tie strength of 0 is only theoretical, because in fact, those ties are not drawn at all.)

The tie strengths in this two-node graph examples are easily to be interpreted as probabilities in a situation of certainty. The probability, that a randomly chosen aspect (say, a specific reference) from a paper is to be accounted for by its single author, is assumed to be 100%. The probability, that the epistemic position of the author is reflected in a randomly chosen paper (out of one) is assumed to be 100%. The probability, that a randomly chosen aspect from a paper is to be accounted for by a person, who is not an author of that paper is assumed to be 0%. The probability, that the epistemic position of the non-author is reflected in a randomly chosen paper she has not authored, is assumed to be 0%.

Authorship ties in the case of uncertainty

When constructing authorship graphs from bibliographic information, we are often confronted with multiple authorships, so that the outdegrees of the author-nodes as well as the indegrees of the paper-nodes might be greater than 1. (This means, that there are multiple authors on one publication and multiple publications by one author.) In this case, we may assume that a randomly chosen aspect from a paper is to be accounted for by one specific author with less than 100% probability.

To account for collaborative authorship and authorship on multiple publications, we seek to assign weights to the edges of the authorship graph. Mathematically speaking, we seek to define a tie-strength function from the set of node pairs (V_a, V_p) to the interval of $[0;1] \subset \mathbb{R}$:

$$s: V_a \times V_p \rightarrow [0;1].$$

If we have no prior knowledge, the most reasonable thing is to assume equally distributed authorship accounts[1]. This resembles the idea of fractional counting as known from the context of citation-based performance indicators, but also for cognitive distance measurement. (Perianes-Rodriguez et al. (2016)). We fractionalize the authorship tie strength by the author count on a publication and the publication count of the author, respectively, because each of the papers contributes likewise to the author's oeuvre.

Definition. [Adjustment of authorship-tie strengths.] *Let paper p_j be authored by author a_i. We define the strength $s_{i,j}$ of the authorship tie between author a_i and paper p_j as the reciprocal product of the indegree of the paper-node and the outdegree of the author-node:*

[1] Expressed in terms of probability theory: A ist the set of all authors in the network, f is the probability of influence of a specific author on a specific aspect of the paper, $s: A \rightarrow [0,1]$ is defined by $f(x) = 1/n$, if x is in the list (with length n) of authors , and by $f(x) = 0$, if x is not in the list of authors. And therefore, $f(A) = \sum_{x \in A} f(x) = 1$.

$$s_{i,j} := \frac{1}{\deg_{out}(a_i)} \cdot \frac{1}{\deg_{in}(p_j)}$$

Please note that $s_{ij} \in [0,1]$ equals 1 if, and only if, author a_i is authoring paper p_j all by himself and it is his only publication.

The assumption of equal distribution does not hold in the light of the theoretical debates on authorship. The meaning of authorship, especially in cases of collaborative publications, is highly researched and disputed in the science studies community (e. g. in Katz & Martin (1997)~~)~~, Laudel (2002), Bozeman & Youtie (2016)), and also seems to be field specific (Johann & Mayer (2018)). Nevertheless, if we assume, that authorship as it is articulated on a paper has at least some meaning, we might assign a suitable probability mass function to the tie strengths, that fits the chosen (social) theoretical angle. As the concept of probabilistic tie strengths is easier to understand with equal distribution, we will stick to it for our model description here.

This does not mean, that the assignment of authorship is ever so easy. Fractionalization of authorship is discussed in the context of fairness of bibliometric evaluation, especially concerning author-level evaluation. While we do not believe, that performance evaluation is a suitable field of application for our model, it should be emphasized, that s is to be understood as a parameter (not a constant) of the hereby proposed model, which has to be chosen and justified in the specific context of research objects and questions, when the model is applied. The tie strength s only opens up a room for theoretically backed adjustments. Let's say, we assume, that articles written in the mother tongue of an author express more of their epistemic position, than others - we could just increase the tie strength to these papers at the cost of the other ones. Let's say, that first authors are more accountable for a paper than last authors - we could adjust the tie strength function accordingly.

Deducing co-authorship tie strength from authorship tie strength

The concept of unweighted co-authorship networks assumes symmetric authorship relations to papers. In the literature, co-authorship is believed to resemble collaboration, which is often interpreted as a sign of epistemic closeness. In such a network, co-authors are connected via the co-authored paper (see Figure 2) and distance is understood as path distance (see, for example, Newman (2004)).

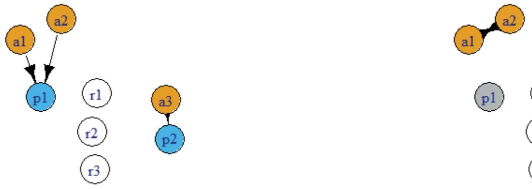


Figure 2. Left: Construction of an authorship network (the information about references is not used). Right: Construction of a co-authorship network (the information about papers is used in an aggregated form, the authorship ties are no longer present.)

In the case of weighted authorship, we need to define the strength of the ties between co-authors accordingly. We might require that the tie strengths between two authors, which are connected through a paper should be less or equal to each of the authors' tie strength to that paper. We use the concept of conditional probability and use the multiplication axiom. Therefore, the strength of the tie between two co-authors a_i and a_k induced by paper p_j is to be defined as

$$r_{i,k}^{j} := s_{ij} \cdot s_{kj}.$$

The overall strength of the tie between those co-authors would then be the sum of the individual ties to co-authored papers and to be defined as:

$$r_{ik} := \sum_{j=1}^{n} s_{i,j} \cdot s_{k,j}.$$

This is coherent with the summing up of branches in a tree diagram.

Bibliographic coupling

To connect authors, who never co-authored a paper, we start with drawing ties between the papers. To construct this paper-paper network, we use the concept of bibliographic coupling, which relies just on the reference portfolios of papers (Kessler (1963)). All of the other properties of the paper (content, title, publisher and authors) are not of interest in this instance of our model (see Figure 3).

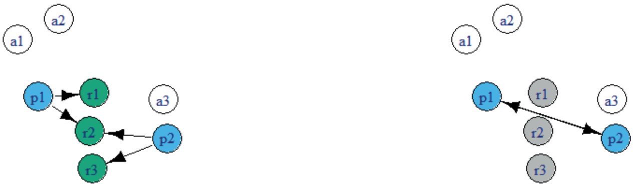


Figure 3. Left: Construction of a paper-reference network (information about authors is not used). Right: Construction of a bibliographic coupling network of papers (information about references is used in an aggregated form, the ties between papers and references are no longer present).

When bibliographic coupling is used in operationalisations of cognitive distance, ties are drawn between papers according to the similarity of their reference lists, if the similarity is larger than zero. Different types of similarities have been used and their appropriateness for different research purposes have been discussed.

The two most relevant similarities in this context are Salton's cosine similarity and the Jaccard-Index (see Leydesdorff (2008)), but other co-occurrence based similarity measures have been proposed and discussed (see van Eck & Waltman (2009)). Salton's cosine similarity relies on a vector space concept, where reference lists are interpreted as vectors in a finite dimensional vector space, and where the cosine of the angle between two reference list vectors is interpreted as similarity. The Jaccard-Index relies on set theory and expresses the shared references as the relation between the cardinalities (i. e. sizes) of intersection and union of two reference lists. In both cases, the similarity function maps to the interval [0,1] with 1 as maximum similarity (identity) and 0 as minimum similarity (non-similarity).

Similarity as probability mass function

We may interpret these two similarities as probability mass functions. The Jaccard-Index gives the probability that a randomly chosen reference from the union of reference lists is occurring in both reference lists. In the case, that both reference lists have the same length, Salton's Cosine gives the probability, that a reference randomly chosen from one of the lists, occurs on the other one.

With no loss of generality, we define here the strength of the paper-paper network tie between paper p_j and paper p_l as similarity $sim_{j,l} :=\sim_{salt} (r_j, r_l)$ of their reference lists r_j and r_l. We normalize the strengths $s_{j,1..n}$ of reference ties from paper p_j to the references $r_{j,1..n}$ by the (common) length of the reference lists: $s_{j,1..n} := \frac{1}{\deg_{out}(p_j)}$. Afterwards, Salton's cosine similarity may be noted as follows:

$$\underset{salt}{\sim} (p_j, p_l) = \sum_{i=1}^{n} s_{j,i} \cdot s_{l,i}.$$

Note: This is very similar to our definition of authorship strengths above.

In the next step, we define the strength $s_{i,k}$ of the author-bibliographic coupling tie between author a_i and author a_k induced by two papers p_j and p_l, they authored. Again, we use the concept of conditional probability and the multiplication axiom to do so:

$$s_{i,k}^{j,l} := s_{i,j} \cdot sim_{j,l} \cdot s_{k,l}.$$

Thus, we have conceptionally collated the author-paper network and the paper-paper network to define an author-author network tie. It covers the authorship uncertainty and the reference list similarity alike.

The normalization of tie strengths from papers to references does not necessarily require simple fractionalization. There are several theoretical angles to the meaningfulness of references and theoretical desiderata (see Moed (2006), Wouters (2018), Rafols (2018)). Also, Schubert et al. (2006) proposed to adjust referencing ties to account for self-citation, and Teplitskiy et al. (2018) have pointed out, that citation might "mean" something completely different. The hereby presented model opens up the possibility to adjust the referencing ties accordingly - to increase the tie strength of "more important" references at the cost of others.

Reference portfolios and author-bibliographic coupling tie strengths

Remember, that our networks considered so far consist only of two authors and two papers with corresponding reference lists. To depict larger networks, we have to extend these concepts again. As we assume that only the reference lists contain relevant information, we want to use from the papers in author bibliographic coupling, we need to build reference portfolios for each author. Therefore, we reduce the network by taking out the papers and defining an author-references relation based on the author-paper relation and the papers' reference portfolios (see Figure 4, left). This is state of the art for author-bibliographic coupling networks as may be seen in Wang & Sandström (2015).

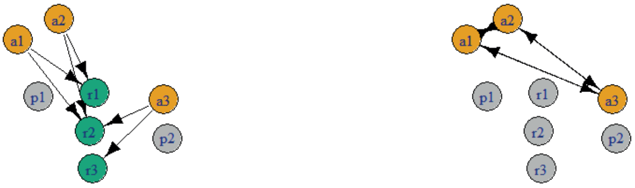


Figure 4: Left: Construction of author-reference portfolios (information about the papers is used in an aggregated form, but the ties from and to papers are no longer present). Right: Construction of an author-bibliographic-coupling network (information about papers and references is used in an aggregated form, but ties from and to papers or references are no longer present).

We use the tie strengths between author a_i and paper p_j and its references $r_{j,1..n}$ to construct ties between a_i and $r_{j,1..n}$ with strengths defined by multiplication (the arrow signifies that $\vec{s}$ is a vector of length n):

$$\overrightarrow{s_{a_i,r_{j,1..n}}} := s_{i,j} \cdot \overrightarrow{s_{j,1..n}}.$$

To construct the strength of a tie from an author to a single reference, we have to sum up all ties induced by all the different papers p_l.

$$s_{a_i,r} := \sum_l s_{a_i,r_l}.$$

(The tie strengths from author to reference induced by one paper are defined multiplicatively. Afterwards, the ties between author and reference induced by all papers together are summed up.)

From the now given bipartite author-portfolio relation, we seek to build an author-author network by taking out the reference portfolios and aggregating their similarity information into tie-strengths (by multiplication and summing up over all references r_x in our network):

$$s_{a_i,a_k} := \sum_x s_{a_i,r_x} \cdot s_{a_k,r_x}.$$

From tie-strength to cognitive distance

So far, we have constructed an author-author-network with weighted ties. To compute a path distance from this type of tie strengths is by far not trivial, because strength and weakness are not to be used synonymous to closeness and distance.

We want to make use of the full topological structure of the network using distance concepts from graph theory. So, pairs of authors, if they are not directly connected in our weighted author-bibliographic coupling network, are connected indirectly if they belong to the same graph component. (If they do not, the distance is infinite.)

We will define the strength alongside a path as the product of the strengths of the path's segments (the edges along a set of nodes), and make use of the multiplication axiom of conditional probability again. The path strength between authors a_i and a_k alongside the path via a single transit author a_t would then be:

$$s_t(a_i, a_j) := s(a_i, a_t) \cdot s(a_t, sa_j).$$

If there are several transit authors, we may denote the path P as an ordered set $P := (a_{t_1}, \dots, a_{t_p})$. The strength of the connection between a_i and a_k alongside path P would then be defined as:

$$s_P(a_i, a_k) := s(a_i, a_{t_1}) \cdot s(a_{t_p}, sa_j) \cdot \prod_{i=1}^{p} s\,(a_{t_i}, a_{t_i+1}).$$

Finding the strongest path

To algorithmically find the strongest path between two authors a_i and a_k in our weighted author-bibliographic coupling network, we have to do a trick. As we know shortest-path algorithms for additively defined paths, we use the logarithmic transition from multiplication to addition for positive real numbers a and b:

$$log(ab) = log(a) + log(b).$$

We make use of the fact, that all our tie-strengths s are positive real numbers and that the logarithmic function is strictly monotonically increasing. Thus, to maximize s_P means to maximize

$$log(s_P(a_i, a_k)) := log(s(a_i, a_{t_1})) + log(s(a_{t_p}, sa_j)) + \sum_{i=1}^{p-1} l\,og(s(a_{t_i}, a_{t_i+1})).$$

We now make use of our model's property, that all $s(a_i, a_k)$ are in [0,1]. It follows, that $log(s(a_i, a_k)) \leq 0$ in any case. Finally, this means, that to maximize $log(s(a_i, a_k))$, we need to minimize $|log(s(a_i, a_k))|$, which is simply a shortest path problem.

There are different algorithms in the graph theoretical toolbox to find shortest paths. Newman (2001) and Brandes (2001) use shortest-path algorithms to compute betweenness centrality in weighted collaboration networks. Both point out, that shortest paths are costly to compute in weighted networks. Two distinct but related problems in algorithmics are to consider here: the all pairs shortest paths problem (APSP-problem) and the single-source shorted paths problem (SSSP-problem). For both, there are algorithms for either the exact or the approximate computation of shortest paths. The range of computational costs is wide. Dijkstra's classic algorithm for the exact computation of APSP is of cubic ($\mathcal{O}(n^3)$) runtime (see Chan (2010)). The classic Bellman-Ford algorithm for the exact computation of SSSP is of linear runtime ($\mathcal{O}(n)$) (see Elkin (2017)). For the APSP, faster algorithms are known in the meantime - Chan (2010) gives a short overview. There is a vital research community working on those problems, e.g., Elkin (2017) provided a new algorithm for the SSSP of sublinear runtime. So, for small networks the computation is feasible, but this may not be the case for large networks.

Definition of cognitive distance and further considerations

We propose to interpret this logarithm of the additively inverse of the strength of the strongest path as cognitive distance. It is not limited to the interval [0; 1], therefore it enables exploration of socio-epistemic networks in high resolution. Also, in networks with equal distribution of edges, this distance can be expected to be approximately normally distributed (see Figure 5). It also fits the requirement, that co-authors are closer than non-co-authors (if all other properties are the same). With the adjusted tie strengths, we account for uncertainty of accountability in the case of multiple authorships and also for the uncertainty of epistemic representativeness of papers for authors with multiple publications.

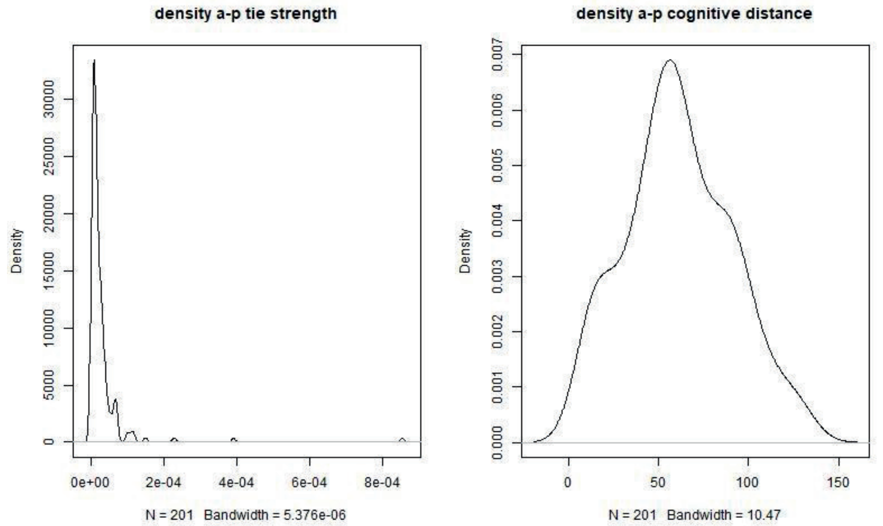


Figure 5. For illustration: density plots of path strengths and corresponding distances (without portfolios) from an authorship network of a population of reviewers in a peer review process. The data was taken from an ongoing project on "Evaluation Practices in Research and higher education."

In our depiction, we implicitly assume, that the 'whole' of a publication consists of 'parts', each of them stemming from exactly one of the authors, and the probability of each part belonging to a specific author is equally distributed between the authors. Also, we assume, that the entirety of the author's oeuvre (her epistemic position) is reflected in her publications and the probability of a specific publication reflecting the epistemic position is equally distributed between the publications. The conditional probability of a specific publication reflecting the epistemic position of a specific author is given as a product of the individual probabilities (following the multiplication axiom for independent choices as used in a tree diagram).

It should be emphasized, that the values of cognitive distance as defined above have no absolute "meaning" - they are only to be used comparatively in the same network. The potential of the hereby presented approach lies in the adjustability of the strength function to account for less uncertainty, if there occurs prior knowledge about the circumstances of paper production and authorship or the meaningfulness of references.

References

Balland, P. A., De Vaan, M., & Boschma, R. (2013). The dynamics of interfirm networks along the industry life cycle: The case of the global video game industry, 1987-2007. Journal of Economic Geography, 13(5), 741-765.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the national academy of sciences, 101(11), 3747-3752.

Bozeman, B., & Youtie, J. (2016). Trouble in paradise: Problems in academic research co-authoring. Science and engineering ethics, 22(6), 1717-1743.

Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.

Broekel, T., & Boschma, R. (2012). Knowledge networks in the Dutch aviation industry: the proximity paradox. Journal of Economic Geography, 12(2), 409-433.

Chan, T. M. (2010). More algorithms for all-pairs shortest paths in weighted graphs. SIAM Journal on Computing, 39(5), 2075-2089.

Elkin, M. (2017). Distributed exact shortest paths in sublinear time. arXiv preprint arXiv:1703.01939.

Havemann, F., & Scharnhorst, A. (2010). Bibliometrische Netzwerke. In Handbuch Netzwerkforschung (pp. 799-823). VS Verlag für Sozialwissenschaften.

Johann, D., & Mayer, S. J. (2018). The Perception of Scientific Authorship Across Domains. Minerva, 1-22.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration?. Research policy, 26(1), 1-18.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American documentation, 14(1), 10-25.

Laudel, G. (2002). What do we measure by co-authorships?. Research Evaluation, 11(1), 3-15.

Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. Journal of the American Society for Information Science and Technology, 59(1), 77-85.

Moed, H. F. (2006). Citation analysis in research evaluation (Vol. 9). Springer Science & Business Media. (pp. 193-220)

Newman, M. E. (2001). The structure of scientific collaboration networks. Proceedings of the national academy of sciences, 98(2), 404-409.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences, 101(suppl 1), 5200-5205.

Perianes-Rodriguez, A., Waltman, L., & van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics, 10(4), 1178-1195.

Ràfols, I. (2018). S&T indicators in the wild: Contextualization and participation for responsible metrics. Research Evaluation, 28(1), 7-22.

Schubert, A., Glänzel, W., & Thijs, B. (2006). The weight of author self-citations. A fractional approach to self-citation counting. Scientometrics, 67(3), 503-514.

Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. (2018, September). Why (almost) Everything We Know About Citations is Wrong: Evidence from Authors. In 23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands. Centre for Science and Technology Studies (CWTS).

Van den Besselaar, P., & Sandström, U. (2017). Influence of cognitive distance on grant decisions. Science, technology and innovation indicators 2017.

Van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology, 60(8), 1635-1651.

Wang, Q., & Sandström, U. (2015). Defining the role of cognitive distance in the peer review process with an explorative study of a grant scheme in infection biology. Research Evaluation, 24(3), 271-281.

Wouters, P. (2014). The citation: From culture to infrastructure. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact, 47-66.

Zhao, D., & Strotmann, A. (2008). Author bibliographic coupling: Another approach to citation-based author knowledge network analysis. Proceedings of the American Society for Information Science and Technology, 45(1), 1-10.

The link between research quality and technology transfer in the Italian Evaluation of Research Quality VQR 2011-2014

Brigida Blasi[1], Andrea Bonaccorsi[2], Carmela Anna Nappi[3] and Sandra Romagnosi[4]

[1] *brigida.blasi@anvur.it*
Italian National Agency For The Evaluation Of Universities And Research Institutes
Via Ippolito Nievo, 35 00153 Rome (Italy)

[2] *andrea.bonaccorsi@unipi.it*
University of Pisa
Largo Lucio Lazzarino, 56122 Pisa

[3]*carmelaanna.nappi@anvur.it*
Italian National Agency For The Evaluation Of Universities And Research Institutes
Via Ippolito Nievo, 35 00153 Rome (Italy)

[4] *sandra.romagnosi@anvur.it*
Italian National Agency For The Evaluation Of Universities And Research Institutes
Via Ippolito Nievo, 35 00153 Rome (Italy)

Abstract

In recent years, the growing interest of Universities in valorization of research activities (technology transfer - patenting and licensing- and academic entrepreneurship) has generated a debate on their impact on scientific knowledge production and a large literature has analyzed the determinants of researchers' engagement with firms, at individual and institutional level. Empirical studies have shown that researchers with high scientific performance impact tend to engage more in technology transfer activities: indicators of scientific impact/productivity have a positive relationship with indicators of engagement in transfer activities, both at the individual and institutional levels and through contract research as well as through commercialization (licensing and start-up creation).

This paper focuses on the link between research quality and valorization of research activities using data from the Italian Evaluation of Research Quality VQR 2011-2014. The estimation results confirm the existence of a positive correlation at institutional level.

Introduction

In last decades universities have been subjected to strong pressure for change, triggering a discussion on the inner nature of academic institutions. Today's university is recast as major player in the knowledge society and it has significant impact acting as economic engine for communities (Florida and Cohen 1999).

Universities' international reach provides an important pipeline for local and national economies to access the global networks of production and circulation of knowledge and their host cities benefits from a great social and economic footprint, through job creation and demand for services. This has produced an increasing emphasis on interacting externally and engaging in new relationships with non-academic domains (Etzkowitz 2003; Slaughter and Leslie 1997), what has been called third mission (Etzkowitz 1998).

New modes of knowledge production have been theorized in which universities underpin innovation and economic progress, embedded in a helix model that involves at first industry and government (triple helix in Etzkowitz and Leydesdorff 2000), then civil society (quadruple helix in Carayannis and Campbell, 2009) and more recently environment (quintuple helix in Carayannis et al., 2012).

Academic institutions and policymakers worldwide provide dedicated incentives at various levels to support the involvement of universities in technology transfer and commercialization of research (Mowery and Nelson 2004), with the result of an increasing propensity towards patenting (Nelson 2004; Stiglitz and Wallsten 1999) and licensing (Thursby et al. 2001), increasing numbers of university researchers engaging in academic entrepreneurship (Shane 2004), and the diffusion of technology transfer offices, industry collaboration support offices and science parks (Siegel et al. 2003).

However, research activity and technology transfer are driven by different motivations and incentives that can generate effects of substitution or complementarity between the two activities (Merton 1968; Mitroff 1974; Mulkay 1976; Dasgupta and David 1994). For example, patenting activity is inspired by commercial and revenues objectives and the monopoly constituted by patents can delay or even interrupt the disclosure of results, while knowledge production is based on openness, i.e. publication and discussion of results within the academic community (Florida and Cohen 1999; Hane 1999; Nelson 2004; Lissoni and Montobbio 2006; Lissoni at al., 2012).

However, universities can benefit from the collaboration with the business sector, especially for access to new funding channels, use of equipment and infrastructure, and opportunities of verification and refinement of theories and discoveries in concrete situations with deep consequences in terms of capacity to support research lines and to generate broader impact (D'Este and Perkmann, 2011). So collaboration can have positive influence on scientific productivity of teams, training of students, career paths of young researchers, especially in the fields of Medicine and Engineering (Bonaccorsi et al. 2013).

On the other hand, valorization of research takes time away from 'blue-skies', and academics are increasingly interested in linking closely science to technology with an entrepreneurial perspective, by commercialization (Clark 1998; Shane 2004; Etzkowitz 2003). Resources, especially researchers' time and equipment, are diverted from fundamental and long term research (and teaching activity) to development processes (Jensen and Thursby, 2002; Dasgupta and David 1994).

Critics have underlined the potentially detrimental effects of 'entrepreneurial science' on the long-term production of scientific knowledge, voicing fears that academic science is being instrumentalized and even manipulated by industry (Noble 1977; Slaughter and Leslie 1997; Krimsky 2003; Kenney 1986). The line between public and private good is crossed and profit and efficiency principles are contaminating the system of norms and values that characterize academia, skewing the research agenda towards more applied objectives (Merton, 1968), impacting on academic freedom (Blumenthal et al. 1986; Behrens and Gray 2001), lowering levels of research productivity among academics (Agrawal and Henderson 2002) and slowing down open knowledge diffusion (Nelson 2004; Rosell and Agrawal 2009; Murray and Stern 2007).

However, this effect is weaker in those sectors where basic and applied research are more entrenched, mainly in the Transfer Sciences (Blume, 1990) – or Pasteur's Quadrant Sciences (Stokes, 1997) – such as biotechnology or informatics, where codification of new knowledge can be achieved through publications or patents, at relatively lower costs for translation from one to the other compared to other scientific fields, and both types of outputs are accepted by the two epistemic communities. Especially in rapidly developing areas such as biotechnology, top scientists excel both as academic researchers and academic entrepreneurs (Zucker and Darby 1996). Over time, universities are demonstrating ambidexterity (Ambos et al. 2008) as positive feedback loops between publishing and patenting activities lead to a hybrid system where the best universities report scientific and technological success (Owen-Smith 2003).

Also the reversed nexus has to be considered, i.e. the influence of high quality scientific research on the engagement on third mission. Many studies converge on showing that researchers with higher productivity and scientific impact are generally more engaged in technology transfer activities (D'Este and Perkmann, 2011; Van Looy et al. 2011; Gulbrandsen and Smeby 2005). The international scientific visibility acts as a lever for reputation and it raises business's interest for partnerships (Bruno and Orsenigo 2003).

However, this effect is weaker for small and medium enterprises since only large firms have absorptive capacity to run a global scanning of academic research, locate excellence centers and collaborate at a distance. In this framework, there is also a convergence with researchers' objectives, not only for financial reasons, but also for the higher expected chances of publication because large firms compete at the frontier and have an international coverage. Instead, SMEs are likely to look for a more targeted and applied know-how, generally less fit for publication, and are more interested in geographical proximity and universities' capacity to support the whole innovation process and train high skilled workers. Once again, the effect of the disciplinary specialization of the institutions is relevant: engineers are more active in industrial partnerships, while life scientists in commercialization of results since research in that field has a direct impact on technology development (OECD, 2013; Laursen K., Salter A. 2004; Callaert et al. 2006).

This paper aims to contribute to this framework investigating the relation between research quality and engagement on third mission at institutional level, following the results of the national assessment exercise Evaluation of Research Quality VQR 2011-2014 carried out in 2016-17 by ANVUR, the Italian Agency of Evaluation of Universities and Research Institutes.

In fact, ANVUR since 2012 has grounded the evaluation of the third mission of universities and research institutes on a broad notion, defining it as *the openness of the university towards the socio-economic context through the valorization and transfer of knowledge* and, on this base, the agency has developed an evaluation method and associated measurement tools within the framework of the two rounds of the national research assessment exercises, VQR 2004-2010 and 2011-2014.

ANVUR's third mission definition encompasses different knowledge transfer channels:

- circulation and networking within relatively permanent organizations based on public-private collaboration, often at regional/local level (intermediaries);
- development, application and testing contractual relations, particularly between industry and academia (third party funds);
- closure into Intellectual Property Rights (patents, plant varieties);
- embedment into scientists-entrepreneurs (spin-off companies).
- creation and management of cultural heritage (cultural goods);
- design and management of education programs for adult population (lifelong learning);
- clinical research and training (registered clinical experimentation, biobanks);
- production of advice, expertise, informed opinion, contributions to controversies, communication of science (public engagement).

In this context only data on technology transfer activities (patents, spin-off companies and third party funds) have been used, specifically number of patents and average revenues, number of spin-off companies and average revenues, total third party funds. These data have been analyzed in relation to research quality indicators, normalized and standardized by fields. Partial correlations between indicators of research and third-mission have been analyzed through the estimation of an empirical regression model in which controls for universities' observable characteristics are included (size, type of university, geographical area, funding, field specialization).

Data

In the research assessment exercise Evaluation of Research Quality VQR 2011-2014, research and third mission performances are evaluated in an autonomous way, so the first step of analysis has been the creation of a unique VQR research-third mission database (VQR_R-TM DB). In this section, these data are described as well as the process of data merge.

Patents data are taken from the European Patent Office (Worldwide Patent Statistical database), while data on entrepreneurial activities are drawn from the Chamber of Commerce data on firms[1]. These data have been subsequently validated and integrated by universities in the Annual Third Mission Form (SUA-TM), and further information on university patents regarding cash revenues registered per year, from licenses, sales and options, have been collected there. Data on third party funds are derived from universities' balance sheets and reported in SUA-TM. All data refers to the VQR evaluation period.

Patents data are available at individual level, hence for each researcher participating in the VQR 2011- 14, it is possible to know her patents productivity and affiliation department.

Spin-off data were collected at the university level and concern only companies active and accredited by the university in at least one year in the VQR period. In spin-off company data on revenues there is a large number of missing values, so only firms that have reported revenues in at least one year in the period have been considered in the analysis.

Data on the third party funds here analyzed are the sum of the university's funds and those of all its departments. Third party funds are considered:

- revenues from commercial activity (including revenues from research and teaching activities carried out on behalf of third parties and from other commercial activities);
- funding from private and public enterprises;
- funding from institutional relations. Revenues from competitive calls are excluded.

Research quality indicators have been computed with the following methods. Data on quality of research outputs assessed in the VQR 2011-2014 were aggregated at university level using normalized and standardized indicators. Only STEM fields scores were considered, since those fields are the most active in patenting, spin-off constitution and third party funding attraction. Therefore, the following indicators have been calculated:

a) two university indicators normalized by Scientific Field (from now on *R Area*) and by Scientific Sector (from now on *R SSD*) weighed for the quota of outputs in the Area / SSD for each university. We define: $v_{j,k}$ as the sum of the scores obtained from outputs in area j (or SSD j) in the University k; $n_{j,k}$ the number of research outputs of area j (or SSD j) in the university k; V_j and N_j respectively the total scores and the number of outputs in Area j (or SSD j) at national level. Hence we define a normalized indicator at University and Area levels that compares University mean score in a specific Area with national mean score in that specific Area: $R_{j.k} = \frac{\frac{v_{j,k}}{n_{j,k}}}{V_j/N_j}$. By summing up all $R_{j.k}$ indicators in each University with a weight equal to the share of outputs in area j (or SSD j) on the total University production in terms of number of outputs (NP_k), we obtain a single University indicator that aggregate the performances in all the Areas (or SSDs):

$$R_k = \sum_{j=1}^{J} R_{j,k} * \frac{NP_{j,k}}{NP_k}$$

b) a standardized indicator (*ISA*) at university and Scientific Sector levels obtained by summing up all the standardized scores ($\mathrm{VS}_{\mathrm{i,k,s}}$) computed with respect to the mean and

[1] This process of data integration from existing databases has been useful for two main reasons: 1) to raise data quality level; 2) to lower statistical burden and costs for universities.

standard deviation for each SSD at national level: $\mathrm{VS}_{\mathrm{i,k,s}} = \frac{\mathrm{VP}_{\mathrm{i.k,s}} - <\mathrm{VP}_{\mathrm{s}}>}{\sigma_{\mathrm{s}}}$. We define Standardized University Score (VS_{k}) the normalized sum of all standardized scores in the University: $\mathrm{VS}_{\mathrm{k}} = \frac{\sum_{\mathrm{s=1}}^{\mathrm{S}} \mathrm{VP}_{\mathrm{i,k,s}}}{\sqrt{\mathrm{NP}_{\mathrm{k}}}}$. By using the normal cumulative function for the standardized university score, we can know the position of the university k in the national distribution of universities with the same staff composition in terms of field specificity. This value multiplied by 100 represents the percentile in which the University stands and indicates the probability that a University k compared with an ideal set of Universities formed by the same number of staff members of university k specialized in the same SSDs, but chosen randomly, obtain a lower evaluation than the one actually obtained.

$$P_{\inf} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{VS_d} \mathrm{e}^{-\frac{x^2}{2}} dx$$

Empirical Analysis

The following analysis aims at verifying the existence of a correlation (and its sign) between the indicators of research quality and technological transfer. The estimates reported here do not address the problem of endogeneity related to reverse causality; it is not possible, in fact, to establish how research and technological transfer act and to define a causal links between the two variables. More in detail, one can argue that good research activity could favor a higher commitment in the activities of technology transfer while, reversely, a more intense third mission activity creates virtuous circles that produce improvement in research. For this reason, the conclusions of this work are aimed at finding simple correlations between the two variables of interests (research and technological transfer).

Table 1 shows the correlation matrix between research quality indicators and technological transfer data at the university level, reporting in brackets the values of the Student's T statistic. Technological transfer indicators shown in the table are: number of patents and spin-offs per capita, average patent and spin-off companies' revenues, total third-party funds broken down by relevant item (revenues from commercial activities, with also the specification on revenues from third party research, funding from enterprises and funding from institutional relations). All the third-party variables are calculated per-capita i.e. by dividing the university totals in the four year of VQR period by the university staff headcount. University research quality indicators reported in the table are those defined in the previous paragraph.

Correlation between research quality indicators and number of per-capita patents is positive and significant: in particular, the correlation coefficient varies between 0.22 and 0.29 with a statistical significance of 10 percentage points for normalized research indicators R Area and R SSD and 5% for standardized research indicator ISA.

A higher positive correlation of 32% and 38% is found between normalized research quality indicators and the number of per-capita spin-offs with a statistical significance of 5% in the first case and 10% in the second case.

Correlations between average patent/spin-offs revenues and research quality indicators are positive but not significant. However, the distribution of patent and spin-offs revenues is very skewed and reveals a high concentration (i.e. a small bunch of patents/spin-offs produce high revenues), hence it is not surprising that correlation is not statistically significant (see Fig. 1).

Correlation between per capita third parties funding and research quality indicators (normalized with respect to Area and SSD) is positive and varies in a range of 33-36%, statistically significant at 1% and 5%. In particular, statistically significant correlations are found between revenues from commercial activities and R Area (at 5%) and between funding from institutional relations and all the research quality indicators.

Table 1: Correlation matrix (pairwise correlation) between research quality and technological transfer indicators at the university level

	R Area	*R SSD*	*ISA*	*Nr. of patents per capita*	*Average revenues per patent*	*Nr. of spin-offs per capita*	*Average revenues per spin-off*	*Per capita third party funds*	*Per capita revenues from commerci al activity*	*Per capita revenues from third party research*	*Per capita funding from enterprises*	*Per capita funding from institutional relations*
R Area	1											
R SSD	0.9839*** (0.0000)	1										
ISA	0.6875*** (0.0000)	0.6938*** (0.0000)	1									
Nr. of patents per capita	0.2897** (0.0062)	0.2758** (0.0093)	0.2223* (0.0373)	1								
Average revenues per patent	0.0602 (0.5774)	0.0827 (0.4435)	0.0466 (0.6661)	-0.0362 (0.7378)	1							
Nr. of spin-offs per capita	0.3197* (0.0120)	0.3756** (0.0029)	0.1048 (0.4214)	0.8725*** (0.0000)	-0.0672 (0.6070)	1						
Average revenues per spin-off	0.0761 (0.5666)	0.0623 (0.6394)	-0.0210 (0.8747)	0.0635 (0.6329)	0.2446 (0.0596)	-0.0407 (0.7596)	1					
Per capita third party funds	0.3629*** (0.0005)	0.3293** (0.0017)	0.1841 (0.0859)	0.0985 (0.3614)	0.0242 (0.8169)	0.3698** (0.0034)	0.3323** (0.0095)	1				
Per capita revenues from commercial activity	0.2137* (0.0456)	0.1960 (0.0672)	0.0266 (0.8059)	0.1422 (0.1864)	0.1518 (0.1440)	0.3268** (0.0102)	0.5112*** (0.0000)	0.6591*** (0.0000)	1			
Per capita revenues from third party research	0.1813 (0.0910)	0.1709 (0.1114)	0.1520 (0.1575)	0.5261*** (0.0000)	0.0949 (0.3631)	0.5936*** (0.0000)	0.2789* (0.0309)	0.5196*** (0.0000)	0.6596*** (0.0000)	1		
Per capita funding from enterprises	0.1650 (0.1244)	0.1272 (0.2375)	0.1061 (0.3253)	-0.0142 (0.8953)	-0.0386 (0.7120)	0.2017 (0.1191)	0.1087 (0.4084)	0.6900*** (0.0000)	0.2687** (0.0088)	0.2704** (0.0084)	1	
Per capita funding from institutional relations	0.3597*** (0.0006)	0.3444** (0.0010)	0.2289* (0.0319)	0.0813 (0.4517)	-0.0802 (0.4425)	0.2447 (0.0574)	0.1179 (0.3697)	0.6474*** (0.0000)	-0.0047 (0.9642)	0.0825 (0.4291)	0.2470** (0.0164)	1

Note: in brackets the statistic values of the Student's T are shown. * p < 0.05, ** p < 0.01, *** p < 0.00

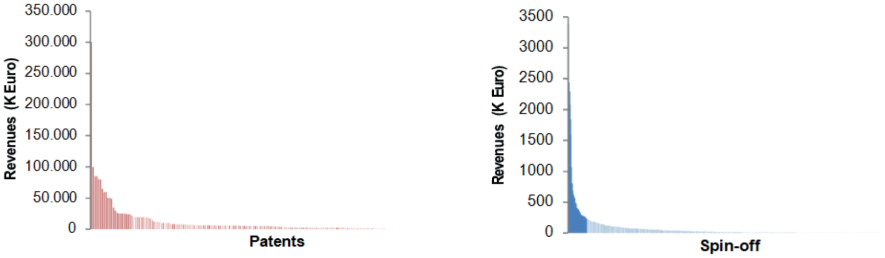


Figure 1: Distribution of revenues of patents and spin-off in the VQR period 2011-2014. (Source: Anvur elaboration on VQR 2011-2014 data)

A regression analysis is proposed in order to study the partial correlations between research quality and technological transfer indicators (those for which we found a significant correlation in Table 1), and to verify the persistence of the correlation after controlling for observable university's characteristics. The regression model estimated is the following:

$$Y_u = \alpha + \beta VQR_u + X_u{}'\gamma + \varepsilon_u$$

Y represents alternatively the number of patents and spin-offs per capita and $\mathrm{VQR_u}$ represents one of the research quality indicators (we have used alternatively the three indicators defined in the previous paragraph - R Area, R SSD and ISA, but the tables report only the specifications in which we use R Area1).

The vector of the control variables X' contains observable university's characteristics such as geographical location (binary variable for university's location in Southern Italy); the type of university (binary variable equal to 1 for Polytechnics and School for Advanced Studies); the legal status (dichotomous variable equal to 1 in the case of state university); a Gini index of heterogeneity of the scientific areas covered by the university, which measures the generalist or specialist character of the university; the presence of a technology transfer office; average funding from enterprises per capita and from commercial activities carried out on behalf of third parties, both calculated in relation to the university's research staff.

Table 2 contains estimates of the model in which the dependent variable is the number of patents per capita. In the table, the relation with our variable of interest, i.e. the indicator of research quality, is positive and significant at 1%. In specification 2 we add university's characteristics and find that the relation with the research quality indicator is no longer statistically significant. The links with university's size and type (Polytechnics and School for Advanced Studies) are not statistically significant, while that with the geographical location is statistically significant at 10%, with the Southern universities producing less patents than universities located in other areas of the country. The presence of a technology transfer office has a positive relation but not statistically significant. With regard to funding from enterprises per capita, there is a positive coefficient statistically significant at 1 percentage level, revealing it as an important driver for patenting: the higher the funding the higher the number of per-capita patents. In specification 3 we include only the variables that were significant in model 2 and the relation with research quality variable remains non-statistically significant. Our estimates state that the relationship between research and patenting is mediated by the effect of funding: universities with good research quality attract more funding and the latter are a decisive variable for patenting. Highest financial resources could, in fact, facilitate the engagement of a university in patenting activities, especially in STEM areas, which use funding for infrastructure and equipment.

Table 2 –OLS estimates at university level. Dependent variable: number of patents per capita.

	(1)	*(2)*	*(3)*
R Area	0.208***	0.0387	0.0416
	(0.0741)	(0.0998)	(0.0888)
South		-0.0823*	-0.0889**
		(0.0452)	(0.0409)
Polytechnics and School for Advanced Studies		0.0430	
		(0.0930)	
State university		-0.0738	
		(0.0624)	
Heterogeneity index		0.273	
		(0.220)	
Revenues from commercial activity		0.000209	
		(0.000383)	
Funding from enterprises		0.132***	
		(0.0438)	0.134***
Presence of Technology Transfer Office		0.0537	(0.0322)
		(0.0572)	
Constant	-0.0905	-0.216	0.0201
	(0.0737)	(0.203)	(0.0833)
Observations	88	83	83
R^2	0.084	0.295	0.265

*Note: Standard Deviation in brackets; *** p<0.01, ** p<0.05, * p<0.1*

Table 3 – OLS estimates at university level. Dependent variable: number of spin-offs per capita.

	(1)	*(2)*	*(3)*
R Area	0.169**	-0.126	-0.0829
	(0.0654)	(0.0778)	(0.0661)
South		-0.0166	
		(0.0190)	
Polytechnics and School for Advanced Studies		-0.0380	
		(0.0508)	
State university		0.278***	0.248***
		(0.0992)	(0.0838)
Heterogeneity index		0.0959	
		(0.131)	
Revenues from commercial activity		0.00133***	
		(0.000411)	0.00116***
Funding from enterprises		0.000861***	(0.000307)
		(0.000192)	0.000641***
Presence of Technology Transfer Office		-0.0319	(0.000144)
		(0.0424)	
Constant	-0.116*	-0.237	-0.183*
	(0.0664)	(0.146)	(0.0979)
Observations	61	61	61
R^2	0.102	0.510	0.477

*Note: Standard Deviation in brackets; *** p<0.01, ** p<0.05, * p<0.1*

Table 3 shows the results of the model in which the dependent variable is the number of spin-offs per capita. In the first specification, in absence of control variables, the research quality indicator is positive and significantly linked at 5% percentage level, but with the introduction of the university characteristics (specifications 2 and 3) the relation is no longer statistically significant. In specifications 2 and 3 state universities coefficient is positive and statistically significant at 1%, hence state universities seem to be more active in terms of business creation than non-state ones. Third party revenues from commercial activities and from enterprises per capita are again positively correlated and significant at 1%. Also in the case of spin-off companies, therefore, the funding attractiveness from third parties, positively affect the propensity to create business.

We also have the possibility to analyze patent data at individual level. Table 4 shows the estimates in which the dependent variable is the number of patents produced in the VQR period by the individual researcher and the explanatory variables are the research quality indicator, the characteristics of the researcher (gender, age, academic position, disciplinary area) and the institutional characteristics (geographical area, university specialization, legal status, type).

Table 4 – OLS estimates at individual level. Dependent variable: number of patents.

	(1)	*(2)*	*(3)*	*(4)*
Normalized score	0.0485***	0.0369***	0.0347***	0.0898***
	(0.00301)	(0.00313)	(0.00315)	(0.0342)
Female		-0.0393***	-0.0391***	-0.0713
		(0.00686)	(0.00686)	(0.0698)
Age		-0.00166***	-0.00168***	0.00121
		(0.000413)	(0.000415)	(0.00400)
Full professor		0.0945***	0.0948***	0.296***
		(0.00883)	(0.00885)	(0.0741)
Mathematics and Computer Science		-0.281***	-0.267***	-0.437*
		(0.0125)	(0.0127)	(0.228)
Physics		-0.203***	-0.191***	-0.435***
		(0.0144)	(0.0146)	(0.121)
Chemistry		-6.76e-05	0.0199	-0.0517
		(0.0132)	(0.0136)	(0.0833)
Earth Science		-0.272***	-0.249***	-0.651**
		(0.0191)	(0.0193)	(0.303)
Biology		-0.183***	-0.162***	-0.391***
		(0.0116)	(0.0120)	(0.0940)
Medicine		-0.241***	-0.220***	-0.480***
		(0.0101)	(0.0106)	(0.0964)
Agricultural and Veterinary Science		-0.234***	-0.212***	-0.510***
		(0.0130)	(0.0134)	(0.126)
Civil Engineering		-0.212***	-0.203***	-0.0776
		(0.0163)	(0.0164)	(0.165)
Heterogeneity index			0.137*	2.022***
			(0.0784)	(0.690)
South			-0.0319***	-0.118*
			(0.00683)	(0.0665)
Research funding (log)			0.0353***	0.208***
			(0.00615)	(0.0574)
Polytechnics/School for Advanced Studies			0.0911***	0.663***
			(0.0222)	(0.179)
State university			-0.0725***	-0.990***
			(0.0196)	(0.217)
Constant	0.0295***	0.306***	-0.137	-1.692**
	(0.00642)	(0.0232)	(0.0931)	(0.849)
Observations	32,616	32,616	32,557	2,359
R^2	0.008	0.048	0.051	0.068

*Note: Standard Deviation in brackets; *** p<0.01, ** p<0.05, * p<0.1*

In specification 1, in which the only independent variable is the individual score normalized by area without further controls, a positive and statistically significant relationship is found. In specification 2, in which personal researcher characteristics are added, the score coefficient continues to be statistically significant at 1% and positive but decreases by 1 percentage point. All the characteristics considered are statistically significant, in particular male, young and full professors are more prolific in patenting and, compared to the field of Engineering, which is considered the reference group, all the areas have a significant and negative coefficient, except for the Chemistry Area. In specification 3 university's variables are included and research quality indicator is still positive and significant at 1 percentage level. Results relative to the individual characteristics are all confirmed and also university's variables turn out to be statistically significant: in particular the relationship is negative for the state and universities located in Southern Italy, while there is a positive relationship for Polytechnics and School for Advanced Studies and research funding per capita (in logs). Finally, in specification 4, the preferred and most complete specification (n. 3) was estimated excluding researchers not active in patenting and it shows that the positive and statistically significant relationship persists even on the productive subpopulation.

The positive relationship between scientific quality production and patenting activity found in our estimates is in line with numerous studies in Europe that affirm that the most active researchers are also the most prolific in patenting. A similar relationship seems to be confirmed also at the institutional level, despite the different legislative frameworks in force in Europe concerning intellectual property financed by public research funds (Callaert et al., 2006)[2].

Conclusion

In this paper we have analyzed the relationship between research quality and engagement on technology transfer has been analyzed. The results confirm the existence of a positive correlation between the quality of research and the valorization processes at institutional level. Estimates confirm a positive correlation between quality of research and patent productivity as well as between quality of research and prolificacy in the creation of spin-off companies at university level.

It also emerged that third party funds are positively correlated with technology transfer. As reported in the literature, universities' engagement into third party research contracts could bring universities closer to the market, allowing the development of appropriate business models and triggering spin-offs creation. At the same time, spin-off companies could offer to universities more opportunities for research on behalf of third parties and produce projects and collaborations.

Finally, the positive correlation between research quality and patenting is confirmed also by analyzing data at individual level and persists in the presence of controls linked to individual and institutional characteristics.

[2] Additional specifications in which we use as control variable alternatively the indicator of research quality normalized by Scientific Sector (SSD) or standardized with respect to the mean and variance of SSD are available on request.

References

Agrawal, A., and Henderson, R. (2002). Putting patents in context: Exploring knowledge transfer from MIT. *Management science*, *48*(1), 44-60.

Ambos, T. C., Mäkelä, K., Birkinshaw, J., and d'Este, P. (2008). When does university research get commercialized? Creating ambidexterity in research institutions. *Journal of management Studies*, *45*(8), 1424-1447.

Behrens, T. R., and Gray, D. O. (2001). Unintended consequences of cooperative research: impact of industry sponsorship on climate for academic freedom and other graduate student outcome. *Research policy*, *30*(2), 179- 199.

Blume, S. (1990). Transfer sciences: Their conceptualisation, functions and assessment. Prepared for the Consequences of the Technology Economy Programme for the Development of Indicators'. OECD, Paris. 2-5 July.

Blumenthal, D., Gluck, M., Louis, K. S., Stoto, M. A., and Wise, D. (1986). University-industry research relationships in biotechnology: implications for the university. *Science*, *232*(4756), 1361-1366.

Bonaccorsi A., Colombo M.G., Guerini M., Rossi Lamastra C. (2014), The impact of local and external university knowledge on the creation of knowledge-intensive firms: Evidence from the Italian case, *forthcoming* in Small Business Economics. *Small Bus Econ*, 43, 261–287

Bruno, G. S., and Orsenigo, L. (2003). Variables influencing industrial funding of academic research in Italy. An empirical analysis. *International Journal of Technology Management*, *26*(2/3/4), 277-302.

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., and Thijs, B. (2006). Traces of prior art: An analysis of non- patent references found in patent documents. *Scientometrics*, *69*(1), 3-20.

Carayannis, E. & Campbell, D. (2009). 'Mode 3' and 'Quadruple Helix': Toward a 21st century fractal innovation ecosystem. *International Journal of Technology Management* 46.

Carayannis, E. & Barth, T., Campbell, D. (2012). The Quintuple Helix innovation model: global warming as a challenge and driver for innovation. *Journal of Innovation and Entrepreneurship*. 1.

Clark, B. R. (1998). *Creating Entrepreneurial Universities: Organizational Pathways of Transformation. Issues in Higher Education*. Elsevier Science Regional Sales, 665 Avenue of the Americas, New York, NY 10010 (paperback: ISBN-0-08-0433545; hardcover: ISBN-0-08-0433421, $27).

Dasgupta, P., and David, P. A. (1994). Toward a new economics of science. *Research policy*, *23*(5), 487-521.

D'Este, P., and Perkmann, M. (2011). Why do academics engage with industry? The entrepreneurial university and individual motivations. *The Journal of Technology Transfer*, 36(3), 316-339.

Etzkowitz, H. (1998). The norms of entrepreneurial science: cognitive effects of the new university–industry linkages. *Research policy*, 27(8), 823-833.

Etzkowitz, H. (2003). Innovation in innovation: The triple helix of university-industry-government relations. *Social science information*, *42*(3), 293-337.

Etzkowitz, H., and Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research policy*, *29*(2), 109-123.

Florida, R., and Cohen, W. (1999). Engine or infrastructure? The university role in economic development. *From industrializing knowledge. University–industry linkages in Japan and the United States,* 589-610.

Gulbrandsen, M., and Smeby, J. C. (2005). Industry funding and university professors' research performance. *Research policy*, *34*(6), 932-950.

Hane, G., 1999. Comparing university–industry linkages in the United States and Japan. In: Branscomb, L.M., Kodama, F., Florida, R. (Eds.), *Industrializing Knowledge: University–Industry Linkages in Japan and the United States*. MIT Press, London, pp. 20–61.

Jensen, R., and Thursby, M. (2001). Proofs and prototypes for sale: The licensing of university inventions. *American Economic Review*, *91*(1), 240-259.

Krimsky, J. (2003): "Small gifts, conflicts of interest, and the zero-tolerance threshold in medicine". *The American Journal of Bioethics*, 3 (3): 50-52.

Laursen, K., and Salter, A. (2006). Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. *Strategic management journal*, *27*(2), 131-150.

Lissoni, F., and Montobbio, F. (2006). Brevetti universitari e economia della ricerca in Italia, Europa e Stati Uniti. Una rassegna dell'evidenza recente. *Politica economica*, (2), 259-381.

Lissoni, F., Pezzoni, M., Potì, B., and Romagnosi, S. (2012). *University autonomy, IP legislation and academic patenting: Italy, 1996-2007* (No. hal-00779750).

Merton, R. K., and Merton, R. C. (1968). *Social theory and social structure*. Simon and Schuster

Mitroff, I. I. (1974). Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists. *American Sociological Review*, 579-595.

Mowery, D. C., Nelson, R. R., Sampat, B. N., and Ziedonis, A. A. (2004). Ivory tower and industrial innovation: University-industry technology before and after the Bayh-Dole Act in the United States. Stanford University Press.

Mulkay, M. J. (1976). Norms and ideology in science. *Social Science Information*, *15*(4-5), 637-656. Nelson, R.R., 2004. The market economy and the scientific commons. *Research Policy* 33, 455–47.

Murray, F., and Stern, S. (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior and Organization*, *63*(4), 648- 687.

Noble, D. (1977). America by Design: Science. *Technology and the Rise of Corporate Capitalism*.

OECD (2013) Science, Technology and Industry Scoreboard, Paris: OECD. Science, Technology and Industry Outlook 2014, Paris: OECD.

Owen-Smith, J., and Powell, W. W. (2003). The expanding role of university patenting in the life sciences: assessing the importance of experience and connectivity. *Research Policy*, *32*(9), 1695-1711.

Rosell, C., and Agrawal, A. (2009). Have university knowledge flows narrowed?: Evidence from patent data. *Research Policy*, *38*(1), 1-13.

Shane, S. A. (2004). *Academic entrepreneurship: University spinoffs and wealth creation*. Edward Elgar Publishing.

Siegel, D. S., Westhead, P., and Wright, M. (2003). Assessing the impact of university science parks on research productivity: exploratory firm-level evidence from the United Kingdom. *International journal of industrial organization*, *21*(9), 1357-1369

Slaughter, S., and Leslie, L. L. (1997). *Academic capitalism: Politics, policies, and the entrepreneurial university*. The Johns Hopkins University Press, 2715 North Charles Street, Baltimore, MD 21218-4319.

Stiglitz, J. E., and Wallsten, S. J. (1999). Public-private technology partnerships: Promises and pitfalls. *American Behavioral Scientist*, *43*(1), 52-73.

Stokes, D. E. (1997). Pasteur's quadrant Basic science and technological innovation. Washington, DC Brookings Institution Press.

Thursby, J. G., Jensen, R., and Thursby, M. C. (2001). Objectives, characteristics and outcomes of university licensing: A survey of major US universities. *The journal of Technology transfer*, *26*(1-2), 59-72.

Van Looy, B., Landoni, P., Callaert, J., Van Pottelsberghe, B., Sapsalis, E., and Debackere, K. (2011). Entrepreneurial effectiveness of European universities: An empirical assessment of antecedents and trade- offs. *Research Policy*, *40*(4), 553-564.

Zucker, L. G., and Darby, M. R. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences*, *93*(23), 12709-12716.

Recognition through performance and reputation

Peter van den Besselaar[1], Ulf Sandström[2] & Charlie Mom[3]

[1] *p.a.a.vanden.besselaar@vu.nl*
Network Institute & Department of Organization< Science Vrije Universiteit Amsterdam (Netherlands)

[2] *ulf.sandstrom@indek.kth.se*
KTH Royal Inst Technol, 100 64 Stockholm (Sweden)

[3] *charlie@teresamom.com*
Amsterdam (Netherlands)

Abstract

As the various disciplines have different forms of social and intellectual organization (Whitley 2000), scholars in various fields may depend less on their peers, and more on other audiences for recognition and funding. Following Merton (1973) we distinguish between *performance* and *reputation* for building up *recognition*. We show that there are indeed differences between the disciplines: in life sciences and social sciences, the reputation related indicators are dominant in predicting the score that grant applicants get from the panel, whereas in the natural sciences, the performance related indicators dominate the panel scores. Furthermore, when comparing within the life sciences the grantees with the best performing non-grantees, we show that the former score higher on the reputation indicators and the second score better on the performance variables, supporting the findings that in life sciences one probably gains recognition over reputation more than over individual performance. We suggest that this may not be optimal for the growth of knowledge.

Introduction

In the Credibility Cycle (CC) (Latour and Woolgar 1986), recognition is the step that follows publications and that precedes money – the resource that enables a new round of research – which then may result in publications and recognition. This is a somewhat traditional view on the research process, as recognition is not only based on publications.

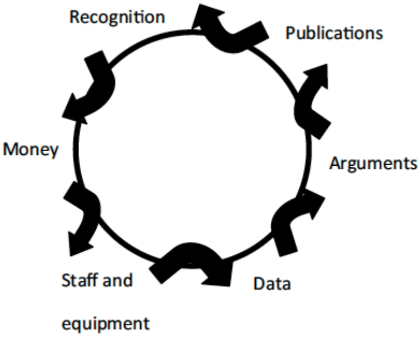


Figure 1: the Credibility Cycle

(Latour and Woolgar 1986)

As the various disciplines have different forms of social and intellectual organization (Whitley 2000), scholars in various fields may depend less on their peers, and more on other audiences for recognition and funding. For the CC this means that the phase "publications" should be

combined with other sources of recognition, such as innovations (e.g., patents), policy reports, and contributions to societal problem solving and to the public debate. More recently, also outputs related to other tasks of scholars, including teaching, community service are claimed to contribute to recognition – the extent to which needs further investigation. Another more recent phenomenon that may modify the credibility cycle is that money (grants) is not anymore solely an effect of recognition and an input for research, but receiving a (prestigious) grant is more and more seen as a performance, and bringing directly additional recognition (Van Arensbergen et al 2014; Van den Besselaar et al 2018).
Publications have many properties that may help to increase recognition. It could be the number of publications (productivity), the number of highly cited papers, the number of citations received (impact), the size and quality of the co-author network, the international nature of the co-author network, the journal impact factor (reputation of the journal), and so on. Furthermore, when recognition comes into play in e.g., grant selection procedures, also other signs of recognition may play a role – which can be found in the CV of the applicant: reputation or performance of the organizations the applicant has worked and/or collaborated with or is going to work, the reputation of the PhD supervisor, and the amount of earlier acquired grants (see above). These contributing factors partly relate to the performance of the applicant and partly to the reputation, a distinction already made by Merton (1973), and also at the level of research organizations and universities (Paradeise and Thoenig 2013).

Research question

Based on the considerations above, we try to find out whether reputation, performance, or both constitute the recognition that leads to winning new grants.

Data

We use data on a large and prestigious European grant program. We have data for 3030 of the applicants, which is 95% of all applicants in that round. We do have data on personal characteristics, the subfield of the proposal, the scores the application got from one of the 25 panels, and the full CV of all applicants. Furthermore, we collected bibliometric data from the Web of Science for (now 60%) of the applicants
For about 1800 of the applicants we downloaded the WoS records, and we checked for different name variants and for different (first) initials. The resulting records were processed by the BMX program (Sandström & Sandström 2009) and then first semi-automatically disambiguated. After that, a manual disambiguation was done. Using the BMX program, the scores for different performance variables were calculated – as mentioned in Table 1.

Table 1. Performance and reputation indicators.

Name	*Description*	*Type*
P-frac	Number of fractionally counted publications.	Performance
Citations(2y)	Field normalized citations, two years citation window.	Performance
Top5%	As above, but now the fields' top 5% cited papers.	Performance
Journal Impact	The field normalized average journal impact factor.	Reputation
Network quality	Median ranking (top 10%) score of linked organizations.	Reputation
Other Grants	The number of other obtained grants by the PI	Reputation

Several other variables could only be retrieved from the CVs of the applicants, and the various data processing steps were done using the SMS platform for data integration and enrichment.[1]

[1] The SMS platform: www.sms.risis.eu.

The extracted organization names were linked to the Leiden Ranking, in order to determine the quality of the host institution (where the project will be done) in terms of the share of 10% highest cited papers. In the same way, the organizations the applicant has collaborated with or has worked are given a rank-score. We use the median of these scores as indicator for quality of the applicant's network. We manually extracted information about other grants of the applicant from the CVs.

For a smaller set (four life science panels) we have not for 60% but for all applicants the bibliometric data. This smaller set is used for one of the analyses, and for those we also have some additional variables as showed in table 2.

Table 2. Additional performance and reputation indicators

Name	*Description*	*Type*
Top10%	As above, but now the fields' top 10% cited papers.	Performance
Co-authors	Average number of co-authors	Reputation
International	Average number of international co-authors	Reputation
Host quality	The ranking (10% PP) of the host institution	Reputation

We now can differentiate between indicators for performance and indicators for reputation. Performance is measured by indicators that say something about the individual researcher's scholarly work, such as number of publications (fractional count), the number of top cited papers, and so on. Reputation indicators are more indirectly related to the work of the scholar, such as the Impact Factor of the journals one publishes in, the ranking of organizations in the network, and the earlier grants. The last column of Table 1 gives the classification.

Method

We use the variables mentioned above to predict the score the applicants get from the panels, using stepwise linear regression. As disciplinary differences are expected to influence what are considered important for accumulating recognition, we do the analysis by discipline. In this version of the paper, it is done at the level of meta-disciplines: (i) life sciences and medicine, (ii) physics and engineering, and (iii) social sciences and humanities. We report here which variables play a role, and which not, but do not go into the numerical aspect of the regression outcome.

Then we look in more detail at the difference between the grantees and the group of best performing rejected applicants.

Findings 1: what predicts the panel score?

In Table 3, we show for the three different domains what variables are included in the stepwise regression outcomes, using the six variables from Table 1. We then compare the emphasis on reputation related aspects versus on the performance-based aspects. Please be aware that this is at the domain level, and that within the various domains the disciplines may differ. This is also the case within the social science and humanities, but as we use Web of Science bibliometric data, the scores for SSH are strongly dominated by economics and psychology. The two latter disciplines are strongly oriented at publishing in journal articles, whereas this is much less the case in the other SSG disciplines, such as literature, history, philosophy, anthropology, and sociology and political science.

What does Table 3 show? All the three reputation indicators are includedin predicting the scores of the life science applications, and only one of the performance-based indicators: top 5% cited papers. The overall score for *emphasis on individual performance* is low: 0.33. For the social sciences and humanities we find the same pattern, but with a different performance indicator:

citations. Finally, physics and engineering show a very different pattern. All the three performance variables contribute to predicting the score, and two of the reputational variables. Interestingly, the *journal impact* indicator does not. The overall *emphasis on individual performance* is high: 1.5.

Table 3. Performance or reputation

Name	*Life sciences*	*Physics and engineering*	*Social sciences and humanities*	*Type**
P-frac		+		Perf
Citations(2y)		+	+	Perf
Top5%	+	+		Perf
Journal Impact	+		+	Rep
Network quality	+	+	+	Rep
Other Grants	+	+	+	Rep
Perf/Rep**	.33	1.50	.33	

Source: Van den Besselaar et al. (2016)
* Perf = performance indicator; Rep = reputation indicator;
** Perf/Rep = number of performance indicators divided by the number of reputation indicators

Findings 2: reputation and performance within the very good group

As showed elsewhere, the selection process of grant panels is not very strong, as the best rejected applicants are in average at least as good as the granted applicants (Bornman et al 2010), and the predictive validity is low (Van den Besselaar & Sandström 2015). For four panels where we have bibliometric data for all applicants, we compare the granted applicants with the set of best performing non-granted applicants. Table 4 presents the results.

Table 4. Granted versus best non-granted: performance versus reputation

Name	*Granted*	*better?*	*Best non-granted*	*F*	*Sign*
Performance					
P-fractional	1.9	=	1.9	0.006	0.940
Citations(2y)	3.05	=	2.88	0.360	0.550
Top10%	1.23	<	2.18	11.54	0.001
PModel*	17.3	<	22.5	2.101	0.150
Reputation					
Journal Impact	2.31	>	2.07	2.455	0.120
Network quality	195.5	=	196.9	0.005	0.946
Ranking host	0.14	=	0.13	0.518	0.473
# co-authors	6.76	=	6.52	0.225	0.636
# international co-authors	1.73	=	1.65	0.610	0.436
Other Grants	2.7	>	1.7	5.413	0.022

* The PModel indicator is explained in (Sandström, Sandström & Van den Besselaar 2019)

As the table shows, the granted applicants score (marginally) significant better on two of the reputation indicators, and equal on the other four than the best-performing non-granted. In contrast, the best-performing non-granted applicants score better than the granted applicants on two of the four performance indicators and equal on the two others.

Conclusions and discussion

The conclusion is that in the life sciences reputation is more important than performance for acquiring funding. For the social sciences and humanities, but mainly for economics and psychology, we observe the same pattern. On the other hand, in physics and engineering the pattern is different and there performance seems to be dominant over reputation.

As future work, we will repeat the analysis at a lower level of aggregation: for the individual disciplines, and we will relate the findings to other characteristics at the field level, among other with the occurrence of gender bias and nepotism. We suggest that fields focusing on performance may be less susceptible to bias and more strongly following Merton's CUDOS norms.

Acknowledgments

We acknowledge support of the ERC grant 610706 (the GendERC project) and EC-H2020 grant 824574 (the GRANteD project).

References

Bornman L, Leydesdorff L, van den Besselaar P (2010), A Meta-evaluation of Scientific Research Proposals: Different Ways of Comparing Rejected to Awarded Applications. *Journal of Informetrics* **4** 211-220

Latour B and Woolgar S (1986). *Laboratory life: The construction of scientific facts*, 2nd ed. London: Sage.

Merton, RK (1973). Recognition' and 'excellence': instructive ambiguities. In RK Merton, *The sociology of science: theoretical and empirical investigations* (pp. 419–437). Chicago: University of Chicago Press

Paradeise C, Thoenig J-C (2013). Academic Institutions in Search of Quality: Local Orders and Global Standards. *Organization Studies* **34** 189–218

Sandström U, Sandstrom, E & Van den Besselaar P (2019) The P-model: An indicator that accounts for field adjusted production as well as field normalized citation impact (in preparation)

Van Arensbergen P, van der Weijden I, van den Besselaar P (2014) Different views on scholarly talent – what are the talents we are looking for in science? *Research Evaluation* **23** 273-284.

Van den Besselaar P, Sandström U (2015). Early career grants, performance and careers; a study of predictive validity in grant decisions. *Journal of Informetrics* **9** 826-838

Van den Besselaar P, Schiffbaenker H, Mom C, Sandström U (2016) *Explaining grant application success: the effect of gender*. Deliverable 6.1 GendERC project.

Van den Besselaar P, Schiffbaenker H, Sandström U, Mom C (2018). Explaining gender bias in ERC grant selection. *STI 2018 Conference Proceedings*, 346-352

Whitley 2000 Whitley, Richard. 2000. The intellectual and social organization of the sciences, 2nd ed. Oxford: Oxford University Press.

Towards a multidimensional classification of social media users around science on Twitter

Adrián A. Díaz-Faes[1], Nicolás Robinson-García[2], Timothy D. Bowman[3] and Rodrigo Costas[4]

[1] *diazfaes@ingenio.upv.es*
INGENIO (CSIC-UPV), Universitat Politécnica de València, Valencia (Spain)

2 *elrobinster@gmail.com*
Delft University of Technology, Delft (The Netherlands)

3 *timothy.d.bowman@wayne.edu*
School of Information Sciences, Wayne State University, Detroit. USA

4 *rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands)
Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University,
Stellenbosch (South Africa)

Abstract

With the advent of altmetrics, digital traces that go beyond the scientific impact can be tracked. Twitter stands as the most appealing platform for their inspection since it gathers academic and non-academic users that discuss a wide-ranging number of topics. This research aims at developing and proposing a fine-grained classification of social media users based on mapping techniques and clustering methods and compare them with other tentative classifications proposed elsewhere. To do so, online activity of over 1.3 million Twitter users is examined, considering both their overall activity on Twitter as well as their interaction with scientific publications

Introduction

The shift from print to digital format has led to new forms of accessing, using and sharing scientific information. These interactions leave digital traces that can be stored and computed, leading the way to a new set of indicators, grouped under the umbrella term of *altmetrics* (Priem et al., 2010; Wouters, Zahedi & Costas, 2018). Altmetrics provide data on events related to the dissemination and consumption of scholarly information such as reading, sharing, downloading, commenting, or recommending scientific literature. Although the myriad of platforms included within altmetrics is wide and heterogeneous, Twitter seems to drive a large portion of this activity (Haustein, Costas, & Larivière, 2015; Thelwall et al., 2015). Furthermore, its wide use among the general public position Twitter as a potential platform in which academic and non-academic actors interact (Robinson-Garcia et al., 2018) and share interests discussing a wide-ranging set of topics (Letierce et al., 2010; Sugimoto et al., 2017).

This research-in-progress aims to identify and characterise social media communities of users interacting with scholarly objects (i.e. scientific publications) generated by researchers. Several studies have examined specific features of users' behaviour within social medial platforms (see Kwak et al., 2010; Brandtzæg, 2010), including a tentative attempt to classify social media users presented by Haustein, Bowman, and Costas (2015). To date, no study has suggested a typology of social media users based on both the overall user's behaviour on Twitter as well as their specific interaction with scholarly objects. This work presents preliminary results of an attempt to develop and propose a fine-grained classification of social media users based on mapping techniques and clustering methods and compare them with other classifications proposed elsewhere.

Data and methods

In this study, the authors built upon recent work that seeks to shift the attention of researchers from *primary* to *secondary* altmetric indicators (Díaz-Faes, Bowman, & Costas, 2019). Primary indicators are indicators derived from the direct count of mentions to scholarly objects within social media platforms (e.g., Mendeley readers, number of tweets). Secondary altmetric indicators are focused on the interactions that occurs between users and scholarly objects as well as their overall online activities. This second generation of altmetrics is in line with recent developments that advocate for an approach based on interactions, rather than direct counts of mentions (Haustein, 2015; Robinson-García, van Leeuwen, & Rafols, 2018).

Díaz-Faes et al. (2019) disclose the four dimensions in which users' behaviour on Twitter around science can be classified through a Factor Analysis. These are the following:

- 'Science Engagement' characterises users based on the extent to which they tweet only about science. Here metrics such as number of (re)tweets or hashtags are included.
- 'Science Focus' mirrors tweeters involvement in science related issues. Here metrics such as share of tweets to papers and the time between the publication of the paper and the tweet of the tweeter gain more relevance.
- 'Social Media Activity' refers to what users share and to which issues are of a greater interest.
- 'Social Media Capital' comprises metrics that mirror user's influence and centrality within the social media realm.

For this work-in-progress, the data sample is comprised of users with at least one tweet linking to a scholarly object occurring between 2011 and 2017 as covered by Altmetric.com. Profiles exhibiting extremely low activity (≤15 total tweets) were removed. The final dataset is comprised of 1,340,695 users accounting for over 14.6 million tweets to papers. Table 1 summarizes the metrics that comprise each dimension.

Table 1. Dimension of social media activity around science (source: Díaz-Faes et al., 2019).

Dimension	Variable	
Science Engagement	Number of (re)tweets to scientific publications	tws
	Number of original tweets to scientific publications	otw
	Number of distinct publications (re)tweeted	tws hash
	Number of (re)tweets containing hashtags	p tw
Social Media Activity	Number of tweets overall	tweets
	Number of likes given	likes given
Social Media Capital	Number of followers	followers
	Number of followees	followees
	Number of lists in which users are listed	listed count
Science Focus	Average length of the titles of the papers tweeted	avg title length
	Average time between the publication of the paper and the tweet of the tweeter	avg days to tweet pub
	Share of tweets to papers	pwts to papers

Since these four dimensions provide a general empirical framework that accounts for users' behavior in the social media realm, a fine-grained classification of social media users can be extracted from their analysis. Given the high skewness of Twitter data and the fact that activity around science is examined within a users' overall behavior, developing a robust classification is by no means an easy task. This work will first present the results from previous tentative classifications (e.g. Altmetric.com; Haustein et al. 2015) and then the authors will present the results of several advanced grouping techniques, such as K-means clustering, Leiden algorithm (Traag, Waltman, & van Eck, 2018) and Archetypal Analysis (Seiler & Wohlrabe, 2013), on the proposed four-dimension schema.

Results and discussion

Altmetric.com classification

Altmetric.com groups Twitter users as researchers, science communicators, practitioners, and members of the public based on keywords in profile descriptions, journals tweeted, and follower lists. 86% of users are tagged as members of the public, whereas the remaining typologies represent a minimum share, being 7% identified as scientists. Figure 2 gives a snapshot of tweeters features and online behaviour. Members of the public gather a heterogeneous group of users with no clear patterns. However, science communicators stand out for their high Social Media Capital and scientists for their high Science Engagement and Science Focus.

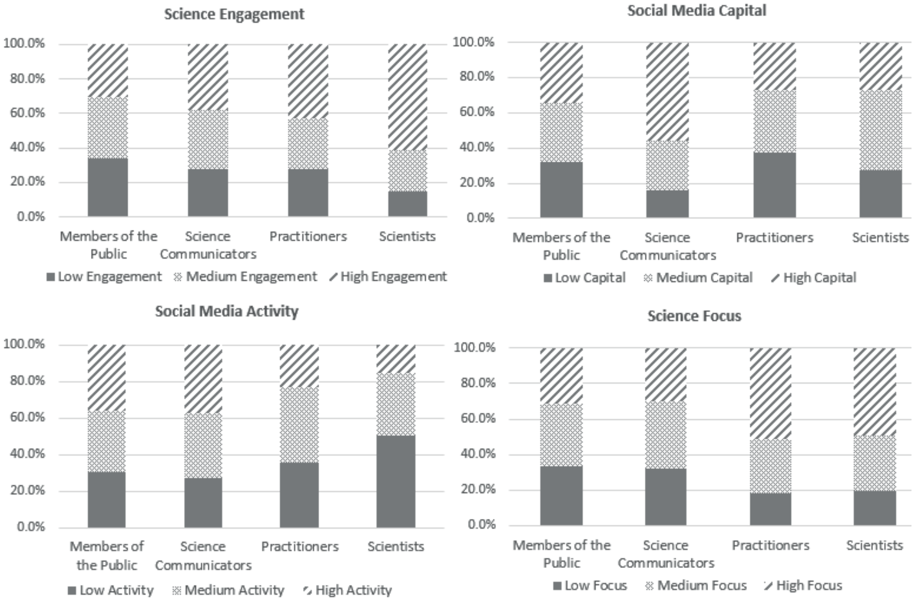


Figure 1. Altmetric.com classification.

Table 2 displays the mean values of (re)tweets to scientific publications, overall tweets (tweets), overall followers (followers), and share of tweets to papers (ptws to papers) for the four dimensions of the framework. The data demonstrates that scientists tweet on average less than all other groups, but have a higher mean average of tweets to papers (4.69%) than all other user types.

Table 2. Mean values for Altmetrics typology of users.

	tws	tweets	followers	ptws to papers
Members of the Public	9.01	8,256	1,133	1.34%
Science Communicators	26.24	6,938	2,204	2.59%
Practitioners	19.44	3,195	903	2.45%
Scientists	30.38	2,266	611	4.69%

Engagement vs. exposure

Haustein et al. (2015) proposed a classification of Twitter users based on two indicators: number of followers (engagement) and dissimilarity between tweet and paper title (exposure). Using these indicators, the authors described four different types of users: brokers (high engagement and exposure); broadcasters (high exposure but low engagement); orators (high engagement but low exposure); and mumblers (low exposure and engagement). This typology was replicated based on the dimensions proposed by Díaz-Faes et al. (2019). Exposure was tested with two of the proposed dimensions: Science Engagement and Science Focus. Important differences were observed based on the dimension that was used (Figure 1). The most apparent difference demonstrates that the share of influencers (highly active and prominent tweeters) is much smaller when exposure is considered as actual involvement (Science Focus), rather than simply tweeting about science (Science Engagement; 19.33 vs. 27.89%).

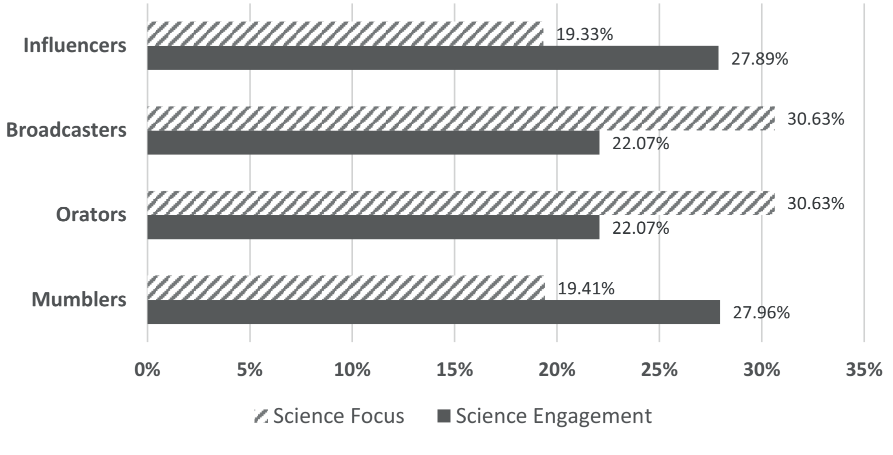


Figure 2. Haustein et al. (2015) scheme based on the four dimensions of user's activity.

In Table 3, the results of examining the classification scheme presented by Haustein, et al. (2015) using Science Engagement and Science Focus categories of the new classification scheme are displayed. These new categories indicate very different results for the categories proposed by Haustein, et al. (2015).

Table 3. Mean values for Haustein et al. (2015) classification schema.

		tws	tweets	followers	ptws to papers
Science Engagement	Mumblers	1.60	4,321.71	180.18	0.48%
	Orators	11.44	2,092.02	122.60	3.62%
	Broadcasters	1.78	11,566.42	2,309.13	0.13%
	Influencers	27.22	11,496.35	1,836.04	2.37%
Science Focus	Mumblers	3.19	4,882.96	173.40	0.68%
	Orators	7.68	2,359.30	142.99	2.62%
	Broadcasters	5.99	12,349.53	1,654.47	0.34%
	Influencers	31.81	10,224.44	2,663.96	3.03%

Concluding remarks

In this research-in-progress the authors present preliminary findings on the development of a classification scheme of Twitter users based on four dimensions, which combine indicators of direct interaction with publications with the overall activity of users in Twitter. This classification scheme is then compared with Altmetric.com's classification scheme categorizing Twitter users based on profile descriptions and the classification scheme proposed by Haustein et al. (2015). These two classification schemes are solely based on altmetric indicators (direct interactions of users with publications) and the latter one can only be compared with two of the four dimensions of the new classification scheme (Díaz-Faes et al., 2019). While important discrepancies were found, Altmetric.com's classification scheme shows little capacity of discrimination (86.3% of the data sample belong to one category), which disregards its use as a benchmark. In the case of the classification scheme proposed by Haustein et al. (2015), the results presented here indicate that two dimensions of exposure (Science Focus and Science Engagement) from the new classification scheme demonstrates different types of exposure, which should be developed further.

Acknowledgments

Adrián A. Díaz-Faes is currently supported by a Juan de la Cierva postdoctoral grant (IJCI-2017-31454) from the Spanish Ministry of Economy and Competitiveness. Nicolas Robinson-Garcia is a Marie Sklodowska Curie Experienced Researcher receiving funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 707404.

References

Adie, E., & Roe, W. (2013). Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, *26*(1), 11–17.

Díaz-Faes, A.A., Bowman, T.D., & Costas, R. (2019). Towards a second generation of 'altmetrics': Characterizing the interactions of Twitter communities of attention around science. 14(5): e0216408. https://doi.org/10.1371/journal.pone.0216408

Haustein, S. (2019). Scholarly Twitter metrics. In W. Glänzel, H.F. Moed, U. Schmoch, M. Thelwall, editors. *Handbook of Quantitative Science and Technology Research*. Springer.

Haustein S., Bowman T.D., & Costas R. (2015). *Communities of attention around journal papers: who is tweeting about scientific publications*. Social Media and Society 2015 International Conference. pp. 1-21. Toronto. Retrieved from https://es.slideshare.net/StefanieHaustein/communities-of-attention-around-journal-papers-who-is-tweeting-about-scientific-publications

Hicks, D., Wouters, P., Waltman, L., Rijcke, S.D., & Rafols I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, 429-431. Comment.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? Proceedings of the *19th international conference on World Wide Web.* ACM. pp. 591–600.

Letierce, J., Passant, A., Breslin, J., & Decker, S. (2010). Understanding how Twitter is used to spread scientific messages.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. 26 October 2010. http://altmetrics.org/manifesto

Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PLOS ONE*, *12*(8), e0183551.

Robinson-García, N., van Leeuwen, T.N., & Rafols, I. (2018). Using almetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy*, 45(6), 815-826. https://doi.org/10.1093/scipol/scy024

Seiler, C., & Wohlrabe, K. (2013). Archetypal scientists. *Journal of Informetrics*, *7*(2), 345-356.

Sugimoto, C.R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037-2062.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLOS ONE*, *8*(5), e64841.

Traag, V., Waltman, L., & van Eck, N.J. (2018). From Louvain to Leiden: guaranteeing well-connected communities. *arXiv preprint arXiv:1810.08473*.

Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. DOI:10.13140/RG.2.1.4929.1363

Wouters P., Zahedi Z., & Costas R. (2019). Social media metrics for new research evaluation. In: W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall, editors. *Handbook of Quantitative Science and Technology Research*. Springer.

Publication trajectory discontinuity – is there gender difference?

Ekaterina Dyachenko[1] and Asia Mironenko[2]

[1] edyachenko@hse.ru
National Research University Higher School of Economics, Moscow (Russia)

2 mironenkoasja@gmail.com
European University at Saint Petersburg, Saint Petersburg (Russia)

Abstract

The study presents an analysis of publication breaks – long periods between consecutive publications in researchers' publication trajectories. Empirical analysis uses data on a sample of Russian medical researchers, but the questions asked and the approach suggested are of wider appeal. We investigate how common the breaks are in publication trajectories, and what is the context of the breaks. Particular focus of the study is on difference between female and male researchers. The issues raised could bring some insights into discussion on gender performance gap in science.

Introduction

There is a shared understanding that in science career domain many gender equality issues have not been solved yet. According to recent OECD report the gap between female and male presence and success in academia remains significant (OECD, 2018). One of the important issues is that women researchers' career develops on average slower than men's. Researchers investigated a number of factors possibly contributing to this, including differences in abilities; in interests and ambitions when choosing a career and pursuing goals ("self-selection"); in work-life balance, in research productivity. Gender biases in hiring, promotion, peer-review, citing, salary and funding decisions have also been studied in detail (Ceci et al., 2014; Van Den Besselaar & Sandström, 2016; OECD, 2018).

The factor which is probably investigated more often than others is performance gap – the difference between male and female researchers' productivity and impact. When comparing productivity measured by number of publications most of the studies found that women are less productive. Ceci with colleagues summarized a number of such studies (covering STEM fields) concluding that performance gap is consistently in favour of men (Ceci et al., 2014). As for the impact/quality of "male" and "female" research measured by citations the results of comparison are not so unidirectional (Ceci et al., 2014; Larivière et al., 2013; Van Den Besselaar & Sandström, 2016).

While some authors compare female and male researchers' performance in order to understand if there any gender bias in fund allocation, hiring or promotion (Van Den Besselaar & Sandström, 2016; Van Den Besselaar et al., 2018), others are interested in possible factors behind the productivity gap. One of the ways to investigate the genesis of this gap is to look at when it appears. In a study of about 400 researchers it was found that there is no much difference in productivity at the early career stage, but it appears over a period of ten years (Van Den Besselaar & Sandström, 2016). Similar trend was observed when a set of 7064 professors at Quebec universities was analyzed (Larivière et al., 2011).

The factors widely suspected as significant are related to family responsibilities. A number of researchers found a statistical relationship between parenthood and research productivity, and how it is different for male and female researchers. However, the reported strength of the relationship varies, and usually authors do not or should not make causal claims (Ceci et al., 2014). Here we argue that analysis of publication breaks, which is not common in bibliometrics, can provide some new insights for studies of research careers. An exploratory qualitative study of women in academia showed that reasons of career breaks vary. They mostly were family-related but not necessarily related to children (Mavriplis et al., 2010).

We use the data from the project in which publication trajectories of Russian cardiology researchers are studied. The work presented here is a spin-off project with the focus on publication trajectories discontinuity. We consider publication record of a researcher as a sequence of events – the sequence of publications and publication breaks. In social sciences sequence analysis methods mostly are used in career and life-course studies (Abbott & Hrycak, 1990; Aisenbrey & Fasang, 2010), but almost never in scientometrics. Although we do not apply here particular methods to which social sequence analysis owes most of its fame (sequence alignment, clustering of sequences), we remain in sequence paradigm when the events are considered in a context of other events.

We ask the following questions: 1) How common are the breaks in publication trajectories? 2) How the breaks are related to the overall performance: do researchers with more breaks perform worse than others? 3) What is the context of breaks: what happens before them and what follows after? All these questions are worth asking beyond the gender context, but here we particularly focus on the comparison between male and female researchers

According to the latest available data from Russian Federal State Statistics Service, in 2016 female researchers accounted for about 40% of all researchers in Russia. The share of women in highest academic positions is much lower. Among 'doctor nauk' researchers (the highest academic degree in Russia) the share of women is 26%, among the full members of Russian Academy of Science – only 5%. Career breaks is an important issue for female researchers in Russia as well as in many other countries. Russia has quite protective family leave policy, but this policy does not solve the problem of unequal career advancement.

Data and Methods

For the ongoing study of publication trajectories we selected all researchers who were awarded PhDs in medical sciences with a specialization in cardiology in Russia in 2005-2006. We used the catalogue of the National Library of Russia to compile the list of these researchers. For 654 discovered cardiology researchers we gathered data on the papers they published in academic journals throughout their careers. We used three sources of data: Scientific Electronic Library[1] – database which covers about 6000 Russian academic journals, Web of Science Core Collection (SCI-EXPANDED, SSCI, A&HCI), PhD theses of the researchers.

The publication record of each researcher was coded with an alphabet of 7 elements:

IT – paper in an international top journal (journal from the top quartile of Journal Impact Factor ranking in any subject category of JCR 2016);
IO – paper in an international non-top journal (any other non-Russian journal);
NT – paper in a national top journal (journal included to RSCI[2]);
NO – paper in a national non-top journal (Russian journal not included to RSCI);
P1 – publication break of 1 year;

[1] https://elibrary.ru/

[2] The Russian Science Citation Index is a collection of Russian journals, consisting of "the most influential scholarly literature in Russia based on citation analysis" [https://clarivate.libguides.com/webofscienceplatform/rsci].

P2 – publication break of 2 years;
P3 – publication break of 3 or more years.

Duration of a break was measured as a number of full years with no published papers. For example, if a researcher published a paper in 2004 and the next one in 2006 we register P1 break. For each researcher the sequence of coded papers and breaks was obtained, which could look as follows: NO NT NO NT P1 IO NT NT P1 IT NT IT.

The choice of data sources was determined by the fact that most Russian medical researchers do not post CVs online. Instead of extracting full publication records from CVs we reconstructed them from three academic databases. To improve the coverage, we plan to add the Scopus database as the fourth source. Preliminary analysis shows though that adding the Scopus does not affect significantly the number of papers discovered.

Some parts of the further analysis were performed with TraMineR, which is a package for R (Trajectory Miner for R), developed in University of Geneva (Gabadinho et al., 2011). The rest required pieces of algorithm developed ad hoc and coded on Python. At this moment we have the data for about half of the sample – the researchers who got PhD in 2005. Thereby, the results and discussion presented further should be considered as preliminary.

Results and discussion

There are 335 cardiology researchers who got PhD in 2005. For 280 researchers we found at least one paper in academic journal. The results below describe this set of authors. Only one quarter of those are male researchers. In total, these cardiologists published 3,868 papers in 493 different journals. The earliest published paper appeared in 1987, the most recent – in 2018. About half of the researchers published after 2006. The number of papers published by the most prolific author is 149, the average number of papers for the sample is 13.8, the median is 5. The share of cardiologists who stopped publishing after getting PhD is higher for women, although Chi-squared test showed no statistical difference between two groups.

Table 1. Characteristics of publication trajectories.

	All researchers (280)	*Female researchers (209)*	*Male researchers (71)*
Average number of papers, by type of journal			
NO	7.1	5.89	10.59
NT	5.8	5.28	7.18
IO	0.7	0.69	0.84
IT	0.2	0.24	0.21
All papers	13.8	12.1	18.8
% of researchers with at least one paper of the following types			
NO	76	75	77
NT	81	82	79
IO	26	26	27
IT	9	8	11
% of researchers with no papers published after 2006			
	48	50	42
% of researchers with at least one publication break			
P1	35	34	38
P2	16	16	17
P3	27	26	28
P1 or P2 or P3	56	53	66

Table 1 shows characteristics of the publication records for the whole sample of cardiologists and subgroups of male and female researchers. Male researchers in our sample turned out to be significantly more productive in terms of number of papers published per researcher, which is in line with majority of studies. Men publish on average more papers in all categories of journals except the most prestigious category, the top international journals[3]. The question is how the difference in the total productivity is structured in time. Does it take a female researcher longer to produce a paper, or does she lose the race because she takes more or longer breaks? According to Table 1, chances of having a break are similar in two groups. The difference is rather unexpected – larger proportion of male researchers have at least one break in publication trajectory. Figure 1 shows more detailed picture of how common are the breaks of each duration in two groups of researchers.

The bars show breaks of duration from 1 year to 13 years (the longest break we observed). The top bar relates to shortest breaks – one total year without a paper. Our hypothesis was that men tend to have relatively short breaks if any, while women more often have long and very long breaks. Figure 1 does not support this hypothesis, as the diagram looks more or less symmetrical.

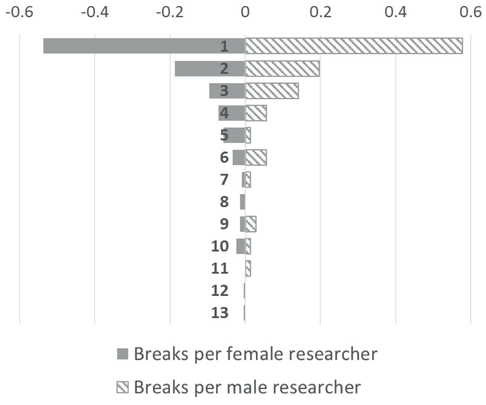


Figure 1. Publication breaks of a certain duration (number of breaks per researcher) .

Our focus of interest is mostly on long breaks – those coded as P3 (3 or more years). Figure 2 shows the distribution of the P3 break starting year for two groups of researchers. We observe similar pattern with the different size of the groups taken into account. The most common start of the long interruption is when a PhD is completed. The difference between groups is in the right tail – the share of female researchers taking a break after 2005 is larger than that of male researchers.

Finally, we were interested in how the long breaks affect the trajectory of a researcher. Does a break go with a setback not only in quantity but also in quality of papers? Here we use the category of a journal as a proxy. For each researcher we compared the level of papers published before a long break (P3) with those published after. To calculate the level we looked at three papers published immediately before P3 and three published after. The level of each paper was quantified according to the following scale: NO paper – 1 point, NT paper – 2 points, IO paper – 3 points, IT paper – 4 points, and then the average was calculated. Table 2 shows that long

[3] The difference is statistically significant only for NO papers (papers in national ordinary journals).

breaks generally do not bring down the level of journals where a researcher publishes, but neither they raise it[4].

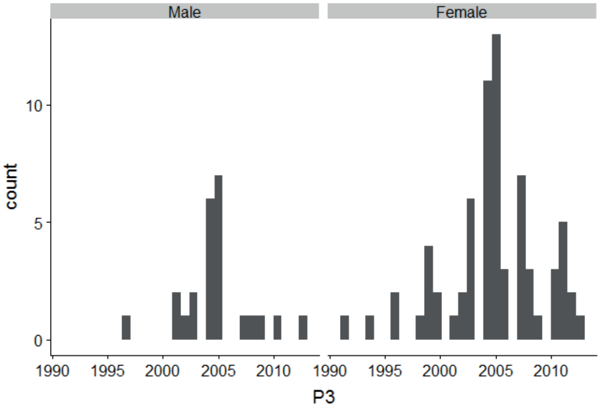


Figure 2. Histograms of the starting year of P3 breaks (3 years or longer), by gender.

We also compared levels of journals for women and men separately. It could have been different situation in two groups, because the reasons behind the long breaks can be different. The suggestion is that for women publication break more often indicates the career break than for men, and the drivers of the breaks can also differ. The comparison of two groups does not reveal different patterns, although there is some difference, statistically not significant, – men tend to gain from breaks more than women[5].

Table 2. Average level of papers published before and after long breaks (scale from 1 to 4).

	All researches (93)	*Male researchers (24)*	*Female researchers (69)*
Level of papers before P3 break	1.62	1.61	1.62
Level of papers after P3 break	1.67	1.69	1.66
Difference	0,05	0.08	0.04

Conclusion

We analyzed the breaks in publication records of cardiology researchers. There could be difference between male and female researchers induced by career breaks related to family leaves and other reasons. Family leave is considered to be a factor shaping male and female career development. They usually happen in the early career stage which in case of academia is the most precarious and the one largely shaping future success (Laudel & Gläser, 2008). The importance of the issue is recognized by many experts and government agencies. At the same time, there is not much empirical evidence on how common career breaks are among female researchers and how they affect careers. At least one study reported statistical evidence of the length of career break affecting females' chances of attaining high academic rank in Scotland

[4] Non-parametric Wilcoxon signed-rank test for paired samples does not show that the difference is significant.
[5] Another thing worth checking is the intensity of publishing (number of papers per year) before and after breaks. We are going to include this into analysis.

(Ward, 2001). For other fields, there is evidence that women are penalized for career breaks both in terms of wages and status, although the scale of the penalty varies (Arun, Arun & Borooah, 2004; Aisenbrey, Evertsson & Grunow, 2009).

For the sample of Russian cardiology researchers we saw that men on average are more productive than women. We wondered whether this difference could be explained by career breaks. Underlying assumption was that career breaks often go with publication breaks. Somewhat unexpectedly, we did not find significant difference between men and women researchers – publication breaks are almost equally common in two groups, the context is not so different either. We assume that similar analysis can reveal differences for a bigger sample, which we are planning to reach. The major advantage of publication trajectories analysis is that it provides a cheaper way to study the dynamics of individual performance than using survey data. An important limitation of this approach is that a publication break is not equivalent to career break. Still, with all the complexity taken into account, the analysis of publication breaks seems promising in context of gender issues and in other topics related to research career studies.

References

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. American journal of sociology, 96(1), 144-185.

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The" second wave" of sequence analysis bringing the" course" back into the life course. Sociological Methods & Research, 38(3), 420-462.

Aisenbrey, S., Evertsson, M., & Grunow, D. (2009). Is there a career penalty for mothers' time out? A comparison of Germany, Sweden and the United States. Social Forces, 88(2), 573-605.

Arun, S. V., Arun, T. G., & Borooah, V. K. (2004). The effect of career breaks on the working lives of women. Feminist Economics, 10(1), 65-84.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. Psychological Science in the Public Interest, 15(3), 75-141.

Gabadinho, A., Ritschard, G., Studer, M. & Mueller, N. S. (2011). Mining sequence data in R with the TraMineR package: A user's guide (for version 1.8).

Larivière, V., Vignola-Gagné, E., Villeneuve, C., Gélinas, P., & Gingras, Y. (2011). Sex differences in research funding, productivity and impact: an analysis of Québec university professors. Scientometrics, 87(3), 483-498.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. Nature News, 504(7479), 211.

Laudel, G., & Gläser, J. (2008). From apprentice to colleague: The metamorphosis of early career researchers. Higher Education, 55(3), 387-406.

Mavriplis, C., Heller, R., Beil, C., Dam, K., Yassinskaya, N., Shaw, M., & Sorensen, C. (2010). Mind the gap: Women in STEM career breaks. Journal of technology management & innovation, 5(1), 140-151.

OECD (2018), “Gender in a changing context for STI”, in OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption, OECD Publishing, Paris.

Van Den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: a longitudinal case study. Scientometrics, 106(1), 143-162.

Van Den Besselaar, P., Schiffbaenker, H., Sandström, U., & Mom, M. (2018). Explaining gender bias in ERC grant selection–Life Sciences case. STI 2018, Leiden, Netherlands. CWTS.

Ward, M. E. (2001). Gender and promotion in the academic profession. Scottish Journal of Political Economy, 48(3), 283-302.

Prediction of Microblogging Influence and Measuring of Topical Influence in the Context of Terrorist Events

Lu An[1,2], Yuxin Han [2], Xingyue Yi [2] and Gang Li[1]

[1] *anlu97@163.com, imiswhu@aliyun.com*
Wuhan University, Center for Studies of Information Resources, No. 299 Bayi Road, Wuhan (China)

[2] *yxhanccnu@163.com, yixingyue@126.com*
Wuhan University, School of Information Management, No.299 Bayi Road, Wuhan (China)

Abstract

The outbreak of terrorist events usually causes great damage to the society and arouses public concerns on the microblog platform. In this study, an influence prediction model of microblogging in the context of terrorist events has been constructed which involves the user, time and content features of microblogging. The Word2Vec and the K-means clustering technique as well as the text sentiment analysis method are used to extract the topical and sentiment characteristics of microblogging respectively. Users are classified into eleven types and the lifecycle of the event is identified. The experimental results show that the accuracy rate of the proposed model reaches 85.7%, which can effectively predict the influence of microblogging. The characteristics of microblogging with high influence are also explored. A method of quantifying the influence of individual microblog topics based on h-index is also proposed and the evolution patterns of topics are further analyzed. The findings of the study can facilitate identification of microblog entries of high influence in the context of terrorist events, identify topics of high influence and understand their evolution pattern, and assist counter terrorism departments in understanding public opinions and decision-making during terrorist incidents.

Introduction

Terrorist events have the characteristics of suddenness and unpredictability. They pose threats to the safety of the public and society. Once or even before a terrorist event occurs, it often receives much attention and the counter terrorism departments need to make decisions immediately, which needs enough information about the event. With the development of social media, microblogging has become the most popular channel of information dissemination and sharing, which can provide valuable clues for terrorist events. Sina Weibo is the most influential microblog platform in China. According to the China Internet Network Information Center (CNNIC), as of the third quarter of 2017, the quantity of monthly active users of Sina Weibo has reached 376 million. When a terrorist event breaks out, the users' information exchange behavior on the microblog platform can reflect the tendency and change of the public's emotions, thoughts and behaviors in the terrorist incident. By observing the information exchange behavior of the public on Weibo during and after the terrorist event, this study uses machine learning techniques to predict the influence of microblogging about the event, and explores the characteristics and evolution of high-influence microblogging.

Related research

Studies on terrorist events generally consist of three aspects, i.e. information monitoring of terrorist events, information dissemination modes of terrorist events, and terrorist sentiment analysis. Zhou (2017) used a time series neural network model to quantify the activity of extremist supporters and predict their likelihood of launching terrorist activities. In the information monitoring of terrorist incidents, Xu et al. (2017) and Kaur et al. (2016) respectively proposed to classify microblog accounts by machine learning classification models to identify terrorism actions. Shaikh et al. (2015) monitored terrorism-related information by adding context features to the SVM model. Reddick et al. (2015) discussed information monitoring and extraction as well as privacy security issues in the context of terrorist incidents.

In the aspect of social media analysis, Williams et al. (2017) confirmed that social media information about terrorist events can provide key guidance information for counter terrorism departments, which is of great significance for starting the event recovery process and cultivating defense capabilities. In terms of public opinion analysis, Jong et al. (2016) found that negative emotions increased significantly after terrorist events. Hampton et al. (2017) found that the antonym word frequency can describe the urgency of the event. It is found that few studies predict the influence of social media information about terrorist events and reveal their influence evolution patterns.

Research methods

In this study, the microblog entries related to the terrorist event "Kunming Railway Station violent attack" on Weibo were used as experimental data. This study collected microblogging data from Sina Weibo by Python-based crawlers. First, the microblogging texts were preprocessed by removing noisy data, segmenting sentences into words and phrases, and removing stop words. Second, the word2vec and K-means clustering techniques were used to obtain the topics of the microblogging texts. The TF-IDF algorithm was used to sort the words in different clusters to obtain the feature words. Then each topic was described by feature words and summarized.

Third, we used the emotion dictionary and rule-based methods to analyze the sentiment of microblogging content. The existing sentiment dictionary was expanded and a reasonable emotional polarity calculation rule was designed. The emotional polarity of each microblog entry was quantified, and the classification result of the emotional polarity was obtained. The formula for calculating the sentiment polarity of a microblog entry is shown in Equation (1).

$$\mathrm{Sco}(Si) = [\sum Sco(emoi)] \times [W(conj1) \times \cdots \times W(conjn)] \times [W(punc1) \times \cdots \times W(puncm)] + [W(sym1) + \cdots + W(symp)] \quad (1)$$

Here, $\sum Sco(emo_i)$ is the sum of the intensities of all the emotional words in the microblog entry, $W(conj_{\mathrm{i}})$ is the weight of the ith conjunction, $W(punc_m)$ is the weight of the jth punctuation, and $W(sym_k)$ is the intensity of the kth emoji. The sentiment polarity of a microblog entry is determined by the combined weight of emotional words and the weights of conjunctions and emoji. As punctuations have modification effects on emotions, multiplication is performed. Emoji can express emotions independently, so they are added and processed. When Sco(S_i) is larger than 0, the emotional polarity of the microblog entry is positive. When Sco(S_i) is smaller than 0, the emotional polarity of the entry is negative. When Sco(S_i)equals 0, the emotional polarity of the entry is neutral.

To analyze the relative importance of individual features of microblogging, several logistic regression models were constructed, each of which omitted a certain feature. The relative importance of a feature was measured by the difference between the influence prediction model omitting the feature in question and the model considering all the features. The relative importance of the ith feature (Weight(f$_i$)) is calculated as Equation (2) shows.

$$\mathrm{Weight}(\mathrm{f_i}) = \mathrm{Ln}(\mathrm{FP_{LR}} + \mathrm{FN_{LR}} - \mathrm{FP_i} - \mathrm{FN_i}) \quad (2)$$

FP_{LR} and FP_i represent the number of low-influence microblog entries which were wrongly predicted by the original model and by the model omitting the ith feature respectively as high-influence ones. FN_{LR} and FN_i represent the number of high-influence microblog entries which were wrongly predicted by the original modeland by the model omitting the ith feature respectively as low-influence ones.

To illustrate the topical characteristics of microblogging, this study proposed a method of measuring topical influence based on the h-index and revealed the evolution pattern of the influence of the microblogging topics. As the lifecycle of the event consists of four phases, i.e. the initial period, the outbreak period, the recession period, and the calming period, the ith topic at the tth phase was denoted as $k_i(Q_t)$, $t \in (1,2,3,4)$. Rank all the microblog entries regarding $k_i(Q_t)$ in a descending order according to their influence, i.e. the sum of retweets, comments and praises. If the influence of the hth microblog entry was no longer than h and the influence of the h+1th microblog entry is lower than h+1, the h-index of the ith topic at the tth phase equals h. Equation (3) shows the h-index of the ith topic in the four phases.

$$h(k_i) = \{h[k_i(Q_1)], h[k_i(Q_2)], h[k_i(Q_3)], h[k_i(Q_4)]\} \qquad (3)$$

In order to explore the characteristics of high-influence microblogging about terrorist events, this study proposed the influence tendency of feature values to study the relationship between feature values and influence of microblogging. First, we calculated the percentage of high-influence microblog entries among all the microblog entries when a feature took a specific value. Second, we calculated the percentage of high-influence microblog entries among all the corpus. Third, we compared the difference between the two percentages, i.e. the influence tendency of the feature value in question. See Equation (4).

$$\text{Impact}(V_i) = \frac{num(V_i|V_c = High)}{num(V_i)} - \frac{num(D|V_c=High)}{num(D)} \qquad (4)$$

Results analysis and discussion

Data source

The four hot topic hashtags “#blessing Kunming#”, “#God bless Kunming#”, “#Kunming violent attack#” and “#Kunming Railway Station hacking incident#”, which are about terrorist event "Kunming Railway Station violent attack", were employed to collect 153,797 microblog entries between March 1 and March 31, 2014. The influence of microblogging was measured by the sum of the counts of forwarding, comments, and praises. The microblog entries with their influence higher than 10 were considered as those of high-influence.

The user, time and content features of microblog and values

To predict the influence of microblogging in the context of terrorist events, the user, time and content features of microblogging were considered and extracted. See Table 1 for the features of microblogging and their value ranges. A logistic regression model was constructed to predict the influence of microblogging

Table 1 Features of microblogging and their value ranges

Features		Value ranges
User feature	Authentication type	Institutional, personal or no authentication
	User’s location	Provinces
	Whether the user belongs to the public security system	Yes/No

		User’s industry	Traditional media, new media, we-media, government agencies, medical institutions and practitioners, university/research institutions and practitioners, enterprises, public welfare organizations, individual group organizations, public figures, or others
Time feature		Lifecycle	Initial, outbreak, recession, or calming period
		Time span	Late night, early morning, morning, noon, afternoon, or evening
Content feature	Text structure	Original/Retweet	Original/Retweet
		Whether the content contains an URL	Yes/No
		Whether the content contains a hashtag	Yes/No
		Whether the content mentioned a user	Yes/No
		Whether the content contains an emoji	Yes/No
	Event keywords	Location keywords	Yes/No
		Time keywords	Yes/No
		Behavior keywords	Yes/No
		Figure keywords	Yes/No
	Text topics	Topical labels	Topic0 – Topic24
	Text sentiment	Sentiment polarity	Positive, negative or neutral

Topic identification and sentiment analysis in the microblogging texts

To obtain a reasonable number of topics that are suitable for later analysis, we first tried to cluster the terms into 20 to 30 topics by the word2vec and k-means method. It was found that the clustering effect was the best, achieving high intracluster similarity and low intercluster similarity when the number of clusters was 25 after several rounds of experiments. Thus, a number of 25 topics were identified. They were further summarized into four categories, i.e. reports of violent incidents and related derivative events, emotional expressions, topic discussions, and new public opinion events.

The emotional polarity of each microblog entry was calculated, considering emotional words, adverbs of degree, negative words, emoji, punctuations and conjunctions. The sentiment of all microblogging entries was showed in Table 2.

Table 2 Sentiment of microblogging entries

Sentiment of microblogging entries	positive	negative	neutral
Number of microblog entries	9517	3667	2391

Evaluation of the microblog influence prediction model

The confusion matrix revealed that the influence prediction model of microblogging in the context of terrorist events can successfully predict 85.7% of microblogging influence overall. In order to examine the performance of the logistic regression (LR) model, we experimented on the same data, employing C4.5 decision tree (DT), Bayesian belief network (BN), naive Bayes (NB), random forest (RF), Multi-layer perceptron (MLP) and support vector machine (SVM) respectively and compared their performance. See Table 2 for the comparison results. It can be seen that the predictive model based on the logistic regression technique had superior performance over other models.

Table 3 Performance comparison among different classification models

	LR	DT	BN	NB	RF	MLP	SVM
Precision	**0.858**	0.856	0.837	0.837	0.836	0.847	0.855
Recall	**0.857**	0.849	0.83	0.83	0.835	0.846	0.849
F-measure	**0.855**	0.847	0.827	0.827	0.834	0.846	0.847
ROC area	**0.919**	0.891	0.9	0.9	0.903	0.903	0.838

Analysis of the relative importance of microblogging features

Using Equation (2), the relative importance of each feature was shown in Figure 1.

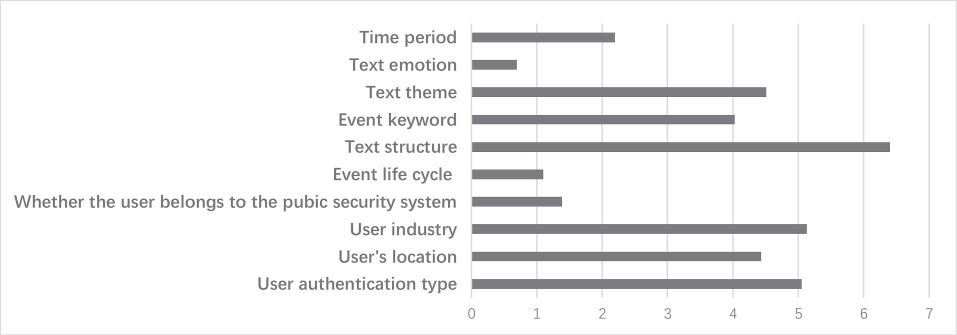


Figure 1 The relative importance of each microblogging feature

The influence tendency of microblogging feature values

Using Equation (2), from the perspective of subject classification, the theme of the event report class has a relatively high high-impact tendency, and the topic discussion class also has a certain high influence tendency. The emotional expression class and the new public opinion event class have low influence tendency as a whole.

Measuring the influence of microblogging topics and their evolution analysis

As introduced in the section of Research methods, the h-index values of each topic at each phase were calculated. In the initial period, topic 0 (The gangsters cut people with knives had the highest h-index value (52). In the outbreak period, topic 0 continued to lead (achieving the h-index value of 190) and topic 11 (Expressing blessing for the Kunming city) ranked second (158). In the recession period, other events occurred, such as Malaysia Airlines lost contact (topic 22 and topic 5). However, topic 23(The suspects were captured) were also salient (the h-index value of 28). In the calming period, topic 1(Case planning and casualties) were still influential (the h-index value of 22).

Conclusion

This study constructed an influence prediction model of microblogging in the context of terrorist events. Seventeen features in three aspects, i.e. user, time and content characteristics were considered and extracted from the microblogging corpus. The Kunming railway station violent attack event was chosen as the investigation case and a total of 153,797 related microblog entries were collected. The influence prediction model based on logistic regression can successfully predict 85.7% of microblogging overall and has superior performance over other six classification algorithms. The text structure, user's industry and authentication type are the first three most important features. The influence tendency of feature values has been proposed and calculated for each feature value. It is found that we-media and individual group organizations in the user's industry feature, institutional accreditation in the user's authentication feature, and "yes" in the feature of whether the user belongs to the public security system tend to arouse high influence. The h-index was used to measure the influence of topics and their evolution patterns were also explored. The findings can help counter terrorism departments effectively and rapidly predict microblogging and topics of high influence and take preventive measures in advance.

Acknowledgments

This research was supported by the Major Project of the Ministry of Education of China (Grant No. 17JZD034) and the National Natural Science Foundation of China (Grant No. 71603189, 71790612).

References

China Internet Network Information Center. "The 41st Statistical Report on Internet Development in China" Retrieved May 2, 2018 from http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/t20180305_70249.htm.

Hampton, A. J., & Shalin, V. L. (2017). Sentinels of Breach: Lexical Choice as a Measure of Urgency in Social Media. *Human Factors*, 59(4), 505-519.

Jong, W., & Dückers, Michel L.A.(2016).Self-correcting mechanisms and echo-effects in social media: an analysis of the "gunman in the newsroom" crisis. *Computers in Human Behavior*, 59, 334-341.

Kaur, K. (2016). Development of a framework for analyzing terrorism actions via twitter lists. *Proceedings of the International Conference on Computational Techniques in Information & Communication Technologies (ICCTICT)*. New Delhi: IEEE, pp 19-24.

Reddick C G, Chatfield A T, Jaramillo P A. (2015). Public opinion on national security agency surveillance programs: a multi-method approach. *Government Information Quarterly*, 32(2),129- 141.

Shaikh M., Salleh N., Marziana L. (2015). Social networks content analysis for peacebuilding application. In: Abraham A., Muda A., Choo YH. (eds) Pattern Analysis, Intelligent Security and the Internet of Things. *Advances in Intelligent Systems and Computing*, vol 355. Springer, Cham, pp 193-200.

Williams, G. A., Woods, C. L., & Staricek, N. C. (2017). Restorative rhetoric and social media: An examination of the Boston Marathon Bombing. *Communication Studies*, 68(4), 385–402.

Xu J, Lu T C. (2017). Automated classification of extremist Twitter accounts using content-based and network-based features, *Proceedings of the 2017 IEEE International Conference on Big Data Workshop*. Boston: IEEE. (pp. 2545-2549).

Zhou Y. (2017). Pro-ISIS fanboys network analysis and attack detection through Twitter data, *Proceedings of the 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. Beijing: IEEE, (pp. 386-390).

Making it personal: Examining gendered personalization patterns of single-authored papers

Gita Ghiasi[1], Maxime Sainte-Marie[2], and Vincent Larivière[3]

[1] *gita.ghiasi.hafezi@umontreal.ca*
[2] *msaintemarie@gmail.com*
[3] *vincent.lariviere@umontreal.ca*
École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal, CP6128, Station Centre-Ville, Montreal, Quebec,H3C3J7 (Canada).

Abstract

Authorial self-mentions as one of the primary forms of scientific communication in sole authorship could convey not only the salient characteristics of disciplinary epistemologies but a representation of the author. The latter is often reflected through the use of first-person pronouns. The objective of this study is to assess disciplinary, national, and gender differences in the use of first-person subject pronouns in sole-authored scientific articles. Following the analysis of all sole-authored articles published between 2008 and 2017 and indexed in the Web of Science, this study reveals that women refer to themselves in the singular form more often than men, which could be indicative of cultural associations of masculinity and authority in the system of science. The results of this study are of great importance to academic evaluative processes where sole-authorship is rewarded as an independent work and author's credibility.

Introduction

Despite the constant decline in numbers (Abt 2007; Barlow et al. 2018; Kuld and O'Hagan 2018), the reward system of science has traditionally considered sole-authorship as a measure of a researcher's ability to work independently (Gasparyan et al. 2013; Moore and Griffin 2006). Due to the dubious tendency of conferring greater credit to individual contributions, sole-authoring has been given special importance in several research review and evaluation systems, impacting promotion, tenure, and funding allocation management (Moore and Griffin 2006; Vafeas 2010). Moreover, it has also been argued that sole-authorship could offer a distinct level of recognition for the author than co-authorship, which could serve a researcher beyond the tenure and promotion (Moore and Griffin 2006) in the scientific system.

As regards to gender, sole-authorship could play a vital role in women's promotion and tenure reception: women are not only less likely to receive tenure when they co-author, but they also receive less credit for their contributions in co-authored papers (Sarsons 2017). Despite this, women are highly underrepresented as sole-authors (West et al. 2013), and female sole-authored papers are subject to lower citation rates than male-authored ones (Bendels et al. 2018; Larivière et al. 2013).

Along these lines, linguistic choices on authorial self-mention could also impose a considerable impact on scholarly communication, because it sheds light on the author's epistemological and social self-reflection by drawing on assumptions authors hold about their role in the research process (Hyland 2003). Gender differences in language use are of special interest here, as they can provide valuable insight into the social aspects of communication (Newman et al. 2008). Particularly, differences in authorial self-mention provide a rhetorical strategy to mirror an author's self-reflection of his/her contribution to a piece of research, presentation, and promotion of his/her knowledge claims, and research credibility (Hyland 2003).

In this regard, the main objective of this paper is to assess gender differences in the use of pronouns in single-authored articles. More specifically this study analyzes the referential meaning and pragmatic function of both singular ('I') and plural ('We') forms of first-person subject pronouns (termed as 'S' and 'P' hereafter, respectively) that researchers of each gender use to promote their scientific claims and themselves. For this purpose, a cross-country and cross-disciplinary analysis of the use of first-person subject pronouns in abstracts and acknowledgments of single-authored articles is first provided, and gender is further factored into these differences.

Methods

Abstracts and acknowledgments of all 1,184,186 single-authored scientific articles (hereafter simply referred to as 'articles') published between 2008 and 2017 inclusively are extracted from the Web of Science (WoS). For each relevant article entry, the following attributes are extracted: article ID, author full given name, acknowledgment, abstract, publication year, and journal name. Discipline assignation is based on the National Science Foundation (NSF) field classification of journals used in the Science and Engineering Indicators (SEI) reports. Contrary to the WoS disciplinary classification, the NSF classification scheme assigns only one discipline to each journal, which prevents multiple counts of articles published in multidisciplinary journals. The relative citation of a paper is measured as the average yearly number of citations received by a paper divided by the average yearly number of citations to all the papers from the same year, in the same discipline and of the same document type. The normalized journal impact factor is defined similarly, considering the IF of the journal in which a paper is published. In this study, top-cited papers and top-impact journals refer to the top 5% cited papers and the top 5% high impact journals.

Following the procedure described in Larivière et al. (2013), gender is then assigned to article authors using universal and country-specific name lists in sequence. In the first step, gender is attributed to all authors based on 1990 US census data, which provides lists associating each given name to the percentage and gender of the population bearing that name. In cases where a name is used for both genders, a specific gender is assigned only if the corresponding gender is assigned ten times more frequently than the other. In a second step, all unassigned author names are then matched with corresponding entries in country-specific name lists, based on the geographical information given by the institutional affiliations of authors. These assignment procedures resulted in the author gender identification of more than 985,721 articles, which corresponds to 96% of total articles in which the full first name of the author is given.

Finally, in order to extract all first-person personal pronoun information contained in both abstracts and acknowledgments, the textual content of both attributes is first grammatically disambiguated using the TreeTagger part-of-speech tagger (Schmid 1999, 2013), as trained on the British National Corpus tagset (Leech et al. 1994). Following this, words segmented into either 'I' or 'We' and tagged as 'PNP' (corresponding to the 'personal pronoun' part-of-speech tag) by the algorithm are counted for each of the two article attributes. Articles are further categorized under 'I' and 'We', when one attribute equals to zero, and the other is equal or more than 1.

Findings

The results show that authors tend to rely more on S in their acknowledgments and on P while characterizing and summarizing their research. This sheds light on the differences in scientific communication when authors describe their research and express a personal statement, in the sense that P is more popular in the scientific and more formal context. However, the use of S increases in highly cited papers, while P is utilized at a higher rate in papers published in high impact journals. This trend highlights differences in rhetorical stance between two different recognition criteria: publishing in high impact journals and attracting citations. S is mostly used in social sciences, humanities, and psychology while P is most popular in mathematics, physics, and engineering (Fig. 2: Left). This corresponds to the findings of (Hyland 2003), who associated these opposite trends in 'hard' and 'soft' sciences to cross-disciplinary differences in the ways that research is conducted and accepted by the scientific community. Among highly prolific countries, S is highly employed among North American countries (Canada and the US), while P is more used in East Asian countries (China and Japan) (Fig. 2: Right). These differences are further scrutinized using the Hofstede's individualism dimension of national culture (Hofstede 1984) (Fig. 3), which reveals a strong correlation between the two.

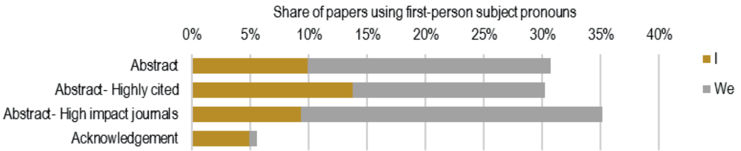


Figure 1- share of sole-authored papers using first-person subject pronouns in their abstracts and acknowledgments

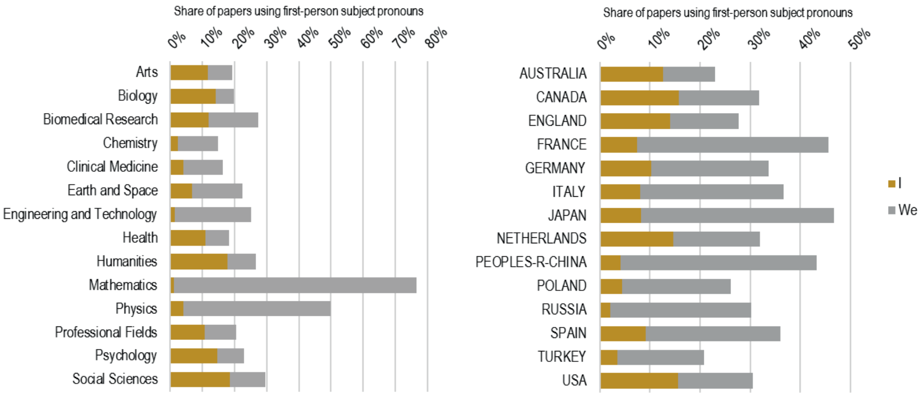


Figure 2- share of sole-authored papers using first-person subject pronouns (left) by filed and (right) by county

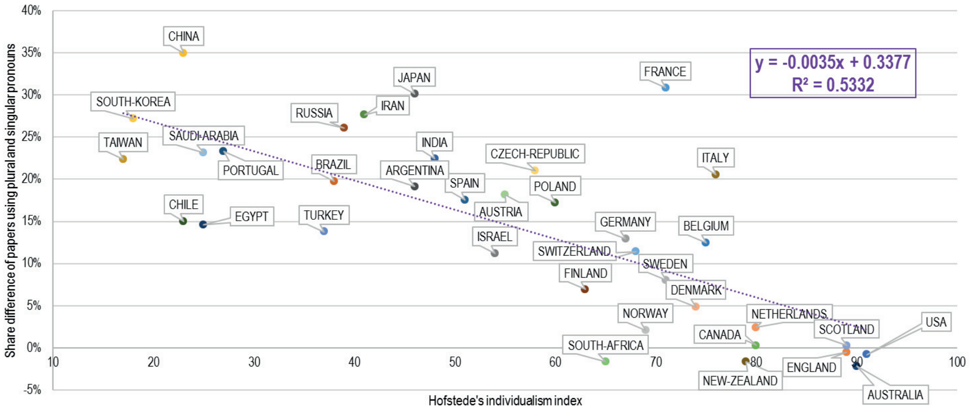


Figure 3- differences in shares of papers using singular and plural pronouns by Hofstede cultural individualism index

Gender Analysis

Considering only articles using S and P, results show that women use S more often than men in their papers and even more so in highly cited papers (Fig. 4). The largest gender differences in the use of pronouns are in the field of social sciences, and professional fields, where women use S more often (Fig. 5), whereas in biomedical research women use P at a higher proportion than men than men.

While considering the national tendencies, gender differences in the use of pronouns in the scientific writing are more pronounced in countries with low power distance and are negligible in societies with high power distance such as China and Russia (Fig. 6 and 7). This could relate to the centralization of authority (Hofstede 1984) and its strong association with social uniformity.

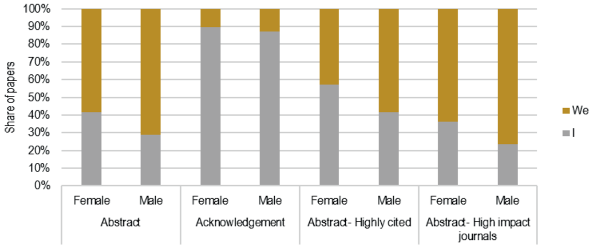


Figure 4- Share of sole-authored papers using singular and plural pronouns by gender

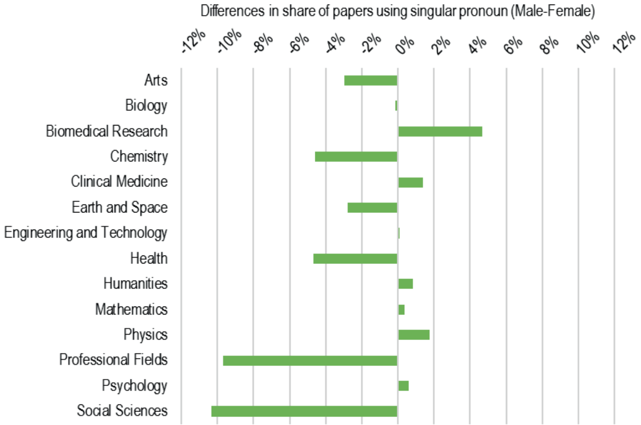


Figure 5- Gender differences in share of papers using singular pronoun by field (Male - Female)

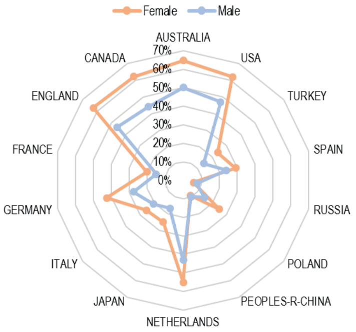


Figure 6- Share of papers using singular pronoun by gender and by country

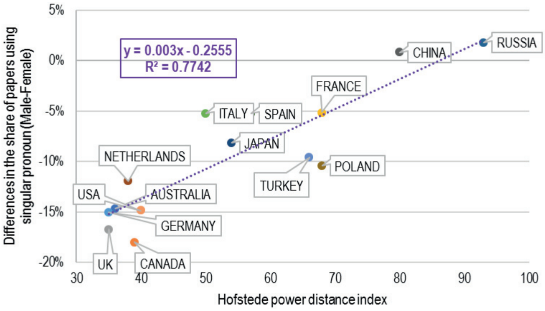


Figure 7- gender differences in the use of singular pronoun by Hofstede individualism index (Male - Female)

Discussion and Conclusion

Results of this study reveal that the use of 'We' generally prevails in the single-authoring context. While 'I' is practiced more often among top-cited papers, articles in top impact journals use 'We' more often. This shows that the use of the plural form might present stronger support for authors to present their research claims for the review process, while the singular person might present a stronger research creditability for a larger audience and the scientific community. Moreover, this study shows that 'I' is used more often in social sciences, humanities, and psychology, while 'We' is most used in mathematics, physics, and engineering disciplines. This is in line with the findings of Hyland (2003), who argued that since hard sciences are experimental, and results are replicable, therefore authors stay objective to their findings and use less intrusive and personal writing style. On the other hand, due to the use of interpretive approach, the level of personal engagement with the readers is important in soft sciences, in the sense that authors are required to present themselves as an informed researcher with a particular point of view to receive credibility for their perspective and research claims.

With regard to gender differences, this study shows that women refer to themselves as "I" more often than men and men use 'We' more often than women and that these differences are highly conspicuous in social sciences and professional fields. This could be indicative of cultural associations of masculinity and authority in the system of science. Because the use of plural pronouns may distance the author from the text, yet exhibit a temporary dominance by conferring the right to speak with authority on authors (Hyland 2001). The indication of less personal intrusion could also refer to the author and his/her possible colleagues and help acceptance of research arguments, based on which the reader assumes that scientific claims of the paper are supported by a research group or community (Zhou 2017). Therefore, one of the possible explanations for larger gender differences observed in social sciences is that women's contribution to papers is the result of an individual endeavor while for men is more of the team efforts.

Finally, this study confirms that the authors' use of singular and plural pronouns in the scientific papers are associated with how a national culture defines its self-image as "I" or "We" and that gender differences in communication style is less evident in countries with power distance culture, as plural pronouns are uniformly practiced to reflect authority.

The results of this study are of great importance to academic evaluative processes where sole-authorship is rewarded as an independent work and author's credibility. Authorial self-mentions as one of the primary forms of scientific communication in sole authorship could convey not only the salient characteristics of disciplinary epistemologies but a representation of the author, upon which research credibility and recognition within the scientific community could be attained. Therefore, gender differences in authorial self-mentions could also address under-recognition of the contribution of women to science.

References

Abt, H. A. (2007). The future of single-authored papers. *Scientometrics*, *73*(3), 353–358. doi:10.1007/s11192-007-1822-9

Barlow, J., Stephens, P. A., Bode, M., Cadotte, M. W., Lucas, K., Newton, E., et al. (2018). On the extinction of the single-authored paper: The causes and consequences of increasingly collaborative applied ecological research. *Journal of Applied Ecology*, *55*(1), 1–4. doi:10.1111/1365-2664.13040

Bendels, M. H. K., Müller, R., Brueggmann, D., & Groneberg, D. A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. *PLOS ONE*, *13*(1), e0189136. doi:10.1371/journal.pone.0189136

Gasparyan, A. Y., Ayvazyan, L., & Kitas, G. D. (2013). Authorship problems in scholarly journals: considerations for authors, peer reviewers and editors. *Rheumatology international*, *33*(2), 277–284.

Hofstede, G. (1984). *Culture's Consequences: International Differences in Work-Related Values*. SAGE.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for specific purposes*, *20*(3), 207–226.

Hyland, K. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of the American Society for Information Science and technology*, *54*(3), 251–259.

Kuld, L., & O'Hagan, J. (2018). Rise of multi-authored papers in economics: Demise of the 'lone star' and why? *Scientometrics*, *114*(3), 1207–1225. doi:10.1007/s11192-017-2588-3

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature News*, *504*(7479), 211. doi:10.1038/504211a

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1* (pp. 622–628). Association for Computational Linguistics.

Moore, M. T., & Griffin, B. W. (2006). Identification of factors that influence authorship name placement and decisions to collaborate in peer-reviewed, education-related publications. *Studies in Educational Evaluation*, *32*(2), 125–135.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), 211–236.

Sarsons, H. (2017). Recognition for Group Work: Gender Differences in Academia. *American Economic Review*, *107*(5), 141–145. doi:10.1257/aer.p20171126

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora* (pp. 13–25). Springer.

Schmid, H. (2013). Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing* (p. 154).

Vafeas, N. (2010). Determinants of single authorship. *EuroMed Journal of Business*, *5*(3), 332–344. doi:10.1108/14502191011080845

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The Role of Gender in Scholarly Authorship. *PLOS ONE*, *8*(7), e66212. doi:10.1371/journal.pone.0066212

Zhou, L. (2017). The Pragmatic Effect of the First Person Plural Pronouns in Single-authored Scientific Articles. In *3rd International Conference on Arts, Design and Contemporary Education (ICADCE 2017)*. Atlantis Press.

Characterizing the Heterogeneity of European Higher Education Institutions Combining Cluster and Efficiency Analyses

Renato Bruni[1], Giuseppe Catalano[1], Cinzia Daraio[1], Martina Gregori[2,1], and Henk F. Moed[1]

[1] *bruni@diag.uniroma1.it, giuseppe.catalano@uniroma1.it, daraio@diag.uniroma1.it, martina.gregori@uniroma1.it, henk.moed@uniroma1.it*

[1]Sapienza University of Rome, Department of Computer, Control and Management Engineering (DIAG), Via Ariosto 25, 00185, Rome (Italy)

[2]Sapienza university of Rome, Department of Mechanical and Aerospace Engineering (DIMA), Via Eudossiana 18, 00184, Rome (Italy)

Abstract

The heterogeneity of the Higher Education (HE) Institutions is one of the main critical issues to address properly the assessment of systemic performance. We adopt a multi-level perspective by combining national (macro) and institution (micro) level data and analyses. We combine clustering and efficiency analysis to characterize the heterogeneity of HE systems (at the national level) exploiting micro level data. We show also the potential of using micro level data to characterize national level performance. The obtained results may provide a quantitative support to identify the higher education institutions that need to be further investigated through qualitative case studies in political science analyses of HE systems.

Introduction

The measurement of academic performance is a relevant issue at the intersection between political science and informetrics. Although the analysis of the performance of Higher Education Institution (HEI) systems is a complex task, there are numerous international comparisons (rankings) of institutions such as Shanghai, Times Higher Education and Leiden Ranking that are published on a regular basis. The heterogeneity of the HEIs is one of the main critical issues to address properly the assessment of performance, in a multi-level (systemic) perspective. There are different sources of heterogeneity, including the mission, the national context, the presence or absence of medical schools, the legal status and the disciplinary orientation and degree of specialization (López-Illescas et al., 2011; Daraio et al., 2011).

We adopt a multi-level perspective by combining national (macro) level data and institution (micro) level data and analyses. We show also the potential of using micro level data to characterize national level performance. In a way we attempt to characterize HEIs accounting for their (i) *Structural* heterogeneity (structure of the national system: systemic factors, e.g. number and types of HEIs that are at place, governance factors), (ii) *Internal* heterogeneity (linked to the type of the production process carried out within the HEIs) and (iii) *Other* heterogeneity sources.

This work presents the first results from a larger project (see Acknowledgements), aimed to study the activities, the performances and the efficiencies of European HEIs. It focuses on a statistical exploration of a series of indicators linking Education, in a *systemic way*, with Research and Innovation. In terms of data analysis, it explores the combination of statistical data from ETER, the European Tertiary Education Register, with bibliometric data obtained from the Leiden Ranking, and with categorizations of national higher education policies obtained from more qualitative studies of national HEI systems. In the project, the existing problems of data availability, quantification and comparability go hand in hand with the need of conceptualization of the performance model before making the analysis (Daraio and Bonaccorsi, 2017). The notion of performance is characterized in a "progressive" way, starting from production ("volume" or extensive variables), going to productivity (intensive or

"size-independent" indicators of production), up to efficiency (combination of outputs/inputs) and more elaborated efficiency models, towards effectiveness and impact (Daraio, 2019).
The present work is organized in two parts. In the first part we tackle the heterogeneity of HEIs calculating country-level statistics based on micro data and analyzing them with qualitative and governance variables. We will call this section the Quali-quantitative analyses.
In the second part of the work, we give an order to this heterogeneity calculating a teaching and research productivity score and providing a cluster analysis that allows us to identify some typologies of HEIs.
The main objective of this work is then to combine clustering and efficiency analysis to characterize the heterogeneity of HE systems (at country level) exploiting micro level data.

Methods

The methods used are K-means (MacQueen J.B., 1967) and DBSCAN (Density-Based Spatial Clustering of Applications with Noise; Ester, M. et al. 1996) clustering approaches, and nonparametric efficiency analysis (Free Disposal Hull, FDH estimation of efficiency scores and a more robust nonparametric estimation in progress, see Daraio and Simar (2007).
K-means is a well-established clustering technique. It aims at partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean (which actually constitutes the centroid of the cluster). The application of this principle leads to a partition of the data space into Voronoi cells. Data are therefore iteratively clustered in n groups of equal variances, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified.
DBSCAN, on the other hand, is a more recent clustering technique but is it one of the most used and cited approaches. The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Therefore, it groups together points that are closely packed together (points with many nearby neighbours), marking as outlier points that lie alone in low-density regions. This set of outliers can eventually be viewed as the last or residual cluster. Due to this density-based approach, the clusters obtained by DBSCAN can be of any shape, as opposed to K-means which assumes that clusters are convex shaped, and the number of clusters cannot be specified in advance.
We estimate the efficiency of universities in producing teaching and research and use the efficiency scores as an additional variable to characterize the groups of universities obtained from the cluster analyses. The DBSCAN cluster analysis lead us to identify three clusters. After that, we run the K-means clustering to characterize the three groups of universities. The combination of the two different approaches was useful to shed some lights on the robustness of the choice done in the K-means approach.

Data

A HEIs performance evaluation analysis, to be as much as possible representative and complete, needs to take into account indicators related to all the different activities carried out in the academic operations, namely teaching, academic research and Third Mission activities (collaboration with industries, patents, etc.).
With the purpose of gather information about the three aforementioned areas, it was made use of different sources. In particular, the following three databases were integrated for the analysis at the micro-level (single institution):

ETER database, *for the information at micro-level (single institution), regarding the TEACHING area*

CWTS Leiden Ranking database, *for the information regarding the academic research*

PATSTAT PATENTS database, *for the information regarding the patents registered*

The list of the variables considered is reported in Tables 1, 2, 3. In addition, a database dedicated to the national regulatory characteristics of European countries was integrated in order to outline part of the heterogeneity present among higher education macro-systems. The considered governance indicators (reported in Table 3) are elaborated in Capano and Pritoni (2019), cover the period 1988–2014 and consider 12 European countries[1]. The governance indicators are in total 24, grouped in 4 dimensions (Regulation, Expenditure, Taxes, Information) and represent the number of government interventions on the observed period in each specific sub-areas. In order to include these indicators in our analysis, we applied few trasformation on data. For each country, the scores per dimension were summed together; next, per dimension a *percentage* score was calculated relative to the total score on this dimension over all countries. Moreover, 3 further indicators have been calculated based on the Capano and Pritoni (2019) data, to try to capture the trends of national government towards a more or less restrictive approach on HEIs system regulation and verify the grade of application of *control measure* on micro-level performance (see Table 3).

In the final dataset, it was necessary to structurally internalise temporal lags between inputs and outputs information. It is well known that a certain period of time has to pass in order to observe effects related to the interventions on academic staff, academic funds, and so on. As it is usually done in the empirical analyses, one year lag to observe effects on academic research publication and two years lag to observe effects on patents applications are acceptable average periods to be assumed. Hence, the data considered refer to the following time ranges: 2011–2014, ETER database (teaching and basic information on inputs); 2012–2015, CWTS data (academic research information)/INCITES database; 2013–2016, PATENTS database.

The teaching outputs (mainly, number of graduates for each degree class) are referred to the same time period of the inputs variables (e.g. academic staff, funds). The choice was driven both by the lack of data of high quality and completeness for years after 2014 and the difficulty in establishing an acceptable lag, due to the different degree classes considered in the analysis. Nevertheless, it has been verified that the annual values assumed by the teaching outputs variable do not vary significantly year by year, in a short range of time.

Furthermore, data on numbers of inhabitants (obtained from OECD and EUROSTAT for the year 2016) are used to draw the possible relation between countries dimensions in population and the HEIs produced output.

The final dataset contains the average variable values over the considered period of each included database; missing values had been excluded from the calculation.

Table 1. Research funding based indicators

Use of metrics in education *(0=NO; 0.5=LIMITED, 1.0=YES)*
Research performance based funding *(0=NO; 0.5=LIMITED, 1.0=YES)*
Use of quantitative formula in research funding *(0=NO; 0.5=LIMITED, 1.0=YES)*
Use of peer review in research funding *(0=NO; 0.5=LIMITED, 1.0=YES)*

We selected all the institutions categorized as universities in the ETER dataset, and for which data are available both on staff, students, graduates, and on publications and citations in the Leiden Ranking dataset. The total number of selected institutions for all ETER countries combined amounts to 664. Nevertheless, due to the presence of missing values on key

[1] The countries considered in Capano & Pritoni (2019) are: Austria, Denmark, England, Finland, France, Greece, Ireland, Italy, Netherlands, Norway, Portugal and Sweden.

variables (namely, academic staff and number of enrolled students) with respect to the cluster analysis procedure, the second quantitative analysis was performed only on a sub-selection of the database composed by 383 HEIs from 22 countries.

Table 2. Definition and source of variables at micro-level

Category (Source)	*Variables*	*Definition*
Cluster analysis variable (ETER, CWTS for Pub_fract)	**Grads_ISCED.5-7/ACADstaff**	Total graduates ISCED 5-7 divided by the Total academic staff (in Full Time Equivalent, FTE); both values represented by yearly averages.
	Pub_fract(av)/ACADstaff	Number of publications (fractional counting) divided by the Total academic staff (FTE); both values represented by yearly averages.
Efficiency analysis	**Mod.Teach.Res.X_ACADSTAFF.FDH**	FDH Inefficiency scores. It may be higher or equal to 1. It is 1 for efficient units, higher than 1 for units that can expand the production of their outputs.
Basic data (ETER)	**Foundation_year**	HEI foundation year.
	Uni_Hospital	Dummy; 1 = presence of an Hospital in the Institution.
	Enrolled_student_ISCED.5-7	Total student enrolled at ISCED 5-7.
	ACADstaff_FTE	Total academic staff (expressed in FTE).
	PhD_intensity_2014	PhD intensity (year of reference: 2014).
	FullProf/ACADstaff_Head	Percentage of full professor on the total academic staff.
	WomenProff_share	Percentage of women on the total number of professors.
	Admn/TOTstaff_FTE	Percentage of administrative staff on total staff (academic plus administrative).
Third mission (Funds ETER)	**Funds_external%**	Percentage of funding from third parties on total funding.
	Funds_third_part/ACADstaff_FTE	Third party funds per academic staff (expressed in FTE).
Specialization (ETER)	**Specialization**	Express the specialization with respect to the disciplinary areas; it is calculated making reference to the Herfindahl index on academic staff. The missing values are filled in with the Herfindahl index on PhD graduates and, in few cases, Herfindahl index on students ISCED 5-7. The values refer to the year 2014.
Research quantity and quality (CWTS, ETER for Acad_staff)	**Pub_top10(av)/ACADstaff**	Number of papers in top 10% (yearly average) divided by the Total academic staff (FTE).
	Pub_in_top10%	Percentage of papers in top 10% (yearly average).
	Pub_international_coll	Percentage of papers with international collaborations (yearly average).
	mnsc_(w-av)_av	Papers mean normalized citations (yearly average, weighted by the number of patent applications).
Third mission (PATENTS, ETER for Acad_staff)	**Patent_application(av)/ACADstaff**	Overall total number of patent applications (yearly average).
	Back_citations(av)/ACADstaff	Number of patents' backward citations (yearly average).
	NPL_av	Number of academic papers citations for patents (yearly average).
	NPL_av/SPA_av	Number of citations from academic papers for each patent (yearly average).

Table 3. Definition and sources of variables at macro-level

Governance (Capano and Pritoni, 2019)	**GOV_Regulation**	Percentage of policy intervention on Regulation [assessment, evaluation and accreditation; agency of assessment, evaluation and accreditation; content of curricula; academic career and recruitment; regulation on students (admission and taxation), institutional and administrative governance; contracts].
	GOV_Expenditure	Percentage of policy intervention on Expenditure [Grants; subsidies and lump-sum funding; targeted funding; loans; performance based institutional funding; standard cost per student].
	GOV_Taxes	Percentage of policy intervention on Taxes [tax exemption; tac reduction for particular categories of students; service-based student fees].
	GOV_Information	Percentage of policy intervention on Information [transparency; certification; monitoring and reporting].
	GOV_Cons_trend	In each country, percentage of regulatory interventions aimed to add *more constraints* respect to the overall regulatory interventions in Regulation.
	GOV_Opp_trend	In each country, percentage of regulatory interventions aimed to add *more constraints* respect to the overall regulatory interventions in Regulation.
	GOV_Control_measures	In each country, percentage of regulatory interventions in the monitoring and reporting, rules on goals in teaching, assessment subjects, respect to the overall regulatory interventions.
System structure (ETER)	**EU_fract_country**	Total enrolled students in the country / Total enrolled student in ETER database (without Turkey).
	NAT_UNI_fract (*number*)	Total number of HEIs of university type in the country / Total number of HEIs of any type in the country.
	NAT_UNI_fract	Total enrolled students in the university institutions in the country / Total enrolled student in HEIs of any type in the country.
	NAT_HEI_fract	Total enrolled students in an institution / Total enrolled students in the country.

Quali-quantitative analyses

Characterizing the heterogeneity of HE systems combining bibliometric indicators, higher education data and Research performance based funding (RPBF) information

This section uses a useful classification of European countries according to whether they have research performance-based funding, proposed by Zacherewicz, Reale, Lepori and Jonkers (Science & Public Policy, 2018, Table 1). This classification is available for 25 countries. Hence, the analyses presented in this section relate to institutions in these 25 countries.

The table by Zacherwicz et al. contains the following information on the research funding system. This system includes Bibliometrics (both Publications, Journal impact based measures, and Citations). As regards "Other formula, elements" it includes indicators on PhD graduates, Patents, Project funding, and Business funding. Finally, it takes into account information from Peer review and Performance Contracts.

The classification in their Table does *not* take into account the factor *time*, although the table's legend gives some additional information about this factor. Funding systems change over time. If a system has been implemented, it takes several years before one can observe any effect at all. Hence, countries that have recently changed their funding system into a performance-based system may not show any effects in the data analysed in this report.

The percentage of top publications (% TOP PUBL in Figure 1) is one of the most frequently used indicators of citation impact. A top publication is a publication of which the citation rate of is among the top 10 percent most frequently cited papers in the subject filed covered by that publication. A country's percentage of top publications is calculated relative to its total publication output. The number of graduates per academic staff is an often used measure of the graduation productivity. The two indicators are probably among the best possible measures for citation impact and teaching performance.

Figure 1 shows a scatterplot of these two variables. Moreover, it indicates whether or not a country has a research performance-based research funding (RPBF) system. The category 'Other' in Figure 1 contains three countries for which PBRF-classifications are unavailable: Germany (DE), Liechtenstein (LI) and Serbia (RS). In Germany, institutional funding of universities is mainly provided at the regional level. As allocation procedures differ from state to state, the authors have not assigned a score to the country as a whole. For a fourth country, The Netherlands, the PBRF table indicates a 'limited PBRF', because in this country 'performance contracts' constitute a determinant for institutional funding.

Figure 1 reveals a rather scattered pattern, showing substantial differences among countries, but there is no sign of a statistical correlation between graduation performance or research impact on the one hand, and RPBF on the other. The next section further quantifies this degree of correlation.

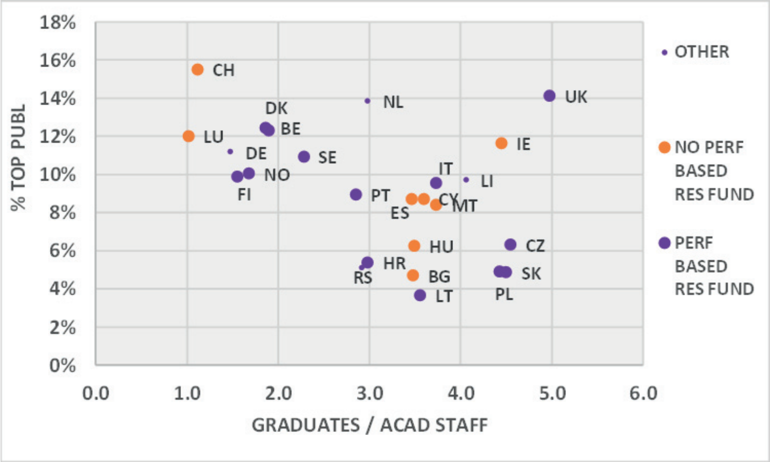


Figure 1. Scatterplot of % Top Publications against Graduates per Academic Staff.

Statistical correlations between 6 key variables

Statistical correlations were calculated pair-wise between the following six indicators:

Total academic staff;

PhD intensity;

Publications per academic staff;

Graduates per academic staff;

Percentage of Top Publications;

Degree of research performance based funding (RPBF);

Total academic staff is size dependent and a good measure of 'size'. The next four indicators are size independent, and measure research intensity, publication and graduation productivity, and citation impact, respectively. The Research performance based funding (RPBF) indicator is derived from Table 1 in Zacherewicz et al. (2018). If this table indicates RPBF, a value of one is assigned; no RPBF corresponds to the value zero. Since there are only nine countries for which data is available both for governance indicators and for the first 5 key indicators, no governance indicators are included in the key set.

Pearson correlations were calculated between each pair of variables. In addition, partial correlations between each pair were calculated, partially out the other four indicators. The number of countries for which data is available for each of the 6 indicators amounts to 25. Table 4 gives results for pairs for which the significance level in at least one of the two computations is above 95 per cent.

Table 4. Statistically significant Pearson and partial correlation coefficients (6 key variables)

Variable 1	*Variable 2*	*Pearson corr.*		*Partial corr.*	
		R	Prob	R	Prob
%Top Publications	Publications per Acad_staff	0.74	0.00	0.82	0.00
%Top Publications	PhD Intensity	0.53	0.01	0.44	0.07
% Top Publications	Total Acad_staff	0.26	0.21	0.61	0.01
%Top Publications	Total grads per Acad_staff	-0.44	0.03	-0.47	0.05
% Top Publications	Research perf. based funds	-0.13	0.58	-0.57	0.01
Research perf. based funds	Publications per Acad_staff	0.12	0.60	0.49	0.04
Research perf. based funds	Total Acad_staff	0.23	0.30	0.48	0.04
Total grads per Acad_staff	PhD Intensity	-0.61	0.00	-0.22	0.39
Total grads. per Acad_staff	Total Acad_staff	0.09	0.66	0.55	0.02

The following observations can be made.

- At the level of countries, citation impact (% Top publications) positively correlates with publication productivity (strongly) and PhD intensity (moderately). It correlates significantly with 'size' (Total academic staff) only if the other factors are partially out. It should be noted that a country's total academic staff is largely determined by demographical factors, for instance, the number of inhabitants.

- Citation impact correlates negatively with graduation productivity. This outcome reveals that, at least at the level of countries, a strong focus on research tends to go hand in hand with a lower graduation performance. It also shows a negative correlation with the degree of research performance based funding – statistically significant only when controlling for the other variables. This outcome is perhaps counter-intuitive. One should keep in mind that the effect of recently implemented RPBF systems may still be invisible in the indicators analysed.

- Apart from its positive correlation with citation impact, PhD intensity correlates negatively with graduation productivity as well (when controlling for the other 4 variables not significant at P=0.05). Figure 2 presents a scatterplot of these two measures. It is hypothesized that this is due to the fact that when a HEI is shifting its orientation towards research, its academic staff puts more efforts in the training of PhD students at the expense of the production of graduate students.

- Interestingly, PhD intensity correlates positively (but weakly) with publication productivity (R=0.35, p=0.09) but their partial correlation is negative (R=-0.25; p=0.34). Since these two correlations are not significant at p=0.05, they are not included in Table 7.

- Apart from the negative correlation with citation impact mentioned above, the degree of research performance based funding (RPBF) correlates positively with publication productivity and total academic staff. But these correlations are only significant if they control for the other four variables in the analysis. The first correlation is in agreement with one would expect to find as effect of RPBF, for the second the current authors do not have an explanation.

- It must be noted that the absolute number of students or graduates is a component in both indicators: it constitutes the denominator in the PhD intensity indicator, and a numerator in the graduation productivity measure. Hence, the indicators are statistically dependent, and a negative correlation between the two is not surprising. This dependence explains the hyperbolic ("f(x)=1/x"-like) left part of the curve in Figure 2.
- Despite the above limitations, and focusing on PhD Intensity, Figures 1 and 2 suggest that substantial differences exist in PhD policies among European countries. The relatively low PhD intensity for Italy and Spain compared to Northern European nations suggests that institutions in these two countries have –at least until recently – given a rather low priority to the foundation of a policy towards the training of PhD students.

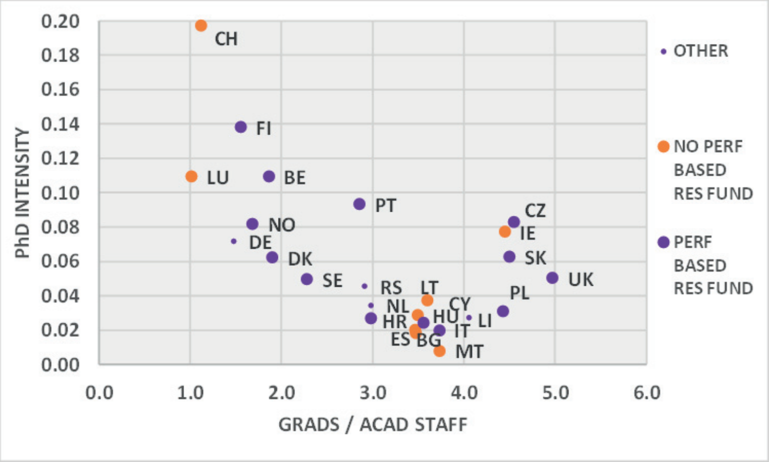


Figure 2. Scatterplot of PhD Intensity against Graduates per Academic Staff

Results from the Cluster and efficiency analyses

The heterogeneity of HEIs exists both across country and within country. Hence, it could be interesting attempt to categorize the HEIs institution regardless their national localization and considering, instead, a specific set of values representing characteristics and performance of each institution with respect to the dimensions of: teaching, research and third mission. The result of such type of analysis could be also used to assess the internal coherence of the national education systems. It could be very helpful to identify institutions to further investigate through case studies.

The variables used to compute the distances for the clusterization are: (i) average publications per academic staff (Pub_fract(av)/ACADstaff; normalized to allow a balanced comparison with the other variable) and (ii) average graduates per academic staff (Grads_ISCED.5-7/ACADstaff). In particular, the results obtained by the K-means (3 clusters) cluster analysis, after the DBSCAN analysis that suggested the existence of three clusters, identify three groups of universities whose main characteristics are outlined in Table 5. We labelled the three groups as: research and teaching oriented (TEAC&RES), research oriented (RES_OR) and teaching oriented (TEAC_OR).

See Figure 3 for an illustration that shows how well the three clusters are spread along the two clustering dimensions. Figure 4 reports the distribution of universities in the three clusters by country.

It appears (see Table 5) that the RES_OR cluster is characterized by the highest number of publications per academic staff (9.57), the highest PhD intensity and the highest proportion of publications in the highly cited journals (0.124), with an average mean normalized citation score (mnsc_(w-av)_av) above the world average (1.16).

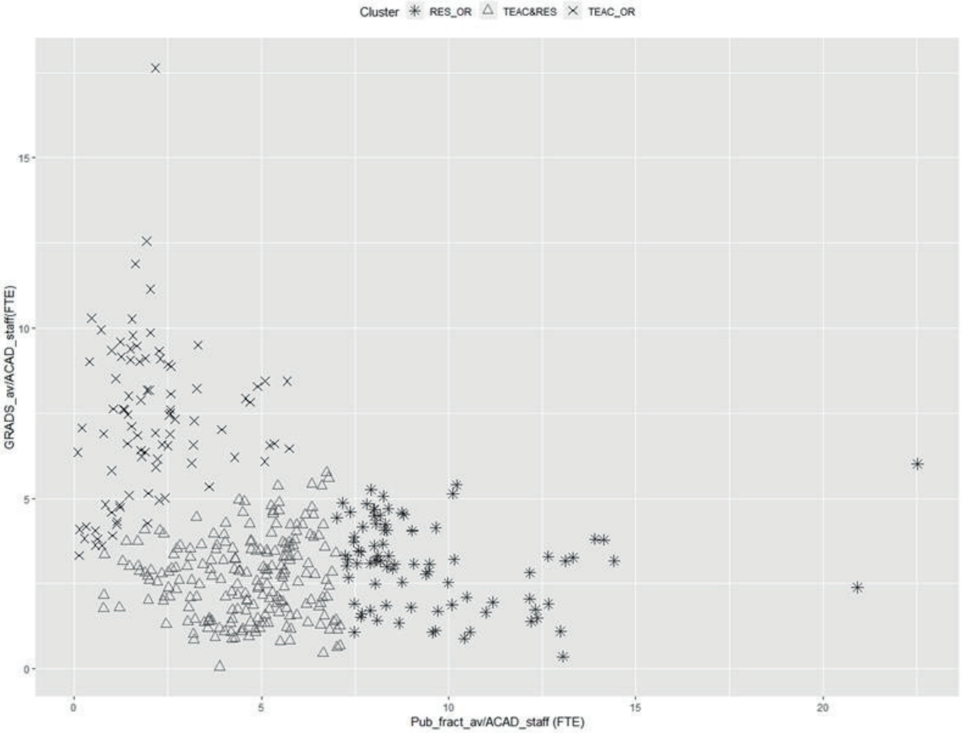


Figure 3. Publications per Acad_staff vs graduates per Acad_staff for the three clusters.

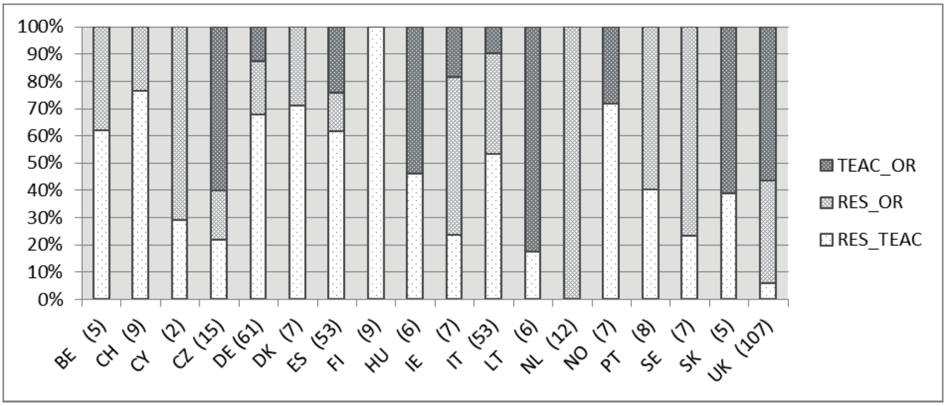


Figure 4. Heterogeneity within countries according to the identified clusters

*** Note: On the X-axe, the *number* in brackets refers to the number of HEIs analysed in each country. Notice that this number ranges from 107 for UK to 2 for Cyprus. Bulgaria, Hungary, Lithuania and Malta were not included because only one observation was available.**

Table 5. Descriptive statistics on the main variables for the obtained three clusters

		TEAC&RES	*RES_OR*	*TEAC_OR*
Cluster analysis variable	Grads_ISCED.5-7/ACADstaff	2.67	3.08	**7.26**
	Pub_fract(av)/ACADstaff	4.61	**9.57**	2.07
Efficiency analysis	Mod.Teach.Res.X_ACADSTAFF.FDH	2.43	**1.67**	**1.64**
Basic data	Foundation_year	1847.84	1785.39	1924.75
	Uni_Hospital	0.531	0.706	0.045
	Enrolled_student_ISCED.5-7	19368.25	21196.18	20143.51
	ACADstaff_FTE	1645.03	1931.45	731.41
	PhD_intensity_2014	0.0652	0.0933	0.0140
	FullProf/ACADstaff_Head	0.1166	0.1491	0.0998
	WomenProff_share	0.1921	0.1943	**0.2760**
	Admn/TOTstaff_FTE	0.4415	0.4797	0.5068
Third mission - Funds	Funds_external%	0.1809	**0.2723**	0.0971
	Funds_third_part/ACADstaff_FTE	30113.47	**60818.98**	23251.66
Specialization	Specialization	0.269	0.261	0.244
Research quantity and quality	Pub_top10(av)/ACADstaff	0.0270	0.0705	0.0105
	Pub_in_top10%	0.0949	**0.1240**	0.0700
	Pub_international_coll	0.5147	**0.5731**	0.4904
	mnsc_(w-av)_av	0.9894	**1.1612**	0.8673
Third mission - Patents	Patent_application(av)/ACADstaff	0.0022	0.0030	0.0008
	Back_citations(av)/ACADstaff	0.0094	**0.0133**	0.0034
	NPL_av	26.76	**43.87**	1.80
	NPL_av/SPA_av	5.63	6.32	2.08
National variables	GOV_Regulation	**8.00**	**6.08**	**4.26**
	GOV_Expenditure	8.81	8.12	8.79
	GOV_Taxes	11.03	11.70	**15.65**
	GOV_Information	**11.62**	9.22	9.85
	GOV_Constraints_trend	0.46	0.49	0.54
	GOV_Opportunities_trend	0.54	0.51	0.46
	GOV_Control_measures	0.33	0.29	0.29
	EU_fract_country	0.0846	0.0917	0.1035
	NAT_HEIs_fract	0.0350	0.0242	0.0187
	NAT_UNI_fract (*number*)	0.5139	**0.5878**	**0.7066**
	NAT_UNI_fract	0.7884	0.8122	0.9198

Interestingly, the RES_OR cluster shows also the highest percentage of funds from third parties (an average of 60,819 euro per academic staff) and the highest intensity of patents per academic staff and patents backward citations, pointing out to the existence of a "Matthew cumulative effect" in place. This means that high quality research is able to attract external funds that are connected to innovative and patenting activities that in turn are self-reinforcing to the scientific activities. On the other hand, we observe that the TEAC_OR cluster is characterized by the production of the highest number of graduates per academic staff (7.26) and presents the highest share of women (0.28) confirming a kind of segregation of women in teaching oriented universities. The TEAC_OR cluster is made, by and large, by institutions belonging to countries with less regulation policies (GOV_regulation is 4.26 against 6.08 of the RES_OR cluster and 8 of the TEAC&RES cluster) and highest policy interventions on Taxes (GOV_Taxes =15.65, against 11 for the other two clusters). Finally, the TEAC_OR cluster is composed mostly by institutions coming from the biggest countries in Europe (EU_fract_country 0.10) and with the highest proportion of universities on the overall number of HEIs (NAT_UNI_fract (number) =0.71, higher than that of the other two clusters).

The TEAC&RES cluster shows instead intermediary values among the two previously described groups.

Finally, it is interesting to note that the average FDH inefficiency score of the group TEAC&RES (2.43) is higher (*i.e.*, they are less efficient) of the inefficiency scores of the RES_OR and of the TEAC_OR groups (around 1.6). We remind to the reader that an inefficiency score equal to 1 means that the institution is fully efficient, so it is producing its outputs (teaching-graduates and research-publications) being on the efficient frontier of its possibilities. On the other hand, an inefficiency score higher than 1 points out to the possibility of improving the production of its outputs given the available resources (or inputs). This result seems to show that the specialization in teaching and in research pays also in terms of efficiency of the overall activities carried out, that is specialized universities, in teaching or in research, tend to have a higher efficiency than those universities that balance research and teaching activities.

Discussion and conclusions

From the analyses carried out in the present work, a rather heterogeneous picture emerges, that does not allow for 'simple' interpretations and conclusions. The statistical findings seem to be broadly consistent with the following observations.

The outcomes most of all reflect the heterogeneity of the European higher education and research system. Large differences exist between countries. The countries are in different phases of their scientific (and economic) development. During the past decade, in several countries, major changes took place in the funding structure and management of HEI, the effects of which are not yet visible in the analyses presented above. A longer term perspective is certainly needed. Therefore, correlations or concordances between quantitative measures on the one hand and more qualitative indicators (such as governance indicators or degree of research performance based funding) on the other hand are difficult to interpret, as they may relate to different time periods.

The results reveal once more the limits and dangers of *one-dimensional approaches* to the performance of HEIs. Analyses dealing merely with one singe dimension, *e.g.*, either research performance or teaching performance, may easily result in unbalanced or even invalid conclusions. As an example, for the teaching-oriented universities, a key part of their performance remains invisible in a purely bibliometric approach. This is perhaps common knowledge. But universities in the process of expanding their research funding and activities may easily show a declining graduation productivity (graduates per academic staff) if an increase in the size of their academic staff is deployed in research, while research output will increase with a delay of several years.

Apart from funding formula, another important aspect of a national HE system is the degree and the modus of quality assessment of research and education. For instance, in the Netherlands, assessment exercises by research discipline (e.g, Physics, Chemistry, Biology) have been conducted every 4–5 years for at least 25 years. Even though the outcomes do not play a formal role in the allocation of government funding of HEI, they do play a role in internal assessment and management processes within HEIs. The prominent position of The Netherlands in several analyses presented above may be at least partly a result of these long lasting and intensive assessment practices.

The combined efficiency analysis and cluster exercises showed the existence of three groups of European universities clearly characterized in their orientation towards teaching activities (TEAC_OR), research activities (RES_OR) or balancing among the two activities (TEAC&RES). Interestingly, the universities specialized in teaching or research show on average a higher efficiency in their main purpose then those oriented to the production of both teaching and research activities.

The obtained results may be useful to identify (select) the HEIs that need to be further investigated through case studies. In this way, our results may provide an evidence-based support to further investigate the heterogeneity of HE systems through qualitative case studies in political science studies of HE.

Acknowledgments

The financial support of the Italian Ministry of Education and Research (through the PRIN Project N. 2015RJARX7), of Sapienza University of Rome (through the Sapienza Awards no. PH11715C8239C105) and of the EU Horizon 2020 RISIS2 Project (grant agreement N° 824091) is gratefully acknowledged. We warmly thank Ed Noyons for providing us the CWTS bibliometric data analysed in this work.

References

Capano, G. & Pritoni, A. (2019). Varieties of hybrid systemic governance in European Higher Education, *Higher Education Quarterly*, vol. 73, n.1, pp. 10–28.

Daraio C. (2018). *Nonparametric Methods and Higher Education*, in: Teixeira P., Shin J. (eds) Encyclopedia of International Higher Education Systems and Institutions. Springer, Dordrecht.

Daraio C. (2019). *Econometric approaches to the measurement of research productivity*, in Springer Handbook of Science and Technology Indicators edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., forthcoming.

Daraio C. et al. (2011). The European University landscape: A micro characterization based on evidence from the Aquameth project, *Research Policy*, vol. 40, pp. 148–164.

Daraio C. & Simar L. (2007). *Advanced Robust and Nonparametric Methods in Efficiency Analysis, Methodology and Applications*, Springer, New York (USA).

Daraio, C. & Bonaccorsi, A. (2017). Beyond university rankings? Generating new indicators on universities by linking data in open platforms, *Journal of the Association for Information Science and Technology*, vol. 68, n.2, pp. 508–529.

Daraio, C., Heitor, M., Meoli, M. & Paleari, S. (2019). Policy turnaround: Towards a new deal for research and higher education. Governance, evaluation and rankings in the big data era, *Higher Education Quarterly*, vol. 73, n. 1, pp. 3–9.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, OR, AAAI Press, pp. 226–231.

López-Illescas, C., de Moya-Anegón, F. & Moed, H. F. (2011). A ranking of universities should account for differences in their disciplinary specialization, *Scientometrics*, vol. 88 n 2, pp. 563-574.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* vol. 1, n. 14, pp. 281–297.

Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2018). Performance-based research funding in EU Member States—a comparative assessment, in *Science and Public Policy*, pp. 1–11 (doi: 10.1093/scipol/scy041).

When Quantity Beats Quality in the Evaluation of Academic Work: An Eastern European Impact Factor?

Florin Nicolae Fesnic[1]

[1]*florin.fesnic@ubbcluj.ro*
"Babeş-Bolyai" University, Center for the Study of Democracy, Minerilor 85, Cluj 400409 (Romania)

Abstract

The use of the Impact Factor (IF) in research evaluation has become ubiquitous. I offer reasons to question the pertinence of such practices. Even if we accept the use of quantitative measures in research evaluation as a necessary evil, there are better alternatives to the Impact Factor. One example is the Article Influence Score (AIS), computed as a weighted average of the number of citations per article, unlike the Impact Factor, which is unweighted, and thus easier to manipulate. I show this through a comparison of the latest IF and AIS scores of journals in the field of Economics, where a regional dimension is immediately apparent: all journals edited in Eastern Europe have very high IF/AIS ratios relative to the median IF/AIS ratio (in statistical terms, all 17 Eastern European journals with a non-zero IF in 2017 are outliers). Thus, choosing one rather than the other can make a dramatic difference. I illustrate this by contrasting the results of evaluations in economics and political science in Romania, an Eastern European country where the current national standards impose the use of the AIS in the former and of the IF in the latter.

Introduction

The Impact Factor (IF) originally emerged as a tool for librarians, helping them in choosing among journals. Eventually, the use of the IF went beyond its primary use; now it is also used as a proxy for the value of individual articles, a practice that has now become ubiquitous in research evaluation. This approach is not uncontroversial, drawing criticism from those who consider that the IF is only appropriate for assessing journals, but not individual papers. Another line of criticism focus on various aspects of the IF, such as the two-year window used for its computation, or the inclusion of self-citations.

In this paper I do not address the first type of criticisms; quantitative evaluation of research is here to stay, at least in the foreseeable future – no matter how unfortunate is this state of affairs. I only focus on a specific aspect of research evaluation, namely, the claim that the Impact Factor is a measure of quality. It is not. Even if we accept the use of quantitative measures in research evaluation as a necessary evil, there are better alternatives to the IF. One example is the Article Influence Score (AIS), computed as a weighted average of the number of citations per article, unlike the Impact Factor, which is unweighted, and thus easier to manipulate. I show this through a comparison of the latest IF and AIS scores of journals in the field of Economics, where a regional dimension is immediately apparent: all journals edited in Eastern Europe have very high IF/AIS ratios relative to the median IF/AIS ratio (in statistical terms, all 17 Eastern European journals with a non-zero IF in 2017 are outliers). Thus, choosing one rather than the other can make a dramatic difference. I illustrate this last claim by contrasting the results of evaluations in economics and political science in Romania, an Eastern European country where the current national standards impose the use of the AIS in the former and of the IF in the latter.

It is true that many, if not most, scientometricians are very critical of the Impact Factor, particularly of its use as an alleged measure of quality. However, this message has yet to reach the broader community of scholars. Even more importantly, perhaps, it has yet to reach the decision-makers in research evaluation – more on this last point in the concluding section.

What is wrong with the Impact Factor?

In a nutshell, the argument that the IF in research evaluation of academic papers is better than its alternatives (particularly those attempting a trade-off between quantity and quality, such as the AIS) goes along these lines:
(i) we need a measure of quality for academic publications (Hoeffel, 1998; Hobbs, 2007);
(ii) the Impact Factor is the best among such measures (Garfield, 2006); this is "because it fits well with the opinion we have in each field of the best journals in our speciality" (Hoeffel, 1998, p. 1225);
(iii) the IF is highly correlated with measures with a qualitative component (such as the AIS); therefore, such measures are redundant (Davis, 2008).

Both (ii) and (iii) are problematic. While my focus in this paper is to offer reasons for questioning the purported position of the IF as the best measure of quality, it is worth addressing the last point as well, even if just briefly. In its original use, as a measure of (alleged) quality, if a librarian is selecting, say, the top 100 Economic journals by IF out of more than 300, the gains by using this approach as a shortcut may be worth the potential exclusion of a handful of journals which are much better (as indicated by their AIS, or another genuine measure of quality) than their IF would suggest and/or the potential inclusion of a handful of journals that are not as good (again, as indicated by their AIS) as their IF suggests. We can call the first type of journals 'good outliers' (or false negatives), while the second type are 'bad outliers' (false positives). Later on, I will demonstrate that the use of the Impact Factor does indeed create both kinds of outliers; indeed, even an almost perfect correlation between IF and AIS (say, 0.97) does not guarantee the absence of outliers (or, for that matter, even extreme outliers).[1]

While librarians assessing the value of hundreds and hundreds of journals may think of the existence of 'underrated' and 'overrated' journals as a mere inconvenience, it is no exaggeration to say that their existence can make or break academic careers. In this context, scholars competing for jobs and grants will typically have just a handful of papers whose value, unlike in the older days, is now judged by the (in)famous IF. Instead of rewarding the best work, the use of the Impact Factor is rewarding strategic behaviour. Inappropriate indicators create perverse incentives (Butler, 2003; Casadevall & Fang, 2014). As a result, "science has taken a turn toward darkness" (Bonell, 2016, p. 56). Before getting to the part where I discuss why this happens and how it happens, as well as the undesirable consequences it leads to, I will first present my data and methods.

Data and Methods

The data I use in the first part of the next section are the latest (2018) Impact Factor and Article Influence scores provided by Clarivate Analytics for the field of Economics. I chose the AIS, rather than one of the other "prestige measures" (Setti, 2013), to show why it is better than the IF as an indicator of quality, for two related reasons. The first is convenience – in its yearly release of scores, in addition to its IF, Clarivate Analytics also provides AIS for each journal, thus, in a sense, acknowledging the AIS as the best alternative to IF. Moreover, it also provides two additional versions of its trademark IF: a five-year version and a modified two-year version (the latter computed without self-citations). Each of these two alternative IF's incorporates one feature of the AIS. If Clarivate Analytics were to provide an IF incorporating both, as well as a weighting of citations based on how prestigious is the journal carrying that citation, such a version of the IF would be very similar to its alternative, the AIS. Thus, my second reason for my choosing the AIS is the fact that, in a way, Clarivate Analytics has already made the choice – it was my default choice.

[1] See Appendix for an illustration. The correlation between IF and AIS in the field of Chemistry Multidisciplinary was, in 2016, 0.97 (N = 162). This did not prevent the existence of extreme outliers.

One important reason for my choice of data for Economics is the large number of journals (n = 341 journals having an IF and AIS greater than zero). Nonetheless, having a large N is merely a bonus. The more important reason is that the field of Economics is very illustrative of the problems of the Impact Factor: it not only has a large number of outliers (many of these are extreme outliers), but the majority is concentrated in just one region, Eastern Europe. The post-Communist countries have 17 journals in this list, and *all* are outliers (an IF/AIS ratio significantly larger than the median IF/AIS ratio of Economics). This latter point is important; a widespread practice among Eastern European scholars (and this is definitely the case in Economics) is to publish primarily in journals edited in their home country. When they do not, they publish in other journals from the same region, leading to the creation of 'citation circles' (Teodorescu & Andrei, 2013), so that many, if not most, WoS-indexed Eastern European journals are "only locally international, but globally national" (Pajić & Jevremov, 2014, p. 276). Economics also provides one of the most remarkable (and worrisome) examples of a journal that, considering its Impact Factor, rose from obscurity to being one of the top journals in the field virtually overnight. In 2009, the Lithuanian *Technological and Economic Development of the Economy* was not present in ISI rankings. In 2010, its first year with a non-zero IF, it was the third journal in Economics (IF = 5.61). Using data from 2008-2009, which provide the basis for computing the 2010 Impact Factor, I show that the aforementioned's journal IF came exclusively from self-citations and citations in other Lithuanian journals. The situation of the other four Lithuanian Economic journals was very similar, their share of self-citations and citations in the other four journals ranging from over 96 to 100 percent.

To show how consequential this is, in the final part of the analysis I compare the dramatic differences that would be observed when publishing an article in *Technological and Economic Development of the Economy* (a 'bad' outlier, or 'overrated' journal) versus *Brookings Papers on Economic Activity* (a 'good' outlier, or underrated journal), depending on whether you are a Romanian academic in Political Science or in Economics. I will use the latest IF and AI scores of the two journals, as well as the Romanian formulas for the evaluation of publications in Political Science (which uses the IF) and Economics (which uses the AIS).

When Quantity Beats Quality in the Evaluation of Academic Work

What follows is, first, a comparison of the latest IF and AI scores in Economics as an illustration of when this happens – i.e., when quantity (the Impact Factor) beats quality (the Article Influence Score). Then I will present the astonishing rise of *Technological and Economic Development of the Economy* as one example of how this happens. I will end by comparing the widely divergent assessment of their work that can result for Romanian political scientists and economists for publishing in the exact same journals, based on whether their scores are a function of IF (in the first case) or AIS (in the second), to show how consequential is this choice.

The "Eastern European" Impact Factor

The central assertion of this paper is that, in spite of what its proponents claim, the Impact Factor is an inappropriate measure of quality. Let us compare the latest IF and AI scores for the 341 Economic journals having non-zero scores for both. Figure 1 presents two boxplots, comparing the distribution of IF/AIS ratios for journals edited in Eastern Europe (n = 17) versus the other countries (n = 324):

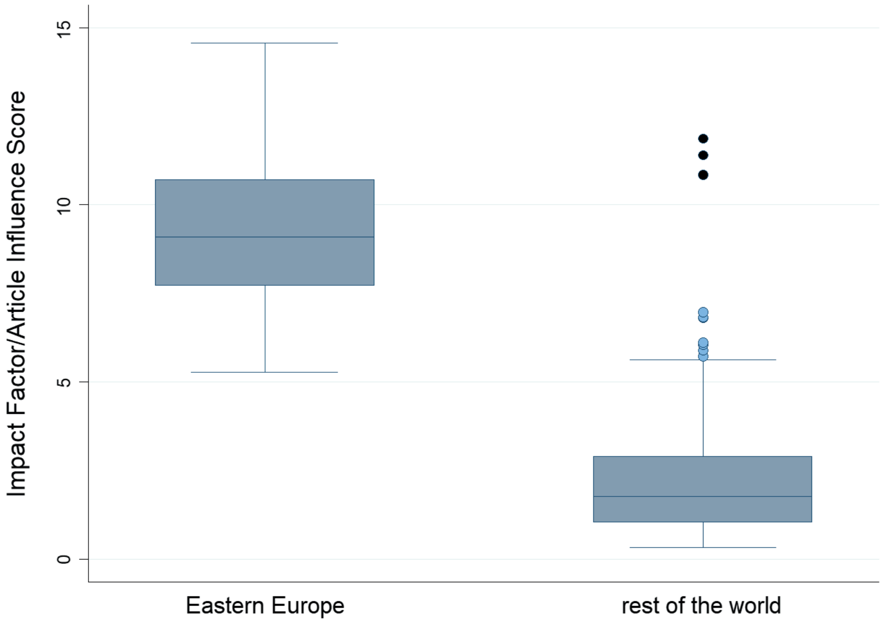


Figure 1. The distribution of IF/AIS ratios in Economics, 2018 WoS edition

The general lesson from Figure 1 is that, for a large number of journals in this field, the Impact Factor is meaningless – or, at the very least, is only appropriate as a measure of popularity, but definitely not as a measure of prestige (i.e., quality). This is especially the case in Eastern Europe, where all 17 Economic journals appearing in the latest Clarivate list are outliers. These results illustrate the dangers of using the IF for measuring the quality of academic publications. In the next section I show an example of how and why we can get in such a situation – i.e., having journals with a much higher IF than expected based on their AIS.

The amazing story of Technological and Economic Development of the Economy

In 2009, the Lithuanian journal *Technological and Economic Development of the Economy* did not have an Impact Factor. In 2010, its very first presence in the list, it was the third journal in Economics by IF (5.61!) Such a jump is, in and by itself, quite problematic. As Setti (2013, p. 233) points out, a good indicator "should not exhibit large fluctuations over a limited time period." An equally important, if not even more so, question, is how was this possible? Where did this impressive IF came from? Figure 2 offers the answer:

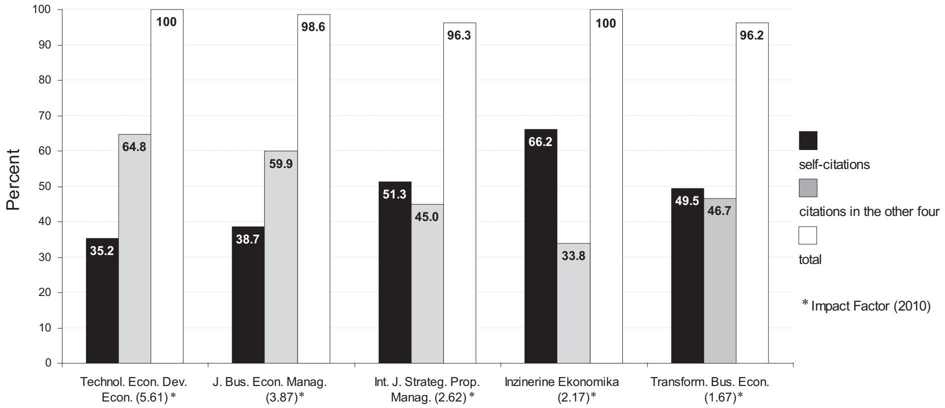


Figure 2. The "Lithuanian" Impact Factor: Share of self-citations and citations in the other four journals for five Economic and Management journals (2010)

In 2010 there was a total of five Lithuanian journals in Economics (four) and Management (one) with an IF greater than zero. Figure 2 shows the share of self-citations for each of the five journals, as well as their share of citations in the other four journals, from the total number of citations in journals with a non-zero IF in 2008 and 2009. The typical (Economic) journal with a high IF has it because it is highly cited by the worldwide community of scholars in the field. The *Technological and Economic Development of the Economy*, like the other four journals, have a high IF because it is highly cited exclusively by Lithuanian economists. This phenomenon is possible because, unlike measures such as the AIS, which are weighted averages – the more prestigious the journal, the higher the weight of being cited in that journal –, the IF is unweighted, i.e., all citations are counted equally. The AIS has what Setti (2013, p. 238) calls the property of "insensitivity to insignificant journals," a property that the IF does not have. The baseball team that wins the World Series is still, technically, only the best baseball team in the US. Nonetheless, given the quality of baseball in that country, calling them World Series champions is still much less of a misnomer than to consider *Technological and Economic Development of the Economy* among the best economic journals in the world, based solely on the assessment of Lithuanian scholars. The next section will offer an illustration of how consequential the use of the IF as a measure of quality can be.

Research Evaluation in Romania: The Impact Factor versus the Article Influence Score

Romania is just one of many countries where the use of quantitative measures in research evaluation is widespread. To give but two examples, the first (and, arguably, the most important) criterion for advancement in academia (say, from *conferenţiar*, which is roughly the equivalent of Associate Professor, to full professor), is to get a certain number of points based on publications. The specific criteria differ from one field to another. Economics is one example of a field where the formula allocation points for each publication uses the Article Influence Score. In Political Science, the formula is based on the Impact Factor. Now let us consider the example of two scholars, an economist and a political scientist. Let us suppose that each scholar is trying to decide whether to publish a paper in *Technological and Economic Development of the Economy* or *Brookings Papers of Economic Activity*. In Table 1 I present the results, assuming the paper is accepted, as a function of the field (Economics versus Political Science) and journal (*Technological and Economic Development* versus *Brookings Papers*).

Table 1. Research Evaluation in Romania: Political Science versus Economics (2018)

	Political Science	*Economics*	Row ratio
Technological and Economic Development	30.0	4.4	6.8
Brookings Papers of Economic Activity	28.5	82.2	0.35
Column ratio	1.05	0.05	

If you are a Romanian political scientist faced with a choice between these two journals, you should be indifferent, assuming (and this might be a bold assumption) that it is equally difficult to publish in either – the reward (score based on the IF of each journal) is virtually the same. However, assuming that you are an economist, the reward (based on the AIS) for publishing in *Brookings Papers* is almost twenty times larger than the score received for a paper published in *Technological and Economic Development*. Obviously, we would expect the scores (and the ratios of the scores) to differ in the two cases. Nonetheless, when not just the scores, but the ratios of the scores differ by an order of magnitude, I think it is reasonable to ask the question whether the two approaches are equally justifiable.[22]

Criticism of the Impact Factor: Much Ado About Something

The literature critical toward the IF may be large; the number of scientometricians who argue against the use of the IF in research evaluation may be substantial. Nonetheless, for the time being, this message has yet to reach a broader audience, whether we think of scholars of other fields or decision-makers. To illustrate this point, I present a selection of "good" outliers (i.e., "underrated" journals) from the latest (as of early June 2019) ISI rankings in the fields of Economics, Sociology, Political Science, Information Science & Library Science, and Chemistry Multidisciplinary (Table 2).

As we can see in Table 2, out of nine journals, two-thirds mention their IF on their home page, and only one of them (the *Quarterly Journal of Political Science*) mentions a measure that is not purely quantitative (its SCImago Journal Rank). For instance, *Political Analysis*, in addition to listing its latest IF, also mentions that it is "25[th] out of 169 [in] Political Science." This is a respectable position already, but if the editors were to use the journal's AIS instead, *Political Analysis* would have been second (or 5th in the SCImago Journal Ranking). Thus, clearly, for the time being, very few "underrated" journals make use of what I would call "bragging rights" – i.e., claiming (for good reason) that they are significantly better than what their IF suggests.

[2] The latest (2018) Impact Factors of the two journals were 3.244 (Technological and Economic Development) and 3.067 (Brookings Papers of Economic Activity), placing them in the 92nd and the 90th percentile, respectively. Their Article Influence Scores were 0.442 (Technological and Economic Development – 39th percentile) and 8.217 (Brookings Papers of Economic Activity – 98th percentile).

Table 2. Do “good outlier” journals make use of their “bragging rights”?

	IF 2017	IF Percentile	AIS 2017	AIS Percentile	Impact Factor[i]	AIS/other[ii]
Economics						
*Econometrica***	3.75	93	11.45	99	0	0
*Brookings Papers on Economic Activity**	3.07	89	8.22	98	0	0
*Quantitative Economics**	1.42	62	3.78	92	**1**	0
Political Science						
*American Political Science Review**	3.25	94	5.70	100	**1**	0
*Political Analysis***	2.59	85	5.36	99	**1**	0
*Quarterly Journal of Political Science***	2.13	75	4.08	97	0	**1**
Sociology						
*Theory and Society***	1.03	44	1.53	90	**1**	0
Information Science & Library Science						
*Information Systems Research***	2.30	72	2.24	98	**1**	0
Chemistry Multidisciplinary						
*Wires Comput Mol Sci***	8.84	88	5.67	96	**1**	0

*The journal is an outlier

**The journal is an extreme outlier

[i] The homepage of the journal mentions its Impact Factor

[ii] The homepage of the journal mentions a measure with a qualitative component (e.g., AIS or Scopus)

Conclusion

Let me repeat Hoeffel's (1998, p. 1225) contention, that "[t]he use of the impact factor as a measure of quality is widespread because it fits well with the opinion we have in each field of the best journals in our speciality." I wonder, if we were to survey all scholars from, say, the top 50 or top 100 Economics departments in the world, asking them to mention the journals they consider to be the best in the field, how many would mention *Technological and Economic Development*? Or, for that matter, how many of them have ever heard of this journal, allegedly (based on its IF), one of the very best in the field? In the light of the analysis presented here I would argue that, unless we want to reward quantity at the expense of quality, we should avoid purely quantitative measures such as the Impact Factor. 'Prestige measures' such as the Article Influence Score, though by no means perfect, do nonetheless a much better job to reward genuinely good research. The use of the Impact Factor in research evaluation is punishing those who focus primarily on research, while at the same time rewarding those who may or may not be first-rate scholars, but are very good at playing the Impact Factor game. In research evaluation, the rules should be designed to make good scholars even better, rather than encouraging them to be increasingly strategic.

References

Bonnell, A.G. (2016). Tide or tsunami? The impact of metrics on scholarly research. *Australian Universities Review*, 58, 54-61.

Butler, L. (2003). Modifying publication practices in response to funding formulas. *Research Evaluation*, 12, 39-46. DOI: 10.3152/147154403781776780.

Casadevall, A. & Fang, F.C. (2014). Causes for the Persistence of Impact Factor Mania. *MBIO*, 5, 1-5. DOI: 10.1128/mBio.00064-14.

Davis, P.M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? *Journal of the American Society for Information Science and Technology*, 59, 2186–2188. DOI: 10.1002/asi.20943.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA – Journal of the American Medical Association*, 295, 90–93. DOI: 10.1001/jama.295.1.90.

Hobbs, R. (2007). Should we ditch impact factors? No. *British Medical Journal*, 334, 569-569. DOI: 10.1136/bmj.39146.545752.BE.

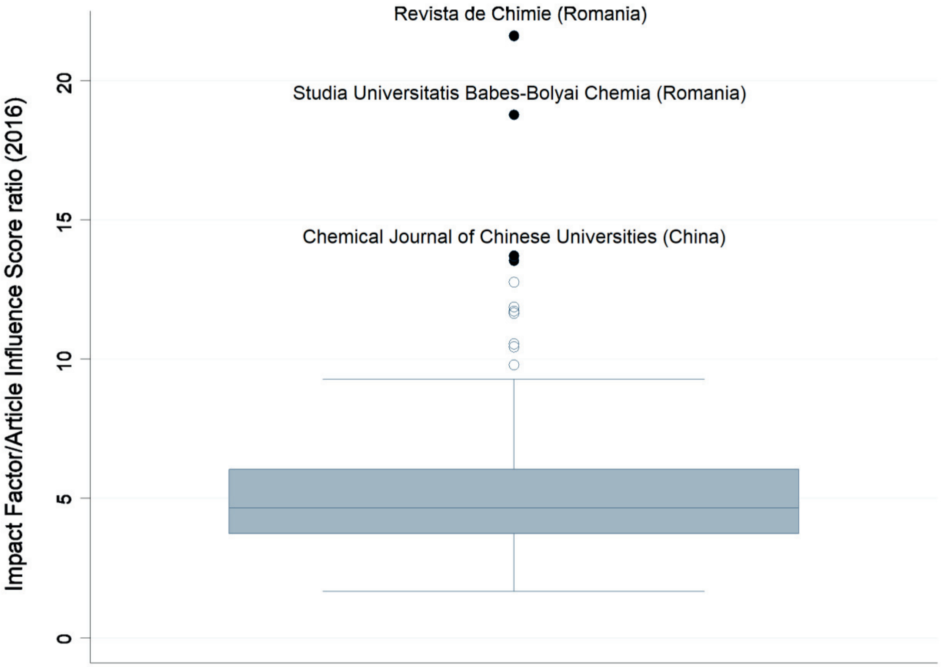
Hoeffel, C. (1998). Letter to the editor: Journal impact factors. *Allergy* 53, 1225. DOI: 10.1111/j.1398-9995.1998.tb03848.x.

Pajić, D. & Jevremov, T. (2014). Globally national - Locally international: Bibliometric analysis of a SEE psychology journal. *Psihologija*, 47, 263–277. DOI: 10.2298/PSI1402263P.

Setti, G. (2013). Bibliometric Indicators: Why Do We Need More Than One? *IEEE Access*, 1, 232–246. DOI: 0.1109/ACCESS.2013.2261115.

Teodorescu, D. & Andrei, T. (2013). An Examination of 'Citation Circles' for Social Science Journals in Eastern European Countries. *Scientometrics*, 99, 209–231. DOI: 10.1007/s11192-013-1210-6.

Appendix



The distribution of IF/AIS ratios in Chemistry Multidisciplinary (2016)

Scholarly communication or public communication of science? Assessing who engage with climate change research on Twitter

Rémi Toupin[1], Florence Millerand[2] and Vincent Larivière[3]

[1] *toupin.remi@uqam.ca*
Laboratory for communication and the digital (LabCMO), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, H3C 3P8 Montréal, Québec (Canada)

[2] *millerand.florence@uqam.ca*
Département de communication sociale et publique and Laboratory for communication and the digital (LabCMO), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal, QC, H3C 3P8, Canada

[3] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal C.P. 6128, Succ. Centre-Ville, Montréal, QC, H3C 3J7 Canada and Observatoire des sciences et des technologies, Centre interuniversitaire de recherche sur la science et la technologie (CIRST) Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal, QC, H3C 3P8, Canada

Abstract

The aim of this research is to build a methodological framework for the identification of users engaging with scholarly productions on Twitter by focusing on their Twitter bios. Based on a corpus of 4 719 research papers, 41 019 tweets and 21 965 unique users engaging with climate change research from 2015 and 2016, we are developing a codebook, by manual and semi-automatic coding of these bios, for the identification for seven types of accounts - 1) Faculty members and students; 2) Institutions and organizations; 3) Bots and automated accounts; 4) Journals and publishers; 5) Communicators; 6) Professionals; 7) Personal. As this work focus on public engagement with science, our focus is on the identification of lay users, defined as those using only Personal expressions in their bios. Preliminary results based on the first iteration of the codebook lead the categorization of 12 415 accounts, 5 949 of them including Personal expressions. However, results also indicate a significant overlap with other categories, especially Faculty members and students (n = 1 782). Future work will focus on refining the codebook for further analysis and manual coding to more accurately measure the precision of these results.

Background information

Twitter has long been considered a suitable platform for the diffusion of research to a broader public as it is used by a wide variety of users, most of them outside the scientific community. As such, it is a significant area of research for the development of indicators that could measure the societal impact of research, altmetrics. However, recent work shows that users engaging with research on Twitter are mostly scholars themselves and that the engagement of users outside of academia is significantly low (Alperin et al., 2019; Bowman, 2015; Côté & Darling, 2018). The direct identification of accounts engaging with research on Twitter remains a key issue as personal information is mostly limited to Twitter bios of 160 characters. It is particularly difficult in regard to so called lay users, as potential expressions allowing for their identification are unspecific at best. So far, studies focused on the identification of accounts maintained by institutions, individuals or automated profiles (bots) by applying codebooks on the information provided in the Twitter bio (Haustein et al., 2016b; Holmberg et al., 2014; Tsou et al., 2015). Recent work also indicate that users may use keywords relating to several categories, making it necessary to develop a methodology through which we are able to assess this overlap (Haustein, 2018). Thus, specific research, focusing on the identification of faculty members, students, communicators, decision makers and lay users, is needed to better understand the engagement with research on Twitter.

Recent technological possibilities, structural incentives – such as access to research funds – and discussions on the societal impact of research have called for a "better" engagement with scientific knowledge outside of academia. This is especially significant in regard to areas of research that build on the impact of research on the public, such as environmental sciences, social sciences, or health or medical sciences (Haustein, 2018). Our study focuses on climate change research where there has been an increase in scientific activity and a significant interest outside of academia in recent years (Haunschild et al., 2016; Haunschild et al., 2019). As such, it provides an ideal context to better understand the public engagement with research and the societal impact of research, focusing on who tweets about climate change research in this case.

Purpose of the study

This work focuses on climate change research to identify who is engaging with research on Twitter by looking at their Twitter bios. It aims to 1) contribute to the discussion about *who* tweets about scientific research and 2) provide a methodological approach for the classification of accounts sharing scholarly productions on Twitter. More specifically, our focus is on users outside of academia – lay users, communicators, decision-makers - to understand the practices and context in which there is a broader engagement with research, engagement that would eventually inform policies. As anthropogenic climate change is currently regarded as a major sociopolitical issue that involves a variety of stakeholders, we assume a higher engagement by users outside of academia than what have been assessed in other disciplines (IPCC, 2014). Our first results indicate that this is the case as, on average, papers are tweeted more and by a higher number of users than for most other fields of study, as shown in Figure 1.

Materials and methods

To investigate the engagement with climate change research on Twitter, we built a dataset of 2015 and 2016 research articles with DOI (n = 4 719) indexed in the Web of Science (WoS) that included the keywords "climate change", "global warming" or "IPCC" in the title. We focused on the title as it is a direct metadata through which we may assess a paper relevance to a particular topic (Thelwall et al. 2013). It also frequently appears in tweets sharing a link to the paper, and so is highly visible to all users. As our focus was on precision rather than coverage, the aim of this query was to retrieve a significant number of papers directly related to climate change, though not all papers. The publication years were chosen as they cover the period before and after the Paris Agreement, a crucial moment for the public understanding of climate change issues (Hopke et Hestres, 2018).

Tweets were collected for all 4 719 articles by cross-referencing the information gathered from WoS with that from the Altmetric database via the Digital Object Identifier (DOI). Altmetric information was gathered through a data dump by the Observatoire des sciences et des technologies. Overall, we collected information for 41 019 tweets and 23 791 retweets sent by and 21 965 unique users linking to 2 620 papers. We then collected metadata about tweeted papers, tweets and user data - Twitter handle, user name, URL and Twitter bio, country information and number of followers were collected for the latter - from Altmetric. Scholarly Twitter metrics - number of papers tweeted, number of tweets, Twitter coverage, Twitter density (i.e., number of tweets per paper) and intensity (i.e., number of tweets per tweeted paper), number of users, user density (i.e., number of users per document) and intensity (i.e., number of users per tweeted document), number of papers retweeted, retweet coverage, share of retweets, retweet density (i.e., number of retweets per paper), retweet intensity (i.e number of retweets per tweeted document) as well as the timespan between first and last tweet, date of first and last tweet - were computed to further describe our dataset following work by Haustein (2018), indicating a significant level of engagement toward climate change research.

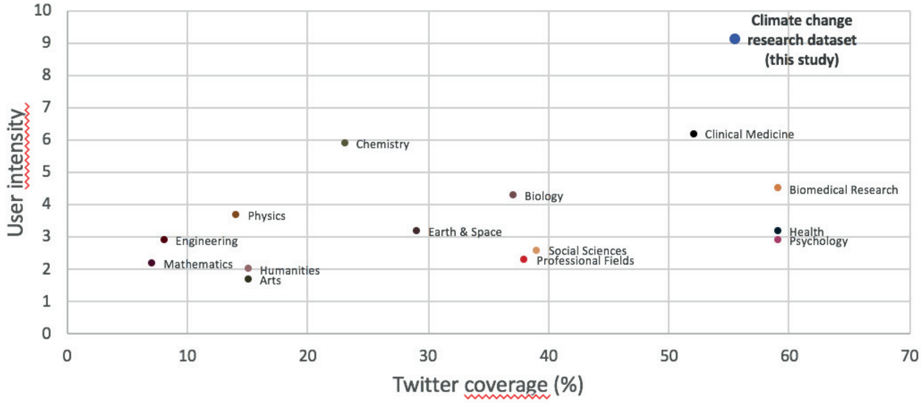


Figure 1: Computed Twitter metrics indicate a significant level of engagement toward climate change research in comparison to other disciplines. Twitter coverage (55% for our dataset) is the share of papers that were tweeted at least once, were as user intensity (9 for our dataset) is the number of unique users who shared tweeted papers at least once. Other disciplines data retrieved from Haustein (2018).

Around 56% of all 4 719 papers form our dataset were shared on Twitter (Figure 1), which exceeds the Twitter coverage of all disciplines (36%; Haustein, 2018) as well as Biology (37%), Earth and Space Science (29%) and Social Sciences (39%), but is comparable for the percentage found in Health (59%), Biomedical Research (59%), Psychology (59%) and Clinical Medicine (52%) (Haustein 2018). This indicate that engagement about climate change research matches that for health and medical sciences, though more comprehensive studies focusing on coverage may be needed. Tweeted papers in our dataset have a higher activity level than for all Medical sciences, as is shown by a higher Twitter intensity (15.7 tweets per tweeted papers for our dataset vs 8.5 tweets per tweeted papers on average for Medical sciences) and user intensity (8.4 users per tweeted papers for our dataset vs 4.5 users per papers on average for Medical sciences) (Haustein, 2018). All other computed metrics are also higher in our dataset, which indicate a significant level activity around climate change research on Twitter. This supports our hypothesis that climate change research is particularly relevant and receives larger attention on Twitter than other fields of research.

To identify and categorize Twitter users, we are building and applying a methodological framework based on expressions retrieval in Twitter bios through an iterative process (Haustein, 2018; Haustein et al., 2016a). Specifically, we are developing a codebook sorting keywords for different types of users, through which we will run scripts, build on the R language, to classify accounts according to how they identify themselves in their bios, going back and forth between the codebook and the results to refine the expressions related to each category (Côté and Darling, 2018; Toupin and Haustein, 2018). Through this process, we will also look at the overlap between categories, mostly in regard to Personal keywords (dad, mom, cat, sports, for example), which we use as a proxy to identify lay users. In this regard, lay users will be those who identify themselves using only Personal keywords. Though a significant level of overlap is expected, it will also allow for the identification of specificities in expressions between categories, further refining the codebook. Iterations will be run until we reach a sufficient level of both coverage (proportion of accounts that are identified) and precision (proportion of accounts falling in the correct classifications).

To assess the precision of our queries and help develop further iterations of the codebook, all Twitter bios from our dataset are manually coded by two researchers (including the main author). Through this process, we are building a manual classification to which we will compare the results of our queries. It also provides us with further information for our codebook, specifically expressions helping us build more refined queries to get a better coverage with the semi-automatic coding. Overall, this should serve as the basis of our methodological framework to further identify who is engaging with scholarly production on Twitter. For the purpose of the study, we excluded all Twitter bios that did not include any information (NULL; n = 2 055) or were written exclusively in other language than English (n = 2 047), which reduce our dataset to 17 837 accounts, though future work will focus on building codebooks for other languages.

Results

The first iteration of our codebook allowed for the classification of 12 415 accounts (69.6%) in seven categories: 1) Faculty members and students (5 651 accounts); 2) Institutions and organizations (3 475 accounts); 3) Bots and automated accounts (335 accounts); 4) Journals and publishers (329 accounts); 5) Communicators and journalists (1 936 accounts); 6) Professionals (473 accounts); 7) Personal keywords (5 949 accounts) (Figure 2). The focus of this work is to provide a way to identify lay users engaging with research in general, and climate change research in particular. So far, we have identified 5 949 (33.35%) accounts that included Personal keywords. However, matching those keywords with other categories indicate a significant overlap. The most significant overlaps are with the Faculty and Students (1 782 accounts), Institutions and Organizations (832 accounts), and Communicators and Journalists (825 accounts) categories. Overlap with the Faculty and Students, and Communicators and Journalists categories are expected as they both allow for the identification of accounts belonging to individuals and indicates that these users identify themselves in more than one way. However, results of the overlap with Institutions and Organizations may indicate that the codebook for these categories is not accurate at this stage. As this is the first iteration of this codebook, future work is necessary to precise the keywords and expressions relating to all categories. This work is currently ongoing with the manual coding of the Twitter bios.

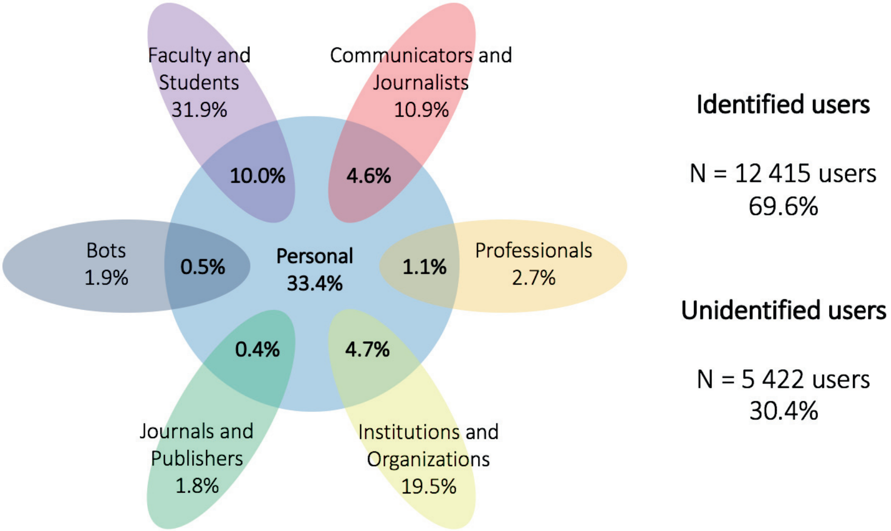


Figure 2: The share of accounts identified through the first iteration of the codebook, with a focus on personal accounts. Percentage of overlap between "Personal" and the other categories is based on the total number of accounts (N = 17 837).

Future work

We are currently working on the manual coding of all Twitter bios in our dataset to provide a basis through which we may assess the precision of the codebook. We will focus on adding new expressions to get a better coverage of our dataset, both by looking at our preliminary results and by adding new keywords through manual coding. We will also look more closely to the overlap between all categories and whether this is the result of expressions that are not specific enough or simply users matching to more than one category, faculty and personal for example. This will help us in developing a methodological framework to identify several types of users engaging with research on Twitter, and social media in general, as well as provide an insight on the participation of lay users to communication of research, specifically climate change research in this study. We also wish to build codebooks for other languages to have a better understanding of who is engaging with climate change research outside of the english-speaking community. Finally, we will improve our categorization framework toward the identification of potential decision makers on Twitter.

Acknowledgments

This research is funded by a SSHRC Joseph-Armand Bombardier Canada Doctoral Scholarship (Grant 767-2017-1329, 2017-2020). This research is produced with the help of the Canada Research Chair on the Transformations of Scholarly Communication, the ScholCommLab, and the LabCMO - Laboratory for communication and the digital.

References

Alperin, J. P., Gomez, C. J., & Haustein, S. (2019). Identifying diffusion patterns of research articles on Twitter: A case study of online engagement with open access articles. Public Understanding of Science, 28(1), 2-18. https://doi.org/10.1177/0963662518761733

Bowman, T. D. (2015). Investigating the use of affordances and framing techniques by scholars to manage personal and professional impressions on Twitter (PhD Dissertation, Indiana University). Retrieved from http://www.tdbowman.com/pdf/2015_07_TDBowman_Dissertation.pdf

Côté, I. M., & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops?. FACETS, 3, 682-694. https://doi.org/10.1139/facets-2018-0002

Haunschild, R., Bornmann, L., & Marx, W. (2016). Climate Change Research in View of Bibliometrics. PLOS ONE, 11(7), e0160393. https://doi.org/10.1371/journal.pone.0160393

Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author keywords and hashtags. Journal of Informetrics, 13(2), 695–707. https://doi.org/10.1016/j.joi.2019.03.008

Haustein, S. (2018). Scholarly Twitter Metrics. In W. Glänzel, H.F. Moed, U. Schmoch, & M. Thelwall (Eds.), Handbook of Quantitative Science and Technology Research, Springer. Retrieved from https://arxiv.org/abs/1806.02201v2

Haustein, S., Bowman, T. D., & Costas, R. (2016a). Interpreting "altmetrics": viewing acts on social media through the lens of citation and social theories. In C. R. Sugimoto (Ed.), Theories of Informetrics and Scholarly Communication (pp. 372-405). Berlin: De Gruyter Mouton.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016b). Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. Journal of the Association for Information Science and Technology, 67(1), 232–238. https://doi.org/10.1002/asi.23456

Holmberg, K., Bowman, T. D., Haustein, S., & Peters, I. (2014). Astrophysicists' conversational connections on twitter. PloS One, 9(8), e106086–e106086. https://doi.org/10.1371/journal.pone.0106086

Hopke, J. E., & Hestres, L.E. (2018). Visualizing the Paris Climate Talks on Twitter: Media and Climate Stakeholder Visual Social Media During COP21. Social Media + Society, 4(3), https://doi.org/10.1177/2056305118782687

IPCC. (2014). Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Intergovernmental Panel on Climate Change. Geneva. 151 pages. Retrieved from https://www.ipcc.ch/report/ar5/syr/

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting Links to Academic Articles. Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics, 17(1), 1–8.

Toupin, R. & Haustein, S. (2018). 2nd Codebook for users categorization – Climate change research. Figshare. https://doi.org/10.6084/m9.figshare.8236598

Tsou, A., Bowman, T. D., Ghazinejad, A., & Sugimoto, C. R. (2015). Who tweets about science? In Proceedings of the 2015 International Society for Scientometrics and Informetrics (pp. 95–100). Istanbul, Turkey.

Variations in citation practices across the scientific landscape: Analysis based on a large full-text corpus

Wout S. Lamers, Nees Jan van Eck and Ludo Waltman

{w.s.lamers, ecknjpvan, waltmanlr}@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

Abstract

Publication full text is a potent new source of data for scientometric research and offers a unique look into how practices of citation differ across scientific disciplines. We present an analysis of citation label styles, the use of cited author surname as an integral part of the narrative of citing sentences, and the use of reporting verbs, across a corpus of 58 million citing sentences. Results are both aggregated over five broad scientific disciplines and visualised across a map of science for greater detail.

The occurrence of author-year style citation labels overlaps with integral citations in social sciences and humanities, but mathematics and computer science uses integral citation without author-year labels, and life and earth sciences shows the opposite pattern. Reporting verbs are slightly more common in social sciences and humanities, especially cognition verbs, which also feature often in mathematics and computer science.

Individual verbs show great differences in their usage across science. Several seemingly similar verbs favour almost entirely distinct areas of research. Overall, we observe remarkably gradual changes in these phenomena across the scientific landscape, suggesting that disciplinary conventions in writing styles overlap and combine in areas between prominent disciplines.

Introduction

Publication full text has long held promise as a potent source of data for scientometric research, containing a wealth of information on the position, frequency and context of citations far beyond what is available in traditional metadata-based citation indices. Yet, collecting full text data, let alone processing it, has long been such a challenging and time-consuming task that research relying on full text data was limited to often manual approaches and smaller data sets covering at most several hundreds of publications. Any analysis at the level of science as a whole was simply not feasible. Instead, research into differences in academic discourse was typically constrained to a limited set of disciplines (e.g. Hyland, 2006).

The recent increase in the availability of machine readable publications in electronic formats is slowly removing these limitations. Boyack et al. (2018) have previously compiled and described large datasets of in-text citations and their immediate context and distributions. In this paper, we build on the data collected during this previous research to present a large scale analysis of variations in citation practices across the scientific landscape. Our analysis is based on 58 million citing sentences from 4 million publications in the Elsevier corpus citing close to 7 million Web of Science (WoS) publications. Specifically, we analyse three features of these citing sentences: the style of the citation label, the use of the cited author's name in the narrative (integral citation), and the use of reporting verbs in the sentence.

Integral citations are citations that explicitly use the cited author's name in the narrative of the citing sentence (Swales, 1990). Hyland (1999) studied this feature in a limited sample and argued that integral citations signify increased emphasis on or increased engagement with the cited work. He also observed that disciplinary differences exist in the frequency of the use of integral citations. We hypothesise that an writer's propensity to use integral citations relates to the occurrence of author-year style citation labels in the author's research area, as a citation can be made integral by essentially lifting a cited author's name out of the parentheses into the main body of the sentence. We will put this hypothesis to the test by comparing these two phenomena across the scientific landscape.

Reporting verbs help communicate a writer's interpretation of a cited author's work and rhetorically frame the cited work for the reader by signalling the writer's stance (Bloch, 2010; Hyland, 1999). Thompson and Ye (1991) provided a framework for the assessment of reporting verbs, distinguishing categories of denotation (research verbs, cognition verbs and discourse verbs) and evaluation (factive, non-factive and counterfactive). The denotation categories are of particular interest to this research, as these categories might reveal areas of science that favour a particular denotation category and a specific type of knowledge production.

Beyond reporting verb categories, we will also investigate the occurrence of individual reporting verbs. Verbs in general are interesting proxies for activities associated with cited research, as demonstrated by their varying occurrence along the IMRaD structure (Bertin & Atanassova, 2014) and potential uses in identifying the function of citations (Bertin, Atanassova, Sugimoto, & Lariviere, 2016). Other research employed verbs to characterise the use of a research object (Li, Rollins, & Yan, 2018). Verbs have also been shown to consistently rank as the most differentiating socio-epistemic features when comparing language use between disciplines (Demarest & Sugimoto, 2015).

In this paper, we describe each of the above features as they occur across the scientific landscape. We rely strongly on visual presentations using VOSviewer (van Eck & Waltman, 2010). We use a base map of the scientific landscape based on WoS publications grouped into 868 scientific fields using the methodology of Waltman and van Eck (2012). Data on citation label types, integral citation rates, and verb usage can be overlaid onto this map. We will discuss our method for retrieving and processing this data from the Elsevier corpus in the next section, and subsequently present our results and conclusions.

Method

We started our data collection by selecting all research articles and review articles from WoS published in the period 2002-2011. This yielded a total of 10,120,981 seed publications. We then used the Elsevier corpus (see Boyack et al., 2018 for earlier anaylsis of this corpus) to retrieve full text records citing these seed publications, resulting in 4,056,996 Elsevier full-text publications with a total of 89,664,777 in-text citations to 6,961,899 of the WoS publications, in 58,160,057 unique sentences.

For each in-text citation, we established if the label was in author-year form. This was done by scanning the label for a sequence of four numbers (representing years) and for strings such as *accepted*, *forthcoming*, *in preparation*, *in press*, *in print*, *in process*, *in review*, *in revision*, *submitted*, *this issue*, *under review* and *unpublished.* An extensive manual inspection of our method of classifying citation labels revealed misclassifications only in rare cases where the original author had made an error in the citation labelling or some transcription error had occurred.

Whether a citation is integral was established by searching the accompanying sentence for the surname of the first author of the cited publication. Only parts of the sentence outside brackets were considered. For the search itself, a regular expression was constructed matching only the surname when it is preceded by the start of the sentence or a non-alphabetic character, and followed by a sentence end or a non-alphabetic character. Both the original rendition of the surname in the cited publication and a variant which forced the capitalisation of the first character were considered (to account for alternative renditions of surnames preceded by a non-capitalised string such as *van Eck*, which are often cited starting with a capital letter).

Extraction of verbs

To extract verbs from sentences, we performed word tokenisation and part-of-speech tagging using the Python NLTK library. In order to ensure a better result, citation labels were replaced with neutral tokens designed to be interpreted as proper nouns. Subsequently all tokens tagged as verbs (VB*) were extracted, except those tagged as gerunds or present participles (VBG) or past participles (VBN). Each extracted verb was then lemmatised using NLTK's WordNet lemmatiser, simplifying them to their dictionary forms. Finally, tokens tagged as verbs that were not recognised by the lemmatiser were discarded.

Final counts of verbs for each area of the scientific landscape were established by combining all the verbs found in the sentences containing in-text citations to the relevant publications. Reporting verbs were identified by combining a list derived from literature (primarily Hyland, 1999) along with an inspection of frequently occurring verbs.

Definition of research areas

We distinguish two different types of research areas at different levels of aggregation – broad disciplines and smaller fields. These are found using the clustering approach introduced by Waltman and van Eck (2012). They devised an algorithm that produces clusters of publications, allowing for the construction of a hierarchical classification system of science at different levels of aggregation. We use the meso-level, consisting of 868 distinct clusters of publications, which we will refer to as scientific fields. For analysis of larger aggregate areas, we use the top level of aggregation consisting of five broad scientific disciplines (*biomedical and health sciences*, *life and earth sciences*, *mathematics and computer science*, *physical sciences and engineering* and *social sciences and humanities*).

In our analysis, we consider context of citations from the perspective of the publications *receiving* the citations. Statistics presented for scientific fields or disciplines are the aggregated features of citing sentences pointing towards those areas of science. This perspective is important, as it allows us to characterise publications by their use in citing publications in upcoming research.

Visualisation

We use a map of clustered WoS publications as a representation of the scientific landscape, visualised using VOSviewer (van Eck & Waltman, 2010). This map, displayed in Figure 1, is based on direct citations among WoS publications spanning 2000-2017, grouped together into the aforementioned 868 scientific fields. Each field is represented as a circle, their sizes correspond to the number of publications in WoS belonging to the field and their colours represent the five broad scientific disciplines. This map will be used in the results section to display the values of selected features for each scientific field by overlaying a colour scale onto the map.

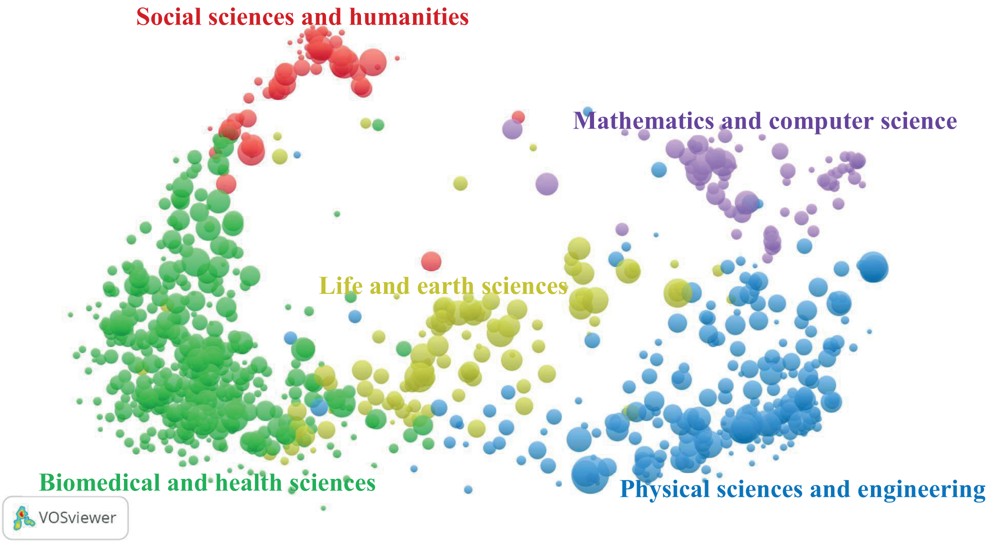

Social sciences and humanities
Mathematics and computer science
Life and earth sciences
Biomedical and health sciences
Physical sciences and engineering
VOSviewer

Figure 1. Base map of WoS publications (2000-2017). Colours represent disciplines. Circles represent 868 scientific fields. Size corresponds to the number of publications in WoS.

Assessing verb occurrence

To assess the occurrence of specific verb groups or single verbs across the scientific landscape, we divide the proportion of the verb (group) of interest (out of all verbs) within a scientific field by the expected proportion across the entire data set:

$$\text{actual/expected ratio}_{verb,field} = \frac{p_{verb,field}}{p_{verb,total}}$$

The benefit of this actual/expected ratio is that it is easy to interpret – a score of 2 means a verb (group) occurs twice as often in a specific scientific field than it does in the entire data set, while a score of 0.5 indicates the verb is half as common (or twice as rare). A downside of this measure is that our visualisations in VOSviewer allow only for linear overlay colour scales. To compensate for this, our visualisations instead use a surprisal score, defined as the base two logarithm of the frequency ratio, resulting in a symmetrical measure where +1 corresponds to a value double as high as expected, while -1 corresponds to a value half as high (or twice as small) as expected:

$$\text{surprisal}_{verb,field} = \log_2(\text{actual/expected ratio}_{verb,field}) = \log_2\left(\frac{p_{verb,field}}{p_{verb,total}}\right)$$

Results

We first present our results for citation labels and integral citation. We then report results for verb groups and individual verbs. For each of these, we first present results at an aggregate level over the five main scientific fields, before moving on to a more detailed description of the features across the scientific landscape.

Citation labels and integral citation

Table 1 presents averages for label style and integral citation rates for the five disciplines. Author-year labels dominate in *social sciences and humanities* and in *life and earth sciences*. Integral citation appears to be prevalent in the *social sciences and humanities*, but contrary to our hypothesis also in *mathematics and computer science*, where author-year labels are less common. On the other hand, *life and earth sciences* shows notably less integral citation despite its inclination towards using author-year style citation labels. This shows that citation label style and integral citation are not as closely linked as we expected, despite clear overlap of the practices in *social sciences and humanities*.

Table 1. Aggregate statistics over five broad scientific disciplines.

Main scientific field	*Citing sentences*	*Author-year labels*	*Integral citation*
Biomedical and health sciences	41,230,270	32.6%	5.9%
Life and earth sciences	15,272,782	76.6%	12.9%
Mathematics and computer science	3,256,527	27.7%	22.7%
Physical sciences and engineering	24,649,515	13.3%	10.1%
Social sciences and humanities	5,255,683	87.2%	22.0%

Figure 2 presents a more detailed look at the use of author-year style citation labels across the scientific field. Changes in the use of author-year style citations appear remarkably gradual, with a notable exception of a handful of clusters at the interface between *life and earth sciences*, *mathematics and computer science*, and *physical sciences and engineering*. These fields appear to predominantly feature tectonophysics (1), planetary science (2) and solar physics (3), which may explain why their use of author-year style citation labels appears to mimic the larger *earth sciences* discipline. Another interesting observation is the absence of author-year type labels at the 'fringes' of the map – on the right hand side, the realms of mathematics (A), particle physics (B), superconductivity (C), organic chemistry (D), and on the left a large region of surgery-related research fields (E).

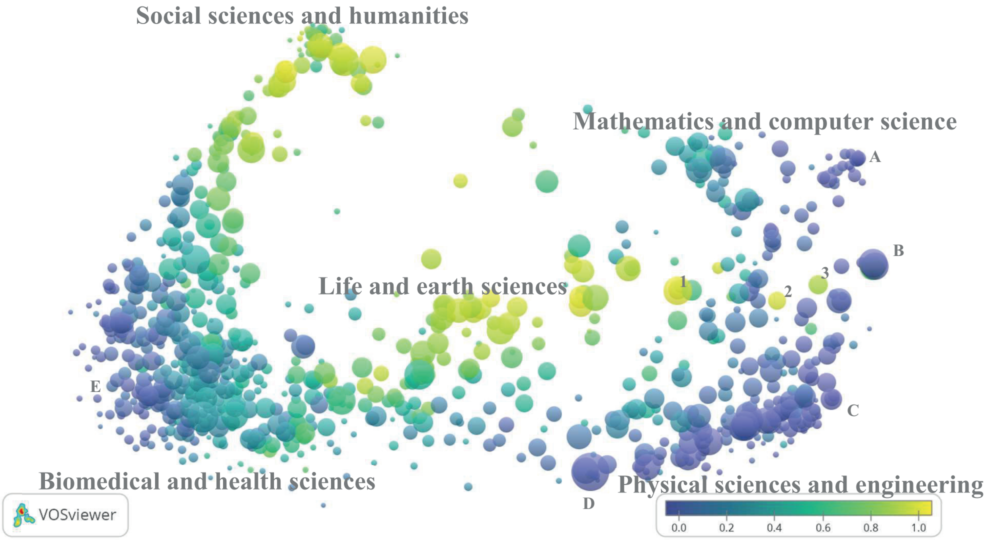


Figure 2. Share of author-year style citation labels.

Figure 3 presents a more detailed look at where in the scientific landscape integral citation happens. The disjunction of the author-year citation labels and integral citation becomes more apparent when comparing with the previous figure. Like the citation label style, integral citations are fairly common in the *social sciences and humanities*, yet they feature far less in the *life and earth sciences*. Planetary sciences again appear as fields where this phenomenon occurs, though this time adjacent fields pertaining to fluid dynamics, heat and mass transfer (5) and material simulation also feature integral citation at higher rates. Integral citations are used relatively frequently in the *mathematics and computer science* fields. They are especially prevalent in finance-related fields (6), statistics (7), and discrete mathematics (8). These findings align in large part with those of Hyland (1999), who found that integral citations occur often in philosophy, sociology, linguistics and marketing (in our *social sciences and humanities* discipline) and also in mechanical engineering, while they occurred less often in electrical engineering, physics and biology.

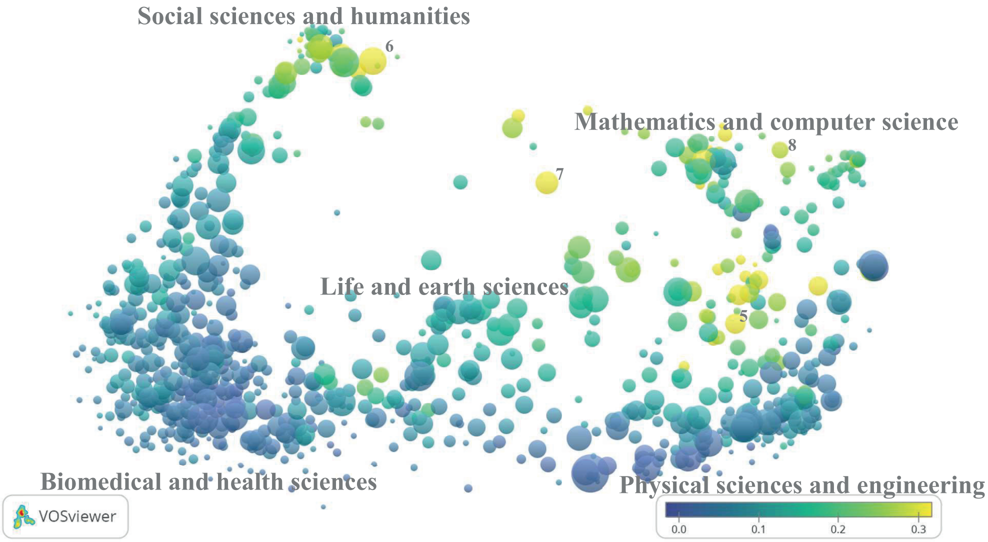


Figure 3. Share of integral citation.

Inspecting the data underlying these distributions, we find that use of the cited authors' last names appears to coincide with explicit acknowledgement of a certain contribution made by the cited authors. For instance, consider these example sentences from the discrete mathematics (8) field:

- *This conjecture was settled by Kathiresan and Amutha [9].*
- *In [9], C. Thomassen showed the following result.*
- *This game was proposed by Fukuyama in 2003 [15,16].*
- *Yang et al. [15] proved that all GHTs are hamiltonian graphs.*
- *Again Thomassen [6] proved the crucial result.*

Or these examples from the heat and mass transfer (5) field:

- *Following Weidman et al. [42], we introduce the new dimensionless time variable t.*
- *This source term definition was also used by Yang [18].*
- *A more complex model was elaborated by Kuhn et al. [2] and Placido et al. [3].*
- *A description of the calculation program is given by Sajjan et al. [10].*

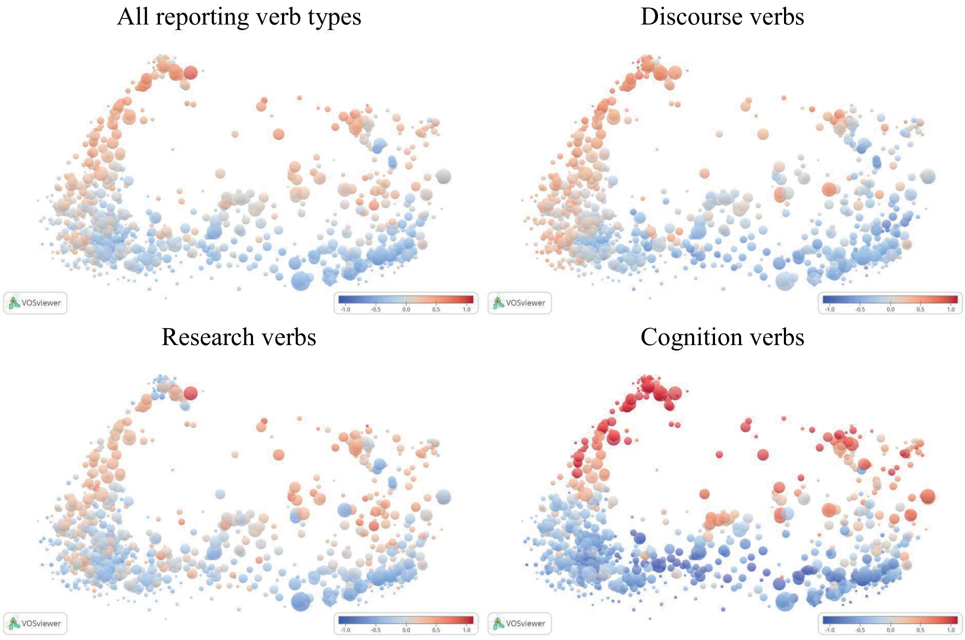
Reporting verbs

Table 2 provides an overview of the occurrence of (types of) reporting verbs in the citing sentences associated with each of the main scientific fields. Statistics are presented both for all citing sentences (all), and limited to only integral citations (int).

Table 2. Actual/expected ratio of reporting verbs in broad scientific disciplines.

Broad scientific discipline	Reporting verbs (all)		Discourse verbs		Research verbs		Cognition verbs	
Types of citation	all	int.	all	int.	all	int.	all	int.
Social sciences and humanities	1.37	0.98	1.40	0.91	1.22	0.97	2.31	1.71
Biomedical and health sciences	0.98	1.08	1.01	1.12	0.97	1.08	0.87	0.74
Physical sciences and engineering	0.91	0.96	0.85	0.91	0.96	1.01	0.79	0.73
Life and earth sciences	1.00	0.98	1.01	1.02	1.00	0.96	0.99	0.90
Mathematics and computer science	1.22	0.94	1.13	0.97	1.20	0.85	1.70	1.71

While it appears that certain disciplines favour the use of certain types of reporting verbs, the effects are fairly small. This holds true especially when we consider only sentences with integral citation. Only cognition verbs show a clear preference for social sciences and humanities and for mathematics and computer science. These results align partially with the results obtained by Hyland (1999), who also showed that the soft sciences use more reporting verbs than the hard sciences, a phenomenon he attributed to the more discursive nature of these disciplines. Hyland, however, found more pronounced differences in the usage of discourse verbs and research verbs.


All reporting verb types

Discourse verbs

Research verbs

Cognition verbs

Figure 4. Surprisal scores of reporting verbs in all citing sentences across scientific fields.

Figure 4 visualises the occurrence of reporting verbs across the scientific landscape. These visualisations confirm the general patterns across the disciplines observed in Table 2. Moreover, they show that reporting verbs decrease in prominence in the southeast of the map – the area of various material science and chemistry topics. Curious outliers may also be observed, for instance the prominence of research verbs in the finance-related field (6 in Figure 3). In this case, the field seems to prominently feature some very specific research verbs such as *see*, *find*, *estimate* and *test*. In cognition verbs, an apparent north-south divide persists, with fields that feature these verb types in other disciplines predominantly located closer to the social sciences and humanities and mathematics and computer science disciplines. Still, the small variations of reporting verb type usage across the scientific landscape (beyond cognition verbs) call for closer examination of specific verbs.

Figure 5 provides visualisations for six common verbs for each of the three reporting verb categories. It is immediately apparent that large differences in the use of these verbs exist, even within the same reporting verb category. See, for instance, the difference in usage between *suggest* and *propose* within the discourse verbs. Both verbs may be used to report on a tentative statement made by cited authors (compare "*Matthews et al. in [20] propose a solution to this problem.*" and "*Corbetta and Shulman [13] suggest that the ventral system might work as an alerting system.*") yet the areas of the scientific landscape where these verbs occur frequently are almost entirely distinct. In this case, this may have to do with the fact that *propose* seems to be used to introduce new methods or approaches whereas *suggest* seems to be used to infer something from observation.

Within the research verbs, *study* and *examine* are another pair of verbs that seem to have a similar meaning yet occur at opposite sides of the scientific landscape (compare "*Leuz et al. (2003) examine earnings management in the international context.*" and "*Ishak et al. [10] studied melting heat transfer in steady laminar flow over a moving surface.*"). No immediately apparent explanation for this difference appears, and the favoured verb might simply be a matter of convention within the disciplines. Many more verbs exhibit use patterns that are strongly concentrated in specific areas of the map. A VOSviewer visualisation for a large number of verbs will be made available at lamers.ws/research.

Conclusion

We have presented a field-level analysis of the use of author-year citation labels, integral citation, and reporting verbs across the science system. Apart from our primary contribution of showing exactly where in the scientific landscape author-year style citation labels, integral citations, and (types of) reporting verbs occur, our results allow us to make two key observations.

First, we have shown that integral citations – those using the cited author's name explicitly as part of the narrative of the sentence – overlap only partially with the use of author-year style citation labels, contradicting our initial hypothesis. Clearly, regions of the scientific landscape exist where author names are regularly used within sentence narrative but seldom within citation labels (mathematics, computer science) as well as the other way around (life and earth science). Large regions do however use both these forms (social sciences and humanities) or neither of these forms (large parts of biomedical and health sciences and of physical sciences and engineering). This suggests that citation label style and the use of integral citation are two separate facets of field-specific conventions of academic writing.

Second, it appears that ex ante defined groups of reporting verbs are far from internally homogeneous, at least as far as the distribution of their member verbs over the scientific landscape is concerned. Only cognition verbs showed a pronounced preference in where in the landscape they are prominent, and all three categories of reporting verbs contain verbs that behave very differently from one another. These findings suggest that these categories of reporting verbs cannot be regarded as monolithic groups, and bottom-up approaches to finding structure in sets of reporting verbs are warranted.

Overall, the features considered in this paper – from broad patterns of citation label styles and integral citation to verb-specific occurrence – show remarkably gradual changes across the scientific landscape. We can observe obvious 'hot spots' in the map that strongly favour the use of a specific verb or feature, as well as 'cold spots' far less inclined to use a certain feature, and regions of transition between them. Our extensive data set of full text citing publications provides ample avenues for future research into these phenomena at a hitherto unachievable level of detail. By exploring these avenues, our ultimate goal is to get a deep understanding of the way scientific publications build on each other and scientific knowledge accumulates.

Limitations

Several limitations of this study have to be acknowledged. First is the weak link between cited work and verb usage in the sentence. By extracting verbs without taking syntax into consideration, there is no guarantee that extracted reporting verbs are used to describe the contribution of cited authors. Instead reporting verbs can be used to refer to the author's own work (e.g. "*Our results indicate that most donors are generally well informed about donation, which is consistent with other research (87–95%) [22,23,45].*"). A potential solution would be to introduce dependency parsing instead of relying on simple part-of-speech tagging, for instance by using the Stanford Parser (Chen & Manning, 2014). Such an approach could allow us to extract those verbs whose subject is the cited author's last names in integral citations, ensuring a tighter link between extracted verb and cited publication. Still, despite this current limitation, we believe our approach does at least provide an overview of activities reported on across the scientific landscape, even if it remains unclear how exactly this relates to the cited works.

Another challenge to be addressed is our limited selection of citation context. Our analysis of citation context was limited to the sentence explicitly containing the citation label. This disregards any implicit reference to the cited paper, and these implicit references might contain more discussion of the cited work and more integral forms. While algorithms to establish the wider relevant context of citations are available (e.g. Athar & Teufel, 2012; Ou & Kim, 2018; Qazvinian & Radev, 2010; Sondhi & Zhai, 2014), we elected not to venture beyond citing sentence as further expansion of our already-vast data set would be challenging. Still, future research would do well to take the wider context of citation into consideration.

A final limitation of this work is the fact that semantically the same verb might mean vastly different things in different contexts, and that these differences should be expected to play a role across the scientific landscape. For instance, consider the verb *develop* in the context of computer science ("*Aickelin and Dowsland [3] developed a genetic algorithm with an indirect representation.*") and rheumatology ("*No patient developed TB or opportunistic infections [42].*"). Distinguishing between different meanings of the same verb is not trivial, though using n-grams following Bertin et al. (2016) may be a starting point.

References

Athar, A., & Teufel, S. (2012). Detection of implicit citations for sentiment detection. *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse* (pp. 18–26).

Bertin, M., & Atanassova, I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. *CEUR Workshop Proceedings* (Vol. 1143, pp. 5–12).

Bertin, M., Atanassova, I., Sugimoto, C. R., & Lariviere, V. (2016). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, *109*(3), 1417–1434.

Bloch, J. (2010). A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers. *Journal of Writing Research* (Vol. 2, pp. 219–244).

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740–750).

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*, *66*(7), 1374–1387.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Hyland, K. (1999). Academic attribution: citation and the construction of disciplinary knowledge. *Applied Linguistics*, *20*(3), 341–367.

Hyland, K. (2006). Disciplinary differences: language variation in academic discourses. (K. Hyland & M. Bondi, Eds.) *Academic discourse across disciplines*, (January 2006), 17–45.

Li, K., Rollins, J., & Yan, E. (2018). Web of Science use in published research and review papers 1997–2017: a selective, dynamic, cross-domain, content-based analysis. *Scientometrics*, *115*(1), 1–20.

Ou, S., & Kim, H. (2018). Unsupervised citation sentence identification based on similarity measurement. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10766 LNCS, pp. 384–394).

Qazvinian, V., & Radev, D. (2010). Identifying non-explicit citing sentences for citation-based summarization. *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, 11–16.

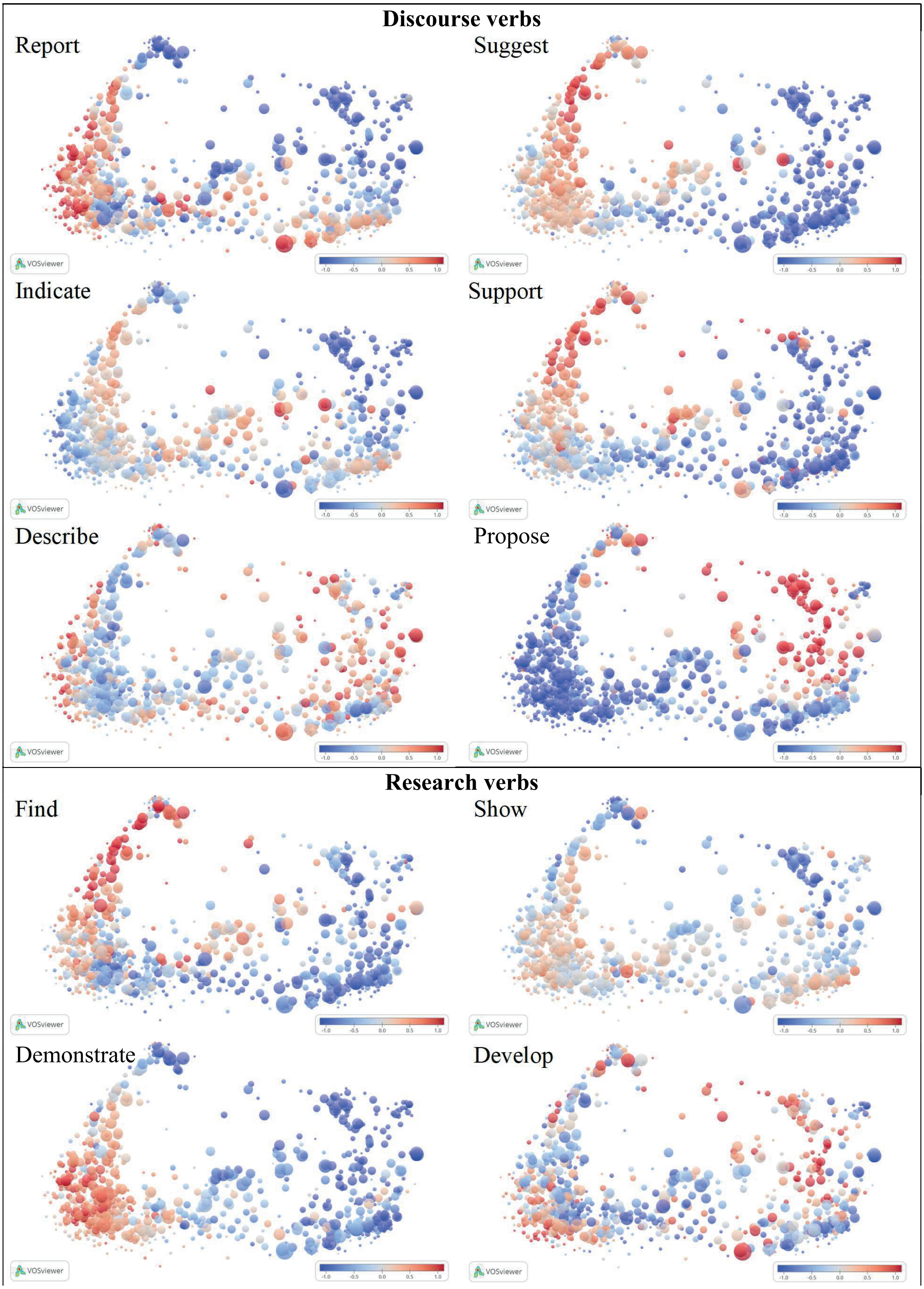
Sondhi, P., & Zhai, C. (2014). A constrained hidden Markov model approach for non-explicit citation context extraction. *Proceedings of the 2014 SIAM International Conference on Data Mining* (pp. 361–369).

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Thompson, G., & Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, *12*(4), 365–382.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.

Appendix


Discourse verbs
Report
Suggest
Indicate
Support
Describe
Propose
Research verbs
Find
Show
Demonstrate
Develop
VOSviewer
-1.0
-0.5
0.0
0.5
1.0

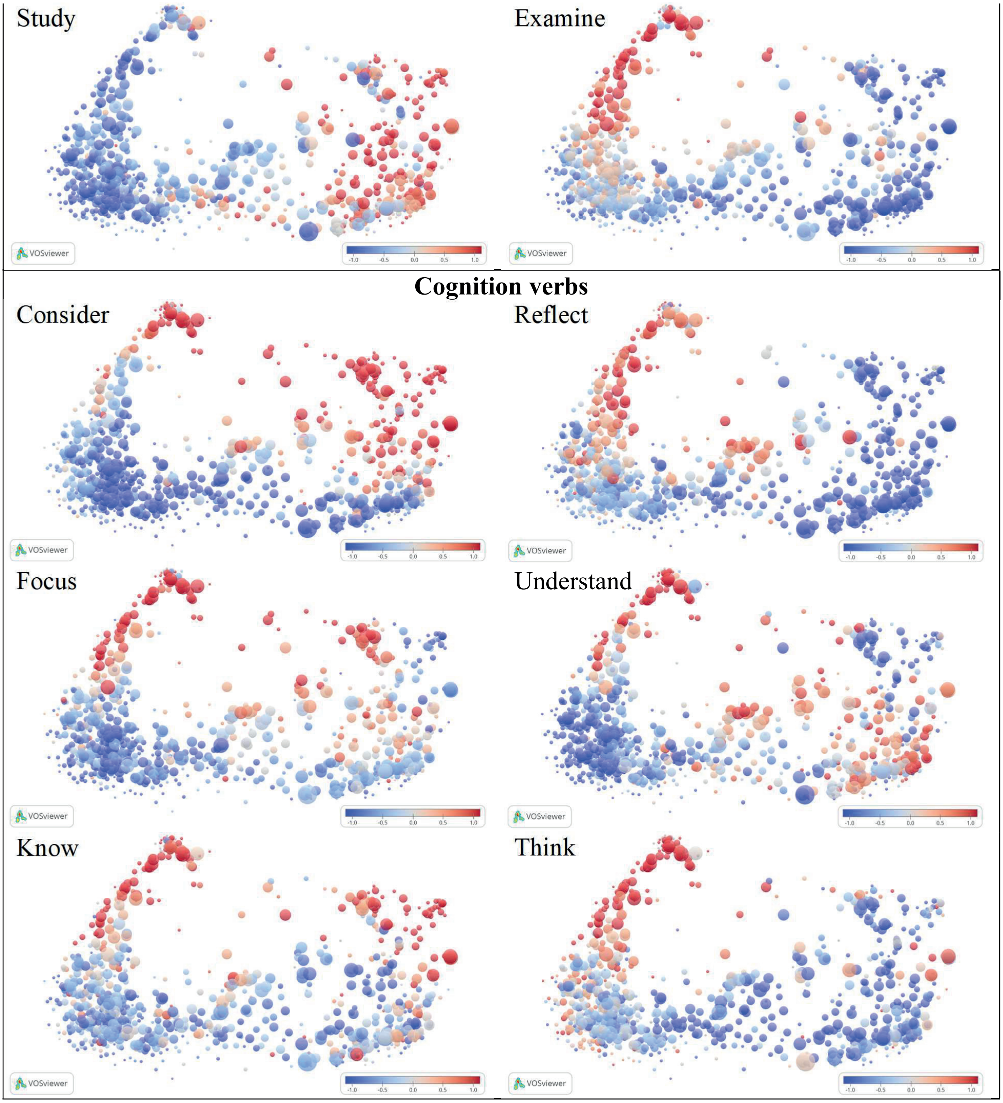


Figure 5. Six common verbs for each reporting verb category.

Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation

Angelo Di Iorio[1] and Francesco Poggi[1] and Silvio Peroni[2]

[1] *{angelo.diiorio|francesco.poggi5}@unibo.it*
Department of Computer Science and Engineering (DISI), University of Bologna, Bologna (Italy)

[2] *silvio.peroni@unibo.it*
Digital Humanities Advanced Research Centre (DHARC), Department of Classical Philology and Italian Studies, University of Bologna, Bologna (Italy)

Abstract

The need for scholarly open data is ever increasing. While there are large repositories of open access articles and free publication indexes, there are still a few examples of free citation networks and their coverage is partial. One of the results is that most of the evaluation processes based on citation counts rely on commercial citation databases. Things are changing under the pressure of the Initiative for Open Citations (I4OC), whose goal is to campaign for scholarly publishers to make their citations as totally open. This paper investigates the growth of open citations with an experiment on the Italian Scientific Habilitation, the National process for University Professor qualification, which instead uses data from commercial indexes. We simulated the procedure by only using open data and explored similarities and differences with the official results. The outcomes of the experiment show that the amount of open citation data currently available is not yet enough for obtaining similar results.

Introduction

Citations indexes are becoming more and more important for evaluating scientific performances of a given research body. For instance, in many countries citation and bibliometric indicators are some of the factors that can be used for assessing individuals or institutions to allocate funding at national level: in Germany the impact factor of the publications is used in performance-based funding systems, in Finland the reallocation system uses bibliometrics and citation indexes as ones of the considered measures, in Norway a two-level bibliometric indicator is used for similar purposes (Vieira et al., 2014a).

Several works analyzed the relation between citation indexes and research assessment procedures. At national level, the relation between bibliometric indicators and the results of the Research Assessment Exercise (RAE) in Britain (Norris and Oppenheim, 2003; Taylor, 2011) or the Italian Triennial Assessment Exercise (VTR) (Abramo et al., 2009; Franceschet and Costantini, 2011) have been investigated. Other studies focused on the assessment of departments (Aksnes, 2003) and research groups (Van Raan, 2006). Just a few works have been made at the individual level (Nederhof and Van Raan, 1987; Bornmann and Daniel, 2006; Bornmann et al., 2008, Poggi et al., 2019), while many analyzed the correlation between indicators and research performances (Leydesdorff, 2009; Franceschet, 2009). Recent works analyzed the correlation between traditional bibliometric indicators and altmetrics by also taking into account quality assessment procedures performed by peers (Nuzzolese et al., 2018; Wouters et al., 2015; Bornmann and Haunschild, 2018).

In this work we focus on the analysis of the Italian National Scientific Habilitation (ASN), a nation-wide evaluation process introduced some years ago by Law 240/2010 (Law dec. 30, n. 240, 2011) for University Professor position recruitment. The ASN is similar to other habilitation procedures already in place in other countries in that it is a prerequisite for becoming a university professor.

The procedure is based on scientific qualification criteria that take into account, among other factors, bibliometric indicators such as the number of citations and the h-index of the candidates. Citation data are taken from commercial databases, as it happens in other countries. One of the reasons is that the open citation indexes are still a few and their coverage is limited (van Eck et al., 2018). This is an issue not only for evaluation procedures but also for research activities on open science, trends and topics analysis, scientometrics and so on. The 'Initiative for Open Citations' (I4OC, https://i4oc.org) has been launched to gather publishers, researchers, and other interested parties and to promote the "unrestricted availability of scholarly citation data". The movement is gaining momentum and making available a lot of free citation data.

This work is part of a larger effort, whose goal is to monitor the growth of these data and their relation with closed ones. We are not only interested in counting open citations but also in exploring their distribution among datasets and domains, and their applicability to evaluation tasks. To the best of our knowledge, this is the first attempt to look at open citations for these tasks.

Here we would like to answer the following research questions:

RQ1. To what extent open bibliographic metadata and open citation data can be used for evaluation purposes today?

RQ2. Which open data would produce results comparable to those of closed ones?

RQ3. Is there any case in which a negative evaluation would turn into a positive one, if open data were used instead of closed ones?

To answer these questions we run an experiment on the Italian Habilitation in the Computer Science domain. The test gave us valuable indications and allowed us to build the overall infrastructure for extending our analysis to other domains.

One of the reasons for starting with Computer Science was the availability of open, complete and well-maintained repositories of articles and publication lists. In fact, our experiment consisted of two phases:

1. computing the indicators proposed by the ASN for all candidates by only taking into account open data. We collected these data from three main sources, namely Crossref (https://www.crossref.org/), DBLP (https://dblp.uni-trier.de/) and COCI (http://opencitations.net/index/coci) that will be introduced in the following sections;
2. comparing the outcome of such evaluation with the official one, whose data were collected from Scopus and Web of Science.

The experiments showed that there is still a quite large gap between open and closed citations and the former cannot yet be used directly for these tasks (RQ1). However, the data about the types of the publications, in particular journals, are comparable with the outcomes of the ANS 2016 (RQ2). Interestingly, we also found a few candidates for which open data would change the evaluation from negative to positive (RQ3).

The paper presents the methods and the results of our experiment and its implications. It is then structured as follows: Section "Background" provides some background introducing both the ASN process and the open citations status. Section "Methods and materials" introduces the sources we used for gathering the metadata and citation data, while Section "Experiments with open data and ASN" explains our experiment in detail. Results and lessons learned are discussed in Section "Results", before concluding and drafting new research directions in Section "Discussion and conclusions".

Background

In order to introduce our experiment we first need to provide readers with some background about the Italian ASN and the I4OC movement.

The Italian National Scientific Habilitation (ASN)

The ASN (Law dec. 30, n. 240, 2011) is a nation-wide research assessment procedure similar to others already in place in other countries. The first two sessions of the ASN took place in 2012 and 2013, followed by other sessions in 2016, 2017 and 2018.
The ASN is meant to attest that an individual has reached the scientific maturity required for applying for a specific role as Associate or Full Professor. Each candidate is bound to a specific Recruitment Field (RF), which corresponds to a scientific field of study. RFs are organized in groups, which are in turn sorted in 14 Scientific Areas (SAs).
The assessment of the candidates of each discipline (RF) is performed by a committee of full professors, which evaluates the CVs submitted by the applicants. The evaluation also takes into account *three quantitative indicators* computed for each candidate.

The ASN introduced two types of indicators: *bibliometric* and *non-bibliometric*. Bibliometric indicators apply to scientific disciplines for which reliable citation databases exist, among which Computer Science, on which we performed our analysis. The three bibliometric indicators are:

A. Normalized number of journal papers;
B. Normalized number of citations received;
C. Normalized h-index.

Since citations and paper count increase over time, normalization based on the scientific age (the number of years since the first publication) is used to compute these indicators more reliably.

The three indicators are computed by ANVUR – the National Agency in charge of the Habilitation process – for each candidate, starting from the data in Scopus and Web of Science. These databases, in fact, contain either the full list of classified publications for each candidate (used to compute indicator A) and the full list of citations received by each article (used to compute indicators B and C). These data were automatically compiled into a CV in PDF, submitted to the committee.

The preliminary step of the ASN consisted of checking, for each candidate, how many indicators exceeded some thresholds. The candidates were required to exceed at least two indicators over three. Exceeding thresholds does not imply that the candidate gets the habilitation automatically but is only an indication for the committee.

Though, this step is the focus of our experiment. We do not analyze the final subjective evaluation of the committee but we compare the ability of each candidate to exceed thresholds when using open or closed data.

The thresholds were computed by ANVUR as well and officially released for each RF. Even in this case, data to compute thresholds were taken from Scopus and Web of Science. In particular, in 2012, the thresholds were defined as the medians of the values computed for all Associate and Full professors already permanent. However, in 2016, the values were established by ANVUR but they did not disclose the algorithm to do that.

Several analyses of the ASN process and results have been carried out by the research community, like the quantitative analyses of ASN 2012 in (Marzolla, 2015) and (Marzolla, 2016), the study on the impact of the ASN on self-citation rate (Scarpa et al., 2018), the analysis on the relationship of the ASN outcomes to the actual scientific merit of candidates (Abramo and D'Angelo, 2015), etc.. Our goal is not to evaluate the reliability of ASN, nor to assess its effects and consequences, but to investigate to what extent such an evaluation could be performed without using commercial citation indexes.

The open citations movement

The first project to introduce for the very first time the open availability of open bibliographic and citation data by the use of Semantic Web (Linked Data) technologies was the OpenCitations Corpus, in 2010, as the main output of a project funded by JISC (Shotton, 2013). However, the availability of open citation data recently changed drastically with the introduction of Initiative for Open Citations (I4OC, https://i4oc.org), in April 2017.

The Initiative was born with the idea of promoting the release of open citation data, and explicitly asked the main scholarly publishers, who deposited their citations on Crossref (https://crossref.org), to release them in the public domain. As a result, now we have several millions of citation data openly available on the Web, a list of important stakeholders – such as libraries, consortiums, projects, organizations, companies, and, in particular, founders (Shotton, 2018) – supporting the movement, several international events (e.g. the Workshop on Open Citations and WikiCite 2018) organised for promoting the open availability of citation data, and several applications (e.g. Bibliography EXplorer (Di Iorio et al., 2015)), projects and datasets (e.g. SemanticLancet (Di Iorio et al., 2017)) have been released so far so as to leverage the open citation data available online. As a result, there is a growing list of publishers that release their citation data in Crossref, and these also include citation data of important Computer Science venues and publishers such as the Association for Computing Machinery (ACM).

Methods and materials

The first step of our analysis consisted in computing the indicators proposed by the ASN for all the candidates in the Recruitment Field *Informatics*[1] by taking into account only open data. To do so, we first collected the curricula of all applicants to the five sessions of the ASN 2016, which have been made publicly available on the ANVUR website for a short period of time. We collected 518 CVs for level I (full professor) and 757 CVs for level II (associate professor). Note that each CV correspond to a single application, and that the same applicant may apply multiple times (i.e. in more than one session) for multiple levels.

The next step consisted in collecting the list of the DOIs of all the publications that each candidate specified in her/his CV, thus excluding all the publications that do not have associated a DOI. This lead us to miss some publications, for instance the workshop articles in the CEUR-WS volumes, which are published without a DOI (though Scopus takes track of them). However, we expect that the loss in term of citations is rather limited, considering that the most relevant works and their extensions usually go to journal articles and conference proceedings papers, which are instead associated with a DOI.

We used two different sources to retrieve the features needed for such computation:

1. DBLP (https://dblp.uni-trier.de): it is a free and publicly available computer science bibliography repository started in 1993 at the University of Trier, Germany. DBLP contained more than 4.4 million bibliographic entries (as of January 2019). We search the candidates by name using the DBLP API, and downloaded the full publication list of each of them. We exploited standard disambiguation techniques and ORCID data to identify candidates.
2. Candidates' CVs: we extracted the text from each CV (which was originally in PDF format), and searched for valid DOIs using a simple pattern matching approach to

[1] Since all the Recruitment Field names have been defined only in Italian, we use here the official English translation provided by the Italian National University Council (CUN), the elected body representing the Italian University System, which is available at https://www.cun.it/documentazione/academic-fields-and-disciplines-list/.

produce the publication list. The DOI system Proxy Server REST API (http://www.doi.org/factsheets/DOIProxy.html#rest-api) have been used to verify the existence and validity of the collected DOIs.

The collected data have been used to produce three publication lists for each candidate: the first contains the DOIs of the publications retrieved from DBLP, the second contains the DOIs of the publications extracted from the CV, and the latter is the union of the DOI of the publications collected from the two sources (where duplicates DOIs have been considered only once). Table 1 reports some basic statistics about these three datasets.

Table 1. Basic statistics about the application submitted for Recruitment Field Informatics at the ASN 2016. For each level we report the number of CVs collected, and the overall number of retrieved DOIs of applicants' publications and, in parentheses, the average number of publications per candidate extracted from (i.) DBLP, (ii) the CVs, and (iii) the union of them. In addition, both DBLP and Crossref were used to retrieve the publication types of all the aforementioned DOIs.

Level	*CVs*	*DOI DBLP*	*DOI CV*	*DOI UNION*
	(in parentheses average DOIs per applicant)*			
Associate Professor	757	31713 (41.9)	31896 (42.1)	36820 (48.6)
Full Professor	518	37728 (72.8)	37793 (73.0)	42375 (81.8)

All the citations related to the DOIs extracted were gathered from COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations (http://opencitations.net/index/coci). This dataset is provided by OpenCitations, a scholarly infrastructure organization dedicated to open scholarship and the publication of open bibliographic and citation data (Peroni and Shotton, 2018) by the use of Semantic Web (Linked Data) technologies. Launched in July 2018, COCI is the first of the Indexes proposed by OpenCitations (http://opencitations.net/index) in which citations are exposed as first-class data entities with accompanying properties, and currently contains 449,840,503 DOI-to-DOI citation links between 46,534,705 distinct bibliographic entities (OpenCitations, 2018).

To date, the majority of the citations stored in Crossref that are not available in COCI comes from just three publishers: Elsevier, the American Chemical Society and University of Chicago Press (Heibi et al., 2019). This is due to the particular access policy chosen by these publishers, since such citation data refer to publications for which the reference lists are not visible to anyone outside the Crossref Cited-by membership. The advantage to access COCI instead of Crossref is that COCI also contains DOI-to-DOI citations that are included in the Crossref 'limited' dataset, which is accessible only to users of the Crossref Cited-by service and to Metadata Plus members of Crossref. Additional information about the way Crossref classifies reference lists is available at https://www.crossref.org/reference-distribution/. The fact that COCI does not contain citations from Elsevier's articles can be a bottleneck to our study, since such publisher manages several of the most important Computer Science journals that are valuable sources of citations to other Computer Science articles.

The source code of the pipeline to collect these data is available as open source at https://github.com/sosgang/asn2016-issi2019, while the data are available on Zenodo and are released with a CC0 waiver (Di Iorio, Peroni and Poggi, 2019).

Experiments with open data and ASN

The core of the experiment consisted of studying the performance of the candidates when using open data instead of closed ones. We repeated the test under three conditions corresponding to the three overlapping sets of DOIs. For the sake of clarity these conditions will be indicated with an acronym from now on:

- **CCV**: DOIs taken from CVs and citations taken from COCI;
- **CDBLP**: DOIs taken from DBLP and citations taken from COCI;
- **CU**: DOIs obtained as the union of the DOIs in CCV and CDLP, and citations taken from COCI.

The following step consists of calculating the three thresholds against which compare our data. Our initial plan was to re-calculate these thresholds as medians of the values of the indicators for Associate and Full Professors, as done for the ASN 2012. We also expected to compute the indicators for each condition. This was not done since ANVUR did not publish the algorithm to calculate the official thresholds in 2016 but only their values. Then we used the official thresholds directly, even if they were calculated from closed (and, potentially, more rich) data. This is not optimal but gave us valuable insights and we plan to do further experiments on the thresholds as discussed at the end of the paper. The current values are shown in Table 2.

Table 2. The ASN 2016 thresholds for the Associated Professor and Full Professor positions.

Role	*#journals articles (A)*	*#citations (B)*	*H-index (C)*
Associate Professor	5	118	6
Full Professor	8	216	8

Then, for each indicator we calculated the percentage of candidates who were able to exceed the thresholds in both our test and the official ASN. We also measured the amount of candidates who exceeded two thresholds over three - and thus were able to continue the process to get the qualification - in both cases. Note that we do not compare the values of the indicators directly, as we expected them to be different, rather their contribution to the habilitation.

Table 3. Two (real) candidates of the ASN accompanied by their values for the three indicators used in the ASN, i.e. number of journal articles, number of citations, and h-index. The number shown refers to those ones retrieved by means of open data and the real

	Open data			*Official ASN data*		
id	*#journal articles (A)*	*#citations (B)*	*H-index (C)*	*#journals articles (A)*	*#citations (B)*	*H-index (C)*
1	15	417	12	17	1144	17
2	8	197	7	33	1939	18

For instance, let us consider the two (real) candidates in Table 3. They both applied for the qualification as Full Professor and exceeded all three thresholds. The values of their indicators were lower when we only took open data into account in both cases. However, candidate #1 was able to exceed two thresholds anyway. The same did not happen for candidate #2. We counted the percentage of these situations to study the relation between open and closed data. The measurements were also repeated on the three datasets CCV, CDBLP and CU in order to get a more precise picture and are fully described in the next section.

Results

We measured the percentage of candidates, for all levels and under all conditions, for which there is agreement between our test and the official ASN outcome. Table 4 summarizes the data on the 517 candidates as Full Professor (Level 1). The three columns correspond to the three conditions introduced CCV, CDBLP and CU. The rows detail each indicator and the overall result (two indicators over three above/below the thresholds).

Table 4. The percentage of candidates as Full Professor who achieved the same result in our open data simulation and the official ASN, for each indicator and under each condition.

Full Professor (518 candidates)			
	CCV	*CDBLP*	*CU*
Overall	59.07%	58.88%	67.95%
Journals	89.77%	89.58%	93.82%
Citations	50.77%	50.58%	58.49%
h-index	59.65%	59.46%	67.76%

Overall, the results on open data are not yet comparable to those on closed ones. In fact, the agreement ranges from 58.88% of CDBLP to 67.95% of CU. The three indicators contributed in different ways to this result: while there was a substantial agreement on indicator A (articles in journals) with a percentage of about 90% for all three cases, the percentages lower to about 50% for the citations (indicator B) and 60% for the h-index (indicator C). It is also worth noticing that the results increase of about 8-9% when considering the union of CCV and CDBLP.

Table 5 shows the results for the applications as Associate Professor. The number of candidates was 757 and the overall agreement was in line with the previous scenario. In fact about 57% of the candidates got the same result in both the evaluations for CCV and DBLP while the agreement grows up to 70.94% for CU. Again, the agreement was very high for the indicator A (journal articles) and the ratio between the three indicators was quite stable.

Table 5. The percentage of candidates as Associate Professor who achieved the same result in our open data simulation and the official ASN, for each indicator and under each condition.

Associate Professor (757 candidates)			
	CCV	*CDBLP*	*CU*
Overall	57.60%	56.80%	70.94%
Journals	80.58%	79.79%	90.36%
Citations	49.14%	48.75%	60.24%
h-index	62.35%	61.56%	73.98%

As expected, the overall trend was that candidates get worse results when only considering open citation data, since the amount of these data was still limited when compared with closed citation data. We also asked ourselves if there are instead candidates whose results improved if the ASN had used open data. It might also happen in fact that DBLP data (i.e. the DOIs it contains) are richer than the corresponding in Scopus and Web of Science, so that some indicators could differ.

To study such aspect we measured the percentage of candidates that exceeded the thresholds with open data but not with the closed ones (and vice versa). We also computed these variations for all indicators and the overall score. Results are summarized in Table 6, under the three conditions CCV, CDBLP and CU.

Table 6. The percentage of candidates who exceeded the thresholds with open data but not with the closed ones (column '+') or vice versa (column '-'). The results are shown for all conditions CCV, CDBLP and CU. The table is split in two mirror-like parts, for c

	CCV		*CDBLP*		*CU*	
	+	-	+	-	+	-
	Full Professor					
overall	0.19%	40.73%	0.19%	40.93%	0.39%	31.66%
journals	1.54%	8.69%	1.54%	8.88%	2.32%	3.86%
citations	0.00%	49.23%	0.00%	49.42%	0.39%	41.12%
h-index	0.19%	40.15%	0.19%	40.35%	0.19%	32.05%
	Associate Professor					
overall	0.13%	42.27%	0.13%	43.06%	0.13%	28.93%
journals	2.77%	16.64%	2.77%	17.44%	3.96%	5.68%
citations	0.26%	50.59%	0.13%	51.12%	0.53%	39.23%
h-index	0.66%	36.99%	0.66%	37.78%	1.06%	24.97%

In a very limited set of cases the open data produced a growth in the performance, with a slightly more evident increment for the indicator A. In general the impact of adopting open data would then be very limited with a few exceptions.

Discussion and conclusions

The results of our experiment are in line with what we expected with some interesting unexpected behaviour. First of all, it is evident that **open citation data are not yet complete to substitute the closed data** used by ANVUR within the ASN in Computer Science. This answer our research question RQ1. It was foreseen considering that several publishers have not released their citation data as open and there is still a gap between the two sets and **some effort is still needed to convince publishers to release their data**. On the other hand, the overall agreement of around 60% is a positive result that makes us optimistic about the possibility of performing some reliable evaluation on these data as well.

The high agreement on the indicator A (journal articles) allowed us to answer the research question RQ2. The classification of the publications is extremely good in open data and we speculate that the fact that the agreement is not full is due to the lack of data instead of their inaccuracy. The **accuracy of open and closed data on the classification of the publication venues (in particular "journals" vs. "non-journals") is comparable**, and counting journals of publication by type can be done reliably.

While the indicator A proved to be stable we witnessed a remarkable improvement of the agreement in the CU scenario compared to the CDBLP or CCV ones. The overall agreement, for instance, goes from to 56% to 70%, and from 63% to 70% on the h-index. We speculate that is a consequence of the nature and the coverage of the two datasets, in relation of the scientific production of the candidates. The DBLP source in fact is very accurate for pure computer scientists since almost all their publications are listed there. On the other hand, DBLP misses articles in close domains, such as bioinformatics and physics. There were candidates that applied for the habilitation in Computer Science even if they are experts in these domains. We are not interested here in the overlap between the two domains, neither on the evolution of the research topics, studied for instance in Osborne et al. (2013) and Salatino et al. (2017), but we noticed that some candidates were penalized by the mono-disciplinary approach of the CDBLP scenario. The CCV condition, on the other hand, produced slightly better results for multidisciplinary experts, with some penalisation for pure computer scientists. The union of these two inevitably had a positive impact on the agreement, that raised up to 70%. Our conclusion is that **combining different sources of open data is a fundamental step to better evaluate candidates**.

A further conclusion that we have drawn from our data is that there are no substantial differences between the simulation on Full and Associate Professors. The overall distribution of percentages and relation between indicators is basically the same, even if specific data are different. Contrarily to other aspects of the ASN in which different behaviours of candidates were pointed out (Marini, 2017) our experiment showed that an evaluation based on **open data works in the same way for either Full or Associate Professors assessment**.

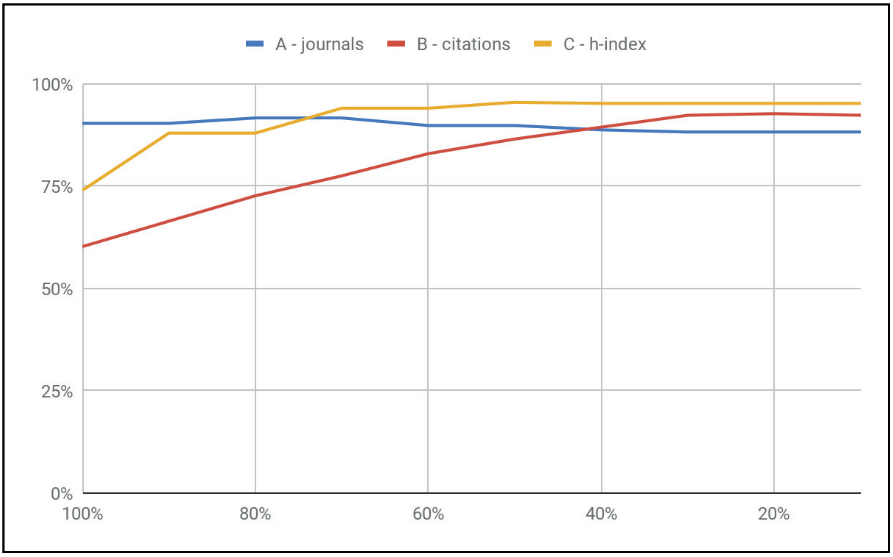


Figure 1. The variation of the agreement between our simulation and the official ASN, when lowering the thresholds. The X axis indicates the ratio between the new simulated thresholds and the official ones; the Y axis indicates the amount of candidates on which *there is* agreement. Data are shown for candidates as Associate Professor under the DBLP condition.

One objection that could be raised on our work is that we used the ASN official thresholds even on open data, that we already knew were less. We are aware that this is not optimal but we had no undisputable algorithm to re-calculate thresholds on our dataset in a consistent way with the ASN procedure, since details about that were not yet published by ANVUR. As discussed earlier, in fact, the ASN thresholds were computed automatically in 2012 and made available by law in the following sessions without further details.

To study thresholds, we artificially tuned them for open data and looked at how the agreement on the candidates, overall and for each indicator, changed when changing these values. Note that we only manipulated our simulated thresholds leaving untouched the official ones. The variation is plotted in Figure 1 for the candidates as Associate Professors in the CDBLP condition. For the sake of brevity we only show this scenario but there are no significant differences with the other ones.

The X axis indicates the ratio between the new simulated thresholds and the official ones; the Y axis indicates the amount of candidates on which there is agreement; the three lines show the behaviour of the three indicators. Thus, the value labelled as 100% on the X axis correspond to the official ASN thresholds; the position 60% correspond to the values 3 (indicator A), 71 (indicator B) and 4 (indicator C) calculated in percentage to the original values 5 (indicator A), 118 (indicator B) and 6 (indicator C).

It is interesting to notice that the indicator A is very stable. This is in line with previous results since there was already high agreement on this indicator and thus the effect of lowering the threshold is limited. Note also that a lower threshold might also reduce the agreement since a candidate might exceed the threshold in the simulated ASN but not on the official one. This is the reason why the indicator goes slightly down.

It is also evident a growth of the agreement when reducing the other two thresholds by 20-30%. This mitigates the limited availability of open data since candidates are given the possibility to exceed thresholds anyway. Note that the agreement never goes up to 100% since lower thresholds might change the performance of some candidates; this also depends on the fact that here we are considering each indicator separately while the agreement is computed on two of them.

In the near future we plan to also study the interaction between these indicators and to put in place optimization techniques to investigate thresholds for open data. Such an analysis in fact is limited since all thresholds are lowered proportionally, while our work highlighted clear differences between indicators caused by the differences between the data sources.

Another activity that we plan for the future is to extend our analysis research to other domains and to dig into the potentialities and limitations of other open repositories, with particularly attention to the field of medicine and open access journals (such as PMC Open Access Subset, https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/). It will also be interesting to see the results of a similar analysis on Google Scholar data, which expose much more resources for each scholar (of different type and authoritativeness) and thus require some pre-processing and filtering.

References

Abramo, G., D'Angelo, C. A., and Caprasecca, A. (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy, 38*(1):206–215. DOI:10.1016/j.respol.2008.11.001.

Abramo, G., & D'Angelo, C. A. (2015). An assessment of the first "scientific habilitation" for university appointments in Italy. *Economia Politica, 32*(3), 329-357.

Aksnes, D. (2003). A macro study of self-citation. Scientometrics, 56(2):235–246. DOI:10.1023/A:102191922.

Bornmann, L. and Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review-A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics, 68*(3):427–440. DOI:10.1007/s11192-006-0121-1.

Bornmann, L., Wallon, G., and Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. *PLoS One, 3*(10):e3480. DOI:10.1371/journal.pone.0003480.

Bornmann, L. and Haunschild, R. (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000 Prime data. *PLoS One, 13*(5):e0197133. DOI:10.1371/journal.pone.0197133.

Di Iorio, A., Giannella, R., Poggi, F., Peroni, S., & Vitali, F. (2015). Exploring scholarly papers through citations. In *Proceedings of the 2015 ACM Symposium on Document Engineering* (pp. 107-116). ACM. DOI: 10.1145/2682571.2797065.

Di Iorio, A., Nuzzolese, A. G., Peroni, S., Poggi, F., Vitali, F., & Ciancarini, P. (2017). Analysing and Discovering Semantic Relations in Scholarly Data. In *Italian Research Conference on Digital Libraries* (pp. 3-19). Springer, Cham. DOI: 10.1007/978-3-319-68130-6_1.

Di Iorio, A., Peroni, S., Poggi, F. (2019). ASN 2016 evaluation with open data. Version 1. Zenodo. DOI: 10.5281/zenodo.2559481

Franceschet, M. and Costantini, A. (2011). The first Italian research assessment exercise: A bibliometric perspective. *Journal of informetrics, 5*(2):275–291. DOI:10.1016/j.joi.2010.12.002.

Heibi, I., Peroni, S. & Shotton, D. (2019). Crowdsourcing open citations with CROCI – An analysis of the current status of open citations, and a proposal. Submitted for publication at the *17th International Conference on Scientometrics and Bibliometrics (ISSI 2019).* Preprint available at https://opencitations.wordpress.com/2019/02/07/crowdsourcing-open-citations-with-croci/ (last visited 7 February 2019)

Law dec. 30, n. 240 (2011). Rules concerning the organization of the universities, academic employees and recruitment procedures, empowering the government to foster the quality and efficiency of the university system (Norme in materia di organizzazione delle università, di personale accademico e reclutamento, nonche' delega al Governo per incentivare la qualità e l'efficienza del sistema universitario), Gazzetta Ufficiale n. 10 del 14 gennaio 2011 - Suppl. Ordinario n. 11. Available at http://www.gazzettaufficiale.it/eli/id/2011/01/14/011G0009/sg. (Accessed 6 June 2019).

Marini, G. (2017). New promotion patterns in Italian universities: Less seniority and more productivity? *Data from ASN. Higher Education, 73*(2), pp.189-205.

Marzolla, M. (2015). Quantitative analysis of the Italian national scientific qualification. *Journal of Informetrics, 9*(2):285–316. DOI:10.1016/j.joi.2015.02.006

Marzolla, M. (2016). Assessing evaluation procedures for individual researchers: The case of the Italian National Scientific Qualification. *Journal of Informetrics, 10*(2), 408-438. DOI: 10.1016/j.joi.2016.01.009

Nederhof, A. J. and Van Raan, A. F. (1987). Peer review and bibliometric indicators of scientific performance: a comparison of cum laude doctorates with ordinary doctorates in physics. *Scientometrics, 11*(5-6):333–350. DOI:10.1007/BF02279353.

Norris, M. and Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation, 59*(6):709–730. DOI:10.1108/00220410310698734

Nuzzolese, A. G., Ciancarini, P., Gangemi, A., Peroni, S., Poggi, F., & Presutti, V. (2019). Do altmetrics work for assessing research quality?. *Scientometrics, 118*(2), 539-562. DOI:10.1007/s11192-018-2988-z

Osborne, Francesco, Enrico Motta, and Paul Mulholland. "Exploring scholarly data with rexplore." In *International semantic web conference*, pp. 460-477. Springer, Berlin, Heidelberg, 2013.

OpenCitations (2018). COCI CSV dataset of all the citation data. Version 3. Figshare. DOI: 10.6084/m9.figshare.6741422.v3

Peroni, S. & Shotton, D. (2018). Open Citation: Definition. Version 1. Figshare. DOI: 10.6084/m9.figshare.6683855

Poggi F, Ciancarini P, Gangemi A, Nuzzolese AG, Peroni S, Presutti V. (2019). Predicting the results of evaluation procedures of academics. *PeerJ Computer Science* 5:e199. DOI: 10.7717/peerj-cs.199.

Salatino, A.A., Osborne, F. and Motta, E. (2017). How are topics born? Understanding the research dynamics preceding the emergence of new areas. *PeerJ Computer Science,* 3, p.e119.

Scarpa, F., Bianco, V., & Tagliafico, L. A. (2018). The impact of the national assessment exercises on self-citation rate and publication venue: an empirical investigation on the engineering academic sector in Italy. *Scientometrics, 117*(2), 997-1022.

Shotton, D. (2013). Publishing: Open citations. *Nature News, 502*(7471), 295–297. DOI: 10.1038/502295a.

Shotton, D. (2018). Funders should mandate open citations. *Nature, 553*(7687), 129-129.

Taylor, J. (2011). The assessment of research quality in UK universities: peer review or metrics? *British Journal of Management, 22*(2):202–217. DOI:10.1111/j.1467-8551.2010.00722.x.

van Eck, N.J., Waltman, L., Larivière, V. & Sugimoto, C. R. (2018). Crossref as a new source of citation data: A comparison with Web of Science and Scopus. CWTS Blog. https://www.cwts.nl/blog?article=n-r2s234 (last visited 26 January 2018)

Van Raan, A. F. (2006). Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3):491–502. DOI:10.1556/Scient.67.2006.3.10.

Vieira, E. S., Cabral, J. A., and Gomes, J. A. (2014a). Definition of a model based on bibliometric indicators for assessing applicants to academic positions. *Journal of the Association for Information Science and Technology, 65*(3):560–577. DOI:10.1002/asi.22981.

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., and Franssen, T. (2015). The metric tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management). London: Higher Education Funding Council for England (HEFCE). DOI:10.13140/RG.2.1.3362.4162.

Identification of technologically relevant papers based on their references

Yasuhiro Yamashita[1]

[1] *yasuhiro.yamashita@jst.go.jp*
Japan Science and Technology Agency, K's Gobancho, 7, Gobancho, Chiyoda-ku, 102-0076 Tokyo (Japan)

Abstract

In this study, two indicators that were derived from references in papers were proposed to characterize papers regarding technological relevance: (1) number of reference papers that obtained citations from patents by the time of observation, i.e., publication years of papers to be assessed (NR-PCP), and (2) number of reference papers authored by the firms' researchers (NR-FP).

Next, the two indicators were applied to papers published in 2001 to assess their performance. The results obtained by the two indicators were evaluated by citations from patents until 2016 in various conditions: scientific field, institutional sector, and period of measurement. Results showed a robustness of both indicators in many conditions. NR-PCP showed better results in most cases than NR-FP, although its recall was inferior to NR-FP for papers in which all references were newer than 1996.

Based on the result that NR-PCP was preferred as an indicator, the rationale of using reference papers cited in the patent by the period of observation (R-PCP) as an indicator was considered based on the papers' potential distance from the border between science and technology, which was obtained from an extended Ahmadpoor's citation network.

Finally, issues to be addressed were discussed.

Introduction

Citation linkages from patents to papers have been deemed as a quantitative indicator of relevance of patented inventions and scientific knowledge since Narin and his colleagues utilized said measure (e.g., Narin & Noma, 1985; Narin, 1991). Although these linkages can be used from both the patent side and the paper side, the measurement conditions of the latter are more severe than those of the former. From the view of patents, these linkages can be observed as soon as the patent documents to be evaluated are published. However, to assess the relevance of the papers to technology, one must wait a long time. Papers' citations from patents (hereafter patent-paper citations, or PPC) are relatively rare and need a longer time to be observed than citations from papers to papers. Today, information on the latest research output is essential in the formulation of science, technology, and innovation strategies. Therefore, the papers' relevance to technology should be grasped as soon as the papers are published.

Many researchers (e.g., Tijssen, Buter & Leeuwen, 2000; Hicks et al., 2000; Fukuzawa & Ida, 2016; Yamashita, 2018) have unveiled features of papers that positively correlate to PPC, such as document type (review), scientific field (life science), institutional sector (firm), and paper or journal impact. Although these features correlate with PPC, they are not suitable for a proxy of technological relevance because they do not designate intuitive relevance to technology (except for the institutional sector) or show bias for papers deemed "relevant to technology." As Gingras (2014) argued, an indicator should "correspond to the object (or concept) being evaluated." Therefore, alternative feature values of papers should be identified to characterize papers regarding technological relevance.

In this study, I attempted to utilize reference information (backward citations). References contain rich information regarding background knowledge on which papers are based. Therefore, they have been used to grasp the characteristics of papers, such as interdisciplinarity (Rafols & Meyer, 2009) and novelty (Uzzi et al., 2013; Wang, Veugelers,

& Stephan, 2017). Consequently, I proposed two indicators based on the papers' references and compare them to show which is preferred as an indicator of technological relevance.

Method and Data

Definitions of indicators

In this study, a simple count of reference papers that were regarded as technologically relevant were adopted for the calculation of indicator values of the papers to be assessed (hereafter "focal papers"). This was based on the idea that the closer the papers' content was to technology, the more technologically relevant knowledge was needed as the papers' background. I focused on the following two features of papers as indicators of technological relevance.

(1) Reference papers cited in the patent by the period of observation (R-PCPs)
If it is assumed that papers cited in patents (PCPs) have singular characteristic as the border between science and technology, then candidates of PCPs, which have the same nature as PCPs, might frequently need them as background knowledge for their research. The number of R-PCPs in focal papers (NR-PCP), an indicator based on R-PCP, was obtained, as shown in Figure 1. Three reference papers obtained citations from patents by the publication year of the focal paper; therefore, the NR-PCP value was 3. Although the period for measuring PPCs obtained by R-PCP could be freely set (as it was a "time-dependent" indicator), it was set to the publication year of the focal papers to determine whether NR-PCP could be applied to new papers.

(2) Reference papers written by the firms' authors (R-FPs).
Papers written by researchers in the firm sector can be assumed to be close to technology in nature, as firms are the core sector of technological development. Therefore, it can be assumed that PCP candidates frequently need the firms' research as part of their background knowledge. The number of R-FPs (NR-FP), an indicator based on R-FP, was obtained, as shown in Figure 2. Two reference papers were (co-)authored by the firms' authors; therefore, the NR-FP value was 2. NR-FP was stable in the time sequence (as it was a "time-independent" indicator).

The result of applying the indicators to actual data was assessed by the rates of focal papers being cited in subsequent patents, as shown in Figures 2 and 3.

Data

Each record of the Web of Science (WoS) that contained papers published from 1981 to 2015 was used as paper data. Papers published in 2001 were assessed to secure an adequate period for observation of both the reference papers and citations from the subsequent patents. References in focal papers were observed for the period between 1981 and 2001 (publication year of focal papers). The citations obtained from the subsequent patents were counted for the period mentioned in the previous subsection. Document type was limited to "article" for both the focal and reference papers. All papers were classified into 22 categories of Essential Science Indicators based on their citations. Papers that contained no references or author affiliation were excluded from the focal papers. In total, 716,584 papers were used for analysis.

For patent data, the 2016 spring edition of Patstat was used. All non-patent literatures appearing in the Patstat were matched to each paper in the WoS. Therefore, the data contained paper citations from patents published until early 2016. The type of intellectual property rights of patents was limited to "patent of invention" (PI) to unify the statistical nature of PPC. The identification of three sectors (firm, university, and public institute) was indispensable for deriving NR-FP indicators and analyzing their effects. It was executed based on a data table to classify the world's organizations that the author and colleagues developed, and organizations not covered in the table were classified based on the keywords shown in Table 1. According to the classification process, of the papers published between 1981 and 2001, 890,451, 9,041035, and 1,654,871 papers which covered both focal and reference papers were attributed to firm, university, and public institute sectors, respectively (including duplication caused by co-authorship across different sectors).

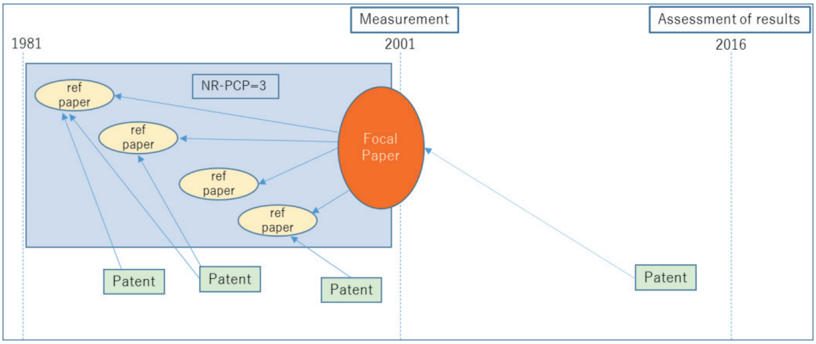


Figure 1. Scheme for measuring NR-PCP

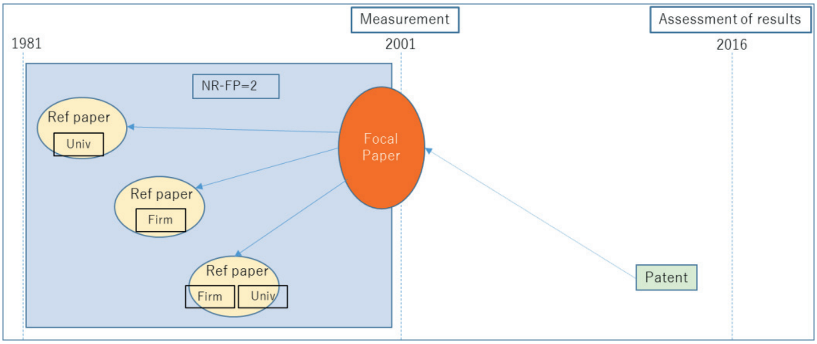


Figure 2. Scheme for measuring NR-FP

Table 1. Keywords for identifying papers in the three sectors

Sector	Keyword or description
Firm (suffix)	INC, LTD, CORP, CO, GMBH, AG, GRP, SA, AB, SPA, PLC, AS, MBH, BV, KG, KK, PC, LLC, NV, OY, LLP, SRL, R&D, KGAA, SRO, LTDA, CONSULTANTS, CONSULTING, CONSULTANCY, CONSULTANT, CONSULT, GROUP, COMPANY, LP, LIMITED, SARL, SAPA, VOS
University	UNIVERSITY, UNIV, UNIVERSITE, UNIVERSITAT, ECOL, ECOLE, ECOLES, MED SCH, COLL, INST TECHNOL, CHU, CHRU, TH, ETH, FAC, GRAD SCH, POLYTECH, POLYTECN
Public Institute	NATL, ACAD, FED, NACL, NATL, MINIST, EUROPEAN, GOVT, BUNDES, CITY, MUNICIPAL, PUBL, PREFECTURAL
Other	Organizations not classified in the three sectors above, such as hospitals, non-profits and unknown.

Comparison of NR-PCP and NR-FP

Threshold values of papers and obtained results

The precision of results increased along with the increase of threshold values of both NR-PCP and NR-FP, as shown in Figure 3. Although the two indicators did not necessarily indicate identical entities, their precision curves were similar (NR-PCP was slightly superior to NR-FP in all ranges of threshold values below 20). If the papers containing at least one R-PCP (or R-FP) were deemed technologically relevant (threshold=1), then NR-PCP and NR-FP marked precision of 0.220 and 0.196, respectively, which were much higher than the rate of PCPs in all papers (background rate) of 0.126. The precision of results from NR-PCP and NR-FP indicators reached 0.521 and 0.509, respectively, when threshold values were set to 20.

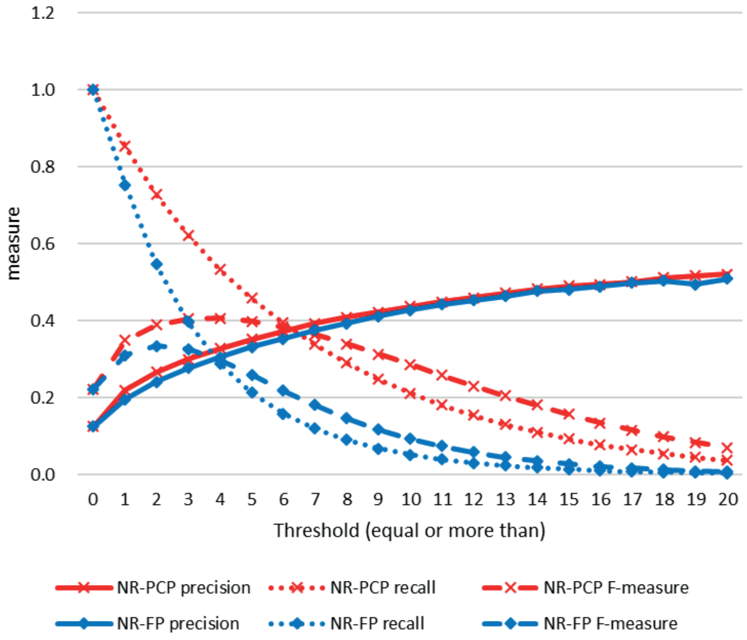


Figure 3. Precision, recall, and F-measure of results by threshold value

As for recall, NR-PCP showed values that were much higher than NR-FP in all ranges of threshold values, as shown in Figure 3. Recall of NR-FP rapidly decreased with an increase in threshold of NR-FP. Relatively low values of recall might be partially caused by the coverage of firms' papers.

Generally, NR-PCP received better results than NR-FP, since F-measures of NR-PCP were much higher than those of NR-FP. The maximum value of the F-measure for NR-PCP was 0.406 (threshold=4; precision=0.328, and recall=0.533), while that of NR-FP was 0.335 (threshold=2; precision=0.242, recall=0.546).

The Spearman's rank correlation coefficient between NR-PCP and NR-FP was 0.525. The inclusion index between the predicted papers was 0.70, and that of the correct answer between them was 0.91 at threshold=1. These inclusion indexes decreased as threshold values increased; nevertheless, inclusion indexes measured at threshold=20 remained at 0.48 and 0.59, respectively. Therefore, both indicators modestly correlated with each other.

Although the two indicators showed effect of prediction on the future PCPs, we should know which of the indicators (NR-PCP and NR-FP) and whole references would be essential factors to predict papers becoming PCP in the future. Therefore, logistic regression analyses which included the indicators and the residual part of references, NR-PNCP (number of referenced papers not cited in patents until 2001), and NR-NFP (number of referenced papers that include no firm's researcher) as independent variables, were executed.

Both NR-PCP (Table 2) and NR-FP (Table 3) showed a positive correlation to the probability of the papers being PCPs by 2016 in all the conditions presented in the tables. NR-PNCP and NR-NFP showed small coefficients; however, the means of NR-PNCP (13.45) and NR-NFP (14.31) per paper were much larger than those of NR-PCP (2.22) and NR-FP (1.36). Therefore, their influences per paper were roughly the same as the indicators.

Although it should suggest that NR-PCP and NR-FP contained essential factors for predicting focal papers being PPCs in the future, the number of residual references showed the opposite tendency each other. While papers tended not to be cited by patents as NR-PNCP increased, NR-NFP showed positive correlation to focal papers being future PPCs. However, in Models 2-b and 2-e, which contained impact factor as one of the independent variables, their coefficients were close to zero (0.0005 and 0.0010, respectively). This suggests that NR-NFP partially contained factors that were positively correlated to the probability of the focal papers becoming PCPs and could be explained by impact factor.

Table 2. Logistic regressions of probability of being PCP by 2016 (NR-PCP)

Variable	Model 1-a (PCP)	Model 1-b (PCP)	Model 1-c (PCP)	Model 1-d (PCP)	Model 1-e (PCP)
NR-PCP	0.1895***	0.1636***	0.1831***	0.1603***	0.1412***
	(0.0008)	(0.0009)	(0.0010)	(0.0009)	(0.0011)
NR-PNCP	-0.0174***	-0.0262***	-0.0158***	-0.0155***	-0.0214***
	(0.0004)	(0.0004)	(0.0005)	(0.0004)	(0.0006)
(intercept)	-2.3005***	-2.3215***	-2.2491***	-2.2592***	-2.3414***
	(0.0063)	(0.0069)	(0.0185)	(0.0106)	(0.0237)
Impact factor	No	Yes	No	No	Yes
Sector	No	No	Yes	No	Yes
Sci. field	No	No	No	Yes	Yes
Observations	716,584	620,018	505,405	579,989	376,007
Pseudo R^2	0.112	0.122	0.119	0.107	0.127

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 1 ' '

Note: In models that included the above variables, cases were restricted to those which matched following conditions:

Impact factor: Papers in which journals had available impact factors in 2001.

Sector: Papers in which all authors were affiliated with either a university, public institute, or firm.

Scientific field: Scientific fields in which NR-PCP were more than a thousand. Multidisciplinary field was excluded.

Table 3. Logistic regression of probability of being PCP by 2016 (NR-FP)

Variable	Model 2-a (PCP)	Model 2-b (PCP)	Model 2-c (PCP)	Model 2-d (PCP)	Model 2-e (PCP)
NR-FP	0.2041^{***}	0.1848^{***}	0.1759^{***}	0.1680^{***}	0.1361|***
	(0.0013)	(0.0014)	(0.0016)	(0.0014)	(0.0018)
NR-NFP	0.0126^{***}	0.0005^{***}	0.0143^{***}	0.0080^{***}	0.0010^{***}
	(0.0003)	(0.0003)	(0.0004)	(0.0003)	(0.0005)
(intercept)	-2.5000^{***}	-2.5648^{***}	-2.4212^{***}	-2.3321^{***}	-2.4231^{***}
	(0.0061)	(0.0068)	(0.0182)	(0.0102)	(0.0232)
Impact Factor	No	Yes	No	No	Yes
Sector	No	No	Yes	No	Yes
Sci. field	No	No	No	Yes	Yes
Observations	716,584	620,018	505,405	579,989	376,007
Pseudo R^2	0.061	0.084	0.066	0.073	0.096

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 1 ' '

Note: condition of observations were the same as Table 2.

Tendencies by scientific field

Whether the two indicators, NR-PCP and NR-FP, work reliably in any scientific field is essential for their actual use. Figure 4 showed the relative precision (precision divided by its background rate) and background rate of each scientific field. All cases of both NR-PCP and NR-FP, except for NR-FP in space science, showed a higher precision than their background rates, although it did not seem to be preferable to apply the indicators to scientific fields in which papers were rarely cited in patents. Recalls of two indicators showed similar tendencies (Figure 5). Scientific fields in which papers were rarely cited in patents showed relatively low recalls for both indicators. In total, NR-PCP and NR-FP indicators covered more than half of the PCPs in 2016 for most scientific fields (NR-PCP and NR-FP covered half of the PCPs in 2016 for 18 and 20 out of 22 fields, respectively); therefore, they should work effectively in most scientific fields.

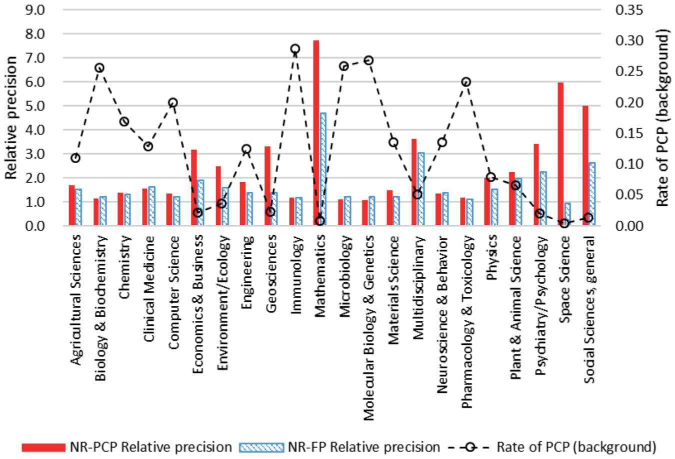


Figure 4. Relative precision and rate of PCP by scientific field (threshold=1)

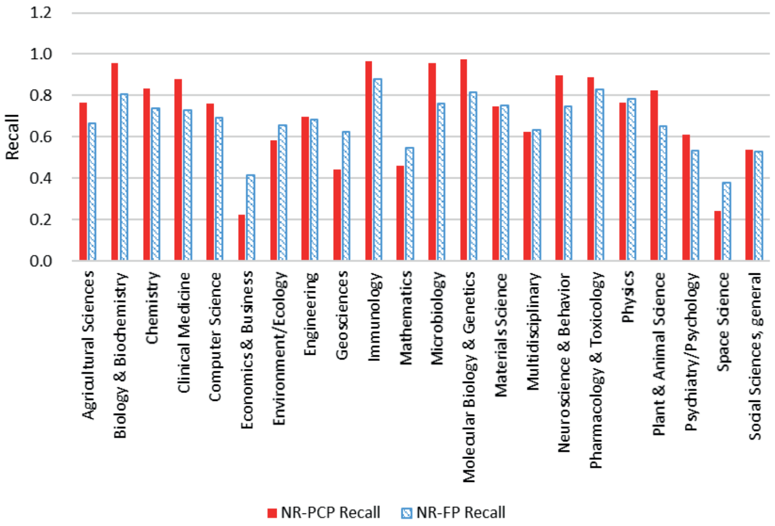


Figure 5. Recall by scientific field (threshold=1)

Tendencies by authors' sector of focal papers

To clarify whether NR-PCP and NR-FP could be applied to any sector, I showed trends in three major sectors: firms, universities, and public institutes (Figure 6). To eliminate the influence of other sectors, papers written by only focal sectors were counted.

All cases except precision of NR-FP of public institutes showed similar tendencies. For public institutes, the number of focal papers in which NR-FP was more than 14 was only 27. Therefore, the decrease of precision of NR-FP was caused by the small number of identified papers. Generally, the indicators showed robust trends in small values of thresholds for all three sectors.

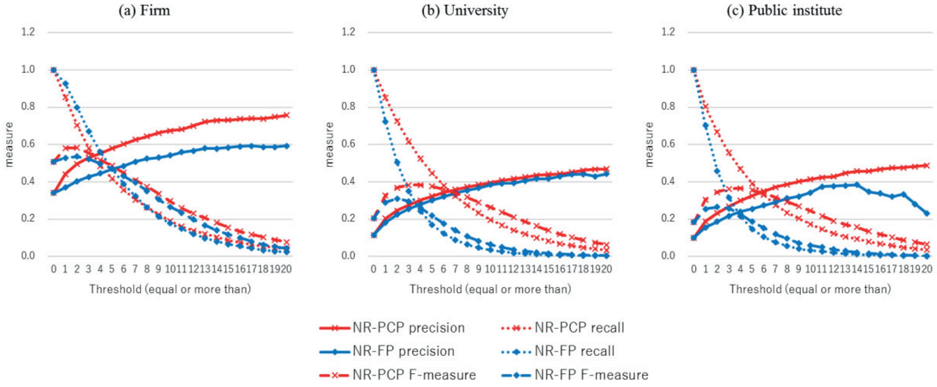


Figure 6. Precision, recall, and F-measure of NR-PCP and NR-FP in the three sectors

NR-PCP and NR-FP showed much better results in the firm sector than the other two sectors; their precision rates reached 0.758 and 0.593 at threshold=20, and their maximum F-measures were 0.582 and 0.537 at threshold=2, respectively. Therefore, both indicators worked effectively for firms' research.

Since NR-FP deemed firms' papers as technologically relevant, it should cover most of the firms' PCPs; otherwise, the firms' papers themselves should be directly used as an indicator. The firms' high recall value (92.6%) for NR-FP at threshold=1 proved the efficiency of this indirect application of firms' papers as an indicator. Even when including co-authored papers with other sectors, the recall value of firms' NR-FP was as high as 90.0% at threshold=1 (not presented).

Influence of newness of references on precision and recall

NR-PCP depends on forward citations of reference papers. It suggests that papers that consist of relatively new references do not tend to contain R-PCP, since all references have a minimal chance of obtaining citations from patents by the publication year of focal papers (i.e., 2001).

Figure 7 showed the extent to which the oldest publication years of reference papers influence precision and recall for predicting future PCPs. While both NR-PCP and NR-FP maintained much higher precision values than background rates for the oldest reference years, recall decreased rapidly in the 1990s, especially for PR-PCP, and PR-FP outperformed PR-PCP after 1996. However, NR-FP also showed a gradual decrease in recall, despite its independence to observation time. This might be caused by the number of references the focal papers contained; the mean number of references of focal papers that contained references in or older than 1998 was 16.1, while that which only contained references from 1997 or before was only 2.75.

Moreover, it should be noted that the rate of the focal papers that only cited papers published between 1997 and 2001 was relatively small (4.6%); and their rate of PCP was 9.1%, which is much lower than that of all focal papers (12.6%). Therefore, the influence of the newness of references seemed to be relatively limited.

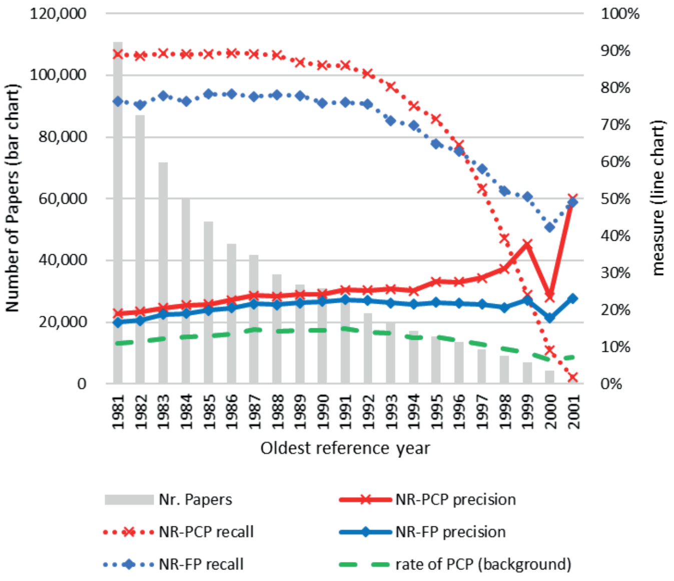


Figure 7. Precision and recall of results by oldest reference year of focal papers

For NR-PCP, securing a few years to observe the PPC of reference papers should improve results. To grasp how the three measures (precision, recall, and F-measure) of the results improved by adding a short observation period, the results measured in the publication year of the focal paper (2001) and two years later (2003) were compared to each other (Figure 8).

Precision was almost identical, while recall improved visibly. The difference in recall increased as threshold increased and reached a peak value of 9.6 points at threshold=7. Maximum F-measure improved slightly, from 0.406 (threshold=4) to 0.420 (threshold=5). The results suggested that an observation period after the publication year of focal papers ensured higher recall.

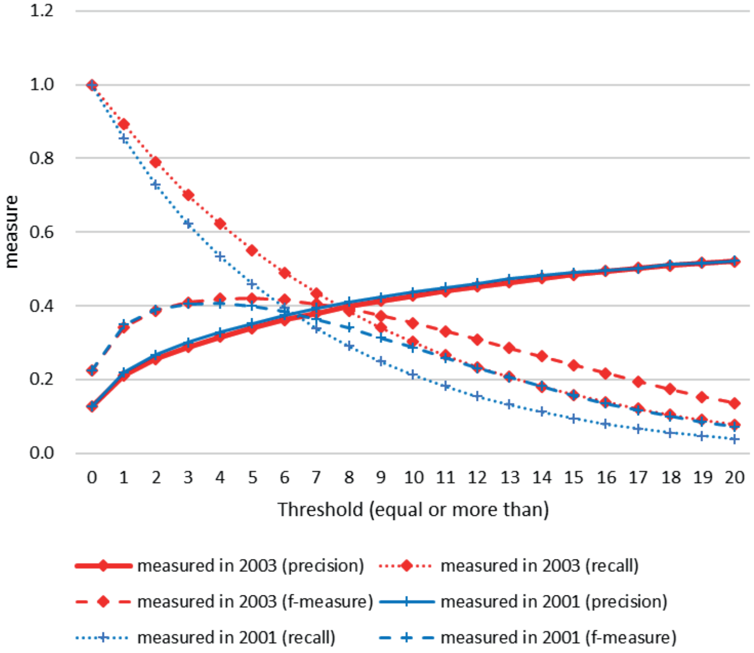


Figure 8. Comparison of results of NR-PCP measured in 2003 and 2001

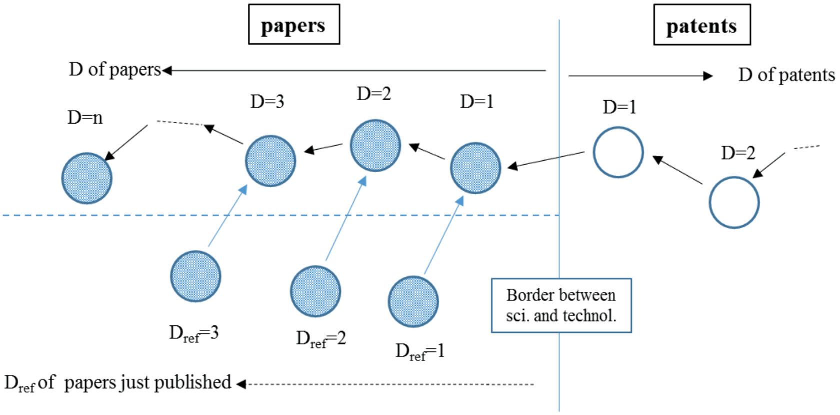
Scrutiny of the results of the comparison

Though NR-PCP and NR-FP generally showed similar tendencies, NR-PCP seemed to be preferred as an indicator in terms of stability and operability. NR-FP needs an appropriate definition and identification of firms. The author checked its robustness under different settings (without an organization classification table or the inclusion of non-profit organizations); however, its results did not surpass those of NR-PCP (not presented). It should be noted that combinatorial use of both NR-PCP and NR-FP (not presented) did not yielded any result of which f-measure exceeded the maximum f-measure obtained by sole use of NR-PCP (i. e., 0.406).

On the other hand, NR-PCP needed only accurate identification of citation linkages, which required no knowledge outside the databases. The time-dependent nature was a shortcoming of NR-PCP; however, its influence may visibly lessen by ensuring a few years of citation measurement.

Consideration of papers' nature in the aspect of D-metric of their references

NR-PCP, which was based on existing PPC linkages, provided some insights into the relationship between science and technology in a citation network of patent-patent, patent-paper, and paper-paper citations. Here, I tried to clarify its conceptual position in the citation network, extending the D-metric defined by Ahmadpoor & Jones (2017). Figure 9 showed the structure of the extended Ahmadpoor's network. The distances 'D' from the border between science (papers) and technology (patents) were defined based on hierarchies using citation linkages in the network. Papers that were attributed to Ahmadpoor's D-metric were illustrated in the upper left part of Figure 9. Ahmadpoor & Jones defined D=1 papers as those cited by patents (i.e., PCP), D=2 papers as those cited by papers of D=1, but not by patents, D=3 papers were cited by those of D=2, but by neither those of D=1 nor patents, and so on.



Note: The author created this figure based on Ahmadpoor & Jones (2017)

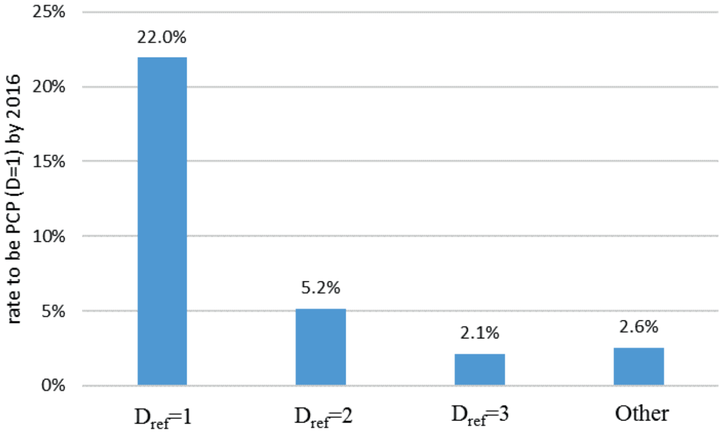
Figure 9. Extended Ahmadpoor's citation network

Papers referencing R-PCPs (PR-PCPs) in this study rarely obtained citations from papers or patents because of their newness; therefore, they did not have any position in Ahmadpoor's original network at the period of their publications. However, their probability of being D=1 papers in the future was higher than papers published in same year which did not cite PCPs, so their potential distance from technology was relatively closer to D=1. Here, I defined another type of distance from the border, D_{ref}, for stratifying recently published papers in Ahmadpoor's citation network. In the case that recently published papers cited D=1 papers, their distance from the border of science and technology was defined as D_{ref}=1. Papers of D_{ref}= n (n>1) were defined as those cited papers of D=n but did not cite any paper of D_m ($1 \leq m < n$), as well.

By attributing D_{ref}-metric to recently published papers, a part of the role of reference papers in the citation network was revealed. Figure 10 shows how the rates of papers published in 2001 to be PCP (D=1) by 2016 were affected by their D_{ref}-metric. It showed that the deeper the D_{ref}-metrics of papers, the less likely they would be classified as D=1 in the future, although the rate of 'other' papers (D_{ref} >3, and the papers themselves or their successors not cited by patents) of being PCP by 2016 was slightly greater than that of D_{ref}=3.

It should be noted that the results contained some truncation noise caused by the limited observation period (21 years). If a paper cited another paper of D=1 published before 1981, its distance from the border was not measured as D_{ref}=1. Therefore, the results shown in this section are rough estimates.

The results suggested that papers could be stratified according to their reference papers' D-metric (i.e., D_{ref}-metric), and that R-PCP was a better predictor of future PCP than papers of D>1. The results also suggest that both forward and backward citations should be considered to understand the mechanism in which PPCs occur, since PPCs typically occurred near existing PCPs (D=1).



Note: The reference papers' D-metric used for the calculation of D_{ref}-metric were measured in the period between 1981 and 2001.

Figure 10. Rate of papers published in 2001 of being PCP by 2016 according to their D_{ref}-metric (Only D_{ref}-metrics 1 to 3 were described)

Discussion and Conclusion

This study attempted to develop two reference-based indicators (NR-PCP and NR-FP) of technological relevance. A comparison of both indicators' behaviors showed NR-PCP's relative advantage in steadily obtaining better results in many cases. To understand the rationale of using R-PCP as a predictor of events that lead focal papers to become PCP in the future, the focal papers' rate of being PCP was analyzed by their potential distance from the border between science and technology (D_{ref}-metric), which was obtained from an extended Ahmadpoor's citation network. However, many issues remained unaddressed.

(1) In this research, each reference paper was regarded as a unit of knowledge, and only a simply counted number of reference papers were adopted as indicators. However, the number of PPCs obtained by reference papers were discarded to simplify the scheme. To improve the performance obtained by NR-PCP, more sophisticated indicators that may consider the number of PPCs that reference papers obtained should be developed.

(2) In addition, the likelihood of whether a specific type of paper was often missed by the indicators should be investigated. One of the possible biases was the presence or absence of the authors' and inventors' self-citations. As Tijssen et al. (2000) inferred for Dutch patents and papers, many self-citation linkages should be in my data. My finding that PPCs tended to occur in the neighborhood of existing PCPs might also imply an influence of self-citation linkages. However, NR-PCP and NR-FP are aimed at measuring the introduction of technologically relevant knowledge to focal papers, not assuming the

extraction of only self-citation linkages. Can the indicators appropriately predict any type of future PCPs, even if there is no self-citation? The extent to which the authors and inventors shared in focal papers, their reference papers, and patents that cited either focal or reference papers should be investigated to verify the versatility of reference-based indicators as NR-PCP.

(3) For NR-PCP, I compared results measured in the publication year and two years after. The results showed that a two-year observation period secured a visible improvement of recall; however, the exact time that this should be measured to obtain reliable results remains unclear. The period of measurement should be determined by considering the imbalance between the immediacy and the validity of the results.

(4) The results were assessed only regarding the future PPC that papers would obtain. Nonetheless, it should only be one of the features that represent the technological relevance of a paper. The indicators should be assessed from various perspectives, such as the co-occurrence of keywords between focal papers and patents, presence or absence of citations from firms' papers, or assessment using correct answers (i.e., papers regarded to have promoted technological innovation).

Acknowledgments

The author used a data table to classify the world's organizations, which the author and colleagues developed with the support of both JSPS KAKENHI Grant Number 23500304 and the research funding program from JST/RISTEX titled "Science of Science, Technology and Innovation Policy."

References

Ahmadpoor, M & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, 357, 583. doi: 10.1126/science.aam9527

Fukuzawa N. & Ida T. (2016). Science linkages between scientific articles and patents for leading scientists in the life and medical sciences field: the case of Japan. *Scientometrics*, 106, pp.629-644. doi: 10.1007/s11192-015-1795-z

Gingras, Y. (2014). Criteria for evaluating indicators. In: B. Cronin & C. R. Sugimoto (Eds.). *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (pp.109-124). Cambridge, USA : MIT press.

Hicks, D., Breitzman, A., Kimberly, S. & Narin, F. (2000). Research excellence and patented innovation. *Science and Public Policy*, 27, 310-320. doi: 10.3152/147154300781781805

Narin, F. & Noma, E. (1985). Is technology becoming science? *Scientometrics*, 7, 369-381. doi: 10.1007/BF02017155

Narin, F. (1991). Globalization of research, scholarly information, and patents – Ten-year trends. The Serials Librarian, 21, 33-44. doi: 10.1300/J123v21n02_05

Rafols, I. & Meyer, M. (2009). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 84, 263-287. doi: 10.1007/s11192-009-0041-y

Tijssen, R. J. W., Buter, R. K. & Leeuwen, Th. N. (2000). Technological relevance of science: An assessment of citation linkages between patents and research papers. *Scientometrics*, 47, 389-412.

Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342, 468-472. doi: 10.1126/science.1240474

Wang, J., Veugelers, R. & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46, 1416-1436. https://doi.org/10.1016/j.respol.2017.06.006

Yamashita, Y. (2018). Exploring characteristics of patent-paper citations and development of new indicators. In M. Jibu & Y Osabe (Eds.), *Scientometrics* (pp.151-172). London: IntechOpen. doi: 10.5772/intechopen.77130

Evaluating Human Versus Machine Learning Performance in Classifying Research Abstracts

Khiam Aik Khor[1], Giovanni Ko[2], Walter Theseira[3], Xin Qing Cai[4], and Yeow Chong Goh[5]

[1] *mkakhor@ntu.edu.sg;* [4]*xqcai@ntu.edu.sg;* [5]*ycgoh@ntu.edu.sg*
Nanyang Technological University, Talent Recruitment & Career Support (TRACS) Office and Bibliometrics Analysis, 50 Nanyang Avenue, Singapore 639798 (Singapore)

[2] *giovanniko@smu.edu.sg*
Singapore Management University, School of Economics, 90 Stamford Road, Singapore 178903 (Singapore)

[3] *waltertheseira@suss.edu.sg*
Singapore University of Social Sciences, School of Business, 463 Clementi Road, Singapore 599494 (Singapore)

Abstract

Machine Learning (ML) methods are now applied to many problems in Scientometrics. Given sufficiently large training datasets, ML can efficiently complete natural language processing tasks such as classifying research abstracts and outputs, which otherwise require extensive manpower. But what are the relative strengths and limitations of ML methods versus human research assistance when training data is limited? Our study compares the performance of 63 student research assistants to that of an ML model. The task is classifying a research grant abstract into one of nineteen scientific funding areas in physical and life sciences defined by the European Research Council. We find that ML models, even trained on relatively small datasets, outperform the average human research assistant. While some research assistants perform at levels just below that of the ML models, the research assistants display lower inter-rater reliability. Crucially, human classification performance and reliability appears fixed over moderate levels of training and task exposure, suggesting that selecting research assistants based on pre-existing ability could be superior to relying on task-specific training. These results suggest ML classification may be superior to human research assistance for natural language processing tasks even when training datasets are limited.

Introduction

Answering many Scientometric research questions requires processing vast amounts of unstructured textual data. Consider the simple problem of describing how research output from a particular country is distributed across scientific areas. When the data is already labelled with scientific research areas, e.g. it is known whether a paper belongs to “Biology” or “Chemistry”, this problem is trivial. But if a new field of science emerges, or if a new label is desired, e.g. “Bioinformatics”, then even a basic description of the data requires processing the entire existing body of research output. Until recently, such research classification problems were addressed by hiring numerous research assistants, or by using rules of thumb based on keywords and metadata. Machine Learning (ML) models, which are capable of learning how to classify data with high accuracy given sufficient “training” data, have the potential to efficiently solve such large-scale data processing tasks.

Accurate and efficient research classification is particularly important for comparing relative research performance across different research institutions, grant programmes or nations (King, 2004). Consider a comparison of the European Research Council’s (ERC) grant funding performance with that of the United States National Institutes of Health (NIH). The ERC has 9 funding panels in the life sciences, while the NIH distributes funding across 27 institutes and centres. Each funding panel/institute/centre specialises in a different area of life sciences research. If the typical life sciences research project funded by the NIH produces more publications or citations than that of the ERC, is that difference attributable to the greater selectivity of the NIH grant funding process, or to differences in the allocation of funding across areas of life sciences research? A fair evaluation requires understanding how research projects funded by one agency – e.g. the "LS3: Cellular and Development Biology" panel of the ERC – map into the funding structure of the other.

In Khor et al (2018) we used ML models to classify and map funded research projects, showing that controlling for scientific funding structure, researchers funded by the US National Institutes of Health outperformed ERC-Starting Grant researchers. However, we were not able to assess whether our use of ML classification affected research mapping accuracy, compared to the conventional alternative of using human research assistants to classify research. This study compares human classification to ML classification, and shows that ML outperforms human classification in accuracy and reliability even when the available training data is relatively limited.

Literature Review

Recent literature has demonstrated the viability of automated classification of research abstracts using ML methods (Yau et al., 2014; Freyman et al., 2016; Sing et al., 2017). However, the literature has so far only explored unsupervised topic discovery and clustering methods rather than the classification methods we are concerned with. Research abstract classification is an application of natural language processing (NLP) ML methods, which can be highly accurate because the language in scientific abstracts is discipline-specific. Even in interdisciplinary research, ML models can identify distinct underlying disciplines (Freyman et al., 2016) and cluster related papers accurately based on their common topics (Yau et al., 2014). However, ML models generally require a large-scale, high quality training dataset to produce high accuracy, and such an extensive training dataset is not always available in all applications.

There has been no attempt at assessing human versus ML classification performance for Scientometric research abstract classification tasks. While there are similar studies in other fields (Simundic et al., 2009; Schumacher et al., 2010) that compare ML classification methods to human classification, those studies evaluate image and object classification, rather than NLP text classification that is more important in Scientometrics. In these studies – which also feature difficult classification problems – human performance tends to exhibit poor reliability, (Schumacher et al., 2010) high uncertainty (Simundic et al., 2009), and lower accuracy than the ML methods. It remains an open question how human performance compares to ML models in the Scientometric context, particularly where training data may be limited.

Methodology

Research Questions

To evaluate human vs ML performance in research abstract classification, we designed a research classification experiment to answer the following questions:

1. How does the classification accuracy and reliability of a typical undergraduate student assistant compare to that of an ML classification model?
2. Does additional selection and training improve the classification performance of these student assistants? And if so, how does it compare with ML classification models?

Study Design

We recruited 63 student assistants for our study. The student assistants were full-time undergraduates at Nanyang Technological University, a highly selective, research intensive university in Singapore. The student assistants had not previously completed research abstract classification tasks. This allows a fair comparison between ML and human classification performance since both the student assistants and the ML model start "naïve" with respect to the classification task. While graduate students or professional researchers could be more proficient at the task, undergraduate student assistants are the most common source of manpower for research projects, especially if budgets are limited.

Student assistants were invited to a classroom for one day, where under “exam” conditions we instructed them to first study a “training” set of research abstracts labelled with their ERC panels, and then to classify another unlabelled “test” set. Each student assistant was randomly assigned one of four “training” sets, while the “test” set used was common to all student assistants. The “test” and “training” sets are stratified random samples from the ERC-StG corpus.

We then invited 32 student assistants with the highest classification accuracies to continue work in a second stage, to investigate training and selection effects. Each second stage student assistant coded two additional “test” sets. The first additional “test” set was administered without further “training” or feedback on performance. The second additional “test” set was administered after revealing the “answers” for their previous “test” sets to the students.

Data

Our corpus consists of the 2523 abstracts of the ERC Starting Grant (ERC-StG) research grants awarded from 2009 to 2016, which is a subset of the data collected for Khor et al. (2018). This corpus is appropriate as all the ERC grant programmes use a common panel classification framework that has remained unchanged since 2008, and the ERC-StG awarded the most grants among the various ERC programmes and distributes grants relatively evenly across all panels.

Machine Learning Classification

We use the support-vector machine (SVM) algorithm as it had the best performance among the ML algorithms used in Khor et al. (2018). As text classification is a natural language processing task, we utilise the bag-of-words model with text frequency-inverse document frequency as our feature importance measure to convert abstracts into matrices of feature importance for the SVM algorithm to process.

Performance Metrics

We measure classification accuracy using precision and recall (Sokolova and Lapalme, 2009). Precision is the proportion of abstract classifications that are correctly predicted, while recall is the proportion of all abstracts that are predicted correctly. Due to the precision-recall trade-off (Buckland and Gey, 1994), we employ the F1 metric – the harmonic mean of precision and recall – as our preferred summary performance measure.

To compare reliability, we use the Kappa statistics (κ), namely that of Cohen (1960) and Fleiss (1971) for pairwise and multiple-classifier reliability respectively. κ has an upper limit of 1, which represents perfect agreement. Cohen’s κ is interpreted as the proportion of potential non-random agreements between the classifiers that are observed in the data, while Fleiss’ κ has no natural interpretation. For a detailed review of interrater reliability, refer to McHugh (2012).

Results and Discussion

Overall Performance Comparison

Overall, the SVM classification algorithm performed better than human classification at classifying abstracts across all “training” sets used (see *Table 1*). Whereas the performance of the SVM classification algorithm increased with larger “training” sets, human classification performance seems to be unaffected by “training” set size. When Fleiss’ κ was calculated for each human classifier group as defined by the “training” set the students were exposed to, κ seems to be centred around 0.38, indicating “minimal” interrater reliability (McHugh 2012).

Table 1. Mean overall performance of ML and Human classification.

"Training" Set	*No. of Abstracts*	*No. of Students*	*F1 (SVM)*	*F1 (Human)*	*Fleiss' κ*
A	380	16	0.631	0.443	0.375
B	380	16	0.682	0.447	0.382
C	190	16	0.612	0.462	0.388
D	190	15	0.493	0.435	0.378

Inter-Rater Agreement

The SVM algorithm trained on the same training set and hyperparameter space trivially returns the same predictions for any given test set, since it is a *statistical* model. Hence, a meaningful measure of ML classification reliability must compare "test" set predictions from multiple SVM models trained under different conditions, principally using different "training" sets. With our available data, we pairwise compare SVM models trained on similarly-sized "training" sets. Cohen's κ for the ML models trained on "training" sets A and B is 0.560 (95% CI = [0.527, 0.592]) and for the ML models trained on sets C and D is 0.460 (95% CI = [0.428, 0.493]). Although these values indicate "weak" agreement, they indicate ML classification has greater reliability than the overall Fleiss' κ for human classification.

Classification Accuracy

We graph average human versus ML classification performance by funding panel in *Figure 1.* The SVM algorithm performs at least as well as human classification across all panels, and better than human in most panels. Human and ML performance tends to be highly correlated (0.912). This suggests the way humans and the SVM algorithm classify research abstracts are relatively similar. To our knowledge, this is not a previously-reported result. One implication is that a "hybrid" classification algorithm that combines the best of human and ML performance is unlikely to generate superior performance compared to ML alone.

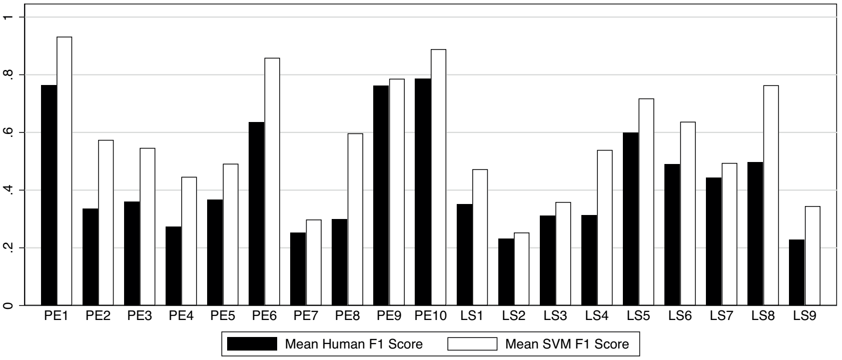


Figure 1. Comparison of Human versus SVM classification in classifying research abstracts.

Comparing the Top Human Classifiers to ML Classification

One concern with our study is that our human classifiers have not been carefully screened for their aptitude to the task. A research project that relied on human classifiers might hire assistants selectively. To compare ML performance against the best human classifiers, we analyse the top quartile of human classifiers by overall accuracy rate, corresponding to 16 human classifiers.

Figure 2 graphs classification performance between the top human classifiers and the SVM algorithm by panel. The performance gap has narrowed, but human classification performance (F1 = 0.496) is still barely comparable to the lowest exhibited performance from the SVM algorithm, in terms of accuracy. There is no evidence that humans can outperform SVM in classifying any panel. The classification performance of humans and the SVM algorithm across panels are still highly correlated (0.925). The top human classifiers are also more reliable (Fleiss' κ = 0.443), and fall within the lower bound of the confidence interval of the Cohen's κ for ML classification when using smaller "training" sets.

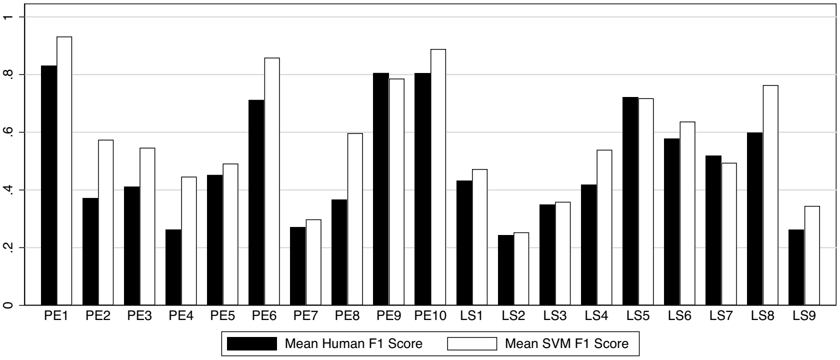


Figure 2. Comparison of top Human versus SVM classification in classifying research abstracts.

Is Human Classification Performance Persistent?

The overall performance of the top human classifiers who classified additional datasets is summarised in *Table 2*. There is no noticeable improvement in classification accuracy or reliability from the first to the second additional dataset. It appears human classification performance is not significantly affected by further "training" or further exposure to the task. Nor is it dependent on the initial "training" set. Hence, we hypothesise that human classification performance is determined largely by the prior knowledge they have of the task and the discipline classification groups rather than task "training".

Table 2. Mean overall performance of top human classification with further exposure.

"Training" Set	*F1 (Set 1)*	*F1 (Set 2)*	*Kappa (Set 1)*	*Kappa (Set 2)*
A	0.486	0.507	0.413	0.383
B	0.511	0.516	0.397	0.374
C	0.535	0.514	0.432	0.392
D	0.511	0.516	0.359	0.403

Conclusion, Further Steps and Limitations

Our results suggest ML models classify research abstracts at least as well as both *untrained* and *"experienced"* human classifiers *on average*. ML models are also more reliable, even when using *relatively small training corpora*. However, our data for calculating inter-model ML classification reliability is limited. We are running further simulations where we generate "training" sets to train our SVM algorithm on in order to generate more "test" set classifications using "new" SVM models. Also, while we have presented results for *average individual* classification performances, we are further investigating classification performances when using *aggregated* classifications for each abstract, i.e. does a "wisdom of the crowd" effect

(Surowiecki, 2004) exist for research abstract classification. Regardless, these results agree with the small human versus ML model literature (Simundic et al., 2009; Schumacher et al., 2010) that also finds automated classification performs at least as well as human classification while maintaining higher reliability.

One caveat is that our "training" set sizes are too small for the SVM algorithm (Joachims, 1998). While this is a desirable feature for our study, as research classification training datasets may be limited in size, ML performance is expected to improve even further as the training dataset grows larger. For simplicity, we assumed that each research abstract can only be classified to one panel, when some research may plausibly be classified to multiple panels. This limitation could be overcome by assigning a panel probability distribution to each research abstract. We also cannot test the hypothesis that human classification performance is largely determined by background knowledge rather than task-specific "training". A further study comparing ML classification models trained on large corpora to expert human classifiers with deep specialist knowledge – such as graduate students and professional researchers – may be appropriate.

References

Khor, K. A., Ko, G., & Theseira, W. (2018, September). Applying Machine Learning to Compare Research Grant Programs. In *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-82.

Sing, D. C., Metz, L. N., & Dudli, S. (2017). Machine learning-based classification of 38 years of spine-related literature into 100 research topics. *Spine*, *42*(11), 863-870.

Simundic, A. M., Nikolac, N., Ivankovic, V., Ferenec-Ruzic, D., Magdic, B., & Kvaternik, M. (2009). Comparison of visual vs. automated detection of lipemic, icteric and hemolyzed specimens: can we rely on a human eye?. *Clinical chemistry and laboratory medicine*, *47*(11), 1361-1365.

Schumacher, J., Zazworka, N., Shull, F., Seaman, C., & Shaw, M. (2010, September). Building empirical support for automated code smell detection. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (p. 8). ACM.

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1), 12–19.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

European Research Council (2018). Starting Grants. Retrieved 15 August 2018 from: https://erc.europa.eu/funding/starting-grants.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378.

Freyman, C. A., Byrnes, J. J., and Alexander, J. (2016). Machine-learning-based classification of research grant award records. *Research Evaluation*, 25(4), 442–450.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427 – 437.

Yau, C.-K., Porter, A., Newman, N., and Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3), 767–786.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Doubleday

King, D. A. (2004). The scientific impact of nations. *Nature*, 430:311–316.

Convergence between rejection citations and X/Y citations across patent offices

Tetsuo Wada[1]

[1] *tetsuo.wada@gakushuin.ac.jp*
Gakushuin University, Dept of Economics, 1-5-1 Mejiro, Toshima-ku, Tokyo 171-8588 (Japan)

Abstract

This paper empirically examines coincidences between X/Y patent citations added by European search reports (ESRs) and those equivalent rejection citations added by the United States Patent and Trademark Office (USPTO) to reject patent applications within the same family, based on a set of triadic application sample. We consider the release timing of European search reports and those of rejection actions by the USPTO for the same family of patent applications. We find that the frequency of rejection (X/Y-equivalent) citation coincidences between the USPTO and the European Patent Office (EPO) according to family-to-family citation criteria increased after the release of European search reports, except for applications with priority in the US. It suggests that the USPTO examiners rely on prior art information collected by the EPO. On the other hand, the results also reveal that the US examiners are less likely to add the same patent citations, when rejections are persistently repeated at the USPTO. This finding provides evidence that US examiners capture spillovers of search efforts from the EPO, whereas repeated office actions at the USPTO imply diversions of negotiation issues away from those at the EPO between examiners and the applicant. The methodology in this paper introduces the novel use of patent examiner citations to compare examiners' citing behavior across jurisdictions.

Introduction

Whereas patent citations have been widely employed as research tools, their own varieties, such as applicant citations and examiner citations, have also recently been studied (Alacer and Gittleman 2006; Criscuolo and Verspagen 2008; Hegde and Sampat 2009; Cotropia et al. 2013). One of the recent research frontiers is the distinction within examiner citations, i.e., whether or not an examiner citation is given as a basis of rejection. That is, US examiners usually indicate specific patents on which they rely as reason of rejections, in particular, for "102" novelty rejections and "103" obviousness rejections. There is no official name for this category of citations, but a recent work (Cotropia and Schwartz, 2018) calls it as "rejection citation" or "rejection patent." This data has been made available and draws attention by the recently released office action database of the USPTO (Lu et al., 2017).

On the other hand, examiners at the EPO and other offices such as the Japan Patent Office (JPO) provide citation category of X or Y. Both citation category X and Y indicate particulary relevant prior art, meaning lack of patentability of application due to the existence of prior arts, while X means "taken alone" and Y means "if combined." At the EPO, search reports are given to indicate patentability information to applicants before rejection actions, according to the EPO guidelines for examination. That is, the objective of the European search report is to discover the state of the art which is relevant for the purpose of determining whether the claimed invention is novel and involves an inventive step. Similarly, at the JPO and other many patent offices, X/Y citations are given on rejection documents when examiners find rejection reasons related with prior arts. In both jurisdictions, X/Y citations indicate clear reasons to reject on the aspects of application related with prior arts. At the EPO, applicants should amend their initial applications to obtain patentability if their initial applications are given X/Y citations in their European search reports. Although the EPO does not usually issue rejection action immediately after the release of search reports, the functions of X/Y citations in European search report is thus equivalent of "rejection citations" in the US, because the EPO examiners employ search report citations if no amendment is made after the issuance of search reports. Therefore, this paper considers US rejection citations and X/Y citations at the EPO as virtually equivalent. This paper then directly compares the two between the USPTO and the EPO at the patent family level.

Background on the gap in patent citation studies

When a patentable invention is internationally valuable, an international family of patent applications is often filed and examined in many jurisdictions. A prior art search for the same invention is conducted separately by different patent offices in each jurisdiction, except for those conducted through international agreements such as the European Patent Convention and Patent Prosecution Highways (PPHs). As a result, separate citations are added to the same family of patent applications in a sequential and/or parallel manner across patent offices. This process of generating citations is different from those in academic paper citations, where a citing document typically confirms the addition of citations to prior art/literature only once, at the time of publication. Thus, by focusing on international family-to-family patent citations across patent offices, we obtain a rare analytical environment in which to examine how a patent citation network expands through additions of new citations to a single citing family of documents, yet by more than one citing entity. In particular, because a family of international patent applications has multiple chances to generate patent citations by different patent offices, and because examiners have access to the results of other offices, we are able to analyze whether office actions across different jurisdictions depend on those of other offices in adding citations, leading to *convergence* of citations. At the same time, we can also evaluate factors that are negatively related to the coincidence of citations across patent offices, implying *divergence* of citations.

There is very limited existing knowledge on whether examiners in a patent office rely on prior citations generated by other offices (Cotropia et al., 2013). When the same patent family is examined sequentially at different patent offices, how can we observe spillovers between patent examination results? This paper proposes to track examiner citations across patent offices. Simply put, we first summarize coincidences of X/Y citations at the aggregate level. Then, we study those patent family citations added by the USPTO for the purpose of rejection (i.e., rejection patent citations, or X/Y equivalents), and examine whether each US rejection citation coincides with X/Y cited families given by European search reports for the same citing patent family, with special attention to the timing of European search reports and office actions by the USPTO. By way of comparing USPTO rejection citations (at the level of international family-to-family citations) with prior art search results conducted by the EPO before and after the release of the European search reports, we can first make a reasonable inference concerning the existence of search result spillovers from the EPO to the USPTO, leading to convergent citations.

Background on X/Y citations and the US equivalents

For analyses, examiner citations for refusals, which are assigned the special categories of "X" and "Y" at the European Patent Office (EPO) and at other offices such as the Japan Patent Office (JPO), have a crucial role. Examiners must indicate specific reasons to refuse patent claims, such as a lack of novelty. Statutory requirements for novelty and inventive step for granting patents are very similar among advanced economies (though there are differences in subject matter requirements in limited areas of technologies). Therefore, the same invention with the same claims and specifications should theoretically invoke very similar grant/refusal responses from patent offices. However, it is known that the decisions regarding patent grants and refusals by the trilateral patent offices (the EPO, the JPO, and the USPTO) often differ with regard to the same content of triadic patent applications, as was found by the Melbourne school (Jensen et al., 2005; Webster et al., 2007; Webster et al., 2014). Behind the discrepancies of grant decisions, there are remarkable discrepancies between patent citations added to the same family of applications by the offices, despite the application of common patentability criteria of novelty and inventive step to generate citations (Wada, 2018). X/Y citations offer a clue for evaluating the mechanism by which examiners at the trilateral offices reach different decisions over the same case. Until recently, direct comparisons between citations for refusals have been challenging, because the US patent system does not provide citation category (such as "X" and "Y") information. However, we have developed a database for equivalent information, which is similar to the recently released office action database of the USPTO (Lu et al., 2017), but with better coverage of international citations (Wada, 2018). By utilizing the database and prosecution history information, we can shed light on how the progress of a prosecution affects the direction of citation generation.

Objectives

This paper established three specific research questions: (1) Is a USPTO office action after the release of European search reports more likely to reflect the same family-level patent citations added by the EPO, compared with pre-ESR (prior to European search reports) office actions at the USPTO? Put differently, do we observe a convergence of patent citations between the two offices? (We focus on unidirectional spillovers from the EPO to the USPTO owing to data limitations). (2) Even if we observe convergence, is it uniform or dependent on pendency length of US prosecution history? (3) Do we find any difference in the level of convergence in relation to the geographical source of a patent application?

Data source

As indicated above, this study takes advantage of a novel large-scale dataset of US rejection patents obtained from rejection documents (the "CTNF" and "CTFR" documents that denote non-final and final rejections, respectively) available as file wrappers on the "Public PAIR" database of the USPTO to compare patent citations employed by examiners as specific reasons for rejections (Wada, 2017; 2018). In other words, by way of approximating citation categories of X/Y for the USPTO, we are now able to measure convergence and divergence of individual rejection reasons used by the two patent offices through family-to-family citations. This idea of international citation comparison was first proposed at the ISSI Wuhan conference by Wada (2017). A similar US X/Y-equivalent database was developed independently by Jeff Kuhn (Kuhn et al., 2017), and yet another comprehensive database has now been made available by the USPTO (Lu et al., 2017), which helped to disseminate the idea of X/Y-equivalent citations. However, neither of these focuses on international citations. International comparisons at the level of family-to-family citations remain unfeasible with those databases. We implemented optical character recognition and natural language processing to extract X/Y equivalents, including international citations, from the file wrapper data at the University of California, Berkeley, by Guan-Cheng Li. We are thus able to evaluate whether the availability of European search reports, which cover most EP citations, influences the prosecution reasons applied by the USPTO through an X/Y equivalent in the US. Combined with the US rejection patent database, the EPO PATSTAT database (Spring 2016) and DOCDB (Backfile 2017 January version) have been used.

The domain of statistical analysis is the set of X/Y citations and equivalents for triadic applications through the PCT (Patent Cooperation Treaty) and non-PCT applications. Triadic patent applications are defined here as EPO DOCDB families that contain all of the EPO, USPTO, and JPO applications recorded on EPO's PATSTAT database. Triadic applications are intended to represent patent applications with high economic value. Although the Chinese and Korean shares of international families have recently increased rapidly, triadic applications offer reasonable criteria for applications in the early 21st century. The citations concern an EPO DOCDB family where only a single DOCDB family ID is observed and where X/Y citations (and equivalents) are added by all of the trilateral offices, representing "twin applications approach." The domain of the study is comprised of 301,186 family citations recorded as X/Y equivalents at the USPTO, found for 43,207 cited triadic families that have single DOCDB family IDs and priority years 2003–2010. Note that we have oversampled applications from Japan (see Table 1 for the composition).

Several caveats should be mentioned in relation to the data. First, all citation data are patent citations, because of the availability of DOCDB family-to-family citations. Thus, the accuracy of international families depends entirely on the DOCDB family table on PATSTAT. Moreover, PATSAT, our primary data source, records non-patent literature in non-standardized formats, so we could not consolidate the same non-patent literature across different records. For this reason, we have only employed patent citations at this time. This is a weakness, although most observed examiner rejection citations concern patents only.

Second, only the dates of search reports at the EPO, not those of post-report examinations, are reliably available at present, even on the EPO DOCDB database, which is the mother database for PATSTAT. In addition, the EPO citation data during the examination phase are incomplete on DOCDB and PATSTAT. Therefore, we do not utilize examination timing information at the EPO. Search report dates for this study are combined, compared, and checked with the EPO DOCDB backfile and are confirmed to be accurate.

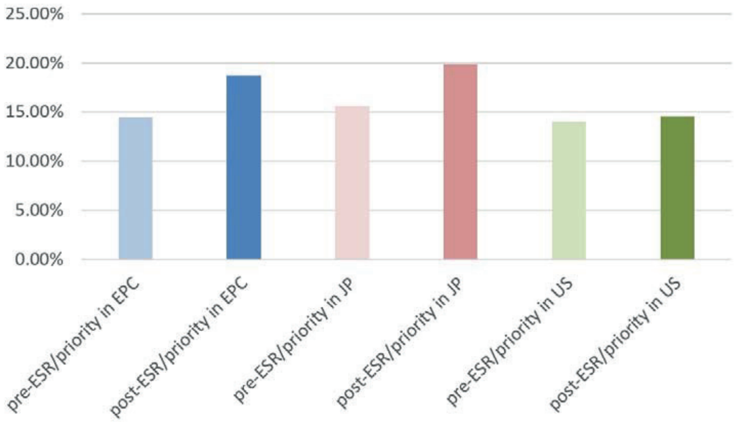
Third, triadic patent applications are defined here as DOCDB families that contain only one recorded DOCDB citing family. Therefore, any divisional or continuation applications that produce more than one DOCDB family ID exclude the family from the sample. As the economic value of a patent application depends on the probability of the use of divisional or continuation applications, this constitutes a bias in the sample selection.

Measurement and simple comparison

A simple aggregate measurement for a particular citing family is the ratio of coincidence of X/Y citations by the European search reports over rejection (i.e., X/Y-equivalent) citations added by the USPTO to the same citing family. To obtain the "EPO–USPTO X/Y family-to-family citation coincidence ratio" for an application, we first list X/Y-equivalent US rejection citations added to the application in chronological order during its prosecution history. The number of repeated rejections in its prosecution history at the USPTO is also recorded. The same patent citation is often used repeatedly in the same prosecution history at the USPTO, so there can be multiple records of the same citation pair with different US office action timing. We take the number of all these X/Y-equivalent US examiner citations for the application as the denominator of the ratio for the family of the application. For each US citation, we obtain a citation mapped onto a DOCDB citation pair from the PATSTAT data. We obtain a dichotomy on whether a citation is coded as X/Y category at the EPO within the same family-to-family citation. When the European search report records citation as X/Y, we define the citation as a coincidental X/Y EPO citation pair with the USPTO X/Y equivalent. Then, we take the number of all coincidental X/Y EPO search report citations for the citing family as the numerator of the ratio.

The ratio equals one if all of the X/Y-equivalent citations at the USPTO are also coded as X/Y at the EPO in the same family. The ratio is zero if none of the X/Y equivalents at the USPTO for an application are recognized as X/Y by the European search report. In summary, this measurement indicates the proximity of a set of rejection citations employed by the USPTO to those X/Y citations indicated by the EPO, in terms of a single family. Figure 1 shows the averages of the citation coincidence ratio over different sets of the sample, comprised of triadic applications with priority in the European Patent Convention (EPC) countries, those with priority in Japan, and those with priority in the US (i.e., geographical sources of applications from each of the trilateral offices). Each of the ratio is calculated according to two stages of US citation timing: pre-ESR and post-ESR. As is evident from the figure, the ratio increases after the release of European search reports, although the effect is not very obvious for applications from the US. A simple interpretation of this would be that the US examiners take advantage of the outcome of European search reports, especially if an application is first made outside the US. However, because this ratio is a simple and aggregate comparison irrespective of pendency length, we need to analyze at more micro-analytic level. With regard to the three basic questions stated above, the answer for the first one seems to be positive, although those of the second and the third are not clear. In particular, applications from the US may have longer pendency on average, which may affect the result.

Figure 1. Average EPO–USPTO X/Y family-to-family citation coincidence ratios



Logit analysis

Methodology: To obtain more micro-level insights, we next focus on the dichotomy describing whether or not a US X/Y-equivalent is coded as belonging to the X/Y category at the EPO as well. By taking this dichotomy as a dependent variable in logit regression, we can analyze correlating factors and their signs. The unit of analysis is a family-to-family citation given at the USPTO as an X/Y equivalent, with office action sequence data and other application-level attributes as explanatory variables. Specifically, let us define y_i as a dichotomy taking a value of one when a family of rejection citation by the USPTO examiners to a triadic application family i coincides with a family of X/Y citation added by the EPO search report. Then, the following model can be estimated assuming that the function F() is a logistic cumulative distribution function. Vectors of explanatory variables are represented by X_j and β is a coefficient vector such as:

$$\Pr(y_i = 1) = F(X_j \beta)$$

Explanatory variables X_j and main predictions: We focus on two key explanatory variables to analyze convergence and divergence of X/Y citations. One variable is another dichotomy, *US_action_after_ESR_date*. It takes a value of one when US rejection citation was given at the USPTO after the release of the European search report for its EP family member application. Because of a "search result spillovers" effect, we predicted that the coefficient would be positive. Along with this "before ESR" and "post ESR" distinction, we also employ another variable of the number of rejection actions at the USPTO. This measures the total number of US rejections for a particular rejection citation within a prosecution history. Spillovers from the EPO search report to the USPTO, if any, should occur only once in a prosecution, since typically one European search report is issued for an application. In contrast, rejection reasons could drift through exchanges of actions (e.g., amendments as responses to past rejections), especially at the USPTO. To incorporate these processes, we employ another key explanatory variable, *US_REJECTION_counts*, which is the number of rejections (non-final and final) in a USPTO prosecution history. Because longer exchanges of rejections and responses mean evolution of bargaining issues in a prosecution, we expect the coefficient for this variable to be negative.

Controls: We employ number of control variables. When an application in a sample is a PCT application and its International Search Authority (ISA) is the EPO, we give a value of one for a dummy variable *ISA_EP*, which means that a family has the EPO as its ISA. The PCT requires that a PCT application should be given an international search report prepared by a patent office. Approximately half of PCT applications from the US chooses the EPO as their ISA, whereas most of PCT applications from Japan relies on the Japan Patent Office for their ISA, and European applicants are required by their rule to ask for search reports from the EPO only. In any case, European search reports are issued for all triadic applications, but their issuance timing tends to become late when an international search report is already issued by another (non-EPO) ISA, and subsequently European search report is issued as supplementary search report. In order to control for the timing difference of search reports, this dummy variable *ISA_EP* is added.

The location dummies for the first priority country, *first_EP*, *first_US*, and *first_JP*, are employed in the full sample estimation. We also controlled for priority years (2003–2010) and the WIPO 35 technology fields of each family. The variable *techn_field_nr_counts* is the number of WIPO technology fields covered by the family, representing the breadth of the technology.

We run logit estimations on the full sample as well as on a sub-sample of applications from the EPC countries, Japan, and the US.

Results: As the first row of the first column (full sample), the second column (applications from Europe), and third columns (applications from Japan) of logit estimation results in Table 1 show, EPO–USPTO X/Y family-to-family citation coincidences are generally more likely to occur after a release of an European search report. That is, we observe positive and significant coefficients for the explanatory variable, *US_action_after_ESR_date,* indicating the convergence of US rejection citations to EPO X/Y citations after the release of European search reports. However, according to the fourth column (applications from the US), the coefficient for applications from the US is not significant. This is in contrast with the result for European and Japanese application sample, where the coefficients are positive, and much larger especially for the EP sample. US examiners may not find the outcome of European search reports for US-based applications as valuable as those for applications outside the US, possibly because examiners are more knowledgeable about prior art concerning local applications. However, we need to interpret the result with caution for the US sample, because, as described below, the distinction between whether or not European search reports are supplementary reports may also affect the result (only US sample can be affected by construction).

The coefficients for *US_REJECTION_counts* are consistent throughout the results and are negative and significant. As predicted, US examiners employ different rejection reasons from those used by the EPO on average, as prosecution takes longer. Therefore, we observe that the longer pendency results in divergence of US rejection citations from EP X/Y citations. The coefficient for the dummy "EPO as an ISA" is positive and significant only for US sample. It may suggest that US examiners find European search reports useful only when the search reports are first reports, not supplementary reports. However, we do not control the number of citations in a family of applications and the ratio of EP-added or US-added citations. If the number of citations are widely different between first reports and supplementary reports, it may affect the result. We need to scrutinize the contributing factor of ISAs further. Meanwhile, none of the region dummies of first priority are significant.

Table 1. Logit regression, dependent variable: coincidence dichotomy between rejection citations at the EPO and USPTO. Unit of analysis: DOCDB family citation pairs.

	Full sample	EP sample (priority in the EPC countries)	JP sample (priority in Japan)	US sample (priority in the US)
n	97,639 (14,795 families)	6,927 (986 families)	56,903 (9,937 families)	33,383 (3,799 families)
us_action_after_ESR_date	0.1020874*** (0.0298779)	1.295876*** (0.4832563)	0.1778124**** (0.0346444)	–0.0453259 (0.0613631)
US_REJECTION_counts	–0.0694963**** (0.0111592)	–0.0849421*** (0.0284)	–0.0831775**** (0.0149259)	–0.0454862** (0.0181918)
ISA_EP	0.3445932** (0.1466986)	0.4396124 (0.3221705)		0.4789017** (0.1940894)
first_EP	–0.0296502 (0.2824401)			
first_US	–0.0962588 (0.2449784)			
first_JP	–0.0063357 (0.2437147)			
techn_field_nr_counts	–0.0618432 (0.1132318)	–0.5136947 (0.371835)	–0.059244 (0.1588454)	0.0251554 (0.180763)
Log pseudolikelihood	–43286.6	–3572.8	–25760.0	–13551.1

Robust standard errors in the parentheses, with clustering of citation families.
Significance level: **** < 0.001, *** < 0.01, ** < 0.05, * < 0.1.
35 WIPO technology field dummies and priority year dummies are included, and citation families are controlled for, but estimated results (including those for constant terms) are omitted for space reasons.

Conclusion

In a summary, we generally find EPO–USPTO convergence of citations after European search report releases, implying the existence of spillovers from the EPO to the USPTO, except for applications from the US. The convergence of citations after the release of European search reports implies that there is benefit in search effort taken by the USPTO. Moreover, we find divergence of citations when prosecution takes longer. Namely, we find divergence of US rejection patent citations from those at the EPO as the process of a prosecution becomes longer, which is typically caused in the US by persistent challenges from applicants appealing repeated rejections. These results imply interdependence between major patent offices with converging and diverging tendencies, which have been found by a novel use of examiner patent citations. The finding suggests that there is benefit in collaborative search mechanisms between patent offices, which have policy implications (such as international search collaborations to reduce discrepancies between grant decisions). Moreover, it has an implication for citation study in general, in that sequential reviews of prior arts with respect to the same citing documents could result in different citation network structures, dependent on the possibility of information sharing between different citing entities. When evaluating international citation impacts via patent citation data, this study warrants a caveat.

This project is an extension of Wada (2018), motivated by the view that examiners are bounded by cognitive limitations. The seemingly wide discrepancies between citation behaviors in the trilateral offices may be attributable to the drifting bargaining issues in prosecution histories, although this conclusion needs further analysis.

Acknowledgments

The author acknowledges financial support by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 18H00889.

References

Cotropia, C. A., Lemley, M. A., & Sampat, B. (2013). Do applicant patent citations matter? Research Policy, 42(4), 844-854.

Jensen, P.H., Palangkaraya, A., & Webster, E. (2005) Disharmony in international patent office decisions. Federal Circuit Bar Journal, 15, 679.

Kuhn, J. M., Younge, K. A., & Marco, A. C. (2017). Patent citations reexamined: New data and methods. SSRN. https://ssrn.com/abstract=2714954.

Lu, Q., Myers, A. F., & Beliveau, S. (2017). USPTO patent prosecution research data: Unlocking office action traits. USPTO Economic Working Paper, No. 10.

Wada, T. (2017). The choice of examiner citations for refusals: Evidence from the trilateral offices. In: Proceedings of ISSI 2017: The 16th international conference on scientometrics and informetrics (pp. 950-957). Wuhan University, China.

Wada, T. (2018). The choice of examiner patent citations for refusals: evidence from the trilateral offices. Scientometrics, 117: 825-843.

Webster, E., Jensen, P. H., and Palangkaraya, A. (2014), Patent examination outcomes and the national treatment principle. RAND Journal of Economics, 45: 449-469.

Webster, E., Palangkaraya, A., & Jensen, P.H. (2007) Characteristics of international patent application outcomes. Economics Letters 95, 362-368.

Using machine learning and text mining to classify fuzzy social science phenomenon: the case of social innovation

Abdullah Gök[1], Nikola Milosevic[2] and Goran Nenadic[3]

[1] *abdullah.gok@strath.ac.uk*
Hunter Centre for Entrepreneurship, Strathclyde Business School, University of Strathclyde, 130 Rottenrow, G4 0GE, Glasgow (UK)
[2] *nikola.milosevic@manchester.ac.uk*
School of Computer Science, University of Manchester, M13 9PL, Manchester (UK)
[3] *g.nenadic@manchester.ac.uk*
School of Computer Science, University of Manchester, M13 9PL, Manchester (UK)
Alan Turing Institute, British Library (UK)

Abstract

Classifying social science concepts by using machine learning and text-mining is often very challenging, particularly due to the fact that social concepts are often defined in a vague manner. In this paper, we put forward a first conceptual step to overcome this challenge. By using the case of social innovation, which has 252 distinct definitions, we qualitatively demonstrated that these definitions group around four different themes where various definitions utilise one or multiple of these criteria in different combinations to define social innovations. We designed an experiment where a database of social innovation projects annotated i) based on an overall understanding and ii) based on a decomposed definition of four criteria. As a next step, we will test the performance of various model specification on these two approaches.

Introduction

We live in machine-learning age. The advent of artificial intelligence and the underlying machine-learning processes is more and more evident in the daily life from transport systems to productions. Similarly, the way natural sciences is conducted now benefits greatly from machine learning. This trend of utilising machine learning has also being increasingly explored in social sciences. However, one particular problem relating to the applications in social science is that most concepts are defined in a comparatively vague manner due to the differential understandings of them in their respective literatures. This makes employing machine-learning challenging as in most cases it requires a well-defined definition of the phenomenon to be classified and/or large amounts of data. This challenge is often attenuated when large amounts of data is not readily available due to the nature of the social phenomenon in question.

In this paper, we propose a conceptual approach to employ machine learning in classifying complex and fuzzy social science concepts. Our approach involves decomposing the social science concept in question to smaller, comparatively more analytically defined components through an extensive qualitative literature review of the differential understandings of the concept.

To test the suitability of our approach, we compare the performance of a machine learning model to classify entities related to the complex and vaguely defined social science concept of social innovation. We implemented two models: one is based on our approach of decomposing the definition of social innovation and another based on the conventional method of classifying entities based on undecomposed definition of social innovation.

The Case of Social Innovation

We use social innovation as a case study to illustrate our approach. Social innovation is broadly defined as "new ideas (products, services and models) that simultaneously meet social needs (more effectively than alternatives) and create new social relationships or collaborations" (European Commission, 2010). Prominent examples include the historical origins of the co-operative movement, hospices, model villages as well as the modern projects such as microfinance, fair trade, the Big Issue, online activism platforms and specific technological solutions that benefit disadvantaged groups such as blind people or refugees . While the most diffused examples of social innovation originates from the Victorian era, it is rapidly growing phenomenon thanks to the increased availability of social media and also the possibility of real-time collaboration through online tools. Social innovation has a huge potential to improve the lives of people where conventional innovation fails the challenge. In fact, social innovation featured heavily in UN Sustainable Development Goals for 2030.

While the importance and the increasing uptake of the concept of social innovation are detailed above, the exact definition of the concept of social innovation is complex and hotly debated in social science and policy literature. Taking its roots from the classics of Karl Marx , Max Weber and Emile Durkheim, the concept of social innovation started being used extensively in 1960s and since then the exact meaning of the concept have been subject of a debate. Edwards-Schachter and Wallace (2017) report that there are at least 252 variations of the concept. This cacophony of the definitions makes any data collection exercise more difficult but it is particularly challenging for a machine learning based approach, which requires a fairly clean and analytical understanding of the subject matter.

To overcome this hurdle, based on Edwards-Schachter and Wallace's (2017) idea, we have conducted a qualitative literature review to establish some common themes in the myriad of definitions. We have established that there are in fact four common themes in the definitions of social innovation (see Table 1 for a summary). While nuances between each of these themes are vastly varying, the broad themes are about social objectives, social interaction between actors or actor diversity, social outputs and innovativeness. However, different definitions include different combinations and different number of these themes (e.g. the EU definition we used above emphasises social objectives and actors interaction).

We used these four common themes in various definitions as distinct criteria in our model. We created four different classifiers for each of these four criterion. This kind of flexible and modular approach not only allows us to add more granularity to a complex concept but also it provides us the flexibility later on to deconstruct and construct any definition.

Table 1. Decomposed Definition of Social Innovation

Element of definition	**Criteria description**
Objectives	The project primarily or exclusively satisfies (often unmet) societal needs, including the needs of particular social groups; or aims at social value creation.
Actors and actor interactions	Satisfy one or both of the following: i. **Diversity of Actors:** Project involves actors who would not normally involve in innovation as an economic activity, including formal (e.g. NGOs, public sector organisations etc.) and informal organisations

	(e.g. grassroots movements, citizen groups, etc.). This involvement might range from full partnership (i.e. project is conducted jointly) to consultation (i.e. there is representation from different actors). ii. **Social Actor Interactions:** Project creates collaborations between ”social actors”, small and large businesses and public sector in different combinations. These collaborations usually involve (predominantly new types of) social interactions towards achieving common goals such as user/community participation. Often, projects aim at significantly different action and diffusion processes that will result in social progress. Often social innovation projects rely on trust relationships rather than solely mutual-benefit.
Outputs/Outcomes	Project primarily or exclusively creates socially oriented outputs/outcomes. Often these outputs go beyond those created by conventional innovative activity (e.g. products, services, new technologies, patents, and publications), but conventional outputs/outcomes might also be present. These outputs/outcomes are often intangible and they might include the following but not limited to: - change in the attitudes, behaviours and perceptions of the actors involved and/or beneficiaries - social technologies (i.e. new configurations of social practices, including new routines, ways of doing things, laws, rules or norms) - long-term institutional/cultural change
Innovativeness	There should be a form of ”implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method”. The project needs to include some form of innovative activities (i.e. scientific, technological, organisational, financial, and commercial steps intending to lead to the implementation of the innovation in question). Innovation can be technological (involving the use of or creating technologies) as well as non-technological. The innovation should be at least ”new” to the beneficiaries it targets (even if it is not new to the world).

Method and Data

We employ the European Social Innovation Database (ESID) in our study. ESID is a comprehensive database of social innovation projects that employs text-mining techniques to collect data about social innovation from the publicly available websites. The methodology used in ESID to populate social innovation projects semi-automatically initially uses currently available databases, lists, case study repositories, and mappings of social innovation projects in order to obtain initial data about social innovations. This phase includes the following steps (see Figure 1 for graphical representation):

1. Compose a list of social innovation sources.
2. Crawl the project description pages from the listed sources.
3. Crawl project websites, if they were available in the social innovation source.
4. Translate the crawled texts to English (if they are not in English).

5. Manually annotate a set of projects. The projects are annotated whether they satisfy social innovation criteria by human coders.
6. Create machine learning models for classifying projects for specific social innovation criteria.
7. Obtain additional features about the project, such as information about organisations involved, location, etc.

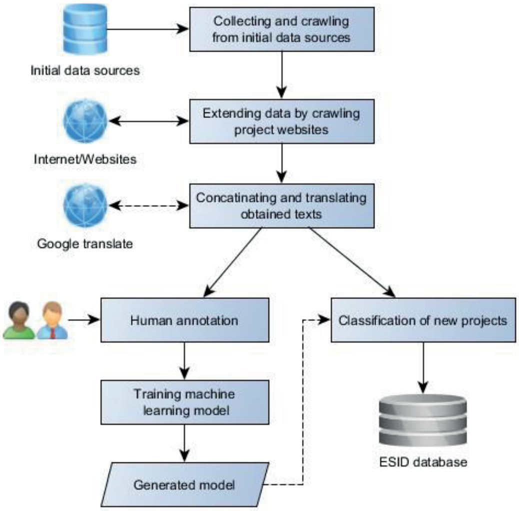

Initial data sources
Collecting and crawling from initial data sources
Internet/Websites
Extending data by crawling project websites
Google translate
Concatinating and translating obtained texts
Human annotation
Classification of new projects
Training machine learning model
ESID database
Generated model

Figure 1. Workflow of a classification of social innovation criteria.

After dropping projects we don't have any available websites or textual information or information less than 350 characters, ESID preliminary contained 3560 projects.

In order to make a data set for supervised machine learning-based approach, we organised a set of data annotation workshops. The annotators were PhD students and research staff whose research is associated with the areas of innovation and social innovation. We created a single document for each project which was a combination of the project description available in the original data source and also the text available on the project websites.

The annotators were asked to annotate sentences that present how a project met defined social innovation as a whole or in terms of the decomposed criteria (objectives, actor interaction, outputs, innovativeness). Annotators were asked to give a score at the document (i.e. project) level for the whole project based on an overall understanding of social innovation as well as based on each of the four decomposed criteria (as presented in Table 1). The document level marks were in the range of 0-3:

- 0 – criteria not satisfied
- 1 – criteria weakly satisfied
- 2 – criteria partially satisfied
- 3 – criteria fully satisfied

Our annotations involved at least two independent annotators (three annotators where there is disagreement between two annotators). We have obtained 728 annotated documents from three annotation workshops out of a total of 3560 projects initially included in the ESID. Of 728 annotations, 120 included annotations based on an overall understanding.

The dataset created during the annotation task was used for training and validation of the machine learning-based approach. The classifier is created for each social innovation criteria (objective, actors, outputs, innovativeness) as well as an overall understanding of the concept.

We have created and evaluated multiple classifiers for each of the criteria. The classifiers that were used were Naive Bayes, decision trees, random forests, long short-term memory recurrent neural networks (LSTM) (Sundermeyer et al, 2012), convolutional neural networks (LeCun et al., 2015) and stacked LSTM and convolutional neural networks (Wang et al., 2016).

The naive Bayes, decision trees and random forests classifiers used bag-of-words language models, with stemmed tokens and excluded stopwords (using Rainbow stop-word list[1]). Also, the naive Bayes, decision trees and random forests used unigram, bigrams, trigrams, and fourgrams as features. The neural network implementations relied on neural language model (Glove embeddings (Peninngton et al., 2014)). Long short-term memory recurrent neural networks (LSTM) classifiers were using a single layer with 100 neurons and a dense layer outputting the class. The convolutional neural network architecture consisted of three layers of convolutional networks with 512 filters in the first layer, 256 in the second and 128 in the third layer. The ensemble architecture consisted of the three-layered described convolutional network whose output was input to LSTM neural network.

Since dataset was not balanced, having more negative instances than positive, we also performed an experiment with balancing data by oversampling the class that had minority instances and adding new negative instances.

Next Steps

As a next step, we plan to construct two different classification models: one based on an overall understanding of the social innovation and the other based on our approach of analytically decomposing the definition of social innovation to four different criteria. We will then be able to compare the performance of these two models to each other and to reveal the added benefit of our approach. We also plan to explore the how the performance difference changes in these two approaches based on specific model specifications.

Conclusion

In this paper, we put forward a first conceptual step to utilise machine learning to classify complex and fuzzy social science concepts. By using the case of social innovation which has 252 distinct definitions, we qualitatively demonstrated that these definitions group around four different themes where various definitions utilise one or multiple of these criteria in different combinations to define social innovations. We designed an experiment where a database of social innovation projects annotated i) based on an overall understanding and ii) based on a decomposed definition of four criteria. As a next step, we will test the performance of various model specification on these two approaches.

[1] http://www.cs.cmu.edu/ mccallum/bow/rainbow/

Acknowledgments

The work presented in this paper is a part of KNOWMAK project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

References

European Union: Social innovation (2017), http://ec.europa.eu/social/main. jsp?catId=1022

Edwards-Schachter, M., Wallace, M.L.: shaken, but not stirred: Sixty years ofdefining social innovation. Technological Forecasting and Social Change 119, 64–79 (2017)

LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)

Sundermeyer, M., Schlu¨ter, R., Ney, H.: Lstm neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association (2012)

Wang, X., Jiang, W., Luo, Z.: Combination of convolutional and recurrent neuralnetwork for sentiment analysis of short texts. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 2428–2437 (2016)

Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

Non-Traditional Indicators for the Evaluation of SBIR-like Programs: Evidence from Brazil

Sergio Salles-Filho[1], Bruno Fischer[2], Camila Zeitoum[3], Paulo Henrique Feitosa[4] and Fernando Colugnati[5]

[1] *sallesfi@unicamp.br*
University of Campinas, Institute of Geosciences, Dept of Science and Technology Policy, Carlos Gomes Str 250, 13083-855, Campinas (Brazil)

[2] *bruno.fischer@fca.unicamp.br*
University of Campinas, School of Applied Sciences, Pedro Zaccaria Str 1300, 13484-350, Limeira (Brazil)

[3] *camilazeitoum@unicamp.br*
University of Campinas, Institute of Geosciences, Dept of Science and Technology Policy, Carlos Gomes Str 250, 13083-855, Campinas (Brazil)

[4] *pfeitosa@usp.br*
University of São Paulo, School of Communications and Arts, Prof. Lúcio Martins Rodrigues Av 443, 05508-020, São Paulo (Brazil)

[4] *fernando.colugnati@ufjf.edu.br*
Federal University of Juiz de Fora, Medicine School, Dept. of Internship, José Lourenço Kelmer Str, 1300, 2nd floor, 36036-330, Juiz de Fora (Brazil)

Abstract

This document reports an ongoing evaluation process concerning the PIPE Program in Brazil. PIPE is a SBIR-like instrument aimed at nurturing Knowledge-Intensive Entrepreneurship (KIE) in the State of São Paulo, Brazil. Our evaluation strategy consists in three overarching steps, namely: (i) data collection from PIPE beneficiaries with projects finished between 2007 and 2016; (ii) data collection for control groups (based on reject proposals); and (iii) online launch of a data platform providing information for past and current projects, with continuous updates of future projects. Envisaged outcomes include the provision of policy insights and the implementation of a culture oriented towards continuous evaluation of the program under scrutiny. Preliminary findings have stressed the pivotal role of managerial and governance practices as determinants of firm-level results, an aspect that is underreported in the literature on evaluation of similar programs. Implications involve the need for related policy frameworks to look more closely into firms' management capabilities and their effects over outputs and outcomes.

Introduction

Knowledge-intensive entrepreneurship (KIE) stands for a socioeconomic phenomenon that drives economic competitiveness and innovative capabilities within the dynamics of economic systems (Ács et al., 2008; Saxenian, 1994). Nonetheless, its systemic nature – often overlooked in favour of individuals and firms – has only recently become an issue of widespread interest among researchers and policymakers (Borissenko & Boschma, 2016; Autio et al., 2014; Zahra et al., 2014; Stam & Spigel, 2016).

The role of policy is paramount in this context, as excessive risks, asymmetric information and absence of well-developed markets for venture capital may get in the way of business development (Pan & Yang, 2018; Lerner, 2002).

In spite of evidences from different assessments, part of the literature in the field has been disputing the actual effectiveness of KIE-promoting policies in face of the observed economic relevance of knowledge-intensive entrepreneurship (Brown & Mason, 2014; Chatterji et al, 2013). Thus, further – and more refined – impact assessments and evaluation exercises on these matters are due.

Based on this background, our proposal is directed towards evaluating the impacts and functioning of the Innovative Research in Small Enterprises Program (PIPE) in the State of São Paulo, Brazil. This program is funded by the São Paulo Research Foundation (FAPESP) to support innovative initiatives in small companies and it resembles in structure and objectives the Small Business Innovation Research (SBIR) program in the United States.

Our framework of evaluation goes well beyond the traditional input-output analysis focused on the plain association between resource allocation and measurement of technological and economic results, usually addressed through intellectual property rights, firm-level turnover, impacts on employment and R&D investment (e.g. Galope, 2016; Inoue & Yamaguchi, 2017).

It includes governance, adoption of professional R&D management tools, networking and individual-level aspects of firms and entrepreneurs, aiming at identifying latent relationships between managerial capabilities and the ultimate results and impacts arising from funded projects.

In this research-in-progress manuscript - and based on a recent field research - we show original findings over the relevance of such indicators as predictors of the performance of funded-firms. This contributes both to the literature of STI impact assessment and technology based entrepreneurship.

Knowledge-Intensive Entrepreneurship: Background and the Argument for Policy

Previous research has identified similarities and differences between the Brazilian PIPE and the US SBIR: main differences in the findings of impact evaluations of both programs were precisely on the volume and importance of private venture capital and the capacity of licensing the technological outputs of funded projects (Salles-Filho et al., 2011).

Several factors are normally pointed as critical to the success of funded firms, as for having previous connections to the target market, having minimal experience in technology commercialization, having previous R&D capabilities, just to mention the most common (Link & Scott, 2012; Connell, 2017).

Accordingly, impact evaluations of SBIR-like programs normally employ indicators able to seize input and output additionality for the variables mentioned above (Link & Scott, 2012).

Although intuitively relevant, indicators for measuring behavioral additionality are not frequently employed (MIoIR, 2015). Amongst these, mechanisms of corporative governance, professionalization of R&D management, and networking and cooperation variables are virtually absent of SBIR-like assessments.

PIPE Program and Research Hypotheses

The PIPE Program (the acronym stands for Technological Innovation in Small Business) is run by FAPESP, the Research Foundation of the State of São Paulo, Brazil. The initiative was created in 1997 to mirror the Small Business Innovation Research (SBIR) program in the USA. The rationale for such approaches to innovation in small, entrepreneurial firms is attached to their role in innovation systems (Salles-Filho et al, 2011).

PIPE is structured along three phases. Phase I consists of initial assessments of technical feasibility. Projects can receive up to USD 60,000 for a 9-month period. Phase II funds the actual development of innovative research proposals up to an amount of USD 300,000 over periods of two years (although justified extensions are allowed). Phase III is headed to support activities between R&D and innovation – FAPESP does not fund directly this Phase but organize calls in cooperation with the Brazilian Innovation Agency (FINEP). Our analysis is oriented towards impacts associated with Phase II.

The definition of small enterprises according to the PIPE criteria includes companies with 250 employees or less. These businesses are obliged to have research units in the State of São Paulo, Brazil, and to demonstrate capacity to leverage internal and external resources from different sources. It is also important noticing that projects are not necessarily supposed to be related to academia – although a large amount of recipients have strong academic backgrounds. This is likely a function of the strong concentration of Brazilian researchers in Higher Education Institutions (Ryan, 2010).

In order to assess the PIPE Program, we derive some insights from related literature. A first aspect to be taken into account concerns the common mistake found in similar programs when it comes to the lack of articulation between the technological side of ventures and their business prospects (Lerner, 2002). In this regard, interventions should focus on 'softer' aspects of innovation dynamics, addressing matters related to managerial support and strategic behaviour of firms (Warwick, 2013). This is so because, in order to become successful, KIE ventures need to evolve in terms of their stocks of resources, capabilities and social capital (Vohora et al., 2004), going well beyond the mere provision of subsidies.

To face this concern, our assessment offers a set of hypotheses based on a decomposition exercise of the explicit objectives of PIPE program and their relationship with managerial practices of firms. We try to accommodate these propositions to allow testing for differences over time in PIPE projects and also to compare these projects with control groups. At the point this manuscript is being written we have already findings for one specific hypothesis:

> "Capabilities associated to structured R&D management and corporative governance enhance SMEs' innovativeness, R&D investment, job creation – both general and higher education positions -, net income, and partnerships."

Method

The methodological approach for impact evaluation of PIPE is multidimensional, looking for input, output/outcome and behavioral additionality. The design is based on the collection of primary and secondary data from grantees - whose projects had started in 2006 and concluded in 2016 – and from rejected projects in the same period. A sample of 400 funded and 2700 rejected projects represents the total population.

Two questionnaires – one for grantees and other for rejected – were built upon an on-line platform. A total of 185 responses for the treatment group and 490 for the control group have been collected and organized in a database. For the purpose of this manuscript 145 answers from the grantees and 85 from the rejected had complete information to test the hypothesis above mentioned.

A quasi-experiment has been designed employing the macthit package of "R". Coarsened Exact Matching (CEM) was employed as matching technique given the small valid sample size. Also secondary data from FAPESP's database has been collected, particularly the ones related to company's and researcher's profiles.

Indicators were defined and organized to cover the following themes:[1]

- Companies and project profiles
- Entrepreneur/Researcher profiles
- R&D Investment
- Financial and Economic data (internal and external market + venture)
- Employment and job creation (total and R&D related)

[1] In the present manuscript we will focus on the preliminary findings related to Governance and management and Partnership and collaboration.

- Intellectual property and technology transfer
- Governance and management
- Partnership and collaboration

A descriptive and a multivariate analysis have been conducted for the funded projects.[2] With respect to the themes of governance and management and of partnership and collaboration, a bivariate analysis was performed correlating these variables with outcome indicators.

Main variables under these themes are:

- Coordinator's background in Business administration
- Explicit R&D&I strategy
- Governance and Compliance formalized
- R&D Project Management formalized
- Partnership with ROs
- Partnership other organisations

Also, multivariate regression models were adjusted for each main outcome indicator and results were compared to the effect estimated after matching. The estimate were model based, providing Average Treatment Effect on Treated (ATT), always performed on the same variable set used in multivariate regression. For counting indicators, quasi-poisson models were used, providing a ratio Treatment / Control effect. For binary responses, the logistic model was the choice, providing Odds Ratios. For continuous economic indicators, like R&D expenditure, log-normal models provides elasticity effect as impact estimate.

Preliminary Results

When assessing the hypothesis above mentioned, initial outcomes drawn from a bivariate analysis identify evidences of a critical role played by governance and managerial capabilities concerning their influence over firms' technological and social-economic outcomes.

An analysis comparing grantees and rejected groups before and after matching reveals PIPE's effects over some variables of interest, namely: R&D expenditures, job creation, job creation in R&D, job creation of higher education positions in R&D, and establishment of partnerships with research and non-research organisations.[3] Main findings are listed bellow.

- Regarding R&D expenditures awardees-group showed positive effects over non-rejected-group only before matching them, and for few segments of industrial manufacturing (Brazilian's standard industry classification).
    - After matching groups no effect over R&D expenditures has been observed.
    - Management variables showed no significant effect over R&D expenditures, neither for treatment nor for control groups.
- Regarding job creation PIPE's companies have showed positive and significant effects for the 3 sorts of jobs above mentioned (total, R&D, Higher Education in R&D).
    - For total job creation although null effects before matching, companies that had adopted management skills showed an increase of circa 41% over total job creation.
    - After matching groups, PIPE's effect showed an increase of around 60% in total job creation.

[2] Data from rejected projects are now being analized.

[3] All input and output variables were measured before inception (average of 2 years) and after completion (average of 2 years) of projects. For rejected proposals we adopted an average of 3 years for completion then getting data before and after using the same rationale.

    - As for job creation in R&D activities, both for tertiary and non-tertiary levels, PIPE's effects have been positive and significant before and after matching groups. Before matching PIPE's companies hired twice more people for R&D than rejected group. Considering the subgroup presenting management skills, this number increased by 3,6 times against non-awardees.
    - After matching, PIPE's effect was three times higher than control group.
    - For higher education jobs in R&D the subgroup of managerial capabilities hired 33% more than PIPE's alone effect. After matching, companies that received PIPE's grants hired twice more than non-grantees.
- Regarding the effects over partnerships, although PIPE's companies present a positive effect of 1,6 and 2,2, respectively before and after matching, managerial variables did not influenced the number of partnerships.

All in all, preliminary findings have revealed positive cum significant relationship between variables of managerial skills and job creation amongst companies funded by PIPE. Besides those positive effects observed over employment, effects on R&D expenditures are not that evident showing significance only for some segments of manufacturing. This apparent contradiction needs more investigation to be clarified.

Final remarks

We have found evidences that managerial capabilities can be deemed as pivotal in a context of entrepreneurs and companies that possess technological capabilities, but often lack the necessary skills to effectively turn them into a competitive business. Hence, by looking into such issues – instead of the dominant input-output approaches – implications for policy must be investigated.

For SBIR-like programs to thrive, additional importance should be given to companies' managerial and governance capabilities. That should include approaches that offer support for the development of management practices and skills at the firm-level throughout the application of resources, as well as *ex-ante* analysis of firms' plans to acquire such capabilities.

Next steps

Data analysis is still being conducted for all other themes and indicators both regarding multivariate and counterfactual analysis. Other hypotheses related to the themes of evaluation are now being tested in order to get to stronger evidences for policy design.

Acknowledgments

Authors acknowledge funding from the São Paulo Research Foundation - FAPESP (Grant #2017/17461-1).

References

Ács, Z., Desai, S., & Hessels, J. (2008). Entrepreneurship, economic development and institutions. *Small Business Economics*, 31(3), 219-234.

Autio, E., Kenney, M., Mustar, P., Siegel, D., & Wright, M. (2014). Entrepreneurial innovation: the importance of context. *Research Policy*, 43(7), 1097–1108.

Borissenko, Y., & Boschma, R. (2016). A critical review of entrepreneurial ecosystems: towards a future research agenda. [Papers in Evolutionary Economic Geography #16.30]. *Utrecht University - Urban & Regional Research Centre.*

Brown, R., & Mason, C. (2014). Inside the high-tech black box: a critique of technology entrepreneurship policy. *Technovation*, 34(12), 773-784.

Chatterji, A., Glaeser, E., & Kerr, W. (2013). Clusters of entrepreneurship and innovation. [Working Paper 19013]. *National Bureau of Economic Research Working Paper Series*.

Connell, D. (2017). Leveraging public procurement to growth the innovation economy - An Independent Review of the Small Business Research Initiative by David Connell; Final Report and Recommendations

Fischer, B., Queiroz, S., & Vonortas, N. (2018). On the location of knowledge-intensive entrepreneurship in developing countries: lessons from São Paulo, Brazil. *Entrepreneurship and Regional Development*, 30(5-6), 612-638.

Galope, R. (2016). A Different Certification Effect of the Small Business Innovation Research (SBIR) Program. *Economic Development Quarterly*, 30(4), 371-383.

Inoue, H., & Yamaguchi, E. (2017). Evaluation of the Small Business Innovation Research Program in Japan. *SAGE Open*, 7(1).

Isenberg, D. (2010). How to start an entrepreneurial revolution. *Harvard Business Review*, 88(6), 40-51.

Lerner, J. (2002). When bureaucrats meet entrepreneurs: the design of effective public venture capital programmes. *The Economic Journal*, 112(477), F73-F84.

Link, A. N., & Scott, J. T.. "Real Numbers: The Small Business Innovation Research Program." Issues in Science and Technology 28, no. 4 (Summer 2012).

MIoIR – Manchester Institute of Innovation Research (2015). A review of the small business research initiative. Manchester, Final Report.

Neto, J., Farias Filho, J., & Quelhas, O. (2014). Raising financial resources for small and medium enterprises: A multiple case study with Brazilian venture capital funds in the cities of Rio de Janeiro and São Paulo. *International Journal of Innovation and Sustainable Development*, 8(1), 77-91.

Pan, F., & Yang, B. (2018). Financial development and the geographies of startup cities: evidence from China. *Small Business Economics*. Forthcoming.

Ryan, M. (2010). Patent Incentives, Technology Markets, and Public–Private Bio-Medical Innovation Networks in Brazil. *World Development*, 38(8), 1082-1093.

Salles-Filho, S., Bonacelli, M., Carneiro, A., Castro, P., & Santos, F. (2011). Evaluation of ST&I programs: a methodological approach to the Brazilian Small Business Program and some comparisons with the SBIR program. *Research Evaluation*, 20(2), 159-171.

Stam, E., & Spigel, B. (2016). Entrepreneurial Ecosystems. [Discussion Paper Series n. 16-13]. *Utrecht University – Utrecht School of Economics*.

Vohora, A., Wright, M., & Lockett, A. (2004). Critical junctures in the development of university high-tech spinout companies. *Research Policy*, 33(1), 147-175.

Warwick, K. (2013). Beyond industrial policy: emerging issues and new trends. [Science, Technology and Industry Policy Paper n. 2]. *OECD*.

Wessner, C. (2008). *An Assessment of the SBIR Program at the National Science Foundation*. Washington, DC: National Academies Press.

Zahra, S. A., Wright, M., & Abdelgawad, S. G. (2014). Contextualization and the advancement of entrepreneurship research. International *Small Business Journal*, 32(5), 479-500.

Eponymy and Delayed Recognition: *the case of Otto Warburg Nobel Prize*

Philippe Gorry[1] and Pascal Ragouet[2]

[1] *philippe.gorry@u-bordeaux.fr*
[1]GREThA UMR CNRS 5113, University of Bordeaux, Av. Leon Duguit, 33608 Pessac (France)

[2] *pascal.ragouet@u-bordeaux.fr*
[2]Centre Emile Durkheim, UMR CNRS 5116, University of Bordeaux, 3 ter Place de la Victoire, 33076 Bordeaux (France)

Abstract

Eponymy is defined as the way to name a discovery from the name of the person who discovered it. This practice well established in science. There is evidence that when an author has been eponymized, the author's original publications will be cited without bibliographic reference. Merton defined this as "obliteration by incorporation". The author's original publications will experience "delayed recognition", not achieving recognition in terms of citations until a few years after their original publication. While this phenomenon has been the subject of a renewal interest in scientometrics, there are few analyses of eponymy in science, and none explored the linked between eponymy and delayed recognition. Through the analysis of "cancer research" field, we identified a case study related to Otto Warburg Nobel Prize pioneering work on the role of metabolism in cancer, today named "Warburg effect". Our results demonstrate that "Warburg effect" as concept suffered from delayed recognition in terms of citation, and that delayed recognition of Warburg's publication is not due to a phenomenon of "obliteration by incorporation". In a general way, our results imply that delayed recognition phenomena should be extended to scientific concept and not limited to a single or a bundle of publications.

Introduction

Eponymy

"Eponymy" is a word of Greek origin made up of epi: over, and noma: name. It means 'giving name to something or naming after'. Eponymy is understood as denominating a phenomenon, law, theory, principle, invention or procedure with the originator's name. In this way, the name of a discovery is derived from the name of the person who discovered or described it in the first instance. The practice of eponymy is well established in science, as well as in non-scientific areas of life according to Merton (1973), there are gradations of eponymy, ranging from the founder of discipline field (Boolean algebra) to scientists who are honored for laws, instruments, constants, and methods (for example, the Lotka law in bibliometrics) (deB. Beaver, 1976). Eponymy has been the subject of little analysis in the scientometrics literature. We can mention the case study on the "Southern blot" (Thomas, 1992) and the one on "Nash equilibrium" (MacCain, 2011). Thomas (1992) traced the citations to a paper published by Southern (1975) which introduced a method to detect DNA fragment using gel electrophoresis and reported that Southern blotting had begun to achieve eponymy within 1.5 years. MacCain explored the phenomena of eponymy of in the literature to the game-theoretic concept of the Nash Equilibrium. She suggested that Nash's papers have experienced some "delayed recognition" explained by Obliteration by Incorporation. At last, a first attempt of eponym quantification in a large volume of full-text publications has been proposed by Cabanac (2014) using semi-automatic text mining approach applied to a corpus of Scientometrics articles.

Obliteration by incorporation

There is evidence that an author has been eponymized when they stop being mentioned separately and/or they are cited in the text without being indexed bibliographically, being fully incorporated into the scientific language of a discipline. A citation without bibliographic reference is a variant of the characteristic phenomenon of cultural transmission that Merton (1995) defined as "obliteration by incorporation" which consists in the fact that an author's discoveries and ideas have ended up being so fully incorporated into the current corpus of canonical knowledge that its source is not indexed, nor even mentioned in any way.

Delayed recognition

Delayed recognition (DR) is a phenomenon where papers do not achieve recognition in terms of citations until a few years after their original publication. In the literature, it is also referred to "resisted scientific discovery", "premature discoveries", "late-bloomers", or "Mendel syndrome". In today scientometrics literature, it is generally called "sleeping beauty", a publication that goes unnoticed for a long time (sleeping period), and then, almost suddenly, is awakened by a "prince" (PR), attracting from there on a lot of attention in terms of citations (awakening period). We will call these papers Delayed Recognition papers to use a gender-neutral name and to be free of its rhetorical limits. Since the definition of Sleeping Beauty, introduced by Van Raan, different quantitative criteria have been proposed in scientometrics to characterized DR papers: average-based criteria, quartile-based criteria and parameter-free criteria. Especially, the calculation of the "Beauty coefficient" (B), a parameter-free index proposed by Ke et al. (2015) has been powerful. DR papers have been identified in numerous research fields such as physics, chemistry, and medical sciences. The reasons for DR pattern of citations (sleeping and awakening periods) may be linked to boundary disciplines work, scientific controversy, hypothesis waiting for experimental proof, paradigm shift, industrial application coming up, and social recognition through Nobel Prize, to quote a few (Van Raan, 2004; Gorry & Ragouet, 2016; El Aichouchi & Gorry, 2018a & b).

Case study

The research reported here is a case study identified through the analysis of a specific research field. We analyzed the whole body of scientific in cancer research (3.9 M publications) and the citation life of those publications looking for DR papers. We identified several cases: one is linked to the Judah Folkman's hypothesis on tumor angiogenesis which have been already documented (El Aichouchi & Gorry, 2018b), and the second is related to the Otto Warburg 's hypothesis on cancer and metabolism, the so-called "Warburg effect". Therefore, we focus our analysis on the citation history of Warburg's publications, the trends of the "Warburg effect" literature and the trajectory of "cancer metabolism" research.

Background

Otto Heinrich Warburg (8 October 1883 – 1 August 1970), was a German physiologist, trained in chemistry (PhD in 1906, Berlin) under E. Fisher (Nobel, 1902) and a medical doctor (Heidelberg, 1911). Between 1908 and 1914, Warburg was affiliated with the Naples Marine Biological Station where he performed research on oxygen consumption in sea urchin eggs. In 1918, Warburg was appointed professor at the Kaiser Wilhelm Institute for Biology in Berlin-Dahlem (since renamed the Max Planck Society) where he pursued his research until the age of 86. Warburg investigated the metabolism of tumors and the respiration of cells, particularly cancer cells, and in 1931 was awarded, the Nobel Prize in Physiology for his "discovery of the nature and mode of action of the respiratory enzyme". Warburg's combined work in plant physiology, cell metabolism, and oncology. He hypothesized that cancer growth is caused by tumor cells generating energy mainly by anaerobic respiration. He edited and had much of his original work published in 1927. Thereafter, Warburg continued to develop his hypothesis experimentally and gave several prominent lectures outlining the theory and the data. Since the "Warburg effect" refer in cancer research to this cellular metabolism (Brand, 2010).

Methods

A complex search query was built in order to isolate the whole publications in "Cancer research". Basically, it associates keywords and synonymous of cancer terms (neoplasm, tumor, tumor, metastasis, malignant), cancer-related terms (oncology, carcinogenesis) or cancer-specific terms (leukemia, lymphoma, melanoma, sarcoma), inspired by the Mesh (PubMed), run in the title and abstract fields as well in the source title field with some adaptations to encompass the whole literature published in "cancer research" journals. The query was optimized by using Boolean operators and regular expressions, and run in Web of Science database. The cancer research literature corpus identified was quite large (3.9 M publications). Therefore, we focus only on "article" and "review" published previously the year 1990 and extracted a corpus of 339 835 references with the help of a web API (https://clarivate.com/products/data-integration/). Then, we calculated the "B" coefficient according to Ke et al. (2015) for papers with more than 1000 citations up to the year 2017 (n=2997 publications) by using a JavaScript. Applying their criteria to their database (n=22 M), Ke et al. found that the top 1,000 DR papers have a B ≥ 317.93. We continued further analysis with DR papers in "cancer research" above this threshold and measured the sleeping period, the awakening year and the year of maximum number citations. Then, we explored the citations profiles of Warburg's publications. Finally, we run two additional search queries: one looking for publications with the string of word "Warburg effect" present in the title or the abstract field, and the second looking for publications on "cancer and metabolism". This last query was built by associating the initial query "cancer research" and the terms "metabolism", "anaerobic", "glycolylation", "glycolytic", and "fermentation".

Results

Because of the large number of publications in cancer research (3.9 M), we focused only on "article" or "review" documents that have been published more than 10 years ago and which have cumulated more than 1000 citations. Then, the calculation of the B coefficient using Ke's equation reveals the existence of possibly DR papers with a significant B > 317.93.

Table 1. Top DR papers in cancer research

Author	Title	Journal	Pub. year	C tot.	B	Awakening year	C. max	T max.
Warburg, O	The metabolism of tumors in the body.	J. of Gen. Physiol.	1927	1005	5974,949	2009	178	2017
Youden, WJ	Index for rating diagnostics tests	Cancer	1950	3221	3042,621	2005	528	2017
Warburg, O	Respiratory impairment in cancer cells	Science	1956	1488	2271,248	2004	170	2017
Slaughter, D	Field Cancerization in oral stratified squamous epithelioma…	Cancer	1953	2079	653,225	1990	122	2016
Warburg, O	Origin of cancer cells	Science	1956	6179	606,994	2006	636	2015
Dukes, CE	The classification of cancer of the rectum	J. of Pathol. & Bact.	1932	1347	446,546	1972	56	1997
Folkman, J	Tumor angiogenesis -Therapeutic implications	N.E.J.M.	1971	6554	335,632	1996	425	2012

Among the top 10 cited articles in cancer research with a B coefficient above the threshold (Tab. 1), we noticed 3 publications authored by Otto Warburg, and one by Judah Folkman. This last publication has already been reported as a DR paper case study (El Aichouchi & Gorry, 2018b). Those Warburg' DR papers are all related to his hypothesis on metabolism and cancer: one has been published prior his Nobel prize, and the 2 others later (1956) in Science journal. The one intitled "Origin of cancer cells" is his most cited paper with more than 6000 citations. Warburg published his first article in 1905 and his last paper in 1970. His scientific work totalized 21,449 citations (31/12/2017) and reached 1096 citations/year in 2017 (Fig. 1). Notably, there is a rebound of the number of cumulated citations counted for all his publications around the year 2006 (Fig. 1: green line).

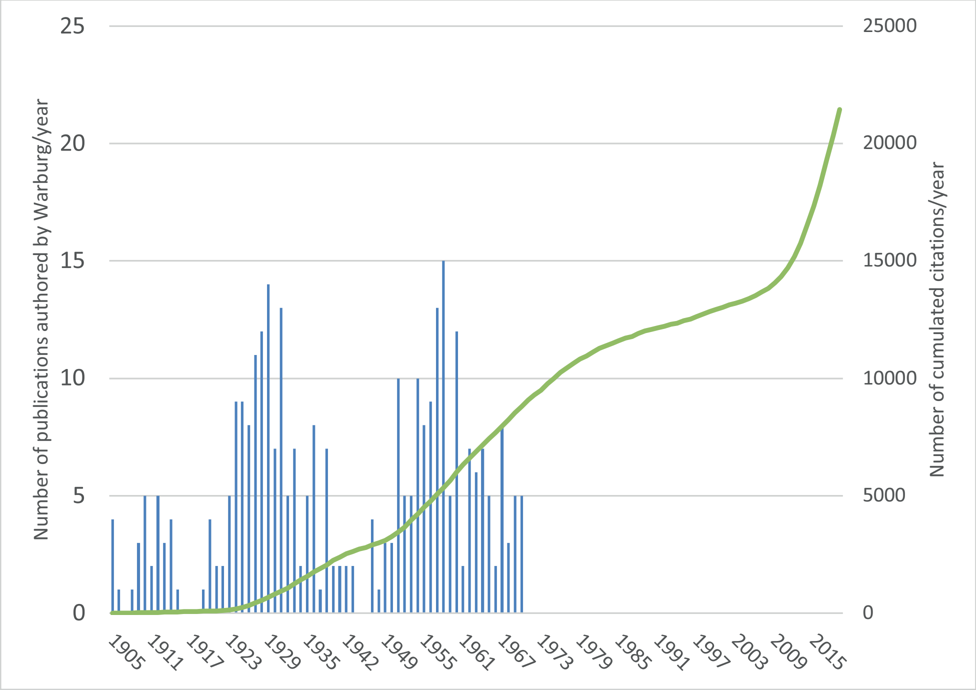


Figure 1. Warburg's publications and citations

Table 2. Journal distribution of Otto Warburg scientific production

Rank	Journal	Number of Publications	%
1	Biochmesche Zeitschrift	107	33.230
2	Zeitschrift fur Naturforschung Part B Chemie Biochemie Biophysik Biologie und Verwandten Gebiete	60	18.634
3	Naturwissenschaften	38	11.801
4	Hoppe Seylers Zeitschrift fur Physiologische Chemie	25	7.764
5	Zeitschrift fur Naturschrift fur Naturforschung Section B A Journal of Chemical Sciences	16	4.969
6	Pflugers Archiv fur die Gesamte Physiologie des Menschen und der Tiere	9	2.795
7	Berichte der Deutschen Chemischen Gesellschaft	7	2.174
8	Science	6	1.863
9	Biochimica et Biophysica Acta	5	1.553
10	American Journal of Botany	3	0.932

During his 65 years' career, he published 323 original articles in various journals (Tab. 2), and collaborated with 55 authors. Many of his publications were published in German in numerous German scientific journals.

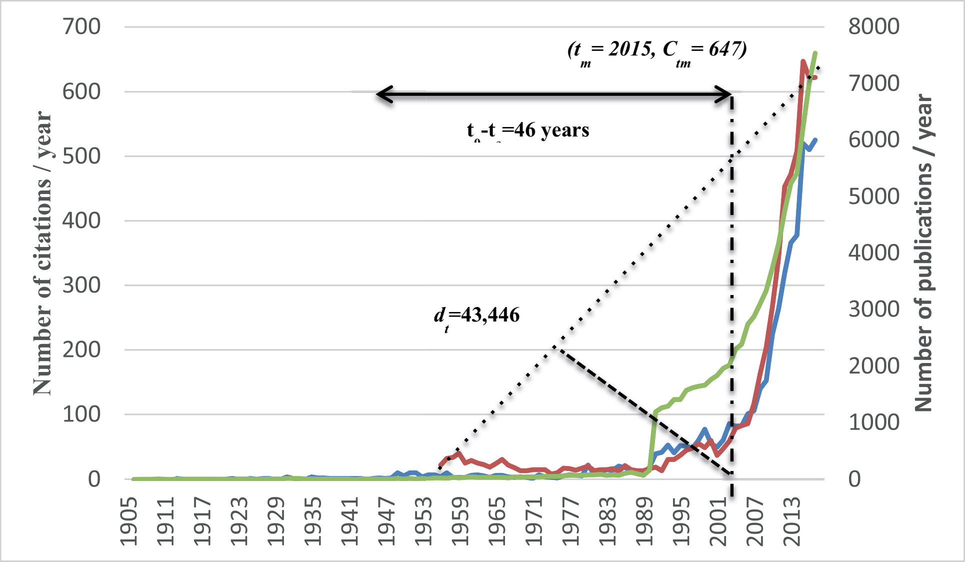


Figure 2. Citations history of Warburg's DR paper compare to Warburg's effect publications and cancer metabolism publications trends

Citations history of Warburg main DR paper is described in Figure 2 (red line). The distance dt defining the awakening time has been calculated as being year 2006, which is also the year witnessing the rise of the topic on Warburg effect (blue line) concomitantly with the explosion of publications on "cancer metabolism" (green line) after a sleeping period of 46 years.

Discussion

Our results explore the extent to which Warburg's hypothesis on cancer metabolism was the subject of a delayed recognition. During his career, he has been frustrated by the lack of acceptance of his ideas, and quoted an aphorism he attributed to Max Planck: "Science advances one funeral at a time". Indeed, citations analysis shows that Warburg landmark papers published in 1927 and 1956 are DR papers which have been ignored by the scientific community for 46 years. Interestingly, the fact that Otto Warburg was the recipient of the Nobel Prize, challenges the idea that academic recognition through honours and awards is sufficient to awake unpopular hypothesis. Moreover, the eponimyzation of his biological discovery after his name should have popularized his hypothesis. Therefore, we could argue as Merton (1995) that Warburg's original publications as been exposed to the phenomena of "obliteration by incorporation". Indeed, "Warburg effect" has been mentioned in numerous textbooks of medical oncology during decades (data not shown). But contrary to the previous studies reported in the literature (Thomas, 1992; McCain, 2011), the eponym "Warburg effect" has also been the subject of delayed recognition. This observation extended the initial concept of DR from paper to scientific concept. This new result is an extension of a previous observation of DR related papers cluster in material sciences (El Aichouchi & Gorry, 2018a). The sleeping period of Warburg's DR papers and "Warburg effect" could not be linked to a lack of academic. Maybe it is related to the controversy raised in the years 1960-70 by Warburg's propositions of preventive nutritional treatment of the cancer at the time when the genetic hypothesis of cancer was emerging. At last, the awakening Warburg's DR papers and the "Warburg effect", remain to be explored in the historical context of the biology of cancer. Interestingly, we could make the assumption that the appearance around the awakening year of new topic around cancer and metabolism is the reason for the rediscovery of Warburg's pioneering work (Koppenol, 2011). This hypothesis needs further analysis through quantitative approach (citations and co-citations analysis) and qualitative approach (semi-directed interview of authors citing Warburg DR papers around the year of awakening).

Conclusion

Our results demonstrate that "Warburg effect" as concept suffered from delayed recognition in terms of citation, and that delayed recognition of Warburg's publication is not due to a phenomenon of "obliteration by incorporation". Somewhat, our results imply that delayed recognition phenomena should be extended to scientific concept and not limited to a single or a bundle of publications.

Acknowledgments

We would like to thanks Edouard Kleinpeter for his computing assistance. This work is supported by a grant "Emergence" from Canceropôle Grand Sud Ouest 2018-2019.

References

Brand, R.A. (2010). Biographical Sketch: Otto Heinrich Warburg, PhD, MD. *Clinical Orthopaedics and Related Research*. 468, 2831–2832.

Cabanac, G. (2014). Extracting and quantifying eponyms in full-text articles. *Scientometrics*, 98:1631–1645

De B. Beaver, D. (1976). Reflections on the Natural History of Eponymy and Scientific Law. *Social Studies of Science*, 6, 89-98.

El Aichouchi, A. & Gorry, P. (2018). Delayed recognition of Judah Folkman's hypothesis on tumor angiogenesis: when a Prince awakens a Sleeping Beauty by self-citation. *Scientometrics*, 116, 385-389.

El Aichouchi, A. & Gorry, P. (2018). Paul Hagenmüller's contribution to solid state chemistry: a scientometric analysis. *Journal of Solid State Chemistry*, 262, 156-163.

Folkman, J. (1971). Tumor angiogenesis: Therapeutic implications. *New England Journal of Medicine*, 285(21), 1182–1186.

Gorry, P. & Ragouet, P. (2016). "Sleeping Beauty" and Her Restless Sleep: Charles Dotter and the Birth of Interventional Radiology. *Scientometrics*, 107, 2, 773 784.

Ke, Q., Ferrara, E., Radicchi, F. & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of National Academy of Science USA*, 112, 7426–743.

Koppenol, W.H., Bounds, P.L., Dang, C.V. (2011). Otto Warburg's contributions to current concepts of cancer metabolism. *Nature Review Cancer*, 11, 325-37.

McCain, K.W. (2011). Eponymy and Obliteration by Incorporation: The Case of the "Nash Equilibrium". *Journal of the Association for Information Science & Technology*, 62, 1412-1424.

Merton, R.K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

Merton, R.K. (1995). The Thomas Theorem and the Matthew effect. *Social Forces,* 74, 379-424.

Wikipedia (2019). Otto Heinrich Warburg. Retrieved May 17, 2019 from: http://en.wikipedia.org/wiki/Otto_Heinrich_Warburg.

Thomas, K.S. (1992). The development of eponymy: a case study of the Southern blot. *Scientometrics,* 24, 405-417.

Van Raan, A.F.J. (2004). Sleeping beauties in science. *Scientometrics,* 59, 467–472.

Warburg, O. (1956). Origin of cancer cells. *Science*, 123, 309-314.

The Nobel Prize in Physiology or Medicine 1931, https://www.nobelprize.org/prizes/medicine/1931/

On the potential for detecting scientific issues and controversies on Twitter: a method for investigating conversations mentioning research

David Gunnarsson Lorentzen, Johan Eklund, Björn Ekström and Gustaf Nelhans

{david.gunnarsson_lorentzen, johan.eklund, björn ekström, gustaf.nelhans} @hb.se
University of Borås, Swedish School of Library and Information Science, S-501 90 Borås (Sweden)

Abstract

In this study, we demonstrate how to collect Twitter conversations emanating from or referring to scientific papers. We propose segmenting the conversational threads into smaller segments and then compare them using information retrieval techniques, in order to find differences and similarities between discussions and within discussions. While the method still can be improved, the study shows that it is possible to collect larger conversations about research on Twitter, and that these are suitable for various automated methods. We do however identify a need to analyse these with qualitative methods as well.

Introduction

The purpose of this paper is to propose a method for mapping issues within conversations on Twitter which in various ways refer to or mention scientific publications. The study builds on work done by Nelhans and Lorentzen (2016), who utilised the Twitter Streaming API to collect tweets including a reference to a digital object identifier (DOI) and the most active users in the collected dataset. By filtering the stream using the combination of search terms and users, they managed to mine conversational threads with references to DOIs. This stands apart from other means of identifying and extracting Twitter conversations that rely on hashtags for identifying tweets, which tend to miss large parts of the conversations. As a second step, this study also takes into consideration that conversations sometimes are divided into segments where new topics emerge. Through the identification of "bifurcations", i.e., parts of threads where Twitter conversations tend to take new directions, segments of a Twitter conversation can be partitioned off and treated as a coherent text to analyse. Such a treatment of parts of the conversation arguably makes it possible to find whether different issues are discussed in different parts of the thread or different perspectives on the same issue can be identified.

Building on the notions of issue mapping as an empiricist digital method for controversy analysis (Marres, 2015) we explore different network analysis-based techniques to identify, segment and measure user interactions conceived of as issues/controversies in Twitter conversation threads. In a paper presenting a checklist for the application of digital methods, Venturini et al. (2018) emphasised issues such as how the platform affords research, to what extent the study object plays out on the studied platform and if we are studying the object as it appears on the platform or if we use it as a proxy (e.g. for the public discourse). Given this, it is important to be aware of the affordances of Twitter at the time of the study. How is it possible to interact and how are conversations presented to the user? And how is it possible for a researcher to collect data? Moreover, we might view the interactions of Twitter as part of public discourse, but not a representation of it. By grounding the findings elsewhere (e.g. Rogers, 2013), we might get closer to a representation of the public discourse.

Through the analysis of Twitter conversations emanating from or including at least one reference to an academic paper, this study aims to further the understanding of the structure and content of Twitter conversations in the context of using them to identify the societal impact of research.

Literature review

The collection of Twitter data in the research literature has been mainly based on either hashtag-based or user-based methods. These methods only use tweets that contain a specific hashtag or keyword to identify the topic or limit the data collection to a set of users. How much of the conversations that are omitted by such methods has to our knowledge only been explored by D'heer et al. (2017) who saw their dataset of 1,719 tweets include 580 non-hashtagged replies and Lorentzen and Nolin (2017) who found an increase of 56 per cent new tweets through the inclusion of non-hashtagged tweets. Although the extent of the missing data will vary from topic to topic, using only hashtag or user-based data collection methods will inevitably render the data incomplete for a full understanding of the actual contents of the discourse.

While hashtags in a tweet can be compared to keywords in a scholarly article (Haunschild et al. 2018), at the same time, replies to, or retweets of other tweets, as well as mentions of a link (e.g. to a DOI) function as "internal" or "external" references, respectively (Haustein et al. 2014), thus corresponding to scholarly references. Mentions of another Twitter user (the @handle) does not have a clear corresponding function but serves both to signal an intended respondent as well as a means for highlighting interaction for this user, who would see an activity indicator in their Twitter interface. These different interactional aspects of the Twitter conversation are used in this study to grasp issues, sometimes in the form of controversies, highlighting both the interactive aspects of Twitter activity around tweets related to published research as their contents.

Collecting and analysing conversations in this sense is not common in Twitter research. Apart from the aforementioned works, Moon, Suzor and Matamoros-Fernandez (2016) found threads in a user-based set collected by Bruns, Burgess and Banks (2016) by following 2.8 million Australian Twitter accounts. Their study focused on conversations emanating from or including the word "uber". They argued that working with larger parts of texts would permit more comprehensive analysis of public opinion around a controversy and that analysis of these conversational threads contributes to a better understanding of social media communication. Another example of analysis of Twitter threads is provided by Zubiaga et al. (2016), who identified several internet rumours and then scraped the Twitter web interface for follow-on conversation attached to given tweets. The threads were then manually annotated as to whether the tweet was a rumour or an attempt to resolve the rumour as true or false.

Within altmetrics and similar areas, the focus has not been on the content and structure of conversations, however, but rather to what extent tweets can be used as a proxy for scientific impact. Even in an article using the term "conversation networks" in its title, Holmberg et al. (2014) explicitly state that it is not the full communication network, but rather the pairwise conversational connections that they study. Focusing on publications produced by Finnish researchers, Vainio and Holmberg (2017) found that those who referred to articles on Twitter "describe themselves more factually and by emphasizing their occupational expertise rather than personal interests". Didegah, Bowman and Holmberg (2018) studied factors behind altmetric scores compared to citations. Of special interest here is what makes a tweet about a research publication successful. Research funding was found to be most important, but journal impact factor and international collaboration also contributed to an increased number of tweets. Discipline-wise, research within medicine, natural sciences, and engineering and technology were more often tweeted than its counterpart within social sciences and humanities.

Nelhans and Lorentzen (2016) in a previous explorative study used a set of conversational threads that mentioned DOI references on Twitter to gain an understanding of the characteristics of the interaction and objects of discussion. Using both quantitative and qualitative methods, the authors characterised the objects of all mentioned DOI during a one-month period on Twitter. It was found that during the collection period articles and reviews from predominantly English-speaking countries at prestigious universities were mainly mentioned. 80 per cent of the mentioned literature was published in the same year as the data collection (which was done during the month of April). Content-wise, mentioned literature was heavily focussed on health, medicine, and the life sciences, as well as a broad range of social science topics, ranging from gender and learning, social media, and artificial intelligence, by way of social medicine and studies of the human condition, suggesting that broad social issues, was at the focus of interest of the tweeting users. In the study, a qualitative analysis of tweet practices was performed by categorising distinct conversational properties as "kinds" based on what was mentioned and how a DOI-referenced article was referenced. Specifically, different modes of discussing the contents, communicating about the context and different conversational practices were identified. In conclusion, it was found that digital object identifier URLs were mainly used for promoting a paper, as a conversation starter or as arguments in a discussion.

To a certain degree, contrasting findings regarding the promotional aspects of tweets were presented by Vainio and Holmberg (2017) who were only able to detect such use of tweets for marketing in the humanities and social sciences. Since that study did not focus on the conversational aspects of tweets, but on the user profiles mentioning highly tweeted articles, different aspects of the conversational practices were not studied. From the above, it is concluded that the study of the conversational properties of Twitter activity around scholarly publications are still in its infancy. The contribution by this study would be directed to social network analysis (SNA) aspects of Twitter conversations by expanding on the thread analysis of the structure and delineation of parts of conversation and identification of issues within the Twitter stream thread.

Method

Data collection was performed through filtering the Twitter stream using keywords and the most active users in the collected dataset, similarly to Nelhans and Lorentzen (2016), but instead of focusing only on DOIs, the present study tracked the keywords "dx doi org", dx.doi.org, arxiv.org, socarxiv.org, researchgate and academia.edu. This means that we did not attempt to collect data related to a particular topic, but rather any potential topic or discipline. Data collection started on August 23 2018 and ended two weeks later. When the data collection was finished, there were tweets in the database replying to tweets that had not been collected, most of them expected to be posted before the data collection started. The IDs of these missing tweets were put in a list and then used to query the statuses/lookup API endpoint, and if the tweet was a reply the ID of the new tweet was added to the list. This procedure added almost 10,000 new tweets, resulting in a total of 29,796 tweets. As tweets are identified by an ID and a reply is denoted by the ID of a tweet replied to, we can then string tweets together as a conversational thread. This procedure yielded a set of threads that varied in length and number of users in highly skewed distributions. The longest thread consisted of 1,458 tweets whereas the mean was 8.79 and the median 3. The largest number of participants was 59, with a mean of 2.7 and a median of 2. As noted in Nelhans and Lorentzen (2016), Twitter threads can take many different forms, including a chain-like, star-like or heavily bifurcated form, meaning that many new interactions could be identified where the discussion takes new directions. For further analysis, we chose two threads including 595 and 1024 tweets respectively, the first involving 30 participants and the second 28.

As few examples of Twitter research on conversational threads exist, a relevant aspect to explore is how to identify metrics for the activities. Such metrics could, for example, include the conversational impact of a tweet. One example of how a metric for statistical analysis of discussion threads was presented by Gómez, Kaltenbrunner and López (2008, p. 652-653): “the h-index h of a post is [...] the maximum nesting level i which has at least h > i comments, or in other words, h + 1 is the first nesting level i which has less than i comments.” However, this metric does not suit the conversational threads found on Twitter, where many threads involve bridges between tweets that spark a reaction from many users. Hence, a thread might start with one tweet replied to ten times (10 tweets at level 2), then one of these tweets is replied to once, and the reply is replied to once (1 tweet at levels 3 and 4), before this subsequent reply triggers a reaction with many replies. Similarly, we cannot rely on the nesting level only. A chain of 100 tweets without bifurcations would end up with a maximum nesting level of 100. In this exploratory work, we propose the identification of relevant threads to study based on the number of bifurcations resulting in at least two branches which include at least a total of 30 tweets.

To partition a conversational thread into segments of sufficient size, we initially identified all the bifurcations in the threads, that is, tweets replied to more than once. For each bifurcation, we then traversed the tree and counted the number of tweets emanating from the bifurcation. Following this step, we had tweet counts representing the number of posts from the point of a bifurcation, to the end of each of its branches. In the next step, we traversed the tree back to the root from the outermost bifurcations containing between 30 and 50 tweets, starting at the part of the thread which included the tweet with the latest timestamp. When reaching a previous bifurcation, we assigned the tweets belonging to that bifurcation a new segment ID if at least 30 tweets had been encountered. Finally, we joined the segments with one or two (depending on the size of the segment) adjacent segments to make the documents more suitable for automated text analysis. For the two examples included in this paper, this resulted in segments of varying sizes. While nine of 16 segments included between 100 and 125 tweets, the lengths of the segments varied from 65 to 157 tweets. The two threads were compared to each other using the cosine similarity measure, and we then focused on the longer thread to find out if it could be feasible to compare the segments within a thread with the rest of the thread. The texts were processed using the Porter stemmer and stop words were removed. In order to illustrate the topicality of the threads and the differences between the least similar segments, we created density maps of the documents using VOSviewer (van Eck and Waltman, 2014). Finally, for a topical analysis of the segments, we used Latent Dirichlet Allocation, LDA (e.g. Blei, Ng and Jordan, 2013). As training an LDA model with few documents, in this case, nine, renders instability, we tokenised the segments into sentences and used the sentences as training documents. From this, the most likely topic for each segment is induced as a list of ten terms ordered by a probability score. As each run results in a different set of terms for the topics, we trained the model in ten iterations and subsequently kept the ten terms most often coupled with each segment across the iterations.

Findings

Both threads are similar in that they both include large bifurcations, and they stretch over a few days, although thread 649 actually starts with a reply to a tweet posted more than three years before. Thread 1282 has a few more hubs which are tweets with many replies, but only a few replies result in larger branches. Topic-wise, the threads are different. The cosine similarity score for the thread comparison was 0.37. Judging by the density maps (Figure 1 and 2), thread 649 is about vaccination and related issues whereas thread 1282 is about learning, teaching and knowledge. The fact that the two threads differ much from each other comes as no surprise considering that the data collection was not restricted to one topic.

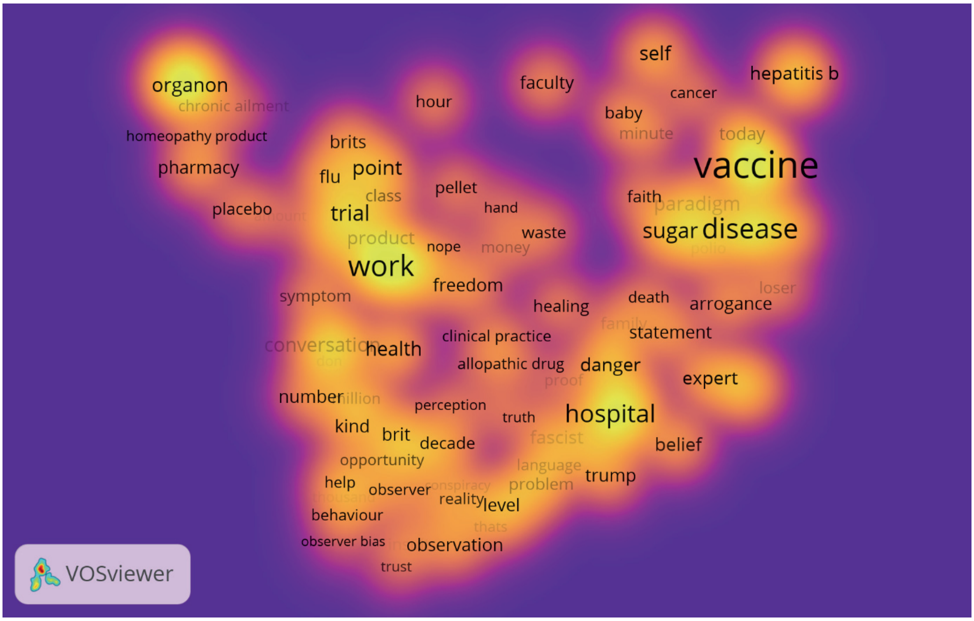


Figure 1. Density map of thread 649.

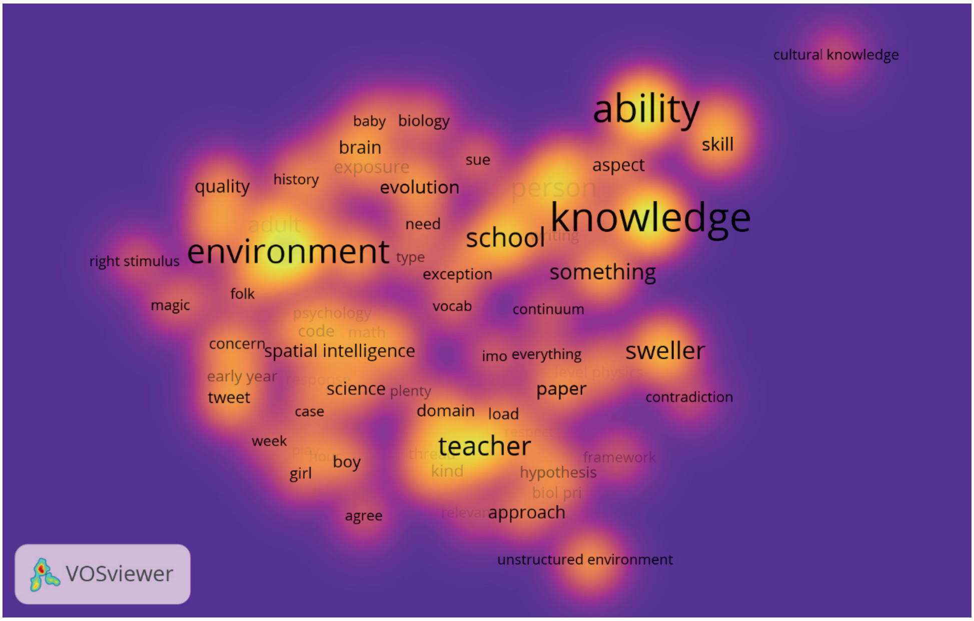


Figure 2. Density map of thread 1282.

We then focused on the longer of the threads for further analysis. The thread included many bifurcations which made it suitable for analysis of the topics within the discussion. Figure 3 shows its structure and the segments as identified by our algorithm. The arrow in the upper right corner points to the first tweet of the discussion.

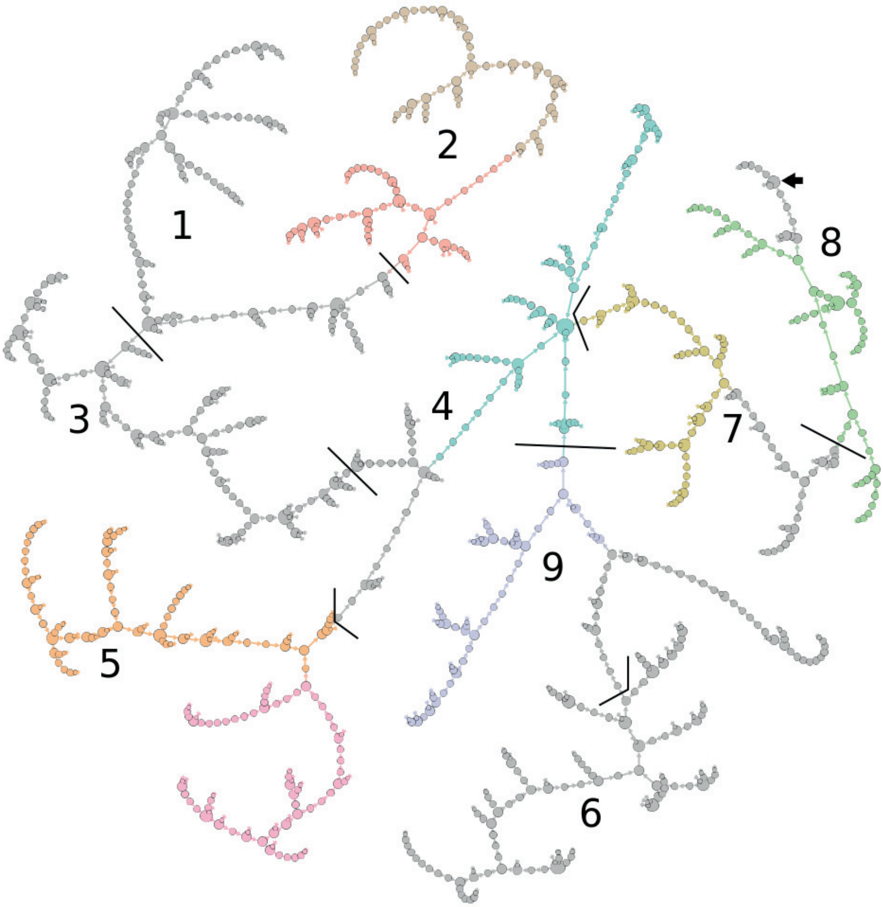


Figure 3. Thread 1282 with its segments. The lines denote where the thread is segmented. Nodes are sized according to the number of replies. The nodes are coloured according to clusters identified by the network analysis software Gephi.

The cosine similarity scores indicated that there were differences between the term frequency vectors representing the segments of the thread (Table 1). When comparing the different segments with each other, similarities were fairly low although they were still higher than the similarity between the threads. That is, the segments in thread 1282 differed less from each other than the two threads did. One hypothesis could be that smaller document sizes contributed to the lower scores as the segments deviated less from the thread when each segment was compared to the rest of the thread. With each segment treated as one document and the other eight segments as another document, the similarity scores were higher, ranging from 0.63 to 0.8.

When considering the top ten terms likely to be representative for each segment, we found numerous words in most of the segments, and it seems like the segments are quite similar to each other. Had the topic model resulted in lists of terms more distinct from each other, they could have been used as labels for the segments, however, in this case it seems as the segments do not differ much from each other according to the LDA model (Table 2).

Table 1. Cosine similarity scores between segments in thread 1282.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Avg. sim.*	*Segment vs. thread*
1	1	0.40	0.41	0.36	0.47	0.44	0.38	0.39	0.35	0.47	0.63
2	0.40	1	0.47	0.37	0.49	0.43	0.43	0.43	0.40	0.49	0.69
3	0.41	0.47	1	0.41	0.50	0.48	0.44	0.41	0.47	0.51	0.73
4	0.36	0.37	0.41	1	0.48	0.47	0.62	0.40	0.42	0.50	0.67
5	0.47	0.49	0.50	0.48	1	0.56	0.50	0.47	0.47	0.55	0.80
6	0.44	0.43	0.48	0.47	0.56	1	0.44	0.37	0.47	0.52	0.72
7	0.38	0.43	0.44	0.62	0.50	0.44	1	0.46	0.46	0.52	0.72
8	0.39	0.43	0.41	0.40	0.47	0.37	0.46	1	0.36	0.48	0.65
9	0.35	0.40	0.47	0.42	0.47	0.47	0.46	0.36	1	0.49	0.70

Table 2. Ten most likely terms for each segment according to the LDA model.

Segment	*Ten most likely terms*
1	think, language, children, learn, teaching, learning, theory, instruction, words, geary
2	think, geary, learn, children, learning, teaching, instruction, child, language, taught
3	geary, think, children, instruction, learning, years, theory, language, reading, way
4	knowledge, ability, explicit, primary, taught, biologically, instruction, think, reading, children
5	think, learning, language, geary, words, see, explicit, instruction, speech, teaching
6	think, language, geary, teaching, child, environment, children, point, speech, spoken
7	geary, think, instruction, reading, explicit, primary, words, said, learn, speech
8	think, instruction, explicit, learning, knowledge, primary, speech, biologically, read, teaching
9	geary, think, learn, teaching, learning, language, children, speech, different, instruction

Words such as “think”, “learning”, “language”, “child”, “children” and the researcher David Geary are occurring at the top end of the terms in multiple segments. This is an interesting finding in itself, as one would expect that the participating user cannot overview the entire thread, but this one stays on the same topic. Incidentally, the segments most similar according to the cosine similarity score are the neighbouring segments 4 and 7. However, the segment that deviates most from the rest of the discussion (1), is also the one farthest away from the start, i.e. the bifurcation with the latest timestamp. The next segment with the lowest similarity with the rest of the thread is the thread start (8). These two segments are also deviating more from the other parts in our segment vs. segment analysis, and they have a fairly low similarity score too. Rather than focussing on the top terms it would be more relevant to focus on the unique words. For example, “biologically” is included in two segments and “environment” in one. These three segments might reveal a different topic than the rest of the thread, if analysed with manual methods, such as quantitative or qualitative content analysis. Seemingly, based on the results of the analysis of these two threads, the method was able to identify what issues were discussed, but as the segments were found to be quite similar, signs of controversies were not detected. What we do not see here, given these analysis methods, is if the thread bifurcates because of some kind of disagreement. Although a bifurcation does not seem to imply a topical shift, other methods might reveal disagreements within the segments, which could be a sign of controversy.

Discussion

We have presented a method for automatic analysis of conversations on Twitter, emanating from or referring to a research publication. We propose dividing a conversational thread into segments where the thread bifurcates. With information retrieval techniques such as the cosine similarity measure and LDA modelling, these segments can be compared with each other. While this measure did highlight differences, a further adaptation for the type of content produced on Twitter is highly recommended. Such adaptation includes the use of a specialised stop word list. Another issue is to train the topic model on a larger body of texts and not just the one thread containing the shorter segments as documents. We would also wish to stress the need for improving the algorithm for segmenting the thread into smaller parts, and an investigation into the optimal size of the segment for automatic text analysis. Considering the article on digital methods and controversy studies by Marres (2015), we conclude that the method presented here is useful for identifying possibly controversial issues as they are discussed on Twitter, but that they then need to be analysed qualitatively or with more sophisticated machine learning methods. For example, a similar approach as the one taken by Buntain and Golbeck (2017), who used a feature-based method for automatic detection of fake news, could be adapted and applied to these threads. Furthermore, standing alone, an analysis of Twitter conversations does not say much more than how people interact on this platform. Grounding the findings in other analyses of other types of conversations is recommended.

While limited, the analysis of Twitter conversations regarding research articles does provide an indication of what type of research a part of the public is interested in, how it is referred to and how it is used as arguments in the discussions. It has for example been found that academic papers also are referred to for promoting ideological views (e.g. Vainio & Holmberg, 2017). We recommend further analysis of this by taking a more comprehensive approach. If focussing on Twitter, we suggest to collect data so that next to complete conversations can be studied, implementing methods similar to those presented here to identify and map possible issues or controversies, and then take the process one step further with an analysis of how the interactions play out and how research is used in the public domain. Particularly of interest would be to investigate the level of disagreement within a branch as well as among the branches.

Finally, we must acknowledge a couple of limitations to this study that should be addressed in future endeavours of this kind. Firstly, the selection of keywords should include "doi.org" as the prefix dx is not needed. Secondly, it is important that the researcher is aware of the presence of bots for further analyses of the discussions, an issue that has been discussed previously (i.e. Haustein et al., 2016; Robinson-Garcia et al., 2017). While we stress that material from bots must be included so that threads are not broken, including bot detection algorithms to present the likelihood that a tweet is posted by a bot would be an important contribution. For example, in an experiment comparing different machine learning algorithms, Haidermota, Mitra and Pansare (2018) concluded that bots seem to be more predictable regarding the timing of the reply to a tweet, while other indicators could be helpful, for example follower counts and usage of URLs. Another interesting option is to make use of the application *Botometer*, previously known as *Botornot* (Davis et al., 2016). If we can identify bots prior to the conversation analysis, then we can also learn more about how bots participate in Twitter discussions, as well as how other users interact with them.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531.

References

Buntain, C., & Golbeck, J. (2017). Automatically Identifying Fake News in Popular Twitter Threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud),* New York, USA, Nov 3-5. IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research. 3 (4–5)* (pp. 993–1022).

Bruns, A., Burgess, J., & Banks, J. (2016). *TrISMA: Tracking Infrastructure for Social Media Analysis*. Retrieved May 27, 2019 from: https://trisma.org.

D'heer, E. et al. (2017). What are we missing? An empirical exploration in the structural biases of hashtag-based sampling on Twitter. *First Monday, 22*(2). Retrieved May 27, 2019 from: https://firstmonday.org/ojs/index.php/fm/article/view/6353/5758.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 273-274).

Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the differences between citations and altmetrics: An investigation of factors driving altmetrics versus citations for Finnish articles. *Journal of the Association for Information Science and Technology, 69*(6) (pp. 832-843).

Gómez, V., Kaltenbrunner, A., & López, V. (2008). Statistical analysis of the social network and discussion threads in Slashdot. Paper presented at the *WWW '08 Proceedings of the 17th international conference on World Wide Web.* Beijing, China, April 21-25 (pp. 645-654).

Haidermota, M., Mitra, i., & Pansare, A. (2018). Classifying Twitter user as a bot or not and comparing different classification algorithms. *International Journal of Advanced Research in Computer Science*, 9(3) (pp. 29-33).

Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2018). Does the public discuss other topics on climate change than researchers? A comparison of networks based on author keywords and hashtags. Retrieved May 27, 2019 from: https://arxiv.org/abs/1810.07456 (pp. 24).

Haustein, S., Bowman, T. D., Holmberg, K., Peters, I., & Larivière, V. (2014). Astrophysicists on Twitter: An in-depth analysis of tweeting and scientific publication behaviour. *Aslib Journal of Information Management*, 66(3) (pp. 279-296).

Haustein, S., Bowman, T., Holmberg, K., Tsou, A., Sugimoto, C., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology*, 67(1) (pp. 232–238).

Holmberg K, Bowman TD, Haustein S, Peters I (2014) Astrophysicists' Conversational Connections on Twitter. *PLoS ONE 9*(8): e106086.

Lorentzen, D. G., & Nolin, J. (2017). Approaching Completeness: Capturing a Hashtagged Twitter Conversation and its Follow-On Conversation. *Social Science Computer Review, 35*(2) (pp. 277-286).

Marres, N. (2015). Why Map Issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values, 40*(5) (pp. 655-686).

Moon, B., Suzor, N., & Matamoros-Fernandez, A. (2016). Beyond Hashtags: Collecting and Analysing Conversations on Twitter. Paper presented at *AoIR 2016: The 17th Annual Meeting of the Association of Internet Researchers.* Berlin, October 5-8. Germany: AoIR.

Nelhans, G. & Lorentzen, D. G. (2016). Twitter conversation patterns related to research papers. *Information Research, 21*(2), paper SM2. Retrieved May 27, 2019 from: http://www.informationr.net/ir/21-2/SM2.html.

Robinson-Garcia, N., Costas, C., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting - about journal articles. *PLoS ONE*, 12(8), e0183551.

Rogers, R. (2013). *Digital methods*. Cambridge, Massachusetts: The MIT Press.

Vainio, J. & Holmberg, K. (2017). Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions, *Scientometrics, 112*(1) (pp. 345.366).

van Eck, N.J. & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring scholarly impact: methods and practice*. Berlin: Springer (pp. 285-320).

Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A reality check(list) for digital methods. *New Media & Society, 20*(11) (pp. 4195–4217).

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLoS ONE 11*(3): e0150989.

Internationally mobile scientists as knowledge transmitters – A lexical-based approach to detect knowledge transfer

Valeria Aman[1]

[1] *aman@dzhw.eu*
German Centre for Higher Education Research and Science Studies GmbH, Schützenstraße 6a,
10117 Berlin (Germany)

Abstract

This paper explores the knowledge transfer of internationally mobile scientists. It builds upon previous work dealing with the development of methods to measure the knowledge transfer of German scientists. Using abstract terms of publications covered in Scopus, this paper proposes a lexical-based approach to identify knowledge transmitters. These scientists are characterized by adapting knowledge of their co-workers during their international stay and transferring it upon return to other German co-workers. Knowledge is operationalized as the co-occurrence of rarely used abstract terms. Knowledge transfer is expressed as the diffusion of these term combinations in co-authorship networks. Due to distinct mobility and knowledge transfer practices among disciplines, the paper considers scientists in Chemistry, Computer Science, Medicine and Physics & Astronomy. The findings suggest that in all disciplines except for Computer Science one in four scientists act as knowledge transmitters between the country of destination and Germany. Results also show that in some disciplines knowledge transmitters are significantly younger than non-knowledge transmitters and more productive after their international mobility.

Introduction

Scientists are knowledge carriers that bear specific skills and talents. The mobility of these knowledge carriers in a globalization process resulting from the opening of national resources in a knowledge-based economy drives innovation and economic growth (Meyer 2001). The circulation of academic human capital can benefit both the sending and recipient countries and increase academic achievements (Musselin 2004).

Internationalization and extended foreign experiences gain increasingly more attention in the discussion on successful career trajectories. Especially at the early-career stage, a stay abroad can play an important role in the advancement of a career. In some disciplines, international mobility is considered as a taken-for-granted academic convention (Bauder 2012) or as an "expectation" in the academic labour market and part of the "excellence" requirement (Morano-Foadi 2005, p. 145). Institutional policies not only expect but also enable international mobility which scientists embrace. In Germany, the *German Academic Exchange Service* (DAAD), the *German Research Foundation* (DFG), the *Max-Planck Society*, the *Humboldt Foundation* and other organizations are the main funders of scientific mobility (Bauder et al. 2016). As an example, the *Emmy Noether Programme* funded by the DFG gives outstanding early career scientists the chance to qualify for the post of professor by leading an independent junior research group up to four years after the completion of the doctorate. The eligibility requirements, however, include substantial international research experience. In contrast, evidence of scientific benefits that international research experience can bring are often missing. With the words of Ackers (2008, p. 420), "it is clear that the quality of the mobility experience is often less important than the fact of mobility".

In the light of this quote, the objective of the underlying paper is to analyse the quality of international mobility by examining the transfer of knowledge as a major benefit of working abroad. Therefore, I propose an automated method that is capable of indicating knowledge transfer of internationally mobile scientists. The method draws on abstracts of scientific publications and a network-based approach to identify knowledge transmitters. This method allows to show to what degree German scientists act as knowledge transmitters, in the sense that they acquire knowledge during their international stay abroad and transfer that knowledge to their home country after return.

The paper builds on recent work on the potential to measure scientific international mobility with bibliometric data alone (e.g. Moed et al. 2013; Conchi & Michels 2014; Halevi et al. 2016; Aman 2018b; Robinson-Garcia et al. 2018) and social science studies on international mobility (e.g. Collins 1974, 2001; Ackers 2005; Bauder et al. 2016). Moreover, this paper extends previous approaches of the measurement of knowledge transfer. Earlier studies relied on citation linkages between publications that assumingly imply a flow of knowledge from the cited to the citing publication (e.g. Van Leeuwen & Tijssen 2000; Hassan & Haddaway 2013). The contribution of this paper is to use bibliometric data alone to identify internationally mobile scientists and to measure their knowledge transfer operationalized by lexical terms instead of relying on dichotomous citation flows. The remainder of this paper is organized as follows: After providing a theoretical background discussing the concept of knowledge, knowledge transfer and scientific collaboration, I describe the data and explain how knowledge transfer is operationalized. The presentation of results is followed by a conclusion and outlook.

Theoretical background

Scientific knowledge can manifest itself in different forms such as intellectual capital, as research processes or products of research activities. Knowledge mainly refers to an individual's personal stock of information, skills, experiences, beliefs and memories (Alexander et al. 1991). This knowledge is idiosyncratic and depends on the personal biography of an individual. Knowledge can be distinguished into declarative knowledge described as "knowing what", or procedural knowledge described as "knowing how" (ibid.). The concept of knowledge is complex and can be explored from various perspectives such as a philosophical, sociological, organizational or technical perspective (Williams & Baláž 2008). A simple distinction of knowledge goes back to Polanyi, describing knowledge as something that can be possessed and transferred in explicit or tacit form (Polanyi 1966). Whereas explicit knowledge can be stored in text, transferred and understood, tacit knowledge is embedded in practices and cannot be easily articulated. This basic distinction of knowledge does not prevent from seeing these two knowledge types as continuous and overlapping. A more sophisticated distinction was proposed by Collins (1993) distinguishing five types of knowledge, of which *embrained* and *embodied* knowledge require a bearer of the knowledge, *encultured* and *embedded* knowledge rely on social and cultural practices and *encoded* knowledge on media to transmit the knowledge.

Whatever definition of knowledge is used, knowledge is embedded in systems of socially constructed signs (Williams & Baláž 2008) and its production and transfer lean on social institutions, shared communications and common interpretations (Gherardini & Nucciotti 2017). Knowledge transfer implies a focused and unidirectional communication of knowledge between individuals. The recipient of the transferred knowledge is supposed to have a cognitive understanding and the ability to apply the knowledge.

Knowledge transfer knows no national borders and geographic mobility plays an important role in the acquisition and recombination of knowledge (Laudel 2003; Gläser 2006). International mobility is also a strategy to access social research networks that enable personal contact and face-to-face interaction which are essential for the transfer of knowledge (Williams & Baláž 2008). Previous work shows that co-location and personal ties facilitate localized knowledge spillovers (Collins 1974; Jöns 2009).

The expectation, the need and the value of mobility differ between disciplines. Whereas some disciplines involve locally contextualized knowledge and therefore tend to be "place-specific"

(Jöns 2007, p. 109), other disciplines tend to make universal knowledge claims and rely on high standardization. Not only the need to access large-scale facilities (e.g. in Physics) can encourage international mobility (Ackers 2005) but also the knowledge type. If the knowledge to be transferred is tacit, geographic mobility can enable personal contact, observation of colleagues and interaction (Collins 1974, 2001).

One major function of international mobility is to gain access to research groups that are crucial units in the production of knowledge (Gläser & Laudel 2001). Research groups bring scientists together to observe, discuss and to combine elements of existing knowledge. Joint research publications resulting from successful integration of internationally scientists into research groups abroad are likely to be indicators of knowledge transfer (Laudel 2002). Co-authorship can facilitate the transfer of tacit knowledge by observation, informal knowledge communicated on demand or formal knowledge codified in publications (Gläser 2006).

Based on these theoretical foundations the method to be presented relies on co-authorship networks. The network-based approach is able to identify knowledge transmitters who act as a bridge between countries. The mobility phase ranges up to three years, because longer stays are associated with the transfer of more complex and tacit forms of knowledge (Edler et al. 2011) and a greater opportunity to interact with scientists in the host country. To account for discipline-specific characteristics in the process of knowledge transfer, the study focuses on four research disciplines.

Data and methods

The data for this study build upon previous work (Aman 2018a) in which knowledge transfer was measured on the basis of similarity, i.e. internationally mobile scientists and non-internationally mobile scientists from Germany were compared in terms of their knowledge base and their similarity towards different types of co-authors.

The identification of scientists relies on Scopus author ID (Elsevier), a powerful algorithm to disambiguate author names. The author ID is supposed to combine all publications of an author under a single ID to handle common first and last names. Previous studies report that Scopus author ID enables reliable author name disambiguation (Moed et al. 2013; Conchi & Michels 2014; Aman 2018b). Mobility is measured by using the address information of publication data. Internationally mobile scientists are identified as those whose affiliation changes from one country to another, whereby the country relates to the geographic location of the institute as stated in publications.

The underlying dataset focuses on scientists who have published the majority of their publications from German institutions within the time period 2007 to 2015. They are referred to as *German scientists* independent of their nationality. I distinguish three time periods: a pre-mobility phase ranging from 2007 to 2009, in which scientists have published exclusively from German institutions; a mobility phase between 2010 and 2012, in which they were abroad, and a post-mobility phase ranging from 2013 to 2015, in which they are exclusively affiliated to German institutions. The choice of the time periods and the limitation to four disciplines with the highest number of scientists diminishes the dataset to 419 German scientists.

To measure the knowledge transfer of these internationally mobile scientists, I work with abstracts of their publications and those of their co-authors in 2007 to 2015 as covered in Scopus. The Scopus data is integrated in a licensed in-house database version at the *Competence Centre for Bibliometrics*[1] facilitating extensive computations.

In a first step, all abstract terms were automatically extracted from abstracts of internationally mobile scientists and their co-authors. The terms extracted are lower case and only consist of alphabetical strings. The publication id, the term extracted and the number of occurrence of a

[1] Competence Centre for Bibliometrics. http://www.forschungsinfo.de/Bibliometrie/en/index.php?id=home

term per publication were determined. In a following step, the term frequency (TF) and the inverse document frequency (IDF) of the abstract terms were computed on the basis of more than 2 million Scopus publications from the years 2007 to 2015. The product of TF and IDF informs about the weight of a term and about the relevance in the database. The higher the score the rarer the term and vice versa.

To operationalize knowledge transfer, I restricted the corpus to abstract terms with a relatively high score. With *corpus* I refer to the set of abstracts of German authors and all their co-authors in 2007 to 2015. The number of distinct terms in the corpus is 68,242 and the IDF score ranges from 0.01 (e.g. the terms *of* and *the*) to 14.65 (e.g. the terms *chloropropionates, dictyophycus, zontivity*). To filter out stop words such as *the, with, from* or *with* that bear no knowledge, I restricted the terms to a score higher than 5.0. Examples of terms with a score of 5.0 in the corpus are: *coli, fourier, lymph* or *semiconductor*. Restricting the overall number of terms in the corpus (852,074) to those with an IDF score greater than 5.0 reduces the dataset to 367,816 terms. Thus, only 43.18% (367,816/852,074) of all terms used in abstracts of the corpus are considered as specific enough to operationalize knowledge transfer. Due to their seldom occurrence they can be traced back to scientists using them. Knowledge itself is operationalized as the co-occurrence of two of these rare terms. To measure the transfer of knowledge, the first publication year of the co-occurrence of two terms was determined. Those two-term combinations are indicative of knowledge transfer which were used by the co-workers abroad between 2007 and 2009, thus, before an interaction had taken place between mobile scientists and co-workers abroad. In addition, these two-term combinations must have been used by the mobile scientists abroad - wherefrom we can derive that the knowledge has been picked up by the mobile scientists and was immediately used in publications abroad that are not co-authored with the previous users of the knowledge.

A further condition is that internationally mobile scientists act as knowledge transmitters between the country of mobility and Germany. Therefore, internationally mobile scientists have to transfer the knowledge they acquired abroad in 2010-2012 to their co-workers in Germany with whom they co-published between 2013 and 2015.

Note that there is a publication delay between working on a paper and its publication ranging from weeks to years. However, the publication delay is assumed to be similar within a discipline and inherent in every year so that the bibliometric data consistently reflect past events.

The identification of co-workers of German scientists relies on the institutional coding that exists for all German institutions (Winterhager 2014). The co-workers abroad are determined as those who are co-authors of the mobile scientists and are affiliated at the same institution abroad.

Findings

This section starts with the illustration of how knowledge transfer is measured. Figure 1 depicts a co-authorship network of scientists who use a two-term combination in the abstracts of their publications. The nodes representing authors are connected if they have co-published independent of time or theme. The closer the nodes the more co-authored publications two scientists have. The larger the node the more publications exist on the topic expressed by a two-term combination.

The edges represent the knowledge transfer from one node to another and the arrows of the edges support the idea that the knowledge transfer is directed. The darker the color the later the year in which the knowledge was transferred from one node to another.

Figure 1 shows that the co-worker abroad transferred knowledge to the mobile scientist who acts as a knowledge transmitter between the co-worker abroad and those with whom he worked upon return in Germany. Thus, we can speak of knowledge transfer because there are nodes in different countries that are connected by a node that acts as a transmitter of knowledge.

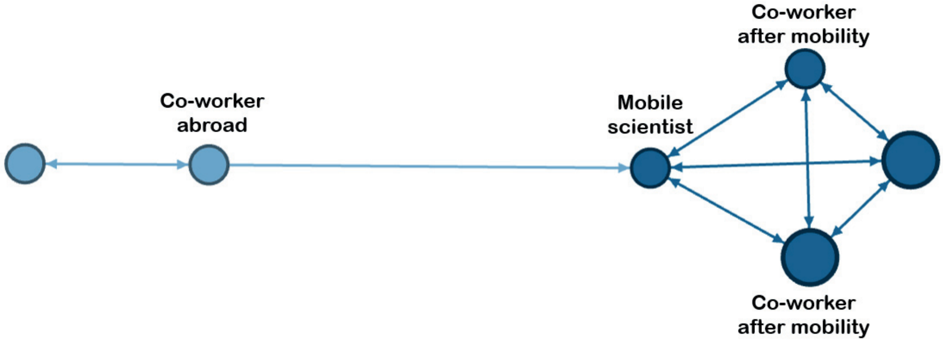


Figure 1. Co-authorship network using a specific two-term combination. The internationally mobile scientist acts as a knowledge transmitter between the co-worker abroad and the co-workers in Germany.

In the following, descriptive results of knowledge transmitters (KT) and non-knowledge transmitters (non-KT) are presented. Table 1 shows the total number of scientists by discipline as well as the amount of KT-scientists and non-KT-scientists. The majority of internationally mobile scientists work in Medicine, followed by Physics & Astronomy, Computer Science, and Chemistry. The share of knowledge transmitters is strikingly similar in all but one discipline. Whereas in Chemistry, Medicine, and Physics & Astronomy one quarter of German scientists function as knowledge transmitters, only a share of 14 percent of computer scientists transfer their knowledge from abroad to Germany.

Table 1. Overview of the number of knowledge transmitters (KT) and non-knowledge transmitters (non-KT) by discipline.

Discipline	No. of scientists	No. of KT	No. of non-KT	Share of KT in %
Chemistry	49	13	36	26.5
Computer Science	50	7	43	14.0
Medicine	195	50	145	25.6
Physics & Astronomy	125	36	89	28.8

The smaller share of knowledge transmitters in Computer Science may be due to the way knowledge is produced in this discipline. Interaction and observation may not be as important as in the other three disciplines and the standardized vocabulary may hamper the identification of knowledge transmitters.

An important role in the process of knowledge transfer plays the country of the international stay. Table 2 provides an overview of the main destination countries of all KT-scientists and non-KT-scientists. Note that a scientist may have been to more than one country within the period 2010-2012. The table reveals that the USA is the most popular destination country across all disciplines. Other important countries of destination are the German-speaking neighbouring countries Switzerland (CHE) and Austria (AUT) and the English-speaking countries Great Britain, Canada and Australia. In general, the order of the countries of KT-scientists vs. non-KT-scientists is similar. It bears mentioning that in Physics & Astronomy, Italy seems to play an important role in the knowledge transfer.

Table 2. Top destination countries of knowledge transmitters (KT) and non-knowledge transmitters (non-KT) in the dataset moving to the country listed. Countries of the same rank are in alphabetical order and may be subsumed in 'Other'.

Chemistry				Computer Science				Medicine				Physics & Astronomy			
KT		*non-KT*		*KT*		*non-KT*		*KT*		*non-KT*		*KT*		*non-KT*	
USA	11	USA	16	USA	6	USA	12	USA	31	USA	54	USA	16	USA	34
AUS	1	ESP	4	CHE	2	NLD	4	GBR	13	GBR	29	ITA	11	CHE	22
AUT	1	AUS	3	FRA	2	AUT	3	CHE	8	CHE	18	CHE	9	GBR	19
BEL	1	JPN	3	ITA	2	GBR	3	CAN	7	AUT	16	FRA	9	JPN	14
CAN	1	FRA	3	AUS	1	CHE	3	DNK	4	CAN	12	GBR	9	ITA	10
Other	2	Other	16	Other	4	Other	19	Other	23	Other	90	Other	32	Other	72

Aside from the country of international stay the career stage may play another role in the transfer of scientific knowledge. Therefore, Table 3 compares the average first year of publication according to Scopus data for internationally mobile scientists who act as knowledge transmitters and those who do not. It is assumed that the first publication in Scopus occurs typically during the time of dissertation or shortly thereafter. The results show that the average first publication year of knowledge transmitters (KT) in Chemistry and Computer Science is 2006. This result is in accordance with the delineation of the dataset, where it is assumed that a pre-mobility phase in 2007-2009 is followed by a post-doc phase abroad (2010-2012) in which knowledge is adapted and transferred to other co-workers upon return in Germany (2013-2015). In Medicine and Physics & Astronomy the research age of knowledge transmitters is on average higher than that of chemists and computer scientists and the standard deviation shows that the age is more dispersed. From that it follows that the mobility phase in which these scientists act as knowledge transmitters occurs at a later stage of career. The table also shows that except for Physics & Astronomy scientists who do not transfer knowledge are on average older than knowledge transmitters. One can infer that a research stay abroad at a later stage of the career does not serve the purpose to transfer knowledge. The last column corroborates the finding that knowledge transmitters in Chemistry, Computer Science and Medicine are significantly younger than scientists who do not transfer knowledge.

Table 3. Overview of the average first publication year (avg. first py) according to Scopus data of knowledge transmitters (KT) and non-knowledge transmitters (non-KT).

Discipline	**Avg. first py of KT (Stddev)**	**Avg. first py of non-KT (Stddev)**	**Two-sided p-value (first_py of KT > first_py of non-KT)**
Chemistry	2006 (1.3)	2001 (4.6)	>0.000
Computer Science	2006 (2.2)	2002 (4.1)	0.004
Medicine	2002 (4.8)	2000 (4.4)	0.005
Physics & Astronomy	2002 (4.9)	2002 (4.7)	0.854

Table 4 compares the productivity of KT-scientists and non-KT-scientists after their international stay (2013-2015). Results show that in each discipline KT-scientists have a higher number of publications than non-KT-scientists. However, the results are only significant for scientists in Medicine and Physics & Astronomy. One explanation for the finding could be that those scientists who transfer knowledge are interacting more with peers which makes them more productive especially in terms of co-authorship. However, additional analyses show that KT-scientists do not have necessarily more publications in the pre-mobility phase (2007-2009), the mobility phase (2010-2012) or the total publication period 2007-2015.

Table 4. Overview of the average number of publications of knowledge transmitters (KT) and non-knowledge-transmitters (non-KT) in the post-mobility phase 2013-2015.

Discipline	Avg. no. of publ. of KT (Stddev)	Avg. no. of publ. of non-KT (Stddev)	Two-sided p-value (No. of publ. of KT > No. of publ. of non-KT)
Chemistry	12.9 (8.3)	10.7 (12.0)	0.478
Computer Science	12.0 (10.1)	7.7 (6.3)	0.311
Medicine	19.5 (13.5)	12.6 (13.5)	0.002
Physics & Astronomy	20.0 (12.8)	11.1 (12.4)	0.001

Another indicator that characterizes KT-scientists and non-KT-scientists is their citation impact. Table 5 presents the average CPP (citations per paper) of papers that were published abroad within 2010-2012. The citation window is open and ranges from the year of publication up to 2015. Only Medicine yields a significantly higher CPP for KT-scientists than for non-KT-scientists. This finding implies that publications using the knowledge adapted abroad are higher cited than publications without evident knowledge transfer.

Table 5. Overview of the average CPP (citations per paper) of publications of knowledge transmitters (KT) and non-knowledge transmitters (non-KT) published in 2010-2012 and cited between 2010 and 2015.

Discipline	Avg. CPP of KT (Stddev)	Avg. CPP of non-KT (Stddev)	Two-sided p-value (CPP of KT > CPP of non-KT)
Chemistry	11.9 (3.8)	12.2 (7.7)	0.840
Computer Science	3.6 (1.4)	3.8 (3.1)	0.729
Medicine	16.9 (19.6)	10.4 (7.9)	0.027
Physics & Astronomy	20.5 (19.9)	16.0 (14.7)	0.218

Discussion and conclusions

Despite the important role that international mobility plays in the careers of scientists, knowledge transfer as one positive outcome has been insufficiently studied. There are still some gaps in our understanding of the contribution of international mobility to the transfer of knowledge and the processes involved. This might be due to limited methods available to trace knowledge transfer.

The underlying paper proposed a lexical-based approach to capture knowledge and used a network-based method to identify knowledge transmitters. To this end, I used bibliometric data of German scientists who have been internationally mobile between 2010 and 2012. The model presented relies on co-authorship networks because these build on social networks which play an important role in the production and transfer of knowledge.

To ascertain that the knowledge transfer can be traced back to international mobility, the model takes the combinations of rarely used abstract terms into account. The idea behind is that rarely used term combinations are adapted by German scientists being abroad and transferred at a later time to co-workers in Germany.

The diffusion of these two-term combinations does not necessarily mean that knowledge transfer has taken place, because the knowledge represented by these term combinations can be also individually acquired without ever interacting. On the contrary, knowledge transfer processes can elude the adaptation of term combinations. However, the use and passing on of two-term combinations representing specific knowledge indicate the transfer of knowledge at least to some degree.

The findings suggest that there are knowledge transmitters among German scientists who adapt knowledge abroad that they transfer to their co-workers in Germany. Surprisingly, the share of knowledge transmitters in Chemistry, Medicine and Physics & Astronomy is similar with one quarter of the scientists. The approximate research age indicates that knowledge is rather transferred in the early career where a stay abroad is associated with a post-doc and intensive research to establish a research trail. The findings also showed that knowledge transmitters in Medicine and Physics & Astronomy are more productive. Their productivity may be ascribable to their overall better performance at top institutions which makes them eminently suitable for transferring knowledge. However, the sample size is rather small and reduces the generalizability of the results.

Although international mobility experience is a prerequisite for higher positions in the careers of scientists, we still know little about the quality of mobility. We cannot expect that mobile scientists go abroad with the specific intention of bringing back knowledge to apply in their home country. However, the findings suggest that international mobility helps to gain new knowledge that is possibly bound to the host institutions abroad. Internationally mobile scientists thus act as knowledge brokers who, through their experience of doing research in more than one country, are able to identify knowledge in one place that can be transferred and applied in another (Coey 2018).

The limitations underlying this study are related to the use of bibliometric data that measure the mobility of scientists, their research age and knowledge transfer only by approximation. Nonetheless, the study can be seen as a step forward in identifying knowledge transmitters and describing the transfer of knowledge. Interviewing the identified knowledge transmitters about their perceived knowledge acquisition and transfer and comparing it with bibliometric findings would enable the validation of the method presented. Apart from international mobility, the method presented can be applied at other levels of analysis, such as the mobility across institutions and sectors or between research fields. It could be also adapted to analyze the spread of diseases in epidemiology studies.

To conclude, the present study provides evidence on the knowledge transfer of German scientists who were internationally mobile. This issue is novel and important as there are no profound methods to measure knowledge transfer. It remains crucial to develop more advanced methods to identify knowledge transmitters and to understand the process of knowledge transfer across borders.

Acknowledgments

The study was funded by the German Federal Ministry of Education and Research (grant number: 01PQ16002). I acknowledge the use of the data from the DZHW Competence Centre for Bibliometrics (grant number: 01PQ17001). I would like to thank three anonymous reviewers for their suggestions and comments.

References

Ackers, L. (2005). Moving People and Knowledge: Scientific Mobility in the European Union. *International Migration*, *43*(5), 99–129.

Ackers, L. (2008). Internationalisation, Mobility and Metrics: A New Form of Indirect Discrimination? *Minerva*, *46*(4), 411–435. https://doi.org/10.1007/s11024-008-9110-2

Alexander, P. A., Schallert, D. L., & Hare, V. C. (1991). Coming to Terms: How Researchers in Learning and Literacy Talk About Knowledge. *Review of Educational Research*, *61*(3). http://journals.sagepub.com/doi/abs/10.3102/00346543061003315

Aman, V. (2018a). A new bibliometric approach to measure knowledge transfer of internationally mobile scientists. *Scientometrics*, *117*(1), 227–247. https://doi.org/10.1007/s11192-018-2864-x

Aman, V. (2018b). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics*, *117*(2), 705–720. https://doi.org/10.1007/s11192-018-2895-3

Bauder, H., Hannan, C.-A., & Lujan, O. (2016). International Experience in the Academic Field: Knowledge Production, Symbolic Capital, and Mobility Fetishism. *Population, Space and Place*, *23*(6). https://doi.org/10.1002/psp.2040

Bauder, Harald. (2012). The International Mobility of Academics: A Labour Market Perspective. *International Migration*, *53*(1), 83–96. https://doi.org/10.1111/j.1468-2435.2012.00783.x

Coey, C. (2018). International researcher mobility and knowledge transfer in the social sciences and humanities. *Globalisation, Societies and Education*, *16*(2), 208–223. https://doi.org/10.1080/14767724.2017.1401918

Collins, H. (1993). The Structure of Knowledge. *Social Research*, *60*, 95–116.

Collins, H. M. (1974). The TEA Set: Tacit Knowledge and Scientific Networks. *Science Studies*, *4*, 165–186.

Collins, H. M. (2001). Tacit Knowledge, Trust and the Q of Sapphire. *Social Studies of Science*, *31*(1), 71–85. https://doi.org/10.1177/030631201031001004

Conchi, S., & Michels, C. (2014). *Scientific mobility: An analysis of Germany, Austria, France and Great Britain*. https://www.isi.fraunhofer.de/content/dam/isi/dokumente/ccp/innovation-systems-policy-analysis/2014/discussionpaper_41_2014.pdf

Edler, J., Fier, H., & Grimpe, C. (2011). International scientist mobility and the locus of knowledge and technology transfer. *Research Policy*, *40*, 791–805. https://doi.org/10.1016/j.respol.2011.03.003

Gherardini, A., & Nucciotti, A. (2017). Yesterday's giants and invisible colleges of today. A study on the 'knowledge transfer' scientific domain. *Scientometrics*, *112*(1), 255–271. https://doi.org/10.1007/s11192-017-2394-y

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften. Die soziale Ordnung der Forschung*. Frankfurt/New York: Campus.

Gläser, J., & Laudel, G. (2001). Integrating Scientometric Indicators into Sociological Studies: Methodical and Methodological Problems. *Scientometrics*, *52*(2), 414–434.

Halevi, G., Moed, H. F., & Bar-Ilan, J. (2016). Researchers' Mobility, Productivity and Impact: Case of Top Producing Authors in Seven Disciplines. *Public Research Quarterly*, *32*(1), 22–37. https://doi.org/10.1007/s12109-015-9437-0

Jöns, H. (2007). Transnational mobility and the spaces of knowledge production: a comparison of global patterns, motivations and collaborations in different academic fields. *Social Geography*, (2), 97–114.

Jöns, H. (2009). 'Brain circulation' and transnational knowledge networks: studying long-term effects of academic mobility to Germany, 1954–2000. *Global Networks*, *9*(3), 315–338.

Laudel, G. (2002). Collaboration and reward. What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3–15. https://doi.org/10.3152/147154402781776961

Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help? *Scientometrics*, *57*(2), 215–237. https://doi.org/10.1023/A:1024137718393

Meyer, J.-B. (2001). Network Approach versus Brain Drain: Lessons from the Diaspora. *International Migration*, *39*(5), 91–110. https://doi.org/10.1111/1468-2435.00173

Moed, H. F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*(3), 929–942. https://doi.org/10.1007/s11192-012-0783-9

Morano-Foadi, S. (2005). Scientific Mobility, Career Progression, and Excellence in the European Research Area. *International Migration*, *43*(5), 133–162.

Musselin, C. (2004). Towards a European Academic Labour Market? Some Lessons Drawn from Empirical Studies on Academic Mobility. *Higher Education*, *48*(1), 55–78. https://doi.org/10.1023/B:HIGH.0000033770.24848.41

Polanyi, M. (1966). The Logic of Tacit Inference. *Philosophy*, *41*(155), 1–18.

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., & Costas, R. (2018). *The many faces of mobility: Using bibliometric data to measure the movement of scientists*. 22.

van Leeuwen, T., & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows. *Research Evaluation*, *9*(3), 183–187. https://doi.org/10.3152/147154400781777241

Williams, A., & Baláž, V. (2008). *International Migration and Knowledge*. Abgerufen von https://www.amazon.de/International-Migration-Knowledge-Routledge-Geography/dp/0415434920

Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014). Institutionenkodierung als Grundlage für bibliometrische Indikatoren. *Bibliometrie - Praxis und Forschung*, *3*(14), 1–22.

Evaluating the evaluators: when academic citizenship fails

Katerina Guba[1] and Angelika Tsivinskaya[2]

[1] *kguba@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[2]*atsivinskaya@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

Abstract

This article questions how the pressure of legal accountability with its emphasis on compliance with proper procedures creates the opportunity for the making the extra-role behavior of academics valuable at the expense of research achievements. We examine the effects of the experiment of the Russian state agency to create the community of experts to assist the state in reducing the number of low-quality higher education institutions. Results based on the bibliometric data from Russian Index of Science Citation available for 554 experts. In the analysis, we use the strategy of comparing the research performance of experts with the research performance of non-experts employed in Russian universities. Results of the empirical analyses indicate that Russian academics whose performance is low in respect with publications and citations in the selective journals are more likely to become an expert engaged in academic citizenship in the form of regulatory activity. In this respect, engagement in citizenship behavior could be considered as a compensatory mechanism (Bergeron 2007) according to which individuals contribute something if they are less able to contribute to core tasks.

Introduction

The effects of performance management and indicators on organizations are a well-developed area of research. Most research considers the effects of professional accountability (Romzek, 2016) with its focus on the outcome control mechanism. National evaluation systems based on outcome control (Ouchi, 1980; Sihag & Rijsdijk, 2018) have the potential to influence knowledge production most. Such implications as goal displacement, task reduction, and potential biases toward interdisciplinary research are well-documented (Rijcke et al., 2015). The effects of behavior control mechanisms have not received much attention even though they are a significant source of transformation for existing hierarchies within academic organizations. This article questions how the pressure of legal accountability with an emphasis on compliance with proper procedures (Romzek, 2016) creates the opportunity for elevating the value of academics' extra-role behaviors at the expense of research achievements. This issue is addressed by analyzing the level of academic achievement required to become an expert and participate in the evaluation of quality assurance of Russian universities.

Russian regulation of higher education is an example of detailed external oversight of performance in compliance with the numerous complicated bureaucratic standards that are only distantly related to the educational process (Romzek, 2016). Increased bureaucracy within universities falls on the shoulders of academics who are supposed to teach and research. At first glance, in spite of bitter criticism, most academics typically respond with some degree of passive disdain. They think bureaucracy needs to be eliminated, but beforehand they spent extra hours producing increasing amounts of senseless paperwork. Moreover, some support the state regulatory actions by participating in inspections of universities, many of which resulted in severe sanctions. Russian academics do not have the choice to be complicit or to be rebellious because the absence of necessary documents means closing the whole organization or a particular department. However, academics do have the option to be an active or passive participant in state regulatory activity. Why have some academics become the self-disciplining enforcers of a system many of them consider absurd? This article applies resource allocation theory to demonstrate that the involvement in academic citizenship behavior depends on people's success at their primary job. Academics less productive in publications and citations are more likely to become the expert engaged in regulatory activity. It seems that the organizational response to increased regulatory pressure was the reliance on academic faculty whose contribution to core activity was less than average but whose willingness to be involved in extra-role behavior was high. Faculty who are more experts in bureaucracy and less in research and education have become a valuable resource for universities that operate under strict legal accountability.

Extra-role behavior of academics

Our starting point is to divide individual performance into two dimensions: in-role performance and extra-role performance. According to Bergeron (2007), in-role performance involves behaviors which are part of the organization's core activity. It includes the primary task requirements for doing what individuals are paid for. Opposite to in-role performance, extra-role performance involves behaviors that could be useful for the organization but are not usually listed in an individual's job description and as a condition of employment (Bergeron, 2007; Farris, 2018; Van Dyne, Cummings, & Parks, 1995). Taking into account that these behaviors are not enforceable, extra-role behavior could be described as voluntary and intentional behavior that is not prescribed by formal job duties. In organizational literature "providing assistance to colleagues, tolerating minor inconveniences, maintaining and promoting a positive attitude" are considered as examples of extra-role behavior (Farris, 2018). Referring to these behaviors as organizational citizenship behaviors is common (Podsakoff et al., 2016).

The majority of prior research was devoted to the impact of the organizational citizenship behavior on the performance of the organization in whole and to the extent it brings bonuses to people engaged in such activity. However, there is a lack of empirical studies exploring the relationship between in-role performance and out-role performance. This question is raised if we recognize after Bergeron as the crucial assumption that individuals have limited time, and time spent on extra-role activity comes at the expense of core task performance (Bergeron, 2007). We suggest that it is necessary to empirically explore this theoretical explication of the trade-off between in-role and extra-role performance. This approach sends to classic paper about the role strain: an individual faces a wide array of the role obligations that sometimes conflict with each other (Goode, 1960). Goode considers the problem of dealing with role obligations from resource allocation perspective. Not only economic resources are scarce but other resources such as time are also limited. According to Goode, the problem is that full dedication to one direction means difficulties in another.

Goode suggests that the individual has some possibilities to overcome the role-strain. For example, the academic could be so useful in fulfilling one obligation that failure in the other will be forgiven. We can imagine a research star prolific in having articles published in Nature, allowing his colleague to overlook the ignorance of teaching duties. In this case, the individual faces some trade-off between different obligations and chooses one at the cost of another. The previous study suggests that spending more time on extra-role tasks may have negative consequences on core task performance (Bergeron, 2007). However, we could question the whole assumption that engaging with the academic citizenship deteriorates the individual-level outcomes. If the causal link is reversed, then individuals with a low level of in-role performance results are more likely to be engaged in academic citizenship because of the proof that they could be useful even with low-level performance in core tasks. It could be that their careers were already damaged when they started to engage in academic citizenship. Recent studies recognize that even if citizenship behaviors were considered initially as prosocial behavior, people could be incentivized by self-serving motives and even negative forces (Bolino et al., 2013, p. 543). Engagement in citizenship behavior has the dark side both in intentions and in consequences for employees and employers (Bolino et. al., 2013).

In this paper, we approach the issue of determining the relationship between core activity and academic citizenship in this context. Bergeron (2007) suggests that there are three types of individual job performance outcomes: performance evaluations, rewards, and career advancement. Each of these types could be explored in relation to the engagement in citizenship behavior. In the context of academic role, core tasks include teaching and research. While it is difficult to evaluate teaching results, it is possible to collect data on research activity regarding both quantity and quality.

Proposition: Individuals who have lower performance in core tasks will engage more in academic citizenship than individuals who have a relatively higher level of performance.

Our primary hypothesis resembles the recent study of faculty service by evaluating the connection between engagement in service and performance in core activities such as teaching and research (Jin, McDonald & Park, 2016). At the same time, we do not accept the assumption that the causal chain starts with service which then affects research productivity. According to our proposition, we expect a similar finding that academic citizenship is negatively associated with faculty research productivity. However, the explanation of this association is adverse: the less productive in research individuals are, the more engaged they are in academic service.

Data and methods

The inspection of universities is conducted by a group of people consisting of a federal inspector and experts who are supposed to represent the academic community. The agency developed a list of 788 experts who are accredited to participate in inspections. One version of the list, in addition to the expert's name, had information about the place of work, position, rank, and degree. This information allows us to identify the experts in the Russian Index of Science Citation (RISC) and collect data regarding publications and citations (554 experts were identified). The nature of the data does not allow us to determine whether all of them actually participated in the inspections or who participated more frequently than others. In the analysis, we use the strategy of comparing the research performance of experts with the research performance of non-experts employed at Russian universities.

Matching as a step of data preprocessing permits decreasing model dependency, researcher discretion, and bias for causal inference (Ho, 2007). Several matching strategies can be applied, and here exact matching was used. We have matched discrete covariates such as sex, year of first publication, university participating in the Russian Excellence Initiative (project "5-100"), and discipline (social science and humanities or other). In our dataset of 554 experts, we successfully matched 542. For matched experts and non-experts, we used as one-to-many matching to allow several non-experts to be matched with an expert through subclusters. For each observation, a subcluster specific weight is calculated for usage in the further models. In comparison with one-to-one matching, we consider this approach as more powerful (Ho, 2007).

Results

First, we analyze the descriptive statistics of raw indicators of research performance. From these results, it is clear that the experts, in comparison with the non-experts, publish more papers in a broad range of journals as well as receive more citations to their works. The median of the number of publications for experts is 32 while for the non-experts the median is almost half the size at 18. The median number of citations for experts is 75 while for non-experts the median is 40. However, if we switch to counting publications and citations for more selective journals, the data indicates different results. First of all, Russian academic faculty (both the experts and the non-experts) rarely publish their papers in journals of high quality as the median of publications in the RISC Core is only 1. In other words, half of the sample published only one paper in journals recognized as journals of high quality. At the same time, if people tend to publish in such journals, they are more likely come from the group of non-experts (the mean and max is significantly higher). The data shows the same results when counting the number of citations from RISC Core journals.

Table 1. Descriptive statistics

Variable	*Min*		*Median*		*Mean*		*Max*	
	expert	*non-experts*	*expert*	*non-experts*	*expert*	*non-experts*	*expert*	*non-experts*
Number of papers in RISC	0	0	32	18	45.65	33.48	320	783
Number of papers in core-RISC	0	0	1	1	5.73	7.29	235	747
Number of citations in RISC	0	0	75	40	222.17	202.45	4951	35904
Number of citations in core-RISC	0	0	3	2	24.11	54.69	1530	32843

Our next step is to recalculate these indicators and analyze shares instead of raw numbers. We also limited our attention to indicators less suitable for gaming. The final list of indicators includes the percentage of papers in the RCSI Core over all papers published by the author, the percentage of citations in RSCI Core overall citations, the weighted-average impact factor of journals where papers were published, the weighted-average impact factor of journals where papers were cited, the percentage of papers in international journals, the percentage of citations in international journals, and the percentage of documents in journals approved by Higher Attestation Commission (VAK). The models are presented in table 2 (see the appendix with more information). Except for the percentage of papers in VAK-journals, all indicators for the experts are significantly lower than for the group of non-experts. The difference is not high for impact-factors but it is noticeable for publications and citations in the RISC core, and papers and citations in foreign journals. On average, an expert publishes 10 percent fewer articles in selective journals compared to a non-expert.

Table 2. Model-based estimations

Variable	*Median*		*Average treatment effect (Δ)*
	experts	*non-experts*	
pct. papers RISC Core	2.78	3.31	-9.69***
pct. citations RISC Core	2.66	4.08	-8.90***

Impact factor published	0.285	0.287	-0.08***
Impact factor cited	0.307	0.308	-0.10***
pct. papers in international journals	0	0	-4.09***
pct. citations from international journals	1.1	1	-4.53***
pct. papers VAK journals	42.05	38.6	1.53

*** p<0.0001 (calculated using permutation test, using random assignment to experts)

Conclusion

We present empirical evidence of negative motives of academic citizenship behavior. Russian academics whose performance is low in respect to publications and citations in selective journals are more likely to become an expert engaged in academic citizenship in the form of regulatory activity. In this respect, engagement in citizenship behavior could be considered as a compensatory mechanism (Bergeron, 2007) according to which individuals contribute something if they are less able to contribute to core tasks. Most research was done under the assumption that engagement in citizenship behavior leads to a lower level of performance in core activity and, as a result, job outcome. Bergeron proposes that the chain of causality could be reversed: "individuals who do not achieve certain outcomes such as task performance (e. g. because of low ability) may switch their focus on OCB as an alternative way to contribute to the work group or organization" (Bergeron, 2007). However, to the best of our knowledge, there is little research taking this assumption seriously. We proposed a research design that allows us to test the nature of the relationship between academic citizenship behavior and core activity.

Why do universities continue to employ people who are on average worse than ordinary faculty? We suggest that universities support this sort of academic citizenship because they consider the experts as insiders of the system who can acquire knowledge on how to successfully get through inspections. Organizations make some effort to increase their stability. Our example includes adding a new role to academics. They are now not only lecturers and researchers, but also experts in paperwork. To be more specific, they are firstly experts in bureaucracy and much less in research and education.

The main limitation, as well as the advantage of this study, is relying on secondary sources that make the research design novel in comparison to surveys usually used for identifying engagement in citizenship behavior. The advantage is relying on more objective information instead of asking people about their motives. The limitation is the availability of data as it does not allow the analysis of the exact time when the individual started to be listed as an expert. As Bergeron suggested, the best way to study the nature of the casualty is longitudinal research but we have data indicates on average two years of being an expert. There is still some possibility that academics did not differ in their research performance at the time of starting the career as an expert but this new responsibility affected further achievement. However, three years seems a rather short period. In this respect, we were close to avoiding the limitations of a cross-sectional design of the study but not close enough.

References

Bergeron, D. M. (2007). The potential paradox of organizational citizenship behavior. Good citizens at what cost? *Administrative Management Review* 32 (4), 1078–1095.

Bolino, M. C., Klotz, A. C., Turnley & Harvey, J. (2013). Exploring the dark side of organizational citizenship behavior. *Journal of Organizational Behavior* 34 (4), 542–559.

Farris, D. (2018). Organizational citizenship behavior in university administrative committees. *Journal of Higher Education Policy and Management* 40 (3), 224–238.

Goode, W. J. (1960). A theory of role strain. *American Sociological Review* 25 (4), 483-496.

Ho, D., Imai, K. & King, G. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15, 199–236.

Jin, M. H., McDonald, B. & Park, J. (2016). Does public service motivation matter in public higher education? Testing the theories of person-organization fit and organizational commitment through a serial multiple mediation model. *American Review of Public Administration* 48 (1), 82–97.

Ouchi, W. (1980). Markets, bureaucracies, and clans. *Administrative Science Quarterly* 25 (1), 129-141.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B. & Bachrach, D. G. (2016). Organizational citizenship behaviors. A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management* 26 (3), 513–563.

Rijcke, S., Wouters, P., Rushforth, A., Franssen, T. & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use –a literature review. *Research Evaluation* 25, 161–169.

Romzek, B.S. (2016). Dynamics of public sector accountability in an era of reform. *International Review of Administrative Sciences* 66 (1), 21–44.

Sihag, V. & Rijsdijk, S. (2018). Organizational controls and performance outcomes: a meta-analytic assessment and extension. *Journal of Management Studies*, online first.

Van Dyne, L., Cummings L. & Parks M. (1995). Extra-role behaviors: in pursuit of construct and definitional clarity (a bridge over muddied waters). *Research in organizational behavior* 17.

The transition cycle measurement to estimate how science impels innovation: A publication-citation analysis of biotech patents

Fang Chen[1], Lili Wang[2], Zexia Li[3], Wu Xiaoyan[4], and Hu Yamin[5]

[1] *chenf@clas.ac.cn*
Chengdu Library and Information Center, Chinese Academy of Sciences, 610041 Chengdu (China)
Department of Library, Information and Archives Management, University of Chinese Academy of Sciences, 100190 Beijing (China)

[2] *wang@merit.unu.edu*
United Nations University – Maastricht Economic and Social Research Institute on Innovation and Technology, 6211AX Maastricht, (The Netherlands)

[3] *lizexia@mail.las.ac.cn*
The National Science Library, Chinese Academy of Sciences, 100190 Beijing (China)

[4] *wuxy@clas.ac.cn*
Chengdu Library and Information Center, Chinese Academy of Sciences, 610041 Chengdu (China)

[5] *yamin369@163.com*
Department of Library, Information and Archives Management, University of Chinese Academy of Sciences, 100190 Beijing (China)

Abstract

Non-patent literature citation has been considered an indicator to measure the contribution of the scientific research to the technological innovation in many science-based technology fields because they are closely related to the original inspiration and theoretical basis of the patent application. This study intended to examine the support of scientific research for technological innovation from the perspective of non-patent citations and measured the average period of transition from research to innovation based on the difference between the patent application year and the average published year of the cited publications. The sample data in US patent grants applied in 2014 and the empirical analysis data in 2015-2017 in biotech field were thoroughly examined in this work. We figured out figured out the difference between the patent application year to the average published year of the cited papers is 6.2, which means that the scientific research published in the second half of 2007 effectively supported the technological innovation in 2014 with a transition cycle of 6.2 years. And we found that the similar characteristics were shown in the medical/therapeutic and industrial biotechnology sub-field with a transition cycle from research to innovation of 6.1 years and 6.7 years in 2015-2017, and in the case of agricultural biotechnology 11.9 years.

Keywords

non-patent literature, science and technology transition, biotechnology

Introduction

The notion that technology springs from scientific base was originally embedded in the 'linear model' of innovation: from basic research through applied research continuing into technology and resultant economic benefit. Publications and patents, being considered as the carrier of scientific research and technological innovation respectively, have attracted a lot of research in recent years. It has been acknowledged that the patent citation part at the end of the patent text is very important because its content relates to a patent application, which includes patents and non-patent literatures cited by an applicant, third party or a patent office examiner. Non-patent citations (esp. the peer-reviewed publications) has been considered a robust indicator to measure the contribution of the scientific research to the technological innovation in many science-based technology fields because they are closely related to the original inspiration and theoretical basis of the patent application (THIJS B., 2006). Narin, F. (1997) conducted a detailed and systematic examination of the contribution of public science to industrial technology by tracing the rapidly growing citation linkage between U.S. patents and scientific research papers and found a rapid rise of science-linked patents, especially in the fields of clinical medicine and biomedical research. Qing K. (2018) presented a comparative study on how biomedical papers are cited by U.S. patents and by other papers over time and observed a positive correlation between citations from patents and from papers. The Normalized Lens Influence Metric that linked the patents and non-patent literatures has been adopted in Nature Index 2017 Innovation and applied to rank the academic influence of global institutions (2017). Jibu M. (2013) used the citations to non-patent literature of patents to analyse the knowledge flows in the pharmaceutical innovation process. This study intends to examine the support of scientific research for technological innovation from the perspective of non-patent citations. We supposed that the time span between the publications cited and the patent applications as the transition cycle of publication to the patent, in this work, as of science to innovation. Therefore, this study measured the average period of transition from research to innovation based on the difference between the patent application year and the average published year of the cited publications.

Data and Methodology

Data preparation

This paper uses patent records in Derwent World Patent Index (DWPI) and Derwent Patents Citation Index (DCPI), which contain patent applications from 44 of the world's patent issuing authorities and citation information. While we only chose the United States granted patent applications in the database because the U.S. patent system is quite representative of the world's technology and we found that this part of the data provided the most standardized and complete information of non-patent citations. According to the OECD (2005)'s definition and classification by the IPC codes for biotechnology, we have extracted the biotechnology patent application records in 2014.

Sample data cleaning

We retrieved a total of 5,643 biotech patent grants by USPTO (United States Patent and Trademark Office) applied in 2014 (retrieved on June 15, 2016) as a sample of data, of which at least one non-patent literatures were cited in 2,473 applications.

In order to conduct analysis and discussions more directly and more targeted, only the research papers published on the journals and proceedings that exemplify the concrete research content are selected and considered to play the role of supporting the technological innovations. Other citations such as the bioinformatics data, legal documents, book series, business and media information cited were cleaned. After repeated cleaning and calibration processes, a total of 2,314 "effective citing" patent applications and 41,280 "effectively cited" papers were finally obtained, as shown in **Table 1**.

Table 1. Non-patent citations performance of biotech patent grants in US, 2014.

Types of data	*Measurement*
All patent applications, P	5643
Patent applications with non-patent citations, P1	2473
Number of non-patent literature cited, n1	48870

Patent applications with research papers citations, P2	2314
Number of research papers cited, n2	41280
Average number of papers cited by patent application	17.8
Percentage of “effective citing” patent applications, P2/P1	93.6%
Percentage of number of papers “effectively cited”, n2/n1	84.5%

Sample data analysis

In order to figure out the average published year of the research papers cited in this dataset, our study calculated the average published year of the research papers cited by each patent application, as of

$$\bar{y} = \frac{\sum_{i=1}^{k} y_i}{k} \quad .$$

The results show that the average published years of research papers cited by each biotech patent grants applied in US in 2014 are mainly concentrated during the period of 1998-2013, as can be seen in **Figure 1**.

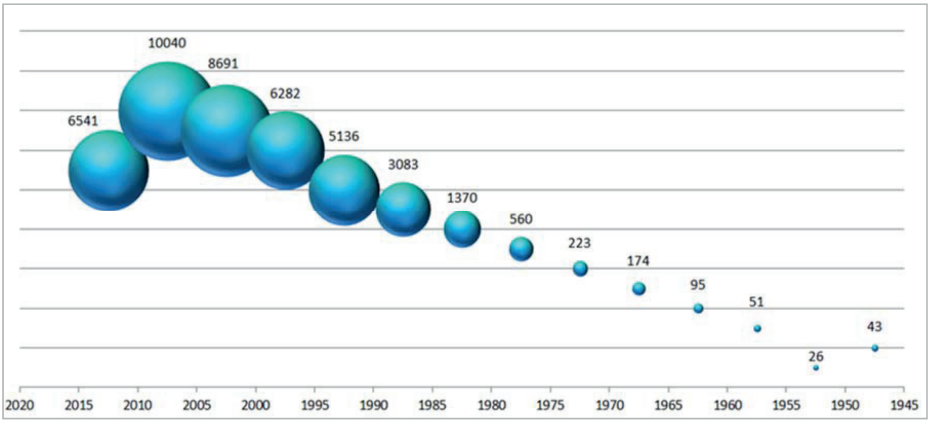


Figure 1. The average published year of research paper cited by each biotech patent applications granted in US, 2014.

While we did not directly calculate the arithmetic mean of the average years in **Figure 1** as the final average year due to the randomness and dispersion of the citation behaviour of every single patent application, which can be imagined and has been detected in our work. Instead, we tried to exclude too discrete data by setting a confidence interval and then calculate a reasonable average year. The data selection principle within the confidence interval is: (1) The published years of research papers cited by a specific patent application not allowed too discrete, and also, (2) the average published year of research papers cited by a specific patent application not allowed too discrete from the final calculated average year.

Suppose in such a confidence interval Π, the number of patent applications in the interval is m, so each patent can be marked as (p_1, p_2, ..., p_m), and the number of research papers cited by them can be marked as (k_1, k_2, ..., k_m), with which published year is (y_1, y_2, ..., y_{km}), and then, the average published year of the research papers cited by each patent application is

$$(\bar{y}_m = \frac{\sum_{i=1}^{k_m} y_i}{k_m}) \quad ,$$

and the standard deviation of average years of each paper cited by this patent application is

$$\sigma_m = \sqrt{\frac{\sum_{\mathrm{i}=1}^{k_m}(y_i - \bar{y}_m)^2}{k_m}}\ ,$$

with the maximum value in the interval is $\sigma_{max,m}$.
At the same time, suppose that the arithmetic mean of the average years for every single patent applications in the confidence interval Π is

$$\bar{y}_\Pi = \frac{\sum_{i=1}^{m} \bar{y}_i}{m}\ ,$$

and the standard deviation is

$$\sigma_\Pi = \sqrt{\frac{\sum_{\mathrm{i}=1}^{k}(\bar{y}_i - \bar{y}_\Pi)^2}{k}}\ ,$$

then the selection criterion for such a confidence interval should be

$$\sigma_{max,m}=\sigma_\Pi\ .$$

In this way, it can be ensured that the data in the confidence interval is not too discrete (σ_Π not too big), nor does it contain data containing too much discrete information ($\sigma_{max,\ m}$ not too big), while ensuring that the confidence interval has a certain width (m is large enough).

Sample data calculation results

For the case of the publication-citation analysis of biotech patent grants applied in US in 2014, we have drawn a histogram of the average published years of research papers cited by the patent applications within the confidence interval (**Figure 2**) and calculated the final average year is

$$\bar{y}_\Pi = 2007.8\ .$$

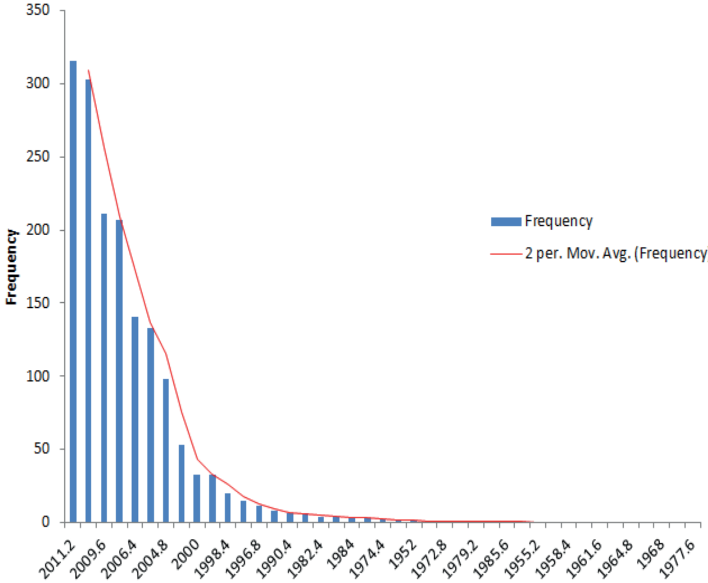


Figure 2. Histogram of the average published years of research papers cited by the patent grants applied in US in 2014, within the confidence interval.

So, in this case, the difference between the patent application year (y_a) to the average published year of the cited papers is

$$\delta_a = y_a - \bar{y}_\Pi = 2014 - 2007.8 = 6.2 \quad .$$

which means that the scientific research published in the second half of 2007 effectively supported the technological innovation in 2014 with a transition cycle of 6.2 years.

Empirical analysis in biotech field

Preliminary analysis and findings for the sample data suggested that the methods of the publication-citation analysis of patents is effective for measuring the transition cycle from the research papers to patent applications and helpful for estimating how science impels innovation. In the empirical analysis work, we applied the methods to a wider range of research assessments. We extracted biotechnology related patent grants applied in US applied during 2015 and 2017 from IncoPat 5.0 Platform (Beijing IncoPat 2018) (retrieved on Jan 30, 2019), and examined their paper cited records using Lens PatCite (Len.org 2018), which is an open platform with the linkage of the patents and non-patent literatures.

Data preparation

In the empirical study, the field of biotech is further classified into three typical biotechnology branches, 1) agricultural biotechnology, including breeding, cultivations, transgenesis biotechnology; 2) industrial biotechnology, including material, food, energy, environmental biotechnology and bioengineering technologies; 3) medical biotechnology, including pharmaceutical, clinical, diagnostic and therapeutic biotechnology.

There were 10000+ patents applications in 2015-2017 have been granted by USPTO (until Jan 30, 2019), with the detailed figures illustrated in **Table 3**. As can be seen that research papers were cited by the majority of patent application, especially in agricultural biotechnology sub-field. To be sure, the number of patent applications is a decreasing trend year by year, this is because there are quite a number of patent applications have not been granted and opened. For the USPTO, average time between application and grant is about 35 months, and even can extend to 44 months. We can also see that average number of publications cited by patent has been increasing year by year. That is, newer patents have more scientific literature citations. While in the case of agricultural biotechnology, the average number of publications cited is smaller relatively.

Table 3. Publications cited by the US patent grants applied in 2015-2017 of three biotech sub-fields

year	***Number of patents; number of patents with research papers cited; average number of publications cited by patent***		
	medical/therapeutic biotechnology	*industrial biotechnology*	*agricultural biotechnology*
2017	2198; 1718 46.1	1607; 1275 41.5	548; 416 17.8
2016	4395; 3711 38.5	3271; 2798 38.6	939; 857 17.0
2015	5260; 4711 33.6	4029; 3615 31.1	944; 849 16.4

Data analysis and comparison

Using the method demonstrated in the previous section, the average published year of research papers cited by the US patent grants applied in 2015-2017 of three biotech sub-fields were calculated respectively and illustrated in **Table 4**, and the standard deviation within the confidence interval were also listed. We found that the narrowest standard deviation occurred in the industrial biotechnology sub-field and the widest in the agricultural biotechnology sub-field. Meanwhile, the average published year of papers cited by the agricultural biotechnology patents could be seen obviously lagging than the cases of the other two sub-fields. That is, the older papers were cited by the patents in the agricultural biotechnology sub-field, and their publication years were also distributed discretely.

Table 4. Average published year of research papers cited by the US patent grants applied in 2015-2017 of three biotech sub-fields

year	*Average published year of papers cited by the patent; standard deviation within the confidence interval*		
	medical/therapeutic biotechnology	*industrial biotechnology*	*agricultural biotechnology*
2017	2010.7 4.00	2009.9 3.21	2004.8 4.31
2016	2010.2 3.39	2009.1 3.50	2003.4 4.36
2015	2008.5 3.16	2008.9 3.16	2004.9 4.09

Then the difference between the patent application year to the average published year of the cited papers in three biotech sub-fields were calculated and illustrated in **Table 5**. We can see that the difference in the medical/therapeutic biotechnology is the smallest, around 6.0, with slight fluctuations in the past 3 years. The difference in the industrial and agricultural biotechnology sub-fields were bigger, particularly in the agricultural biotechnology sub-field. According to the average difference we can see the similar characteristics in the medical/therapeutic biotechnology and industrial biotechnology sub-field with a transition cycle from research to innovation of 6.1 years and 6.7 years. While in the agricultural biotechnology, the transition cycle is 11.9 years.

Table 5. The difference between the patent application year to the average published year of the cited papers in three biotech sub-fields

year	*The difference between the patent application year to the average published year of the cited papers*		
	medical/therapeutic biotechnology	*industrial biotechnology*	*agricultural biotechnology*
2017	6.3	7.1	12.2
2016	5.8	6.9	12.6
2015	6.5	6.1	11.0
average	6.1	6.7	11.9

Results and Discussions

A method to examine the support of scientific research for technological innovation from the perspective of non-patent citations were studied in this work. With the assumption that the time span between the publications cited and the patent applications as the transition cycle of publication to the patent as of science to innovation, this study measured the average period of transition from research to innovation based on the difference between the patent application year and the average published year of the cited publications. The sample data in US patent grants applied in 2014 and the empirical analysis data in 2015-2017 in biotech field were thoroughly examined in this work and we can see that,

(1) The average number of publications cited by patent has been increasing year by year. That is, newer patents have more scientific literature citations.

(2) The similar characteristics were shown in the medical/therapeutic biotechnology and industrial biotechnology sub-field with a transition cycle from research to innovation of 6.1 years and 6.7 years.

(3) To some extend, the case of agricultural biotechnology is different with the cases of medical/therapeutic biotechnology and industrial biotechnology sub-field, showing a lagging transition cycle from research to innovation of 11.9 years. Yet we can hardly arbitrarily regard there can be huge obstacle for the transition, because after all, the patent quantity in the agricultural sub-field is relatively small.

On the basis of the results and findings of the empirical analysis we could propose suggestions and recommendations to the policy makers and researchers that, such work like planning and roadmapping the biotechnology related scientific research could be more forward-looking, e.g., it is necessary to pay attention to basic research in the next 6-7 years of technical transformation and the application prospects and trends, for it could help planning for the medium- and long-term development goals, and help for the deployment and evaluation of disciplines, institutions and talents in biotechnology related research fields.

Some deficiencies may remain in this study. Patent reference motivation and behaviour is complicated. So, we could not take the calculation results and conclusions for the extension of regularity and extrapolation. For example, in this work, we did not distinguish the references from the inventor or from patent examiner, and did not analyse the citing motivation within the text analysis. We believe such approaches could provide more evidences for the detailed identification and evaluation work. At the same time, although we analysed the USPTO granted patent in this work, but we did not consider the quality of different patents, and also did not consider patent text differences from other states or organizations.

In the future, this work could be furthered and extended. In this work we only discussed patents in biotechnology field, while did not discuss the circumstances in other subjects such as chemical, electronics, computer science and technology and so on. Due to the possible disciplinary differences and different technical evolution characteristics, there could be some different results. So more extensive empirical work could be helpful to show the more knowledge about this topic.

References

Narin, F., Hamilton, K.S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy*, 26(3), 317-330.

Meyer M . Patent Citations in a Novel Field of Technology — What Can They Tell about Interactions between Emerging Communities of Science and Technology?[J]. Scientometrics, 2000, 48(2):151-178.

Qing, K. (2018). Comparing scientific and technological impact of biomedical research. *Journal of Informetrics*, 12(3), 706-717.

Nature Index 2017 Innovation, *Top 100 institutions by Lens score*, Retrieved Nov 1, 2018 from: http://www.natureindex.com/supplements/nature-index-2017-innovation/tables/top100-institutions-lens

OECD, 2005, *A Framework for Biotechnology Statistics*, Retrieved June 1, 2016 from: http://www.csdl.tamu.edu/DL94/paper/kling.html.

Callaert J , Looy B V , Verbeek A , et al. Traces of Prior Art: An analysis of non-patent references found in patent documents[J]. Scientometrics, 2006, 69(1):3-20.

Jibu M, Osabe Y., Börner K. Knowledge Flows and Delays in the Pharmaceutical Innovation System. ISSI 2015. 2015.

MESH classification of clinical guidelines using conceptual embeddings of references

Johan Eklund, David Gunnarsson Lorentzen, and Gustaf Nelhans

{johan.eklund, david.gunnarsson_lorentzen, gustaf.nelhans} @hb.se
University of Borås, Swedish School of Library and Information Science, S-501 90 Borås (Sweden)

Abstract

In this study, we investigate different strategies for assigning MeSH (Medical Subject Headings) terms to clinical guidelines using machine learning. Features based on words in titles and abstracts are investigated and compared to features based on topics assigned to references cited by the guidelines. Two of the feature engineering strategies utilize word embeddings produced by recent models based on the distributional hypothesis, called word2vec and fastText. The evaluation results show that reference-based strategies tend to yield a higher recall and F1 scores for MeSH terms with a sufficient amount of training instances, whereas title and abstract based features yield a higher precision.

Introduction

This paper builds on previous attempts to use a combination of cited references as text for machine learning purposes to develop new means for combining text "mining" with scientometric citation-based methods. (Eklund & Nelhans, 2017) Such an approach, if rendered stable, would provide means for developing a joint methodology, between different specialities in information science and to broaden the scope of analysis of citation-based data.

In previous studies, we used the Latent Dirichlet Allocation (LDA) algorithm (see e.g. Blei, 2003) for topic modelling of references, using an approach in which references are treated as "words" and reference lists as "sentences" (or documents) of such "words". We demonstrated that the topical structure of document collections could be studied using a combination of citation network properties and text-based methods.

In this study, we take a somewhat different stance by generating semantic representation vectors of the cited documents treated as semantic compositions of the MeSH terms assigned to the cited documents. The ultimate objective of this approach is to be able to classify a heterogeneous set of non-standardized documents, hence the choice of clinical guideline documents that are extracted from different sources, covering many different languages, having different sets of metadata content, but all containing a matched list of citations. The techniques introduced are generic and are applicable to other kinds of professional and policy documents.

Our approach answers a simple question that could be framed as

- RQ 1. How can we in a meaningful way classify a set of documents for which we do not have access to their texts, but instead their sets of cited references?

Furthermore, we want to evaluate our method by comparing it to a more traditional approach using representation vectors containing term weights of words appearing in titles and/or abstracts.

- RQ 2. How well do reference-based feature vectors perform as compared with a text-based method in a collection of bibliographic records containing titles and abstracts?

Word embedding and semantic composition

The distributional hypothesis of semantics (see e.g. Sahlgren, 2008) entails the notion that words that tend to occur in the same contexts are also semantically related. The idea of generating representations of words that capture their co-occurrence patterns and yield semantic representation vectors has been investigated for several decades in the fields of computational linguistics, information retrieval, and document classification. Among the early methods generating semantic representation vectors based on co-occurrence data can be mentioned Hyperspace Analogue to Language (Lund & Burgess, 1996), latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), and random indexing (Kanerva, Kristoferson, & Holst, 2000).

In recent years a new generation of word embedding methods based on neural network learning has emerged. One of those models is commonly called *word2vec* and was published by Mikolov, Chen, Corrado, & Dean (2013). This algorithm utilizes a shallow feed-forward neural network to implement two different approaches for encoding word context, called continuous bag-of-words (CBOW) and continuous skip-gram respectively. With regard to the CBOW model, the objective is to predict the target word, given an input consisting of a set of context words. Conversely, the objective of the continuous skip-gram model is to identify a set of context words, given an input word. The state vectors of the hidden layer obtained through back-propagation are subsequently used as representation vectors for the vocabulary pertaining to the training collection of documents.

Bojanowski, Grave, Joulin, & Mikolov (2017) present an extension to the continuous skip-gram model based on incorporating morphological information in the learning of word representation vectors. This is accomplished by associating each word with a set of character *n*-grams. For example, the 3-grams contained in the word (up to the word boundaries) *smoke* are *sm*, *smo*, *mok*, *oke*, and *ke*. A representation vector for each character *n*-gram is learned separately and finally, a representation vector for each word in the vocabulary is established by computing the sum of the representation vectors for the *n*-grams associated with the word. A potential advantage of this method is that it facilitates learning of representation vectors in collections with many infrequent words, and consequently only a few instances available for learning. Since this algorithm is implemented in the fastText library (FastText, n.d.) for text classification and representation learning, it is named the *fastText* algorithm in this paper.

In collections yielding sparse feature vectors, term representations that captures co-occurrence patterns may also yield document representations that are related by means of conceptual dimensions, rather than by term dimensions. The traditional vector space model as presented by Salton, Wong, & Yang (1975) is based on the mechanism of assigning document vectors as linear combinations of a sequence of term weights and the orthonormal basis of the Euclidean space. In other words, each document is represented as a weighted sum of a set of mutually orthogonal (unit) term vectors. In this study, we treat each reference as the semantic composition of the MeSH terms assigned to that reference. By the *semantic composition* of terms, we denote structures that are essentially lists of words, such as phrases and sentences. According to Blacoe & Lapata (2012), representational models for structures have received less attention than the semantic modelling of single words. Two basic approaches for generating representation vectors of semantic compositions are by means of vector addition, and elementwise multiplication (the so-called Hadamard product). In line with one of the strategies investigated by Blacoe & Lapata (2012), we then generate representation vectors for every reference document r_j as

$$\mathbf{r}_j = \sum_{\forall k_i \in d_j} \mathbf{k}_i$$

where $\mathbf{k}_i$ is the representation vector of term k_i. In other words, we treat the list of MeSH terms assigned to a document as the semantic composition of those terms. Likewise, we treat each clinical guideline g_k as the semantic composition of the references contained in g_k and produce a corresponding representation vector as

$$\mathbf{g}_k = \sum\nolimits_{\forall r_j \in g_k} \mathbf{r}_j$$

Using such additive semantic representation models facilitates accumulative and distributed computations of the available training data, since addition is a commutative as well as associative operation.

Methodology

In total 6 different strategies for the representation of clinical guidelines were investigated and compared:

A1. Title
A2. Abstract
A3. Title combined with abstract
R1. MeSH terms represented by binary vectors containing 0's in all positions except one unique position containing the number 1; so-called one-hot encoding.
R2. MeSH terms represented by word embeddings (n = 300) generated by the word2vec algorithm.
R3. MeSH terms represented by word embeddings (n = 300) generated by the fastText algorithm.

A dataset consisting of 285 bibliographic records of clinical guidelines has been used. These records have been enriched by MeSH terms and abstracts downloaded from PubMed through the Entrez Programming Utilities API. The guidelines were selected as to ensure that MeSH terms and abstracts are available for each guideline. The complete reference list identifiable by PMIDs were mined from the full texts and meta data, including MeSH terms, were collected for each cited reference. The titles and the abstracts have been preprocessed using the *tm* package for R (Feinerer, Hornik, & Meyer, 2008) by converting the texts to lower case, removing punctuation marks and numeric characters and filtering out common English stopwords. All the details of the classification experiments such as feature selection, training, and cross-validation have been implemented using the *mlr* package for R (Bischl et al., 2016).For each experiment involving title, abstract, or a combination of the two, the texts have been assigned document vectors by means of tf-idf weighting. Among many possible term weighing strategies available, we have opted for the *tfx* strategy as described by Salton & Buckley (1988), which amounts to tf-idf weighting by means of an unnormalized tf component, an ordinary idf component, and no vector length normalization.

For the reference-based strategies as well as the strategy using only words in guideline titles, a chi-square metric (see e.g. Zheng, Wu, & Srihari, 2004) was used to rank and select *k* terms for each target class. After some experimentation, it was decided that the value *k* = 100 yields comparatively good performance and was therefore used for the experiments.

The classification algorithm used for the experiments is *logistic regression*, which is a binary classification algorithm modelling the log-odds of class probabilities as a linear function of the document features. Previous comparative studies using logistic regression for text categorization (Zhang & Oles, 2001; Zhang, Jin, Yang, & Hauptmann, 2003) have shown a performance comparable to state-of-the-art algorithms like the support vector machine (SVM). Since the training of logistic regression classifiers does not require the tuning of hyperparameters (unlike, for instance, the SVM algorithms) it was selected for this study. An L2 penalty term has been used to regularize the regression coefficients and prevent overfitting.

Evaluation

For the evaluation of the classifier performance we have used stratified 3-fold cross validation with 100 iterations to obtain stable performance figures. The use of stratified cross-validation was employed to ensure that all folds contain positive examples, which is a critical factor for classes with only a few examples. The evaluation measures used are precision, recall, and the F1 score. In order to compute the average F1 score (which itself is the harmonic mean of precision and recall) for the repeated cross-validation evaluation there are several possible procedures that could be considered. Forman & Scholz (2010) discuss three different definitions:

1. $F_{avg} \coloneqq \frac{1}{k}\sum_{i=1}^{k} F_i$ where F_i is the F1 score for each fold.
2. $F_{\mathrm{pr,re}} \coloneqq 2 \cdot \frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$ where Pr and Re are the average precision and recall respectively over all folds.
3. $F_{\mathrm{tp,fp}} \coloneqq (2 \cdot \mathrm{TP}) / \, (2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN})$ where TP, FP, and FN are the sum of the true positives, false positives, and false negatives respectively over all folds.

Based on experimental evidence, Forman & Scholz (2010) find that the $F_{\mathrm{tp,fp}}$ strategy is the least biased estimator of the F1 score for cross-validation (especially for datasets with a high degree of class imbalance) and this strategy has also been selected for this study.

Twelve MeSH terms have been used as target terms for the evaluation of each strategy. These terms were chosen on the basis a selection of topics (pregnancy and birth, dietary supplements, smoking, cardiovascular diseases, and mental health), as well as having a sufficient amount of instances in the training set.

Findings

In tables 1-3 we present the performance score (average precision, average recall, and average F1 respectively) for each MeSH term and feature strategy.

Table 1. Average precision and recall for the selected terms and the six feature strategies.

term	*title*		*abstract*		*title+abstract*		*binary*		*word2vec*		*fastText*	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
Pregnancy	0.83	0.52	**0.93**	0.69	**0.93**	0.70	0.92	**0.89**	0.79	0.85	0.83	0.86
Infant, Newborn	**0.56**	0.34	0.51	0.23	0.53	0.24	0.55	**0.48**	0.38	0.37	0.42	0.40
Infant	0.56	0.22	**0.95**	0.25	0.92	0.24	0.59	0.37	0.46	0.40	0.50	**0.46**
Pregnancy Outcome	**0.82**	0.47	0.75	0.43	0.73	0.41	0.49	0.46	0.43	0.45	0.46	**0.51**
Dietary Supplements	0.76	**0.79**	**0.87**	0.52	0.88	0.55	0.73	0.61	0.59	0.65	0.59	0.60
Vitamins	**0.68**	**0.54**	0.66	0.39	0.66	0.39	0.60	0.44	0.40	0.45	0.38	0.42
Smoking Cessation	0.79	**0.85**	0.84	0.62	0.80	0.67	**0.90**	0.76	0.77	0.64	0.74	0.62
Smoking	0.35	0.12	**0.97**	0.38	0.91	0.36	0.46	**0.44**	0.29	0.30	0.28	0.31
Stroke	0.71	0.36	0.78	0.38	**0.80**	0.42	0.50	0.41	0.52	**0.54**	0.50	0.46
Cardiovascular Diseases	0.99	0.67	**1.00**	0.71	**1.00**	**0.71**	0.42	0.34	0.43	0.48	0.40	0.42
Depression	0.02	0.09	0.00	0.00	0.00	0.00	**0.28**	**0.27**	0.16	0.25	0.14	0.22
Anxiety	**0.74**	**0.35**	0.00	0.00	0.00	0.00	0.34	0.34	0.16	0.30	0.17	0.30

What is clearly noticeable in table 1 is that the precision score tends to be higher for the feature strategies based on title and abstract respectively. This in turn indicates that those types of metadata fields tend to contain information that is more focused on the content of the guidelines than the corresponding references.

The average recall for the different strategies, also presented in table 1, follows a different pattern than what can be observed with regard to the average precision. The reference-based strategies tend to yield a higher recall than the title and abstract based strategies, in particular for the more frequent MeSH terms in the dataset. This is an indication that the use of references and their corresponding MeSH terms tends to increase the coverage of the induced classifier.

Table 2. Average F1 score for the selected terms and the six feature strategies.

term	*title*	*abstract*	*title+abstract*	*binary*	*word2vec*	*fastText*
Pregnancy	0.64	0.79	0.80	**0.90**	0.82	0.84
Infant, Newborn	0.42	0.32	0.33	**0.51**	0.37	0.41
Infant	0.32	0.40	0.38	0.46	0.43	**0.48**
Pregnancy Outcome	**0.60**	0.55	0.53	0.48	0.44	0.48
Dietary Supplements	**0.78**	0.66	0.68	0.66	0.62	0.59
Vitamins	**0.60**	0.49	0.49	0.51	0.43	0.40
Smoking Cessation	0.82	0.72	0.73	**0.82**	0.70	0.68
Smoking	0.18	**0.54**	0.51	0.45	0.30	0.29
Stroke	0.48	0.51	**0.55**	0.45	0.53	0.48
Cardiovascular Diseases	0.80	0.83	**0.83**	0.38	0.45	0.41
Depression	0.03	0.00	0.00	**0.27**	0.19	0.17
Anxiety	**0.47**	0.00	0.00	0.34	0.21	0.22

The average F1 score for the different strategies is displayed in table 2. A slight advantage can be noticed for the title and abstract based features when taken as a group of strategies. If we instead compare the individual strategies, we find that the title based and binary reference-based strategies display a comparable performance with 4 top scores each. There is, however, no clearly discernible trend with regard to the top score and the number of guidelines indexed by each respective MeSH term. However, a natural question arises in connection with the observed results, namely what the correlation is between the performance of each strategy and the number of instances available in the dataset for each MeSH class.

Table 3. Rank correlation between the number of instances of MeSH terms and performance scores.

	title	*abstract*	*title + abstract*	*binary*	*word2vec*	*fastText*
precision	0.26	0.22	0.35	0.76	0.69	0.81
recall	0.25	0.35	0.34	0.73	0.57	0.69
F1	0.26	0.31	0.32	0.81	0.61	0.81

In table 3 we presented the rank correlation, as measured by Spearman's rho, between the number of instances of each MeSH term in the dataset and the average performance scores for each strategy. There is a clearly discernible difference between the title and abstract based strategies versus the reference-based strategies, in the sense that the reference-based strategies tend to perform better when more instances are available.

Discussion

McJunkin (1995) argues that title words should be considered when performing keyword-based searches. While controlled vocabularies, such as the Library of Congress Subject Headings, are useful when specific entry terms or word order is not known, the author states that subject-rich terms in titles could be used favorably for keyword-based searching since these are often more current than established controlled vocabulary. (ibid.) The results of the present study indicate that the use of title words alone provides a classifier performance with regard to precision that is comparable to that of words appearing in abstracts. What can also be observed is that reference-based features tend to yield a higher recall, but there is no clear evidence in this study that the use of semantic word embedding yields a more performant representation of the content of the references. This may be explained in terms of the limited amount of data used for producing word embeddings together with the observation that models utilizing neural network training, such as word2vec, generally need large training sets to perform well (Mikolov et al. 2013).

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531.

References

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... & Jones, Z. M. (2016). mlr: Machine Learning in R. *The Journal of Machine Learning Research, 17*(1), 5938-5942.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 546-556). Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391-407

Eklund, J., & Nelhans, G. (2017). Topic modelling approaches to aggregated citation data. Presented at the *22nd International Conference on Science and Technology Indicators*, Paris, September 6-8.

FastText. (n.d.). Retrieved February 08, 2019, from https://fasttext.cc/.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explorations, 12*, 49-57.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 22, No. 22).

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203-208.

McJunkin, M. C. (1995). Precision and recall in title keyword searches. *Information Technology and Libraries, 14*(3), 161.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics, 20*, 33-53.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information retrieval, 4*(1), 5-31.

Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter, 6*(1), 80-89.

Dependence modeling of bibliometric indicators with copulas

Tina Nane[1] and Ashni Bachasingh[2]

[1]g.f.nane@tudelft.nl

[2] ashnids@icloud.com
Delft University of Technology, Dept of Applied Mathematics, van Mourik Broekmanweg 6, 2628 XE, Delft (The Netherlands)

Abstract

Researchers' academic output is frequently quantified by size and quality, and citation impact is frequently used as a proxy for quality. Bibliometric indicators of research output are often included in the periodical evaluations of researchers. In this article, we model the dependencies between several bibliometric indicators beyond the common correlation coefficients. We first investigate the behaviour of the correlation coefficients on different ranges of the distribution and emphasize the (lack) of tail dependence. Investigating the correlations for the division *Social* Sciences unveils intricate relationships between the indicators that emphasize the necessity of more sophisticated dependence modelling tools. We therefore propose copulas in order to account for complex dependency structures of pairs of indicators of 3574 researchers from Quebec. Bivariate parametric copulas that best fit the data are chosen and evaluated with respect to a goodness of fit test. Even though the performance of the parametric copulas is modest for the division *Social Sciences*, the methodology has an undoubtable merit and other parametric families or nonparametric copulas should definitely be further investigated.

Introduction

Dependencies are ubiquitous in bibliometrics (e.g., van Raan, 2006; Hagedoorn & Cloodt, 2003). To identify and appropriately characterize and quantify dependencies is essential to any multivariate statistical analysis endeavour. Pearson and Spearman correlations have been the standard approach to measure and model dependencies in citation analysis. While Pearson correlation measures linear dependence, Spearman correlation unveils the existing co-monotonicity between any two sets of observations. A nonparametric measure of dependency, which is both a measure of strength, but also of the direction of association between variables is the Kendall's rank correlation coefficient.

Bibliometric data might contain ties, e.g., a given proportion of researchers' publications have zero citations. Kendall's rank correlation coefficient has more variants which account for ties, and tau-b is arguably the most commonly used. Numerous studies employ Kendall's tau-b coefficient, for example to consider the inherited dependencies in analysing the effects on relative performance with respect to citation impact (Colliander & Ahlgren, 2011), on studying if the academic productivity is correlated with the well-being at work (Torrisi, 2013), or when comparing researchers' citation indicators when using different databases (Torres-Salinas et al, 2009). Another variants of the Kendall's tau-b coefficient have been used; see, for example, Waltman et al. (2011).

All correlation coefficients are single numbers which aim to describe the existing dependencies between pairs of bibliometric indicators. Nonetheless, this representation has limitations and frequently does not capture the complexity of the dependencies among indicators. To the authors' best knowledge there has not been yet a further and in-depth analysis, that is, beyond correlation coefficients, of the dependence structure between bibliometric indicators. This paper intends to fill this gap and employs the copula function to model the dependence structure between various bibliometric indicators at the researcher level.

Copulas are functions that account jointly for univariate marginal distributions with a dependence structure, in order to represent joint distributions. The literary idea of a copula arose in the 19th century. This was based on the multivariate cases of non-normality. In 1959 Abe Sklar (Sklar, 1959) first employed the word copula in a mathematical or statistical sense in the theorem which now bears his name and which is reproduced in the following section. The theorem describes how the joint distribution can be specified in terms of the marginal distribution and the copula function. The notion of copulas became increasingly popular at the end of the nineties. By this time, researchers in the applied field of finance discovered the notion of the copula. This lead to a wealth of investigations about copulas, with a special focus on the applications of the copulas.

The motivation of this paper lies in modelling the dependence structure between several publications and citation indicators of researchers. We capture these dependencies with the help of parametric copulas. Besides introducing a technique which guides through modelling dependencies between bibliometric indicators, this papers also aims set the grounds for complex multivariate data analysis.

Dependence measures

Correlation coefficients

Correlation is by far the best known dependence measure. We distinguish between linear correlation and measures of rank correlation. The Pearson correlation or the product moment correlation depicts the linear correlation between two random variables and has been historically the most popular measure of dependence. The popularity lies in the ease of calculation and manipulation under linear operations. Another reason is that linear correlation is a natural measure of dependence in multivariate normal distributions, since the correlation coefficient completely defines the dependence structure of this distribution. However, this is not always the case, as marginal distributions together with the Pearson correlation cannot determine the joint distribution in general (Joe, 2014). Furthermore, linear correlation is not preserved by copulas. Which means that two pairs of correlated variables with the same copula can have different correlation coefficients.

Spearman or rank correlation measures the monotonic rather than linear dependence and is given by

$$\rho_S\,(X,Y) = \rho(F(X),G(Y)),$$

where ρ is Pearson's linear correlation; F and G denote the marginal distribution of X and Y. The rank correlation accounts for the monotone relationship between X and Y. The Spearman correlation is invariant under monotonic transformations and, more importantly, is a non-parametric measure of dependence.

Let (x_i, y_i) and (x_j, y_j), with $i, j = 1, \dots, n$, denote n observations from the vector (X, Y) of continuous random variables. We say that (x_i, y_i) and (x_j, y_j) are concordant if $x_i < x_j$ and $y_i < y_j$ or if $x_i > x_j$ and $y_i > y_j$. Likewise, we say that (x_i, y_i) and (x_j, y_j) are discordant if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$. The Kendall's rank correlation is given by

$$\rho_\tau(X,Y) = \frac{C-D}{C+D},$$

where C is the number of concordant pairs and D is the number of discordant pairs. Kendall's rank correlation or Kendall's tau measures the ordinal association between X and Y and, as reflected by the definition, is based on the concordance and discordance of data pairs.

Kendall's tau is regarded as more robust, hence less sensitive to outliers and the p-values, calculated when testing the null hypothesis that Kendall's tau is zero, are more accurate with smaller sample sizes. Alternatively, Spearman's rho is more sensitive to outliers and discrepancies in data. Nonetheless, in most situations, the interpretations of Kendall's tau and Spearman's rank correlation coefficient lead to the same inferences due to similarities.

Tied data

Bibliometric data often contain many ties; e.g., publications or set of publications that receive no citations, researchers that do not collaborate internationally, etc. We therefore need to handle the data and their measure of dependence carefully. For example, the tied data need to be converted into rank for computing the Spearman rank correlation. For the Kendal's coefficient of concordance ρ_τ, a variation has been proposed, denoted as Kendall's tau-b coefficient

$$\tau_B = \frac{C - D}{\sqrt{N_1} \cdot \sqrt{N_2}},$$

where N_1 is the number of data pairs not tied in the first variable and N_2 is the number of data pairs not tied in the second variable.

Copula function

Let H be the joint distribution function of the random variables X and Y with margins F and G. Then there exists a copula C such that

$$H(x, y) = C(F(x), G(y)),$$

for all x, y in the domain of F and G. The result is known as Sklar's theorem and illustrates how the joint distribution can be represented in terms of the marginal distributions and the copula function. The copula is "a function that links a multidimensional distribution to its one-dimensional margins" (Sklar, 1959). The copula C is uniquely determined on the range of F and G, and is therefore unique if the marginal are continuous. Since the distribution of X and Y are uniform, a copula is a joint distribution of two unit uniform random variables.

Copulas or copulae provide a much richer description of dependence than the correlation coefficients. Different dependence structures can result in the same degree of association. The copula focuses on capturing in which subset of the support of the distribution is the association the strongest or the weakest. There are many parametric copula families that describe the dependence structure between random variables. Some important bivariate parametric families include Gaussian, student-t, Archimedian, such as Clayton, Gumbel, Frank, etc. The Archimedian copula families have one copula parameter. Figure 1 below depicts simulations from different copula families, for which the Kendall's tau is 0.5.

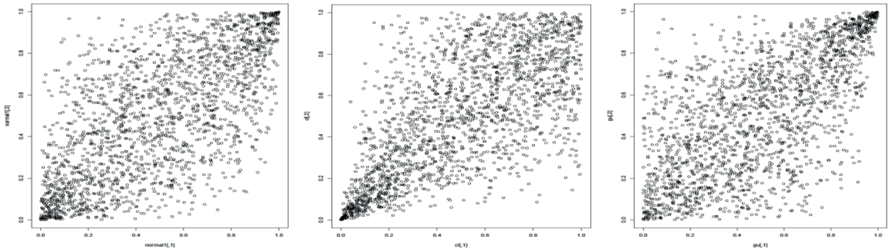


Figure 1. Simulation of 2000 observations from a Gaussian copula (left), a Clayton copula (centre) and a Gumbel copula (right), for which the Kendall's tau is 0.5.

Figure 1 shows how the same Kendall's tau coefficient can be obtained from different dependence structures. The Gaussian or normal copula exhibits symmetry, which suggests that Kendall's tau is 0.5 throughout the entire span of the distribution. The Clayton and Gumbel copula exhibit asymmetry, which suggest a stronger tail dependence. The two variables appear to be more closely in the corners of the graph rather than in the centre. Clayton copula shows a left tail dependence, therefore the correlation between the two variables tends to increase in the lower tail of the distribution. Similarly, Gumbel copula shows a right tail dependence, suggesting that high values in the distribution of X tend to correlate stronger with high values in the distribution of Y.
Furthermore, the Spearman correlation coefficient is not preserved by copulas, that is, two pairs of random variables (X_1, X_2) and (X_3, X_4) with the same copula might have different product moment correlations. Nonetheless, Kendall's correlation τ is constant for any given pairs of random variables with the same copula.

Other parametric copulas are used in practice. For example, the cumulative distribution function of the Tawn type 1 copula is given by

$$C(u, v) = \exp[\log(uv)\ \mathrm{A}[\frac{\log(u)\log(v)}{\log(uv)}]$$

For $u, v \in (0,1)$ and where $A = (1-\alpha)t + \{\alpha^r t^r + (1-t)^r\}^{1/r}$, $\alpha \in [0,1]$ and $r \geq 1$ are the copula parameters. Moreover, copula families can be rotated by 180 degrees, which and these are referred to as survival copulas. A survival copula is given by

$$C(u, v) = u + v - 1 + C(1-u, 1-v).$$

Figure 2 exhibits simulations from a Tawn copula type 1 (left), as well as a survival Tawn copula type 1, with parameters $r = 5.5$ and α=0.5. Notice how the strong tail dependency is rotated by 180 degrees for the survival copula.

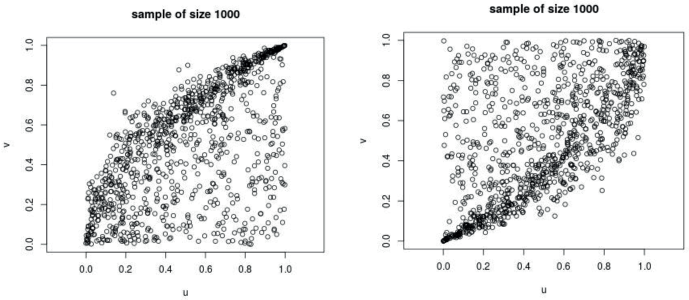


Figure 2. Simulations from a Tawn copula type 1, with parameters 5.5 and 0.5 (left) and a survival Tawn copula type 1, with the same parameters (right.)

Fitting copulas

Similar to fitting parametric distributions, one can employ fitting parametric copulas. There are numerous copulas proposed in the literature (Nelson, 2006; Joe, 2014), while some of the best known copulas are grouped into families such as Gaussian and Archimedian. The package VineCopula in R has more than 35 copulas implemented and has been used to fit parametric copulas in this study. A nonparametric approach can also be employed when fitting copulas. The empirical copula is the analogue of the empirical distribution function and is a non-parametric copula. It is typically used in goodness-of-fit tests for copula, given its commendable asymptotic properties. The test compares the empirical copula with a given parametric copula derived under the null hypothesis.

First, the data are transformed into normalized ranked data, also referred to as pseudo-observations. The pseudo-observations are used for copula fitting and for simulating from a given parametric copula, due to Sklar's representation theorem. Each parametric copula is fitted by using a maximum likelihood estimation for the copula parameters. An Akaike Information Criterion (AIC) is then used to select the best fitting parametric copula. The questions remains though, is the best fitting parametric copula a good fit? A goodness of fit test based on the empirical copula will attempt to answer this question.

Similar to the correlation coefficient, we need to account for the ties in our data also when fitting parametric copulas and testing their goodness of fit. The goodness of fit test mentioned beforehand are under the assumption of continuous marginals. In other words, the assumption made is that ties occur with probability zero. Kojadinovic and Yan (2010) propose to simulate pseudo-observations by randomly breaking the ties. The randomization does not change the results qualitatively, that is the parameter estimate is not effected by the randomization. The authors stress that ignoring the ties in the computation of the pseudo-observations leads to the rejection of numerous of well-fitting copulas.

Data

We explore the dependence measures described in the previous section on bibliometric data at researcher level. The data used for this analysis contain bibliometric information of 3574 Canadian scholars in Quebec, and is part of a larger dataset. The original data set has been used in other studies, e.g. Gingras et al., 2008; Costas et al., 2015. The restricted dataset has been used in this form for predicting the age of researchers using bibliometric indicators (Nane et al, 2017).

Each researcher in the dataset has published at least one article in Web of Science (WoS) between 1980-2012 and the citations of their publications have been recorded until the end of 2014. Our data set provides us the following indicators for each researcher:

- P: The total number of publications in WoS a researcher published between 1980 and 2012
- MCS: Mean citations of all P publications;
- $MNCS$: The normalized mean of all citations of all P publications; the normalization is done with respect to field and year of publications;
- PP_TOP_PROP: The percentage of P publications which are in the top 10% mostly cited papers in their field, per publication year;
- PP_INT_COLLAB: The percentage of P publications which are international collaborations.

The descriptive statistics for our dataset are included in Appendix 1. Each researcher in the dataset is assigned to one of the 9 divisions. The division determines a disciplinary field of activity of the scholar, which is based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP) developed by the U.S. Department of Education's National Center for Education Statistics (NCES).

Accounting for researchers' assignment to division in the analysis is quite valuable; nonetheless, running the analysis for all divisions far exceeds the space limitations of this paper and will be deferred to a later manuscript. Only one division, namely *Social Sciences* will be included in the present analysis. This division includes 500 researchers and is the fourth largest represented division in the dataset.

Main results

We first investigate the Spearman and Kendall’s tau-b dependence measures between the bibliometric indicators of the researchers in the dataset. Though the bins created by splitting the data set according to their quartiles contain ties, the standard deviation of these ranked bins do not equal zero. So Spearman's rank correlation is computed without difficulty, while the Kendall’s tau-b accounts for the presence of ties.

Table 1. Spearman (S) and Kendall’s tau-b (K) correlation coefficients for the bibliometric indicators for all the data, before the 25th percentile (Q1), between the 25th percentile and 50th percentile (Q1-Q2), between the 50th percentile and the 75th percentile (Q2-Q3), in the 90th and in the 95th percentile.

V_1	V_2	*Correlation*	*All Data*	$\leq Q1$	$(Q1, Q2)$	$(Q2, Q3)$	$\geq Q3$	$\geq 90th$ $percentile$	$\geq 95th$ *percentile*
P	MCS	S	0.5	0.16	0.17	0.14	0.2	0.25	0.19
		K	0.36	0.12	0.12	0.1	0.14	0.17	0.13
P	MNCS	S	0.31	0.13	0.09	0.1	0.15	0.28	0.16
		K	0.22	0.1	0.06	0.07	0.1	0.19	0.11
P	PP_TOP_PROP	S	0.38	0.19	0.08	0.11	0.13	0.28	0.15
		K	0.27	0.15	0.05	0.07	0.09	0.19	0.1
P	PP_INT_COLLAB	S	0.33	0.22	0.05	0.04	0.06	0.16	0.17
		K	0.23	0.18	0.03	0.02	0.04	0.11	0.12
MCS	MNCS	S	0.71	0.53	0.22	0.24	0.58	0.59	0.67
		K	0.55	0.39	0.15	0.16	0.42	0.43	0.49
MCS	PP_TOP_PROP	S	0.65	0.29	0.15	0.19	0.44	0.34	0.24
		K	0.49	0.22	0.11	0.13	0.3	0.23	0.16
MCS	PP_INT_COLLAB	S	0.49	0.22	0.11	0.13	0.3	0.23	0.16
		K	0.36	0.33	0.11	0.02	0.19	0.14	0.16
MNCS	PP_TOP_PROP	S	0.87	0.42	0.43	0.47	0.6	0.44	0.19
		K	0.72	0.33	0.31	0.33	0.43	0.31	0.13
MNCS	PP_INT_COLLAB	S	0.3	0.29	0.08	0.05	0.09	0.06	-0.06
		K	0.21	0.2	0.05	0.03	0.06	0.04	-0.04
PP_TOP_PROP	PP_INT_COLLAB	S	0.29	0.23	0.11	0.11	-0.02	-0.26	-0.41
		K	0.22	0.21	0.07	0.08	-0.01	-0.19	-0.33

According to Table 1, the correlation coefficients are quite dispersed. This shows that the correlations are not constant on the entire domain of the variables, but vary within interquantile ranges. Furthermore, we investigate the correlation in the right tail of the distributions, that is, for the 90% highest observations (90th percentile) and 95% highest observations (95th percentile). Relatively high correlations for these percentiles suggest a thicker tail. The correlation coefficients of these bins will be of help with the copula selection. It can be observed that some pairs of indicators are more correlated in the upper part of the distribution, such as (P,MCS) and (P,MNCS), while other pairs are more correlated in the lower part, such as (MNCS, PP_INT_COLLAB) and (PP_TOP_PROP). Despite the fact that the two normalized citation indicators, MNCS and PP_TOP_PROP are overall highly correlated, they are not so much correlated in both tails, that is for both low and high corresponding values. It is quite notable that high values of both normalized citation indicators do not correlate with international collaboration. Moreover, researchers with a very high percentage of publications in the top 10% of their field seem not to collaborate internationally, hence the negative correlation between PP_TOP_PROP and PP_INT_COLLAB for the 90th and 95th percentile.

Generally, the variables are less correlated in the middle of the distribution, that is between the first quartile (25th percentile) and the median, and also between the median and the third quartile (75th percentile). The question is, of course, how much of this correlation heterogeneity can be described by accounting for researchers' field. We now consider the correlation measures between the indicators within *Social Science* subset of researchers and investigate the differences with respect to the correlation measures for all the data. The results are depicted in Table 2.

Table 2. Spearman (S) and Kendall's tau-b (K) correlation coefficients for the bibliometric indicators for *Social Science*, before the 25th percentile (Q1), between the 25th percentile and 50th percentile (Q1-Q2), between the 50th percentile and the 75th percentile (Q2-Q3), in the 90th and in the 95th percentile.

V_1	V_2	*Correlation*	*All Data*	$\leq Q1$	$(Q1, Q2)$	$(Q2, Q3)$	$\geq Q3$	$\geq$ **90th** *percentile*	$\geq$ **95th** *percentile*
P	MCS	S	0.49	0.17	0.11	0	0.22	0.17	-0.31
		K	0.35	0.13	0.08	0.002	0.15	0.12	-0.21
P	MNCS	S	0.3	0.22	0.08	-0.05	0.08	0.02	-0.34
		K	0.21	0.17	0.05	-0.04	0.05	0.01	-0.23
P	PP_TOP_PROP	S	0.39	0.23	0.06	-0.01	0.11	0.05	-0.32
		K	0.28	0.2	0.04	-0.02	0.07	0.02	-0.22
P	PP_INT_COLLAB	S	0.3	0.32	0.1	0.1	0.07	-0.14	-0.02
		K	0.21	0.26	0.07	0.06	0.05	-0.09	0.01
MCS	MNCS	S	0.77	0.59	0.27	0.47	0.55	0.69	0.58
		K	0.6	0.44	0.19	0.33	0.4	0.51	0.43
MCS	PP_TOP_PROP	S	0.71	0.21	0.18	0.37	0.49	0.56	0.29
		K	0.54	0.17	0.13	0.26	0.34	0.41	0.22
MCS	PP_INT_COLLAB	S	0.37	0.37	-0.04	0.13	0.20	0.25	0.24
		K	0.26	0.28	-0.02	0.1	0.14	0.19	0.14
MNCS	PP_TOP_PROP	S	0.84	0.3	0.39	0.48	0.71	0.52	0.55
		K	0.7	0.24	0.28	0.34	0.54	0.38	0.44
MNCS	PP_INT_COLLAB	S	0.36	0.34	0.02	0.09	0.13	-0.04	-0.33
		K	0.26	0.23	0.02	0.07	0.08	-0.04	-0.25
PP_TOP_PROP	PP_INT_COLLAB	S	0.33	-	0.18	0.12	0.14	-0.19	-0.12
		K	0.25	-	0.13	0.08	0.11	-0.13	-0.07

The first notable difference is the significant number of negative correlations, some of which are non-negligible, in the upper tail. Nonetheless, the correlation between PP_TOP_PROP and PP_INT_COLLAB is weaker in the tail for researchers in *Social Science* as compared to all researchers. Comparatively, the correlation between MNCS and PP_INT_COLLAB is higher in the right tail, indicating a strong negative relationship. Alternatively, the correlation between MNCS and PP_TOP_PROP is stronger for researchers in *Social Science* researchers.

Finally, the missing correlation coefficients in the first interquantile range of PP_TOP_PROP are due to the fact that all indicators are zero.

The tail dependencies are sometimes not so straightforward to read from a scatterplot. As an example, consider the scatterplot of MCS and MNCS in Figure 3 (left). Considering the corresponding distribution values, hence transforming the variable to standard uniform can represent a visual aid, as exhibit in Figure 3 (right).

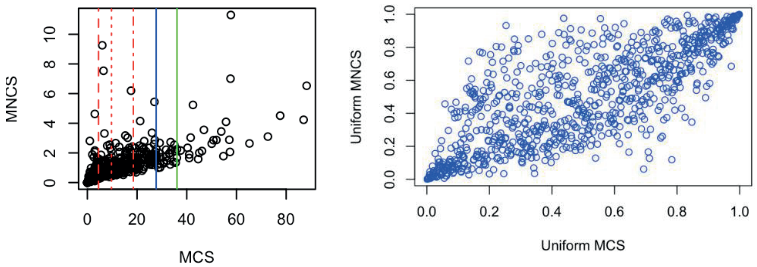


Figure 3. Scatterplot (left) and pseudo-observations (right) of MCS and MNCS for *Social Science.* The lines (from left to right) denote the first quartile (Q1), median, third quartile (Q3), 90th and 95th percentile.

As already observed from Table 1, MCS and MNCS are stronger correlated in the tails, whereas in the middle of the distribution, the correlation is weaker. This is nicely depicted in Figure 3 (right), which plots the pseudo-observations. This graph is typically the first visual step in fitting a parametric copula. Pseudo-observations from six pairs of indicators are included in Figure 4 below.

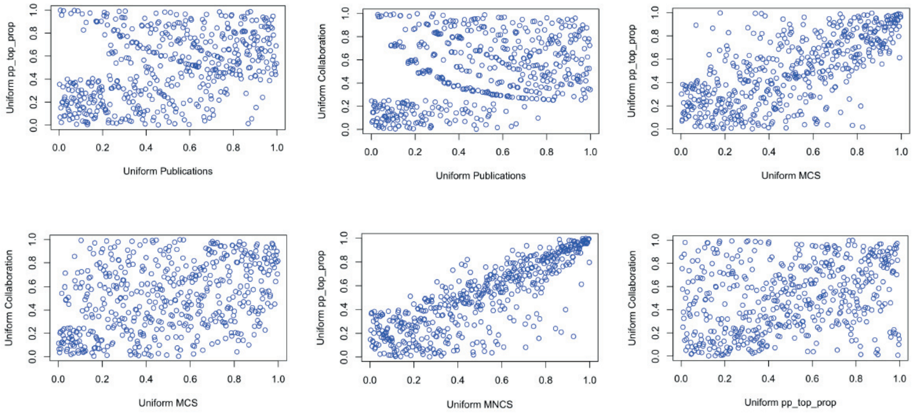


Figure 4. Scatterplot of the pseudo-observations in the division *Social Science* for different bibliometric indicators.

Once a parametric copula family is selected, the parameters of the copula are estimated using a maximum likelihood approach. Observations can be simulated from the resulting parametric copula and compared to the pseudo-observations. For the indicators MCS and PP_TOP_PROP, for all the data, the graph of the pseudo-observations (left) and simulated observations (right) are presented in Figure 5.

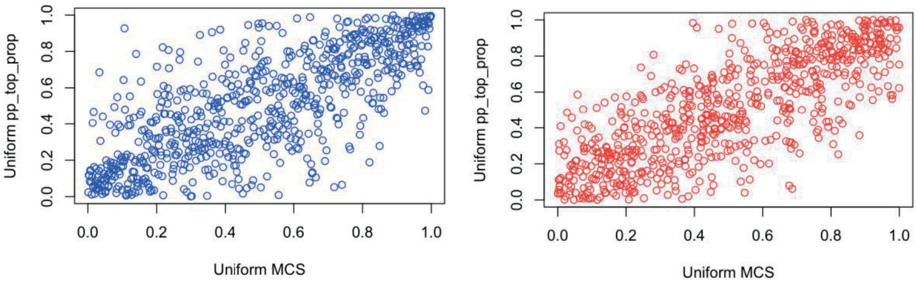


Figure 5. Pseudo-observations (left) and simulated observations (right) from Frank copula, for all data.

We have chosen to fit a Frank copula to the pair (MCS, PP_TOP_PROP) and the estimated parameter is 6.83. Visually, Frank copula seems a reasonable fit, since it manages to capture the lower and especially the upper dependence. Nonetheless, performing a goodness of fit test reveals a p-value of 0.0002, which rejects the hypothesis of a good fit. This mostly has to do with the fact that, even though the tails are modelled well by the copula, the middle part of the distribution is not well captured.

As mentioned in the previous section, the issue of ties needs to be dealt with before fitting copulas and evaluating their fit. Not accounting for ties affects both the selection of copulas, as well as the result of the goodness of fit test. To acknowledge the ties and to illustrate the method of Kojadinovic and Yan (2010), we propose an example for the overall researchers data using variables PP_TOP_PROP and PP_INT_COLLAB, in Figure 6.

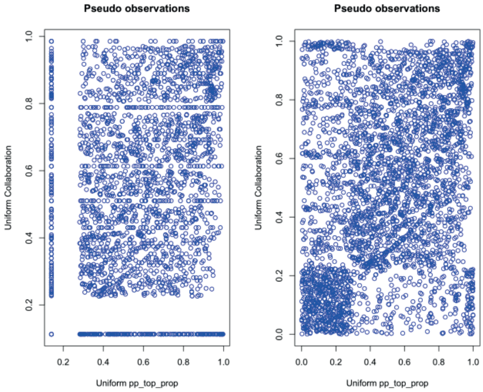


Figure 6. Pseudo-observations of PP_TOP_PROP versus PP_INT_COLLAB, for all data, without accounting for ties (left) and when accounting for ties (right).

The ties are especially visible for researchers that do not collaborate internationally, and those whose publications do not belong in the top 10% most cited papers in their field, which are represented by the lines along each axis. When the ties are randomized, we notice the correlation (concentration) among small values of PP_TOP_PROP and PP_INT_COLLAB. This is also noticeable in the correlation coefficients in Table 1.

We finally present the results of fitting parametric copulas in the division *Social Sciences*. The results for all pairs of bibliometric indicators have been summarized in Table 3. The analysis has been performed in R, by using built-in functions in the package VineCopula. The BB8 copula, also known as Joe-Frank copula is described in Joe (1997). For each pair, all the available parametric copulas have been fitted and their corresponding AIC computed. The parametric copula with the lowest AIC has been reported, along with its maximum likelihood parameters. The Kendall's tau-b correlation coefficient of these copulas are also included. Note the slight differences with respect to the empirical correlation coefficients in Table 2.

Table 3. Output of copula selection for the division *Social Sciences* when ties are randomized.

V_1	V_2	Copula	Param1	Param2	Kendall tau-b	AIC	p-value
P	MCS	Survival Tawn 1	1.88	0.55	0.31	-160.9	0.07
P	MNCS	Survival Tawn 1	1.75	0.4	0.23	-92.46	0.20
P	PP_TOP_PROP	Survival BB8	2.71	0.73	0.26	-79.58	0.00025
P	PP_INT_COLLAB	Survival BB8	1.75	0.95	0.24	-87.28	0.00025
MCS	MNCS	Survival Tawn 1	3.36	0.8	0.59	-609.3	0.27
MCS	PP_TOP_PROP	BB8	5.21	0.72	0.49	-306.9	0.0022
MCS	PP_INT_COLLAB	BB8	2.01	0.88	0.25	-81.57	0.0004
MNCS	PP_TOP_PROP	Survival BB8	5.37	0.92	0.65	-648.6	0.0037
MNCS	PP_INT_COLLAB	Survival BB8	2.29	0.81	0.25	-77.27	0.0357
PP_TOP_PROP	PP_INT_COLLAB	Frank	2.15		0.23	-54.33	0.0107

Each parametric copula selected using the AIC criterion has been tested for goodness of fit. The corresponding p-value is included in the last column of the table. A p-value smaller than the significance level of 0.05 advices the rejection of the goodness of fit hypothesis. We notice that for only two copulas, namely of the pairs (P, MNCS) and (MCS, MNCS), the obtained p-value is larger than the significance threshold.

Discussion and conclusions

Dependence is the standard approach in undertaking any multivariate data analysis and statistical modelling. Understanding the existing dependencies within a dataset is an essential step and no shortcuts should be undertaken. The dependence structure can be described more complexly than by using single correlation estimates, which is by making use of copula functions. Copula functions emphasize the need to explore existing dependencies relationships in bibliometrics beyond correlations.

A sample of 3574 Quebec researchers has been used to model copulas. The dataset includes a heterogeneous set of researchers, covering 9 divisions, and who have a wide range of bibliometric performance. The correlations between researchers' bibliometric indicators are quite spread, and are measured by the Spearman rho and Kendall's tau coefficients. The interquantile ranges of the indicators' distributions reveals that the correlations change, sometimes significantly, with respect to the overall correlation coefficients. These differences emphasize, for example, strong tail dependences, such as for the two well know field-normalized citation indicators MNCS and PP_TOP_PROP. The middle part of the distribution is nonetheless weakly correlated, suggesting that the dependence between the two indicators is more complex. This kind of dependence can be well modelled with copulas and certain parametric families can model this dependence structure.

This study revealed interesting insights into the dependence structure between bibliometric indicators. Accounting for researchers' division unfolded even more diversity of the dependence structures. Due to space limitation, we only addressed the division *Social* Sciences. An extended version of the paper will account for the differences across divisions. The question is, of course, if the Quebec researchers provide a representative sample and if the findings can be generalized to other researchers or research communities. While certain parametric copulas fitted well the data, we notice the limitation of the current parametric copula families in fitting bibliometric data for *Social Science* division. Despite the modest fit of the parametric copulas for this division, the fit for other divisions seem more promising and encouraging. Nonetheless, it also reveals the need for a nonparametric approach in the modelling. Finally, this work has considered only bivariate copulas. Multivariate copulas generalize the concept and should be definitely explored in the future.

References

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics*, *5*(1), 101-113.

Costas, R., Nane, T., & Larivière, V. (2015, June). Is the Year of First Publication a Good Proxy of Scholars' Academic Age?. In *ISSI.*

Gingras, Y., Lariviere, V., Macaluso, B., & Robitaille, J. P. (2008). The effects of aging on researchers' publication and citation patterns. *PloS one*, *3*(12), e4048.

Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance: is there an advantage in using multiple indicators?. *Research policy*, *32*(8), 1365-1379.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC.

Joe, H. (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

Kojadinovic, I., & Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, *34*(9), 1-20.

Nane, G. F., Larivière, V., & Costas, R. (2017). Predicting the age of researchers using bibliometric data. *Journal of informetrics*, *11*(3), 713-729.

Nelson, N.R. (2006). *An introduction to copulas*. New York: Springer. 2nd edition.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, *8*, 229-231.

Torres-Salinas, D., Lopez-Cózar, E. D., & Jiménez-Contreras, E. (2009). Ranking of departments and researchers within a university using two different databases: Web of Science versus Scopus. *Scientometrics*, *80*(3), 761-774.

Torrisi, B. (2013). Academic productivity correlated with well-being at work. *Scientometrics*, *94*(2), 801-815.

Van Raan, A. F. (2006). Statistical properties of bibliometric indicators: Research group indicator distributions and correlations. *Journal of the American Society for Information Science and Technology*, *57*(3), 408-430.

VineCopula R package https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). On the correlation between bibliometric indicators and peer review: reply to Opthof and Leydesdorff. Scientometrics, 88(3), 1017-1022.

Appendix 1. Descriptive statistics for the dataset.

	P	MCS	MNCS	PP_TOP_PROP	PP_INT_COLLAB
Mean	26.99	17.53	1.35	0.13	0.29
St. dev.	36.89	35.13	1.68	0.16	0.27
Min	1	0	0	0	0
Max	777	1550.5	47.33	1	1

Performance of Research Teams: results from 107 European groups

Ulf Sandström[1], Peter van den Besselaar[2]

[1] *ulf.sandstrom@indek.kth.se*
KTH Stockholm (Sweden)
[2] *p.a.a.vanden.besselaar@vu.nl*
Vrije Universiteit, Amsterdam (the Netherlands)

Abstract[1]

This paper investigates what factors affect the performance of research teams. We combine survey data about the team with bibliometric data about the performance of the team. The analysis shows that teams with a few PIs perform better than single PI teams – of course controlling for team size. On the other hand, gender diversity does not have an effect on performance. The good news is that gender objectives can be realized, without any performance problem.

Introduction

A long discussion exists on the role of (gender) diversity on the performance of research teams. Teams gather resources and use them to publish papers. In that process of changing inputs into outputs, there are a number of factors that might have an impact on efficiency. In this paper, we include a considerable number of team characteristics, based on a survey among members of some 100 teams, and on bibliometric data covering those members. The resulting set of data is a rich source of analysis. We investigate whether gender diversity has a role, whether context factors like type of organization or national research systems have a role, whether team composition would change the output levels, and we look into team dynamics.

Team diversity is a complex question and it is not obvious how to do research concerning this topic. A useful definition is provided by Jackson et al. (2003): "We use the term diversity to refer to the distribution of personal attributes among interdependent members of a work unit." Both general similarities are of importance for a group of people working in non-routines task but at the same time, there is a need for cognitive diversity in order to bring in different perspectives for solving the immediate problems. One example of this is gender diversity: this specific dimension (also of political interest) has been debated for quite some years. Some authors claim that gender diversity would lead to more efficient use of scarce resources as well as to increasing the scope and impact of research (e.g. Bear & Wolley 2011; Nielsen et al 2017). Research has been targeted at different levels: individuals, e.g. increasing the number of women researchers; on organizations, e.g. developing and implementing gender equality plans; on research projects, e.g., integrating the gender dimension in research to increase quality and relevance. However, as today's research to a large extent is based on collaboration in teams, the effect of gender diversity on research can probably be expected to be part of the effect of team composition, team culture, and team dynamics.

One of our aims is to examine claims that gender diversity is beneficial to science and research, using data on gender diversity in research teams and on research performance at the team level. We focus on four aspects of research performance: field adjusted volume of publications (total

[1] We acknowledge the Gedii project partners for their contributions, especially in collection of data and the construction of survey questions. The research was funded by the Gedii project (author 1) and the GRANteD project (author 1 and 2).

production) as well as productivity per senior, and field normalized citation-scores (totals and per senior) as a marker of impact. In general, we take a variety of team characteristics into consideration: team composition, team dynamics, team culture, and a variety of dimensions related to gender diversity.

The paper is organized in the following way. *First*, we provide an overview of research on team research and small groups, a body of work that has examined how to make teams more effective. *Second*, we outline our methodological approach. We provide an account of how the bibliometric datamining and data analysis was conducted. Also, the variables used are described in more detail. *Third*, we discuss the results of our models including team characteristics and research performance. We conclude by a discussion of the significance of our findings. Last but not least we discuss the limitations of our work and suggest further research avenues.

Research teams and team science

In the analysis we identify three different types of teams: 1) the PI-led group with a number of post doc's and PhD's; 2) the research teams with about 2-4 seniors and assisting personnel including post docs and PhD's; 3) team science based on large-scale initiatives or consortia's of research teams (Stokols et al. 2008). In this paper we intend to focus on the two former types of research teams and will try to keep the third, "team science", questions aside as they have a slightly different focus: "Most science teams and larger groups are geographically dispersed, with members located across multiple universities or research institutions" (NRC 2015).

Why do researchers team up in more or less stable groups? An increased need for channels for knowledge flow between scientists is the main cause why collaboration is becoming more important over time since decades ago (Adams et al. 2005). In that situation there are two different *strategies* that can be applied for research teams: First, to keep the group small (PI-led, one senior) and collaborate on papers nationally and internationally. Second, another strategy would be to join forces with a small number of qualified seniors who hold complementary competencies/skills and at the same time similar views and cognitive capabilities. To find these attractive components in long-distance collaborations can be costly, and, of course, to find these colleagues for co-location (same city) might include search costs and need for mobility of staff.

Many factors of this type have an influence on team performance, as a recent review of team science (Hall et al. 2018) shows, but there is a low level of precision; and not at all clarified how to operationalize the categories. De Saá-Pérez et al. (2017) points at two different paradigms in a research group or team research: on the one hand similar-attraction and on the other and the cognitive resource diversity (c.f. Hurley 2005). The former accentuates that similarities within the group might promote mutuality and cohesion within the group and that diversity would spur conflict and tensions. The latter paradigm, cognitive diversity, underscores that there should be unique combinations of cognitive resources brought together by team senior members. As often is the case there is need for a balanced approach, which here would translate to a combination of similarity and diversity but in different respects (ibid.). The similarity could be related to bio-demographic dimensions and diversity to the cognitive task-related issues, but there are many other possible combinations.

Vague concepts like the ones mentioned here are often of the type that they hardly can be operationalized, as they include parts that can be interpreted from several perspectives. This has spurred contradictory findings (van Knippenberg & Schippers 2007). We propose for the "balanced version" of diversity to use the concept of inwardness ("internal network density") which is the ratio between the number paper fractions from the own team in relation to all fractions within the total paper network. With this concept, we cannot distinguish between similarity and/or diversity but we can build an indicator which is sensitive to the capability of the group to contribute to larger parts of papers without the help from outside. Probably, this demands some of the features from the similarity side, to ease the collaboration, but it also makes it necessary to understand in the team how different competencies contribute to the cognitive tasks the team is working on. Hence, the more the groups have of inside competence and capability the more the team produces by itself, but that is dependent on a similar vision and a similar overall background of understanding within the specialty that is under investigation, as well as a willingness to collaborate with others outside of the team.

Connected to this inwardness is another important factor: seniority which gives the basis for richer research experience, better cognitive resources, and levels of prestige; all factors instrumental for the production of knowledge. Seniority might also be needed for even more advanced levels of research productivity (Hinnant et al. 2012).

There is quite convincing evidence that team size has an effect, mostly in the form a 'critical mass' thresholds and taking the form of an inverted U-curve-pattern: productivity seems to rise with increasing size of team up to about six or eight persons, above which there is usually little or no extra gain per researcher (Von Tunzelman et al. 2003; Verbree et al. 2015; De Saá-Pérez et al. 2017)

Several other bio-demographic variables affect research performance and they almost inevitably intervene to complicate or twist any simple relationship between size and efficiency. The age structure, of both individuals and institutions, is often found to be relevant to research performance. (Von Tunzelman et al. 2003; Verbree et al. 2015). Career phases also matter. Although smaller groups produce less output about a more restricted range of research topics, they are easier to manage— especially for less-experienced group leaders—and the coordination costs to organize scholarly communication and collaboration among group members are much lower (Carayol and Matt 2004, 2006; Heinze et al. 2009).

These team composition factors are complemented by factors consisting of the *context* for team activities. That could be the larger organization type (university, public research institute, company etc.), but also the fact that research groups are embedded in a national science system which affects important aspects of research and differ between national systems, as do governance and regulations which in turn influence the performance levels (Bonaccorsi and Daraio 2005).

Team dynamics concerns time-dependent factors like average team tenure (time in the team) reflecting staff mobility, the share of temporary contracts, and age of the personnel. These are quite straightforward variables and it is easy to see the influence of time for developing a common team culture and a similar cognitive understanding for how and with what different members can contribute to the team production. Kozlowski and Ilgen (2006) show that social interaction, sharing of perspectives, and collective sense-making might have a considerable effect on team performances. We introduce a number of variables for *team culture* in order to test whether these are important for team performance, and we keep in mind that these factors also coincide with aspects of team composition.

One topic in the literature is the role of gender diversity on team functioning and performance. On the whole, there is no consensus yet from this body of work on whether gender diversity has a positive effect on group performance or not. Nielsen et al. (2017) describe several experimental research items on team science pointing towards a positive productive team mechanism in problem-solving due to diversity. However, as Campbell et al. (2013) make clear, there are several and quite differing results when it comes to actual performance. In fact, many results find a negative effect on performance (Bowers et al 2000; Stewart et al. 2006; Webber et al. 2001).

Where a positive relationship is found, this is attributed to gender diversity allowing for more complex tasks to be solved, for more collective intelligence to support the team, or for more social sensitivity. Recent studies based on topic modeling and text-mining indicates that gender diverse groups seem to have a higher propensity to search outside of the box (see Nielsen et al. 2017 for refs). Likewise, Joshi (2014) has done compelling research showing that recognition of team member competence differs in male-dominated compared to groups dominated by the other sex. These results and reasoning are transposed and used in policy documents. For example, the European Commission states that "*European research still suffers from a considerable loss and inefficient use of highly skilled women"* (e.g. European Commission 2012 page 12) that hinders both the quality and relevance of research. Calls are made for the potential of women in the scientific workforce to be taken on-board and set in motion, as a strategy to make the most of hitherto unused competent researchers.

Furthermore, it remains an issue how to measure the performance of teams. Here we focus on the scholarly output and measure this using bibliometric data – which is at the team level an accepted way of performance measurement (Hinnant et al. 2012). However, what indicators should be used is still open for discussion: here, we introduce an innovative method for this which takes care of differences in means of production between areas (Koski et al. 2016; Sandström & Wold 2015).

From all the factors that may influence team performance, we use in our view the most important subset: we hypothesize that team performance is influenced by some contextual variables, by team composition, and team culture, by team dynamics, and by gender diversity in teams. Several variables we use are new, such as the overall indicator for gender diversity, but also the variable "coverage of needed skills", i.e. inwardness. The latter refers to how complete the team covers skills needed, and could be seen as measuring the level of independence from the environment. The variable "team dynamics" measures the stability of team membership. Not included in our model are the processes leading to the establishment of teams. However, we use "objective" data based on author address collected from the bibliometric files per individual/team in order to categorize the team as co-located (in the same city) on the one hand and nationally or internationally dispersed. This together with information on cohesion (coverage of skills needed for article production) we would guess that set of data we use constitute a more trustful and comprehensive data set than survey data only on meetings and communication within the team.

This leads to the following model:

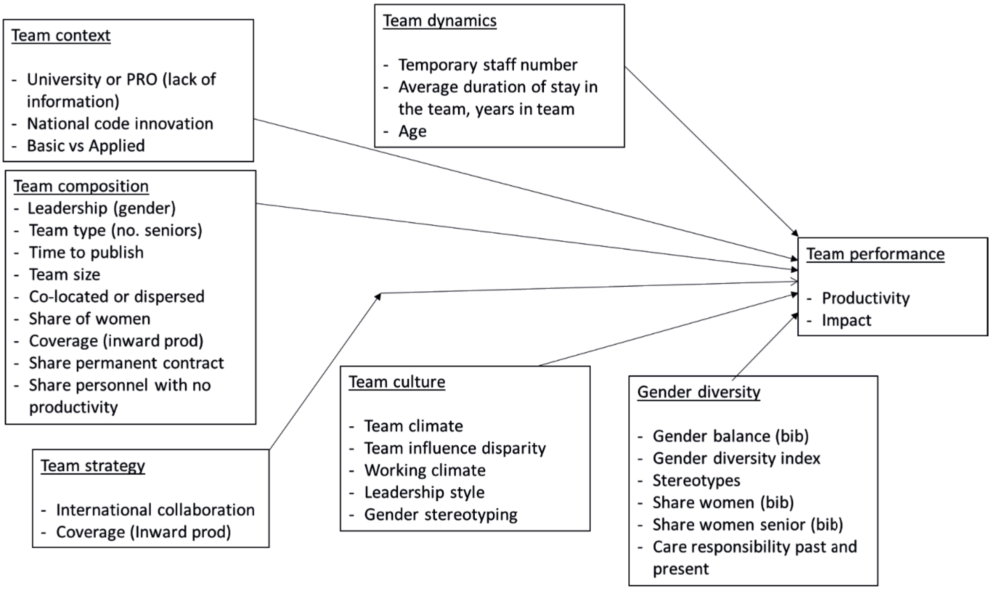


Figure 1: The model with its different components.

Methods and data

The initial sampling strategy was to focus on Transportation Studies and Biomedical Engineering. We retrieved from the Web of Science (WoS) all authors who have at least three publications in the period from 2011 until 2016 within the two subject categories. Based on this information all their publications in WoS were retrieved in order to achieve information about the e-mail address for potential research group leaders. These procedures were the major part of the recruitment process, and this takes us immediately outside of the two starting subfields, actually all fields are activated in this study. The survey is based on responses from seventeen EC countries. Altogether, 159 teams participated in the survey, representing 1,357 individuals, but due to applied rules for selection, there are 107 research teams representing 1,272 individuals in the final round for analysis. Due to missing values, 94 teams were used for the analysis.

The sample of research teams which participated in the surveys is based on self-selection, with oversampling of transportation and medical engineering, and of highly productive researchers. So, the sample consists of research-intensive productive teams. There were four contacts with the teams: First to ask them to participate, and to provide a list of team members. Second a survey among all the members of the teams, with a response rate of 77%. The survey included scales measuring gender stereotyping, diversity climate (Settles et al. 2006), team influence (Curşeu and Sari 2015), team climate (Anderson and West 1998), and team leadership (Berger et al. 2011). In addition, items regarding team communication (Pinto and Pinto 1990), care responsibilities, mentoring, working conditions, science communication activities, as well as acquired funding, were included. Thirdly, the bibliometric data were collected and sent to the contact person for validation. Fourthly, a short survey was sent to the team leader asking for information regarding the team such as its founding year, co-location of members, working methodology, size, and gender composition.

This study is, to our knowledge, one of the few that have taken the opportunity to combine survey data with advanced bibliometric performance indicators. Our design is similar to Verbree et al. (2015), but that study uses the research leader as an informant for the group which may affect the quality of the measurement.

Variables in the model

We have chosen to use four *dependent variables* in the analysis: 1) *Production* which covers the last 5 years of publications from the whole team no matter where their publications were produced. Five years to facilitate for a comparison between groups no matter when they were started or when members joined the team; 2) *Productivity* which is in this case FAP (Koski et al. 2016) divided by the Number of bibliometric seniors; 3) *Impact* is the influence over subsequent literature based on FAP figures but with Percentile Model applied (for further explanation see Sandström & Wold 2015), and. 4) *Impact per senior (bib).* More on the calculation of the respective indicators are given in a later section (see the headline: Dependent variables – the performance indicators). In this paper, team role has not been taken into consideration. This implies that everyone that was assigned membership at the moment of initial recruitment (which includes administrative personnel, master students etc.) affect the averages for all team level data from the survey. The following *independent variables* were used:

Team composition is measured through various variables. 1) *Team type* is based on number of senior researchers with a bibliometric profile over at least five years. We distinguish two types, namely (i) single leader teams; (ii) teams with 2-4 leaders. 2) *Coverage of skills* needed also named *inwardness* and measured as the share of publication fractions within the group related to all publications (bibliometric variable). 3) *Mean time available to publish*/patent (based on the survey). 4) *Team size* (provided in initial contact). 5) The percentage of women in the team (average of the scores of the surveys). 6) The share of senior team members.

Team culture has several dimensions, measured through the team member survey: 1) *Team climate*. 2) *Team influence (*disparity). 3) *Working environment climate*. 4) *Leadership style*. 5) *Gender stereotyping*.

Team dynamics is hard to establish using cross-sectional data but we use a few proxies. 1) Number of staff with a *temporary* contract. 2) Average *duration of stay* in the team. 3) Average *age* of team members.

Team context may affect performance, such as the 1) the *national context* – here measured as a dichotomous variable separating high performing and less high performing science countries. 2) The type of *organization* they are embedded in (University versus Public Research Organizations). 3) The level of *co-location* with three values: co-located, nationally dispersed, and internationally dispersed. 4) Applied vs. basic research, based on the categorization of ScienceMetrix (http://science-metrix.com/?q=en/classification).

Gender diversity is measured in several ways. 1) Firstly gender balance in team membership, using the Blau index (Blau 1977). 2) Secondly, a composite indicator was developed: the Gender Diversity Index (GDI), covering gender processes along seven diversity aspects: age, education, care responsibilities, marital status, team tenure, seniority and type of contract (Humbert & Günther 2018).

Method

Because of overdispersion of the dependent variables, we use from generalized linear models the negative binomial version. Multi-collinearity can represent a significant source of error in modeling. An indication for this is the strong change in regression coefficients when including new variables in the model – which is not the case. We then use the independent variables block by block (see Figure 1). After having done this, we keep from each block the (marginally) significant variables (p <0.20) for the final model. See Table 4 for the results.

Results

Team production: Production increases with teams that include 2-4 seniors that have a bibliometric profile covering several years of the period. Having a single PI-led team is clearly negative for production. This can be understood in terms of complementary skills in the team which is accentuated by the inward factor. Hence, in most cases, we find that higher levels of inwardness, greater coverage of needed competencies in the team makes it easier to publish papers, and not being dependent on colleagues from outside. Being able to do more inside the team increases paper production to higher levels. Overall production becomes better.

Regarding the variables **Gender Diversity Index** and Gender Balance, it turns out that both variables have a weak or no significant effect. Other factors are stronger and more important for the production figures. Among them, we should not forget the time devoted to publishing or patenting activities. As reported by the respondents this has a significant effect on production. Of course, this is a trivial result but might also be understood as a validation of the survey instrument as such. That production becomes higher with more senior members is likewise trivial, here it serves as confirmation of the overall design.

Team productivity: Productivity per senior is a more relevant indicator of performance. We have worked from the hypotheses that seniority is a crucial factor and that there is a need for seniors with a consecutive production of papers in Web of Science, i.e. in international scientific journals. However, the implication is not that more junior personnel also contribute in the same way to productivity. As shown in Table 4 this, team size is significantly negative, thereby indicating that the theoretical prediction seems to be corroborated. Having more personnel in the learning process might take down per senior productivity.

Team type is for productivity a significant *and* positive variable. Teams with two, three and four seniors have a significantly higher output per person. These groups might have more efficient team communication and a better understanding of the common objectives which translates into productivity. Again, team coverage of needed competencies (inwardness) is significant and contributes strongly to productivity. Unfortunately, we do not have communication patterns between seniors as a question in the survey, and therefore we use inward as an indicator for several processes of that type.

Surprisingly, *team culture* and *team dynamics* seem unrelated to productivity. Another slightly negative influence is the percentage of women in the team. It is not a significant factor but this result goes well together with a long series of research and this productivity gap can actually be traced back to the 1930s according to Cole and Zuckerman (1984). Since the 1930's it seems to be the case that women produce two-thirds only of what men produce in general (p.221). These results were confirmed by Xie & Schauman (1998) for the period 1960's to the 1980's. Recently Van den Besselaar & Sandström (2017), based on a Swedish dataset, and showed that these patterns were still at hand also after the millennium: "women are vastly underrepresented in the group of most productive researchers". Looking into the details we find that 80% of the highly productive individuals in the sample (having more than FAP points five times over the normal) are male researchers.

Table 4. What influences research performance?

	Production	Productivity	Impact	Impact/senior
	Beta	*Beta*	*Beta*	*Beta*
(Intercept)	7,200***	6,047***	8,660***	8,014***
Org type- Univ	0,16 (ns)			
PRO	0,17 (ns)			
Other	0[#]			
Nation high	0,22 ns			0,471**
low	0[#]			0[#]
Team Type 1	-1,1***	-0,9***	-1,09***	
Type 2	0[#]	0[#]	0[#]	
Geography Co-located	0,1 (ns)			
Distributed	0[#]			
Orientation-Applied				
Basic				
Team size		-0,04**		
Senior(bib)				
Time to Pub	0,38***	0,352**	0,612***	0,378**
Inward Pub	4,05***	2,711***	2,822***	1,393
Women share		-0,05 ф	0,082 ф	-0,595 ф
PowerInfluenceDisparity				
Team Climate				
Leadership Style				
Working Climate				
Gender Balance	0,18 (ns)	0,086 (ns)	0,63 ns	
Gender Diversity Index	0,04 (ns)	0,408 (ns)		
Stereotypes		0,367 (ns)	0,101 ns	
Temporary staff	0,03 ф		0,02 ф	0,046 ф
Yrs in team	0,00 ф	0,015**	0,001 ф	
Age personnel				
Experience in area(yrs)				
Observations	86	86	86	86
Model fit(score/df)	0,923/61	0,946/88	1,096/71	1,289/87
AIC	1558	1505	1825	2040

Variables not included in this table are found (marginally) significant in the analysis per analytical dimension. Plus all gender diversity-related variables (also when not at all significant in the analysis per dimension).

# When dummy variables are applied redundant variables are set to zero.

Significance levels ns = non-significant, ф < .20, *<.10, **<.5, ***<.01%

Note: Based on dataset Müller et al. (2019).

Team Impact: The so-called percentile model applied here takes citation impact into account. This indicator expresses the total production and impact in one score. The pattern that evolves from Table 4 is very much the same as the one based on the Productivity indicator. However, there are differences and there are two factors that should be stressed in this context: First, team size is not significantly related to impact, and clearly the female share of the team is as a factor unrelated and not significant. The latter result is also in line with results from the literature. Van den Besselaar & Sandström (2017) showed that women are not among the high profiles of impact, but in the regions where women have the same number of papers the share of top papers are about the same as for men, i.e. no difference in impact and there are a several of investigations that have given evidence on this point.

Impact per senior: When impact is investigated as impact per senior (bib) a slightly different result emerges. The model is statistically weaker, and fewer variables are significant. Findings show that team type no longer influences the average number of impact points, but overall the same pattern emerges again concerning what factors that have an effect on citation performance. Female share is negatively affecting the performance, inward is only marginally operative and time for publication is strongly and positively significant. Interestingly, the variable "Nation" now comes back as a strong factor; highly innovative countries – according to the EC indicators - seem to produce more influential research which is taken up and discussed in subsequent work by colleagues. In that respect, the Nordic countries and north-western Europe seems to perform better than southern countries. Here we should remind ourselves about the self-selection effects for the representability of the sample. This is important as several of the groups from the second round happen to be high performing groups.

Conclusions and discussion

The overarching research question for this paper is whether gender diversity makes a difference in the shaping of research teams performance? To answer that question three types of data were used: 1) a list of members of research teams, 2) survey data from these groups (77% response rate); 3), bibliometric data covering all (100%) members listed as participating in the teamwork.

We developed a model and consisting of different variables that could be hypothesized to affect performance: we divided the variables into different blocks representing a multitude of aspects of team research that we have found to be of interest based on literature studies. The model includes team context, team composition, team culture, team dynamics, and gender diversity. The latter has been a general point of departure for the project in the context of which this paper has been written. However, *the gender variables seem to be without significant effect*, other variables are more important for the dependent variables which cover production, productivity, and impact. This holds for the GEDII composite gender diversity indicator – which may due to the problematic nature of composite indicators – but also for other gender-related variables.[2] One gender relevant variable, well known from earlier research, negatively affects productivity in research teams: the share of women in the team.

What is important for creating successful research teams? Our findings indicate firstly that teams with a few experienced researchers with seniority tend to have greater production of papers than teams of single PIs. Such combinations are crucial for producing high-level output. Secondly, we found that team composition matters. When the senior team members complement each other in terms of skills and resources (and there are components of similarities, e.g. in the understanding of objectives, of how to handle research questions), the effect on performance is positive. Basically, this is what we find with the coverage or inward factor.

The sample did not include many groups with an interdisciplinary composition of competencies (measured as publication activity in different ESI fields), but many teams do have variations of specialization, so-called "small IDR". For example, in a team focusing on materials for dentistry, one of the senior members had a main specialization in materials

[2] We tested this (not reported here).

science, but a second one in dentistry, whereas another senior member had most publication in dentistry, but material science as the second specialization. Probably this can be interpreted in the following way: the constellation is built on a common understanding of a research problem and competencies are gathered around this problem, some with a more emphasized grounding in the first ESI field and some with a concentration in the second ESI field. They can do a lot of papers together; they have a mutual understanding and they share a vision (Heinze et al. 2009). Thirdly, also other trivial characteristics matter, e.g. time for writing publications should not be forgotten.

The next question is what factors are marginally or not important for production and productivity? From earlier research, we expected that gender diversity wouldn't be a game changer, but we wanted to know whether it would be possible to find something new using an innovative composite indicator. This is not corroborated. Are there other indications of change? We have a high number of female group leaders (34 out of 107), so that may be on its own already indicating a change in teams: what is sometimes called "transformational leadership" (Jeong & Choi 2015). The Gedii project developed a composite indicator which covers many dimensions, not only gender diversity but also functional and educational diversity. We might have to wait for more detailed studies but the results from the analysis are that these factors are of less importance both for the production as well as for the productivity. So, gender diversity does not result in higher performance, but on the other hand, it also does not imply lower performance, despite significant performance differences at the individual level between female and male researchers.[3] For impact, influence over subsequent research, we find that basic structures are still more important than diversity measures.

Also, the type of organization does not seem to matter. Independently of whether the team is at universities or public research organizations (institutes) they seem to be able to produce the same level of output per senior as long as there is room for an orientation towards research activity. Whether the team is co-located or dispersed over a country or internationally is of low importance for scholarly performance. Another finding is that the orientation towards basic research or applied research is not at all important for production and productivity. Not even the factors in the team dynamics block, such as age, number of temporary staff and team tenure, seem to be of determining performance. Finally, team culture variables did not have a significant effect.

In all, we conclude with the following: For productivity and citation impact there is one thing that is crucial and that is to construct research teams that have the capacity to combine and use different competencies in a creative way.

Limitations and future research

1) The survey is a cross-sectional analysis at a certain moment in time (2017), but the publication and citation analysis build on more "historical" data from 2012 up until 2016, for those we know were in the team in 2017. So, we perform a cross-sectional analysis, where one would prefer longitudinal data: do changes in team characteristics precede changes in performance. On the other hand, performance and team characteristics are expected to be rather stable during a short period, as is considered here. The same holds for the bibliometric performance, as good groups more easily attract good people than not so good groups. Furthermore, the performance data build on a fairly long period of time as we need to get rid of

[3] If gender diversity does not influence team performance, it may influence within team performance differences. For example, in more gender balanced teams, performance differences between male and female researchers may be smaller than in other teams.

the year-to-year fluctuations on an individual level. That also makes that quite a high proportion of the team members do have bibliometric contributions. Finally, half of the sample would qualify as Top10 % best researchers in Sweden, i.e. the best 4,700 researchers out of 47,000 in total. That the less good groups would change considerably in the next few years is possible but definitely not expected.

2) We do not control for which phase of development the research group is going through at the moment of data gathering (2017). Any type of team was considered as interesting and we did not investigate when the group was actually started and how their funding and financing had developed over time. This is a factor that eventually could supersede many of the factors controlled for in this paper. Entrepreneurial research leaders with new funding arrangements available create by themselves an atmosphere of creativity and productivity. So in future studies, these aspects might be taken into account.

References

Adams JD, Black, GC Clemmons JR; Stephan PE (2005). Scientific teams and institutional collaborations: Evidence from US universities, 1981-1999. *Research Policy* **34** (3): 259-285.

Anderson, N., & West, M. A. (1998). Measuring climate for work group innovation: development and validation of the team climate inventory. *Journal of Organizational Behavio*, **19**, 235–258.

Bear JB & Wolley AW (2011). The Role of Gender in Team Collaboration and Performance. *Interdisciplinary Sciece Reviews* **36** (2): 146–153.

Bell, ST; Brown, SG; Colaneri, A & Outland, N (2018). Team Composition and the ABCs of Teamwork. *American Psychologist* **73** (4): 49–362 http://dx.doi.org/10.1037/amp0000305.

Berger, R., Romeo, M., Guardia, J., Yepes, M., & Soria, M. A. (2012). Psychometric Properties of the Spanish Human System Audit Short-Scale of Transformational Leadership. *Spanish Journal of Psychology* **15** (1) 367–376.

Blau PM (1977). *Inequality and Heterogeneity*. New York: Free Press.

Bonaccorsi A & Daraio C (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics* **63** (1): 87–120.

Bowers CA, Pharmer JA, & Salas E (2000). When member homogeneity is needed in work teams: A meta-analysis. *Small Group Research* **31** (3): 305-3–27.

Campbell, L. G., Mehtani, S., Dozier, M. E., & Rinehart, J. (2013). Gender-heterogeneous working groups produce higher quality science. *PloS One* **8** (10), e79147.

Carayol N & Matt M (2004). Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy* **33**(8):1081-1102.

Cohen SG & Bailey DE (1997). What makes teams work. Group Effectiveness Research from the Shop FIoor to the Executive Suite. *Journal of Management* **23** No. 3.239-290.

Cole JR & Zuckerman H (1984). The productivity puzzle: persistence and change in patterns of publication of men and women scientist. *Advances in Motivation and Achievement* **2**, 217-258.

Curşeu, P. L., & Sari, K. (2015). The effects of gender variety and power disparity on group cognitive complexity in collaborative learning groups. *Interactive Learning Environments* **23** (4), 425–436.

De Saá-Pérez P, Díaz-Díaz NL, Aguiar-Díaz I, Ballestreros-Rodríquez JL (2015). How diversity contributes to academic research teams performance. *R&D Management* **47**, 2: 165-179.

European Commission (2013). *Research and Innovation Performance in EU Member States and Associated Countries: innovation union progress at country level*. <ec.europa.eu>.

European Commission (2012). *Communication from the Commission to the European Parliament, the Council and The European Economic and Social Committee and the Committee of the Regions: A reinforced European Research Area Partnership for Excellence and Growth* (European Commission, Brussels).

Heinze T, Shapira P, Rogers JD & Senker J (2009). Organizational and Institutional Influences on Creativity in Scientific Research. *Research Policy* **38** (4):610-623.

Hinnant CC, Stvilia B, Wu SH, Worrall A, Burnett G, Burnett K, Kazmer MM, Marty PF (2012). Author-team diversity and the impact of scientific publications: Evidence from physics research at a national science lab. *Library and Information Science Research* **34** (2012) 249–257.

Humbert AL & Günther E (2018). *Measuring gender diversity in research teams: methodological foundations of the Gender Diversity Index.* Deliverable D3.2 <www.gedii.eu>.

Hurley SK (2005). The Compositional Impact of Team Diversity on Performance: Theoretical Considerations. *Human Resource Development Review* **4**, 2 June: 219-245.

Jackson SE, Joshi A & Erhardt NL (2003). Recent research on team and organizational diversity: SWOT analysis and implications. *Journal of Management* **29**: 801–-830.

Joshi A (2014). By whom and when is women's expertise recognized? The interactive effects of gender and education in science and engineering teams. *Administrative Science Quarterly* **59** (2), 202-239.

Jeong S & Choi JY (2014). Collaborative research for academic knowledge creation: How team characteristics, motivation, and processes influence research impact. *Science and Public Policy* **42**: 460–473

Koski T, Sandström E & Sandström U (2016). Towards field-adjusted production: Estimating research productivity from a zero-truncated distribution. *Journal of Informetrics* **10** (4): 1143–1152.

Kozlowski SWJ. (2012). Groups and teams in organizations: Studying the multilevel dynamics of emergence. In A.B. Hollingshead and M.S. Poole (Eds.), *Research Methods for Studying Groups and Teams: A Guide to Approaches, Tools, and Technologies* (pp. 260–283). New York: Routledge.

Müller J, Busolt U, Callerstig A-C, Guenther EA, Humbert AL, Klatt S, Sandström U (2019). *GEDII Survey on Research Teams Dataset.* https://doi.org/10.5281/zenodo.2545196.

Nielsen MW Alegria S, Börjeson L, Etzkowitz H, Falk-Krzesinski HJ, Joshi A, Leahey E, Smith-Doerr L, Woolley AW, & Schiebinger L (2017). Gender diversity leads to better science. *PNAS* **114** (8):1740-1742. (opinion paper).

NRC (2015). Committee on the Science of Team Science; Board on Behavioral, Cognitive, and Sensory Sciences; Division of Behavioral and Social Sciences and Education; National Research Council; Cooke NJ, Hilton ML, editors. Washington (DC): National Academies Press (US); 2015 Jul 15.

Sandström U & Sandström E. (2009). The Field Factor: towards a metric for Academic Institutions. *Research Evaluation* **18** (3) 243–250.

Sandström U & Wold A. (2015). Centres of Excellence: Reward for gender or top-level research? In Björkman, & Fjaestad (Eds.), *Thinking ahead: Research, funding and the future*. Stockholm: Makadam Publ. (pp. 69–89).

Settles, I. H., Cortina, L. M., Malley, J., & Stewart, A. J. (2006). The climate for women in academic science: The good, the bad, and the changeable. *Psychology of Women Quarterly* **30** (1), 47–58.

Stewart GL (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management* **32** (1): 29–54.

Stokols D, Hall KL, Taylor BK & Moser RP (2008). The science of team science: overview of the field and introduction to the supplement. *American Journal of Preventive Medicine* **35** (2 Suppl), S77–89.

Van den Besselaar P & Sandström U (2017). Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLoS ONE* **12**(8): e0183301. https://doi.org/10.1371/journal.pone.0183301.

Van Knippenberg D & Schippers MC (2007). Work group diversity. *Annual Review of Psychology* **58**, 515–541.

Verbree M, Horlings E, Groenewegen P, Van der Weijden I, Van den Besselaar P (2015). Organizational factors influencing scholarly performance: a multivariate study of biomedical research groups. *Scientometrics* **102***,* 25-49.

Webber SS & Donahue LM (2001). Impact of highly and less job-related diversity on work group cohesion and performance: A meta-analysis. *Journal of Management* **27**: 141–-162.

Von Tunzelmann N Ranga M, Martin B, Geuna A (2003). *The Effects of Size on Research Performance: A SPRU Review*. SPRU Report, Sussex, U.K.

Wuchty S, Jones BF, Uzzi B (2007). The increasing dominance of teams in production of knowledge. *Science* **316** (5827):1036–1039.

Xie Y & Shauman KA (1998). Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *American Sociological Review* **63** (6): 847-870.

Are migrant inventors more productive than native ones?

Julien Seaux [1], *Stefano Breschi* [2], *Francesco Lissoni* [3], *Andrea Vezzulli* [4]

[1] julien.seaux@hotmail.fr
GREThA UMR CNRS 5113 – Université de Bordeaux ; DiECO - Università dell'Insubria, Varese

[2] *stefano.breschi@unibocconi.it*
ICRIOS – Università « L.Bocconi », Milan

[3] *francesco.lissoni@u-bordeaux.fr*
ICRIOS – Università « L.Bocconi », Milan ; GREThA UMR CNRS 5113 – Université de Bordeaux

[4] *andrea.vezzulli@uninsubria.it*
ICRIOS – Università « L.Bocconi », Milan ; DiECO - Università dell'Insubria, Varese

Abstract

We contribute to the literature on migration and innovation by comparing the productivity of foreign (Indian) and native ICT inventors in the United States, as measured by the number of patents filed and the number of citations received. We stress that Indian inventors are often more mobile in the destination country, and this could be cause and consequence of higher productivity. We have control not just for migration motives such as education, access to the labor market and cohort effects, but also for internal mobility. We do so by exploiting a rich database that merges classical information on inventors with social media. We compare migrant inventors to mobile and non-mobile natives, finding the former to be more productive. We also find those migrant inventors who entered in the United States while working for the same company perform better than migrants who changed of company or entered for education reasons.

Introduction

Highly skilled workers are the stars of today's knowledge economy. Their entrepreneurial and innovative spirit are stimulating productivity gains and economic growth. They make exceptional direct contributions, including breakthrough innovations. In this process, the mobility of skilled workers, within and across national borders has become strategic to enhance productivity.

Highly skilled international migration raised in recent decades (Docquier & Rapoport, 2012). In United-States, 19% of the tertiary educated population was foreign-born in 2013 and in certain fields such as science, technology, engineering, and mathematics (STEM) more than 30% of graduates were foreign-born (Ruggles, et al., 2010). The direct contributions of migrants in the host country are nowadays well known. Several hints of the contribution of high skilled migrants in the host country have been found for United-States. Compared to a foreign-born population of 12% in 2000, 26% of U.S. based Nobel Prize recipients from 1990-2000 were migrants (Peri, 2007) and founders of 25% of new high-tech companies with more than one million dollars in sales in 2006 (Wadhwa, et al., 2007). Stephan and Lavin (2001) show that migrants are over-represented among members of the National Academy of Sciences and the National Academy of Engineering, among highly cited authors. Migrants contribute to the host country patenting activity. Kerr (2007) shows that the share of U.S. patents awarded to U.S. based inventors with Chinese and Indian names account for 12% of the total in 2004. Nevertheless, studies comparing the foreign-born inventors' productivity with the native's ones are still scant.

We contribute to the literature on migration and innovation by assessing whether migrant inventors are more productive than the native's ones. Even though the determinants of the inventors' productivity are nowadays well known (Hoisl, 2007) (Hoisl, 2009) (Latham, et al., 2011) (Zwick, et al., 2017), research comparing migrants and natives' inventors' productivity are still scant. We do so, controlling for the inventors' intra-country mobility experience (changing of company) for both the natives in their home country and the Indians migrants at destination. Finally, we are breaking down the different types of migrants according to the channel of entrance in the destination country.

Besides the previous cited literature, one of the main reasons for the scarcity of individual level studies comparing the productivity between migrants and natives, is the unavailability of appropriate data, especially focusing on inventors. Bibliographic and procedural data such as databases provided by patent offices, don't suffice to represent the most important individual's determinants of productivity and to distinguish between the different channels of entrance in the destination country. For this purpose, we build a new and original database, mostly focused on US-resident inventors, that allows us to retrieve information on education, labor market activity and patenting activity of both migrant and native inventors. This paper exploits a panel of 40.806 inventors, from which 36.010 are United-States natives and 4.796 are Indian's migrants, observed during the period between 1969 and 2016.

This paper is organized as follows. We first present, section 2, the data alongside some descriptive statistics, followed by the model specification, estimation and the results. And, section 3 summarized the results and draws some conclusive remarks.

Data source and sample

In this section, we discuss the methodological approach adopted to test our hypotheses, the main one is to test whether migrants' inventors are more productive than natives? And provide some descriptive statistics from a new database on US-resident inventors, which includes information on their mobility patterns within the US and abroad. The database results from matching 424.497 public LinkedIn profiles, associated to employees of large Semiconductor and ITC companies active in the US, with inventor data, as found on patent documents filed by the same companies from 1950 to 2016.

We restrict our attention to inventors who have been active at least once in the US (that is, we do not consider inventors appearing on patents filed by the selected companies, but never with a US address). The ultimate goal is to enrich the inventor information one can retrieve from patent data (address at the time of the patent, name of applicant, identity of co-inventors, and other patent contents) with information on the migrant vs native status of the inventors (plus, for migrants, their country of origin and year of entry in the US), as well as information on education and labor market experience.

In order to assign a country of origin to inventors, we exploit both the information from LinkedIn (such as the country where the earliest education levels have been attained, the individual's native language, and any useful biographical detail) as well as further information such as:

- the inventor's nationality, as reported on a subset of USPTO patent applications filed according to the PCT procedure before 2011 (Miguelez & Fink, 2013).
- the results of name analysis, based on the combination of statistics on the ethno-linguistic origin of names and surnames from the IBM-GNR dataset (Breschi, et al., 2014) as well as additional linguistic analysis (Tyshchenko, 1999).

Our data also allows us to track a substantial part of the inventors' careers, most notably their mobility before the first and after the last patent filed (the two dates coincide for the vast number of inventors with just one patent over their lifetime). This solves one of the major limitations of previous studies on mobility and productivity of inventors from patent data, which were able to track the mobility only for those inventors with at least two patents, based on differences in the addresses reported in one and the other documents (Hoisl & de Rassenfosse, 2014) (Hoisl, 2007) (Hunt & Gauthier-Loiselle, 2010). Another advantage of the new dataset is that it allows us to identify and classify different types of mobility such as:

1) *Education mobility, University mobility*: when the inventors move during their education path, including the different universities in which they have studied.
2) *Labor market mobility, Company mobility*; when inventors move during their professional career (including patenting activity) and the number of different companies where they have worked.

3) Using 1) and 2) we can distinguish if the inventors move for education, employment (in changing or staying of company) or even patenting reasons.

Finally, the longitudinal nature of the data will allow us to deal with some endogeneity issues and try to establish a causal link between our independent variables of interest and productivity. The final panel consists of 40.008 inventors for which we have education, labor market and patenting information.

Descriptive statistics

Table 1 presents the general descriptive statistics. The sample consists of 40.806 inventors, 88.15% of which are natives and 11.85% are migrants. The sample is composed mainly by males (88%) aged between 25 and 77 (with an average age of 38 in 2002). This first statistic shows that our sample has a population of younger inventors with respect to other studies (Trajtenberg, 2005) (Hoisl, 2007) (Hunt, 2004) (Kerr, 2008), which had a sample of older inventors (on average). Concerning education characteristics, 65% of the migrants stayed in the country of destination after having completed their education. The level of education is equally distributed, 32% have a bachelor, 36% a master and 31% have a PhD or an MBA. The sample is composed mostly by engineers (63%), managers (17%) and almost 6% of University scholars. The other 12% are split between inventors working as head or founder of a company (CEO, President, Director…) or others (Accounting, Finance or Human resources).

Table 1: Sample general characteristic (N=40.806)

Variable	Mean	S.D.	Min	Max
Migration status				
% Native	0.88		0	1
% Migrant	0.12		0	1
- Education motives	0.65		0	1
- Labor motives	0.35		0	1
Within company	0.19		0	1
Across company	0.81		0	1
Mobility				
Total nbr. Interfirm mobility made	1.33	1.56	0	38
% Interfirm mobility	73.09		0	1
Other characteristics				
Total nbr. of patents made	13.36	20.4	1	443
Total nbr of citations received	290.56	553.2	0	14746
Age of the inventors in 2002	38.19	7.51	25	77
% Gender (1 = male)	0.88		0	1
Level of education				
- Bachelor	0.32		0	1
- Master	0.36		0	1
- PhD/MBA	0.31		0	1
Title				
- Engineer	0.63		0	1
- Manager	0.17		0	1
- Company's head	0.07		0	1
- Scholar	0.06		0	1
- Founder	0.02		0	1
- Others	0.03		0	1

Besides migration status, information about mobility after education indicate that 73% have changed company at least once, and 19% of the migrants changed country while working for the same company. Finally, on average each inventor has filed, on average, 13 patents all among his/her career (or until 2016), with an average of 290 citations received.

Table 2 provides the main variable of interest separated for Natives (column 3) and Migrants (column 6) representing 88.15% and 11.85% of the whole population. Then, we decompose the status of migration for two different kinds of mobility: stayers and movers. Columns (1) and (4) represent the native/migrant population that never changed company while columns (2) and (5) show the native/migrant population that at least have changed company once.

As Table 2 shows, changing of company, compared to migration, is a more common phenomenon in our sample, explaining an important number of movers for both the natives, 53% and migrants, 65%. Comparing the education level, we notice that migrants (column 6) are, on average, more educated than natives (column 3): 39% of the migrants have a master's degree and 41% a PhD, while 36% of the natives have a Master and only 26% a PhD. This preliminary result suggests a positive selection of migrants due to a higher level of education. This preliminary result suggests a positive selection of migrants due to a higher level of education. As shown in the previous section, we are using two different measures for inventor's productivity: the number of patents and the number of citations. First, we can observe that natives are, on average, significantly less productive than migrants: a native file on average 13.28 patents along with his/her career while a migrant produces 14.16. Second, movers are, on average, significantly more productive than stayers for both the native (12.14 vs 13.71 patents) and the migrant population (13.31 vs. 14.39 patents).

We also observe a significant difference between native movers (13.71) and migrants that changed of company (14.39). Observing the number of citations received, another intriguing result comes out: we observe that natives perform, on average, better than migrants: this could be explained by a better network quality of the natives' movers than their migrants' counterpart.

Table 2: Natives and Migrants broken down by Inter-mobility status (N=40.806)

		Natives			Indians' Migrants	
	Stayers	Movers	All	Stayers	Movers	All
	(1)	(2)	(3)	(4)	(5)	(6)
No. of inventors	12442	23568	36010	1386	3410	4796
Total nbr. of patents	12.14	13.71	13.28	13.31	14.39	14.16
Total nbr of citations	240.4	310.9	291.5	232.6	294.9	281.5
Gender (1 = male)	0.87	0.90	0.89	0.80	0.85	0.84
Level of education						
- Bachelor	0.36	0.33	0.34	0.09	0.11	0.11
- Master	0.37	0.36	0.36	0.44	0.43	0.43
- PhD/MBA	0.26	0.30	0.28	0.47	0.45	0.46
Title						
- Engineer	0.67	0.61	0.63	0.59	0.59	0.59
- Manager	0.18	0.16	0.17	0.22	0.19	0.19
- Company's head	0.05	0.09	0.07	0.05	0.07	0.06
- Scholar	0.06	0.06	0.06	0.12	0.09	0.10
- Founder	0.004	0.02	0.02	0.003	0.02	0.02
- Others	0.03	0.03	0.03	0.01	0.02	0.02

Model specification

In Table 4, we are estimating the impact on productivity for being a migrant or a native with a random effect OLS model for the two dependent variables of interest: the number of patents and the number of forward citations, according to the following model:

$$Productivity_{it} = \beta_0 + \beta_1 MIG_i + \beta_2 CMOV_{it} + \beta_3 X_{it} + \beta_4 X_i + \epsilon_{it} \quad (1)$$

In Table 5, we are then decomposing the migration status into 3 channels of entrance in the destination country: when the inventor enters during his/her study; in changing of company or in staying in the same company.

$$Productivity_{it} = \beta_0 + \beta_1 EDUC\_MIG_i + \beta_2 COMP\_MIG_i + \beta_3 WITHIN\_MIG_i + \beta_4 CMOV_{it} + \beta_5 X_{it} + \beta_6 X_i + \epsilon_{it} \quad (2)$$

We consider patent production as a direct measure of inventor productivity. The yearly number of patents that the inventors contributed to invent between 1969 – 2016, is our quantity measure (NPATit). Patent count was used in different studies investigating inventor's performance and mobility (Hoisl, 2007) (Breschi & Lissoni, 2005). Patent yearly forward citations are used to measure the importance and value of the innovation, our quality measure (NCITit). Gambardella et al. (2006) have shown that the number of citations received by a patent is a good proxy for the value of a patent. However, measures based only on the number of forward citations have some limitations (Hall, et al., 2001). For example, large firms might have larger portfolios of citing patents compared to smaller companies and universities, affecting the number of citations that their patents receive by self-citations. Furthermore, citations cannot be made to or by inventions that have not patented and so, underestimating the importance of some of them.

One of the main contributions in this paper is to compare the difference in productivity between migrant or a native inventor, controlling for other observed individuals' characteristics. We assess that an inventor is a migrant (MIGi) when his/her country of origin is different from is actual working country, here U.S., furthermore the migration status is not changing over the time for our individuals, observing them only in United-States. Hence, we are not considering possible patent filed by migrants in their origin's country. Furthermore, we break down the status of migration by 3 channels of entrance in the country of destination. An inventor can enter in the destination country for education reasons and, after that, enter in the local labour market (EDUC_MIGi). An inventor can migrate while already working for a company in his country of origin and migrated when moving to another company in the destination country (COMP_MIGi). Finally, an inventor can migrate within the same company, using the multinationals' branches (WITHIN_MIGi) existing between his country of origin and the country of destination.

We are identifying inter-firm mobility (CMOVit) (a move from one firm to a different one) by simply counting the number of firms for which the inventor has worked up to year t, minus one. Due to the originality of our data referring the labor market path of each inventors, we are able to measure the inter-firm mobility without using the assignee referred chronologically in the list of inventor's patents as previously used by the existing literature (Hoisl, 2007) (Hoisl, 2009) (Latham, et al., 2011). Hence considering the mobility performed by an individual before his first patent made. Finally, it is important to note that we are using inter-firm mobility only as a control variable in order to control for another form of mobility than migration. In fact, a migrant is considered as mobile per se, while a native can be either mobile or not. Additionally, we hypothesize that a migrant tends to be more mobile than a native in the destination country since has already paid an important opportunity cost in changing of country.

We are controlling for the inventor position in the company as an explanatory factor for such productivity or inactivity during an inventor's career ($TITLE_{it}$).
To consider, the increasing importance of highly skilled migration toward the United-States in the last decades, we are controlling, as suggested by Borjas (1985), for cohort effects (COH_i). Finally, to consider of the widely developed literature on positive selection of mobile highly skilled workers, we are controlling for individual skills, approximated by the level of education ($EDLEV_i$).

Estimation

The dependent variables for our productivity equations (Table 4) are the inventor's per year number of patents and the yearly number of forward citations received by these patents. Due to the excess of 0, the variability and range of each variable, we first apply a logarithmic transformation of the dependent variables and estimate the coefficients with ordinary least squares. As a first robustness check, we then estimate the coefficients using the Maximum Likelihood Estimator (MLE), that fully maximizes the likelihood of the random-effects model, instead of the classical Generalized Method of Moments (GMM). As a second robustness check, we then estimate the coefficients using the non-transformed dependent variables with a negative binomial regression model, finding overdispersion (see LR test Table 4, Panel A and B, column 5) for both the number of patents and the number of forward citations. Due to the invariant characteristic of our main variable of interest, migrant or native (MIG), we are using random effects models and we address heteroscedasticity issues by using robust standard errors.

Discussion of the Results

This section presents and discusses the empirical analysis on how the migrant-native difference in productivity changes when controlling for several inventor's characteristics such as previous mobility across companies, level of education, cohort, position in the current company and year fixed effect dummies ($Year_t$). Additional tables are available on request, here we present the main results and robustness checks.
In Table 4, we present the estimation results of the effect of the migrant dummy on the log number of patents, and log number of forward citations.
The addition of the position in the company have a negligible impact on the migrant advantage on productivity, which is now 8% for patents production and 28.1% for the number of citations. Moreover, the inventor's position in the company ($TITLE_{it}$) contributes to decrease the productivity gap between migrants and natives. Thus, we control for cohort effects to consider variations in the individuals characteristics that change over time with the length of their work experience. Controlling for cohort drastically reduce the migrant-native productivity gap from 8% to 5.4% concerning the patenting productivity and from 28.1% to 19.2% for the citation one, supporting the findings of Borjas (1985) in a more moderate way. Then, we add the inter-firm mobility covariate ($CMOV_{it}$) and its quadratic form in order to investigate the effect of mobility on productivity for inventors moving more than once. Differently from the previous literature (Hoisl, 2007) (Hoisl, 2009) (Hoisl & de Rassenfosse, 2014) we find a negative impact of mobility on the productivity proxies: -1.5% on patents and -1.6% on citations with a quadratic effect, suggesting that inventors moving more than once recover the loss in productivity associated to the first move.

Table 3: Migrants vs Natives productivity

	GLS (1)	MLE (2)	NEG. BIN. (3)	GLS (4)	MLE (5)	NEG. BIN. (6)
MIG	0.045*** (0.008)	0.033*** (0.006)	1.064*** (0.013)	0.163*** (0.021)	0.134*** (0.006)	1.11*** (0.011)
CMOV	-0.017*** (0.003)	-0.022*** (0.002)	0.951*** (0.005)	-0.020*** (0.008)	-0.035*** (0.002)	0.957*** (0.004)
CMOV*CMOV	0.001*** (0.001)	0.003*** (0.001)	1.007*** (0.0001)	0.004*** (0.001)	0.011*** (0.001)	1.006*** (0.001)
EDLEV	0.034*** (0.003)	0.022*** (0.002)	1.041*** (0.005)	0.083*** (0.008)	0.053*** (0.002)	1.062*** (0.004)
Observations	253,239	253,239	253,239	253,239	253,239	253,239
Nbr Inventors	40,806	40,806	40,806	40,806	40,806	40,806
R–squared	0.032			0.045		
Year, TITLE, COH	Yes	Yes	Yes	Yes	Yes	Yes
LR test sigma_u=0		3014***			2123***	
Wald test			4172***			6924***
LR test			3862***			1052***

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

This table estimates the effect of being a migrant on inventors' productivity with individual characteristics. Standard errors appear in parenthesis and are clustered at the inventor level. Column (1),(2) and (3) show the effect on the number of patents produced yearly per inventor, measured as Log(1 + $NPAT_{it}$) for columns (1) and (2) and as $NPAT_{it}$ for (3). Column (4),(5) and (6) show the effect on the total number of citations received yearly per inventor, measured as Log(1 + $NCIT_{it}$) for columns (4) and (5) and as $NCIT_{it}$ for (6). All columns are estimated with a random effect model, with a maximum likelihood estimator (MLE) for column (2) and (5).

This difference from the existing literature can be explained by the different measures used for mobility. In fact, unlike the previous studies and thanks to our newly developed data, we can measure inter-firm mobility even before the first patent, thus contributing to overcome one of the main limitations of the previous studies on mobility of inventors. Furthermore, to the best of our knowledge, this is the first contribution analyzing, separately, both the effect of migration and inter-firm mobility on inventor's productivity. Finally, we add the education covariate ($EDLEV_i$) to account for the difference in skills between natives and migrants. Our findings are in line with the existing literature on positive selection of migrants (Hunt, 2004), since, controlling for education, reduces the migrant-native gap, keeping a migrant advantage of 4.5% for the patent production and 16.2% for the number of citations received.

Next, we decompose the variable of migration (MIG_i) by channel of entrance in the destination country. Doing so, we are giving first evidence on how different kind of migration channels impact on the migrants' productivity. Table 5 presents the coefficients of the three migration dummies that distinguish the migration channels.

Table 4: Migrant channels vs Natives productivity

	GLS (1)	MLE (2)	NEG. BIN. (3)	GLS (4)	MLE (5)	NEG. BIN. (6)
EDUC_MIG	0.048*** (0.011)	0.035*** (0.007)	1.060*** (0.016)	0.168*** (0.027)	0.135*** (0.021)	1.107*** (0.014)
COMP_MIG	0.037** (0.015)	0.033** (0.011)	1.064*** (0.024)	0.137*** (0.039)	0.119*** (0.021)	1.095*** (0.020)
WITHIN_MIG	0.032 (0.021)	0.025 (0.015)	1.082** (0.036)	0.190*** (0.057)	0.159*** (0.046)	1.177*** (0.033)
CMOV	-0.017*** (0.003)	-0.022*** (0.002)	0.951*** (0.005)	-0.019** (0.009)	-0.035*** (0.006)	0.957*** (0.004)
CMOV*CMOV	0.001*** (0.001)	0.003*** (0.001)	1.007*** (0.001)	0.004*** (0.001)	0.009*** (0.001)	1.006*** (0.001)
EDLEV	0.033*** (0.003)	0.022*** (0.001)	1.041*** (0.005)	0.083*** (0.008)	0.053*** (0.007)	1.062*** (0.004)
Observations	253,239	253,239	253,239	253,239	253,239	253,239
Nbr Inventors	40,806	40,806	40,806	40,806	40,806	40,806
R-squared	0.032			0.045		
Year, TITLE, COH	Yes	Yes	Yes	Yes	Yes	Yes
LR test sigma_u=0		3014***			2163***	
Wald test			3845***			1032***
LR test			4172***			6927***

*** p<0.01, ** p<0.05, * p<0.1

This table estimates the effect of being a migrant decomposed by channel of entrance in the United-States on inventors' productivity with different combinations of characteristics. Standard errors appear in parenthesis and are clustered at the inventor level. Column (1),(2) and (3) show the effect on the number of patents produced yearly per inventor, measured as Log(1 + $NPAT_{it}$) for columns (1) and (2) and as $NPAT_{it}$ for (3). Column (4),(5) and (6) show the effect on the total number of citations received yearly per inventor, measured as Log(1 + $NCIT_{it}$) for columns (4) and (5) and as $NCIT_{it}$ for (6). All columns are estimated with a random effect model, with a maximum likelihood estimator (MLE) for column (2) and (5).

We observe that migrants entering in U.S. for education motives[1] are performing better in terms of patent production with respect of the natives, while migrants with a host country education have an increase in productivity of 9.7%, migrants entering in U.S. when changing of company have an increase of 6.3% and when entering in U.S. within the same company of 6.4%. We find similar results when measuring productivity by the number of citations received columns 4,5 and 6. Migrants entering in U.S. for education reasons have an increase in their productivity of 32.6%, while 22.7% for migrants entering by changing of company, and, 26.6% for the ones migrating within the same company. When we include the inventor's position in the company as control variable, we observe a decrease in the impact of migration on the inventor's productivity for both the number of patents and citations. Adding the cohort effect, we observe a loss in significance for both the inter and within company migration, with still migration for education reasons having the highest impact on the inventor productivity (6.3%). Column 4,5 and 6, the coefficients of migration for education motives

[1] Hence getting one degree and entering in the host labor market

and within the same company increase and their impact on the inventor's productivity quality is respectively of 21.2% and 21.4%. Adding the inter-firm mobility covariate ($CMOV_{it}$) that strengthens the positive impact of all three migration dummies for both the Panels, with a loss in significance for the impact of migration within the same company on the number of patents filed. Finally, when adding the education covariate ($EDLEV_i$), we find an intriguing result: we are observing that migrants entering in the destination country within the same company seem to outperform the others form of migration. This result seems to be robust and even stronger when using a Negative Binomial model (column 3 and 6). We explain this result with the possibility of a sort positive selection mechanism performed on migrants within the same multinational firm. In fact, the selection is performed on a case to case basis, and consequently firms can observe the characteristics of an individual during his working experience in his home country. Then, after selection at home, the best employees will be more likely to be re-located at destination in U.S., where the research in ICTs is at the forefront, due to the massive expenditures in R&D and the presence of the best collaborators-colleagues and the best equipment at their disposal.

Conclusion

Our analysis showed that migrants are more productive than natives both in quantity and in quality, even when occupying similar positions within their companies. This productivity gap between natives and migrants is partly due to migrants' better education. Unlike the previous empirical literature findings, we observe that inter-firm mobility has a negative impact on their productivity. This different result can be explained by the unique features of our dataset and have also a theoretical grounding. In fact, while the previous literature on mobility and productivity of inventors were considering the individuals' mobility only between the first and last patent applications, we can extend this window of observation by observing inventors' inter-firm mobility from the beginning of their career. Furthermore, this result is coherent with other theoretical findings describing that companies tend to retain the best employees, hence the inventors leaving their company may not be as skilled as the stayers. In fact, an ongoing selection is occurring since an individual is hired: the company can recognize the prior unobservable characteristics of their new employees and make decisions of his importance for the company, in promoting him and/or reallocating him. We find a first evidence of this process on the migrant's population by decomposing them with respect to their channel of entrance. Indeed, we have shown that migrants are, on average, more productive than natives, while there is heterogeneity in productivity, depending on the migrant's channel of entrance. We observe that migrants entering in U.S. through the own multinational's branches are performing better in quantity and quality than migrants entering in U.S. for education reasons or when changing of company. This finding is another proof on the ongoing selection mechanism occurring within the company, where in that case, the ones that will be reallocated in another country will be the ones positively selected.

Acknowledgments

We would like to acknowledge Valerio Sterzi, Valentina Di Iasio and three anonymous referees for their comments that contribute to enhance the quality of this paper.

References

Borjas, G. J., 1985. Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of labor Economics,* pp. 3(4), 463-489.

Borjas, G. J., Bronars, S. G. & Trejo, S. J., 1992. Self-selection and internal migration in the United States. *Journal of urban Economics,* pp. 32(2), 159-185.

Breschi, S. & Lissoni, F., 2005. Mobility and social networks: localised knowledge spillovers revisited.. *Annales d'economie et de statistique..*

Breschi, S., Lissoni, F. & Tarasconi, G., 2014. Inventor Data for Research on Migration and innovation: A Survey and a Pilot. *WIPO.*

Docquier, F. & Rapoport, H., 2012. Globalization, brain drain, and development.. *Journal of Economic Literature,* pp. 50(3), 681-730..

Gambardella, A., H. D. & Verspagen, B., 2006. The value of patents. *In EPIP Conference, Munich.*

Hall, B., Jaffe, A. & Trajtenberg, M., 2001. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools.. *NBER Working paper ,* Volume No. 8498.

Hoisl, K., 2007. Tracing mobile inventors—the causality between inventor mobility and inventor productivity. *Study of Inventors,* pp. 65-119.

Hoisl, K., 2009. Does mobility increase the productivity of inventors?. *The Journal of Technology Transfer,* pp. 34(2), 212-225.

Hoisl, K. & de Rassenfosse, G., 2014. Knowledge fit and productivity gains from mobility. *Druid Society Conference.*

Hunt, J., 2004. Are migrants more skilled than non-migrants? Repeat, return, and same-employer migrants. *Canadian Journal of Economics/Revue canadienne d'économique,* pp. 37(4), 830-849.

Hunt, J. & Gauthier-Loiselle, M., 2010. How much does immigration boost innovation?. *American Economic Journal: Macroeconomics,* pp. 2(2), 31-56.

Kerr, S. P., Kerr, W., Özden, Ç. & Parsons, C., 2017. High-skilled migration and agglomeration.. *Annual Review of Economics,* pp. 9, 201-234..

Kerr, W. R., 2007. The Ethnic Composition of US Inventors. *Harvard Business School Working Paper,* pp. No. 08-006..

Kerr, W. R., 2008. Ethnic scientific communities and international technology diffusion. *The Review of Economics and Statistics,* pp. 90(3), 518-537.

Kerr, W. R., 2018. *The Gift of Global Talent: How Migration Shapes Business, Economy & Society..* s.l.:Stanford University Press..

Kerr, W. R. & Lincoln, W. F., 2010. The supply side of innovation: H-1B visa reforms and US ethnic invention.. *Journal of Labor Economics,* pp. 28(3), 473-508.

Latham, W., Le Bas, C. & Volodin, D., 2011. Value of invention, prolific inventor productivity and mobility: evidence from five countries, 1975-2002.

Miguelez, E. & Fink, C., 2013. Measuring the international mobility of inventors: A new database. *World Intellectual Property Organization-Economics and Statistics Division.*

Peri, G., 2007. Higher Education, Innovation and Growth. *In Giorgio Brunello, Pietro Garibaldi and Etienne Wasmer eds. Education and Training in Europe, Oxford: Oxford University Press..*

Ruggles, S. et al., 2010. Integrated Public Use Microdata Series (IPUMS): Version 5.0. *Minneapolis: University of Minnesota.*

Stephan, P. E. & Levin, S. G., 2001. Exceptional contributions to US science by the foreign-born and foreign-educated. *Population research and Policy review,* pp. 20(1-2), 59-79.

Trajtenberg, M., 2005. The mobility of inventors and the productivity of research. *Conference presentation, NBER Summer Institute.*

Tyshchenko, K., 1999. Metatheory of Linguistics.

Wadhwa, V., Rissing, B., Saxenian, A. & Gereffi, G., 2007. Education, entrepreneurship and immigration: America's new immigrant entrepreneurs, Part II.

Zwick, T., Frosch, K., Hoisl, K. & Harhoff, D., 2017. The power of individual-level drivers of inventive performance. *Research policy,* pp. 46(1), 121-137.

Altmetrics - on the way to the "economy of attention"? Feasibility study Altmetrics for the German Ministry of Science and Research (BMBF)

Dirk Tunger[1]

[1] *d.tunger@fz-juelich.de*
Forschungszentrum Jülich GmbH, Project Management Jülich, Center of Excellence Analyses, Studies, Strategy, 52425 Jülich (Germany)

Abstract

Altmetrics is still under development and testing. They are still far from making a regular contribution to quantitative science evaluations in the near future. But: Altmetrics represent communication, which is very important in science and which increasingly goes beyond scientific journals. This paper contains the results of a feasibility study on Altmetrics on behalf of the German Ministry of Science and Research (BMBF), highlights application maturity and expressiveness and gives an overview of possible application scenarios.

Introduction

With regard to the communication of research within the scientific community and beyond into society, the Altmetrics approach is controversially discussed. The introduction of so-called alternative metrics (Altmetrics) is at the centre of the current debate as to whether the focus on classical bibliometric indicators in the Internet age still reflects the true impact of research work. In the course of this discussion, the term "Altmetrics" was introduced as a collective term for alternative indicators that takes into account the perception of web-based communication outside the traditional peer review process. It becomes visible who quotes, discusses or forwards scientific publications in national press, social media, policy documents and other web-based sources and who deals with publications within and outside the scientific system.

Altmetrics Research

Since the introduction of the term Altmetrics by Priem et al., the Altmetrics community can look back on about eight years of research on this topic. On the one hand, "the visibility and presence of Altmetrics is quite impressive" (Haustein, 2016a), because it is used by many scientific publishers as a marketing tool, several hundred publications on the subject have already been published, an own journal has been introduced and in the meantime even an Altmetrics conference is being held. On the other hand, there is a lack of a uniform definition and consensus on what Altmetrics is used for and what conclusions can be drawn from it (Haustein, 2016b; Franzen, 2017; Butler et al., 2017).

The Altmetrics Attention Score is currently used by many scientific publishers and institutions as a marketing tool in the form of the so-called "Altmetric Donut. The donut has been implemented on the websites of the journals Nature and Science, as well as in the repositories of the Universities of Cambridge and Zurich. The composition of the Attention Score is based on an algorithm that adds up the attention of scientific output in the various sources differently weighted.

This trend is viewed critically in science (Franzen, 2017). A simple summation of counts to a single metric (composite indicator) is "problematic" (Meier & Tunger, 2017a; Meier & Tunger, 2017b; European Commission, 2017a). In an overall view, the attention score does not reflect the impact of scientific achievement, but is suitable for filtering out publications that generate a high degree of perception in the media (Warren et al., 2016; European Commission, 2017b).

Tension between altmetrics and bibliometrics

Due to the fact that the base communities are the same, there is a certain tension between altmetrics and bibliometrics. Both (sub-)disciplines are intended to fulfill the same purpose, to generate a picture of scientific impact, but based on different influencing factors. Almost like a reflex, the two fields are often set in relation to each other, compared, or set up as an either/or selection.

In contrast, within the community itself, there is a general consensus that both disciplines complement each other instead of one excluding the other (Wouters, et al., 2015). Altmetrics are not intended to replace the peer review process or bibliometrics; rather, they should be viewed as a second opinion (Butler et al., 2017) and a “new perspective on communication by and about science in social media” (Tunger et al., 2017). A report by the expert group on altmetrics on behalf of the European commission also argues for classical bibliometrics that they “offer complementary approaches to evaluation” together with alternative metrics (Wilsdon et al., 2017). The expert group furthermore sees potentials for including a wider audience beyond the closed science system and for collecting information considerably faster than with conventional metrics. Furthermore, the idea of this approach is not limited to conventional scientific publication formats but offers the perspective of making data sources such as software and data sets accessible (e.g., as part of research data management).

The big difference between bibliometrics and altmetrics is the aspect that scientific publications are the traditional and indispensable main output of science. Thus, bibliometrics measures something that is at the center of the scientific reward system. The communication of science to society—that is, what is measured by altmetrics—is not part of the scientific reward system as yet. Creating incentives and expanding this reward system at this point would likely lead to increased use of social media by science and thus also strengthen altmetrics (Tunger et al., 2018).

Attention as a currency in science

It can be assumed that a scientist publishes not only because of the progress of knowledge, but also to enhance his reputation: he does not necessarily have to publish much, but with his publications he has to achieve the highest possible perception in order to achieve the best possible reputation: For every scientist, it is an expression of recognition if his or her work is perceived, assessed as relevant and quoted by a colleague. This applies both to the classical publishing process and to publications on the web: "In the media society it is no longer enough to be rich, you also have to be prominent" (Franck, 1996).

Franck calls this development the "economy of attention" (Franck, 1996). Although this approach cannot be applied identically to science, many scientists also try to achieve a certain degree of familiarity or prominence in the specialist community in order to strengthen their own position. This can also be described by the term "visibility": Anyone who has something to say cannot avoid it. In social media, one goes beyond the pure specialist community and communication in the science system and appeals to a wider audience. The more media society and science move closer together, for example through the use of social media in science, the more the maxim described by Franck is transferred to science.

Results of Feasibility study Altmetrics

This Section represents the main part of the study and comprises the key results from independent quantitative data analyses and qualitative expert interviews. The quantitative analyses combined with a workshop with data partner Altmetric.com form the basis of the interviews. Impulses and ideas from the interviews were echoed in the subsequent talks and reflected in the fine-tuning of the data analysis.

Quantitative data analysis

The data analysis presented below not only makes it possible to evaluate the use of altmetrics in research policy based on literature and qualitative analyses but also to verify these analyses by means of concrete assessments of available data. The complete Web of Science publication years 2013–2015 were matched with the data basis of Altmetric.com. To this end, the Web of Science data basis was requested from the local database instance of the Competence Centre for Bibliometrics, which is the basis of all analyses in this section. The advantage of this data basis is not only in its local availability but also the existing unambiguity of affiliations. This permits analyses to be conducted on the level of science organizations, similar to the annual pact monitoring indicator report (Mittermaier et al., 2017).

Each analysed year (2013–2015) featured around 1.6 million publications (which have a DOI) in WoS. These publications registered in WoS represent close to 70 % of the entire publication output of these years and form the basis of our subsequent investigations. There were no restrictions in terms of document types in WoS, meaning that the entire data basis was analysed. Matching the WoS data to the data basis of Altmetric.com revealed that the percentage of WoS publications on Altmetric.com rose from 33.4 % in 2013 to 42.2 % in 2015 (see Tab. 1). This means that the proportion of publications for which altmetric data are available is drawing ever closer to the 50 % mark. A logical conclusion is that the significance of scientific publications on social media is growing and therefore also the opportunities for, interest in, and necessity of analysing these data in a meaningful way. At this point, it must be noted that questions concerning, for example, the impact of science on society have so far not been answered using bibliometric methods. This is where altmetrics come in and might lead to new opportunities.

Tab. 1: Number and proportion of DOIs in WoS and on Altmetric.com (2013–2015)

	2013	2014	2015
WoS publications with DOI	1,586,101	1,625,593	1,635,465
Publications with Altmetric.com feedback	529,392	596,484	690,535
Proportion	33.4 %	36.7 %	42.2 %

The uneven distribution of the original publications across the feedback of the analysed altmetric data set means that distortions may occur in the representation of science organizations. This is comparable to different citation rates in various bibliometric disciplines. While bibliometrics corrects this by means of normalized indicators, such a model is not yet conceivable in altmetrics since no indicators or corresponding interpretation have been determined to date.

A differentiated consideration, according to disciplines, reveals potential distortions in multimedia resonance. Engineering sciences are generally less active on social media while this proportion is very high in medical science compared to other disciplines. This is shown clearly in Fig. 1: The distribution of DOIs for the year 2013 (proportion of WoS DOIs) is shown in red, based on the allocation of publications to WoS subject categories and subsequent aggregation using a classification according to the main disciplines. The respective proportions of altmetric resonance (proportion of Altmetric.com DOIs) are depicted in blue. The disciplines are allocated according to the underlying scientific publication and the allocation is absolutely comparable to the proportions of DOIs in WoS.10 The statistical population is formed from all WoS publications from the year 2013 that have a DOI as well as the resulting proportion of feedback with corresponding data in the data basis of Altmetric.com. Multiple classifications can lead to values over 100 % when added up.

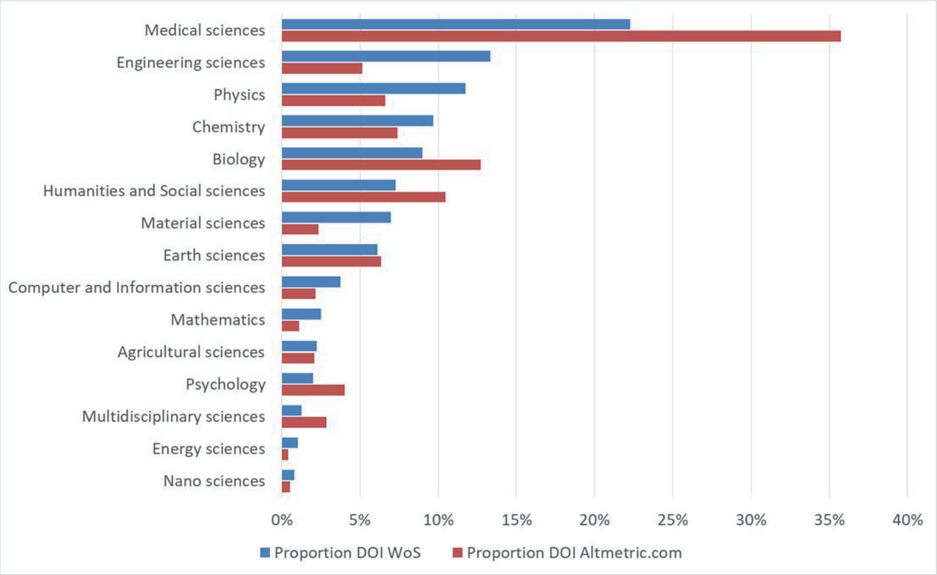


Fig. 1: Comparison of the proportions of the disciplines in WoS (upper line) and on Altmetric.com (lower line, based on DOIs, 2013); sorted by the proportion of DOIs in WoS

In addition to medicine, the humanities benefit greatly from altmetrics. While this discipline has a relatively low proportion in WoS, the proportion of publications mentioned on Altmetric.com is higher. This result also reflects the perception of Hammarfelt (2014). The observation that DOI coverage varies between disciplines was confirmed by the analyses of Altmetric.com and the University of Cambridge. Some disciplines (e.g. engineering sciences) are rarely discussed on the social media platforms covered. This reveals parallels to the discipline-specific distribution of output and citations in WoS, which are described in more detail by Haustein and Tunger (2013). Mechanisms of the news values theory also underlie this observation.

Qualitative statements based on interviews

The following results are based on five guideline-supported interviews and a two-day workshop with data partner Altmetric.com. In selecting the interviewees, particular attention was paid to covering heterogeneous perspectives of scientific discourse as well as the user side. This selection represents the subject area from different points of view.

Interviewee perspectives of altmetrics

The individual perspectives compiled by means of an exploratory approach divert from each other with regard to the interviewees' estimates of the validity and applicability of altmetrics. Within the scope of the interviews, however, sufficient overlap was achieved to gain a comprehensive overall picture from the various points of view. For illustration and summary purposes, the interviewees are arranged according to their estimate of the significance and application maturity of altmetrics in the figure below.

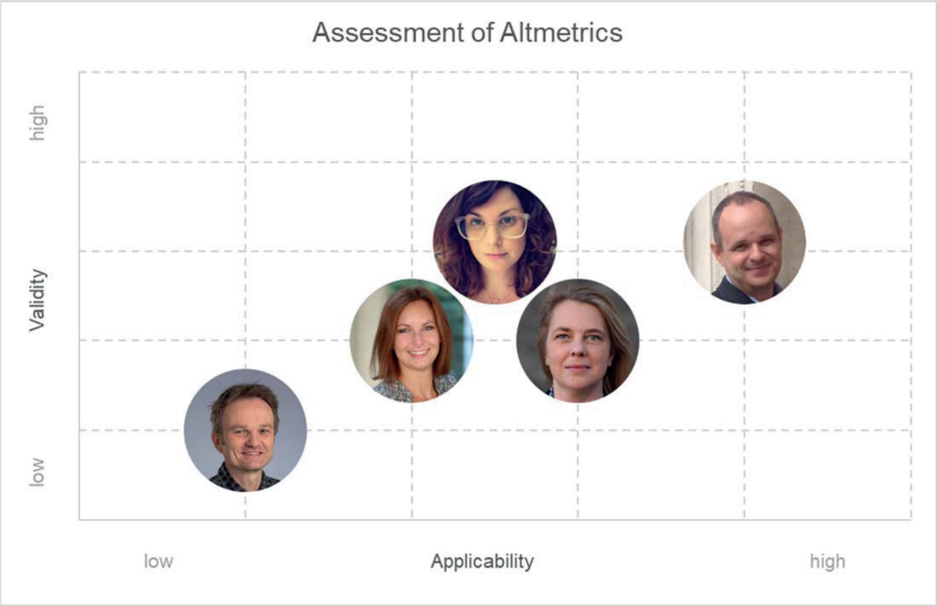


Fig. 2: Significance and application maturity of altmetrics (arranged by the authors)

From left to right: Lutz Bornmann, Isabella Peters, Stefanie Haustein, Martina Franzen, and Jürgen Wastl. Brief profiles of the interviewees can be found in Tunger et al., 2017

Significance of altmetrics

In summary, it should be noted that the opinions on significance differ less than the opinions on application maturity. The significance is estimated to be in a low to medium range. Isabella Peters explicitly emphasized that "high expectations have been consolidated [with regard to the developmental state]." The initial euphoria in the field, focusing on the far-reaching potential – including measuring the social impact and performance evaluation of science – seems to have abated. A multitude of scientific investigations have contributed to this trend, introducing a wide range of problematic issues concerning the significance of altmetrics.

There was seeming consensus that altmetrics should not be seen as an alternative to bibliometrics; instead, they represent a new perspective on the communication of and about science in social media. Perception and "popularity" are emphasized in this context. In contrast, the scientific quality or excellence is reflected poorly, as just one factor amongst many, which only partly has a positive correlation with perception. This contradicts the principle of bibliometrics, which is based on an inherent and peer-review-based approach to evaluating science.

Comparisons between bibliometrics and altmetrics can thus be considered inappropriate. Several interviewees mentioned the need for other [science-reflective] disciplines, such as science sociology or philosophy (cf. Franzen), and in-depth analyses of the motivations underlying social media activities. This view matches the perception expressed by Altmetric.com, which explicitly emphasized that the data basis reflects only the perception, and therefore represents an initial starting point for more thorough analyses. How significant the data are, however, can only be determined in a subsequent step.

On the one hand, several interviewees stressed that the "instruments used in bibliometrics (normalization etc.)" can be transferred in a targeted manner (Bornmann). On the other hand, the bibliometric focus in the analyses was criticized because altmetrics are more of a "window into another world beyond the citation system and the science community" (Peters) and should be used as such. Although the peer-review process remains central to science, altmetrics only cover "what is not visible for bibliometrics" (Haustein). Against the backdrop of current research projects, whose main objectives are comparisons of bibliometric analyses with altmetrics, for example using Mendeley counts, it should also be questioned what added value could thus be created (Franzen, Haustein). On the basis of the "ample data" (Bornmann), the objective is to specifically achieve communication beyond that within the science system.

Application areas in research policy and science management

The association with research policy and science management also represents the primary pillar in the interviews. Furthermore, guiding principles are addressed with regard to the extent to which, and the manner in which, politics can and should support developments. A key to gaining relevant insights in the long run is primarily based on the extent of the experience that can be exploited by this application.

Application maturity of altmetrics

In contrast to the significance of altmetrics, the expert opinions differ more widely between each other with regard to their application maturity. To some extent, this can be attributed to the more widely differing expectations: should altmetric characteristics be a purely quantitative indicator, or do they represent a starting point for qualitative analyses? Furthermore, the fields of application are very wide-ranging and also include marketing activities which currently have less significance for research policy.

Against the backdrop of these heterogeneous perspectives on the topic, there is, however, a consensus concerning one key issue: altmetric characteristics cannot currently be interpreted as stand-alone and quantitative indicators. In particular, the interviewees agreed unanimously that altmetrics do not represent a scientific data basis, which is a prerequisite for evaluating science. Lutz Bornmann also hypothesized that it is the responsibility of science to advise against such applications. With regard to control effects, Isabella Peters also stresses that "no one [...] [should] receive funding because his post was (re-)Tweeted 5,000 times".

Performance cannot be assessed using such conclusions (Haustein), although altmetrics can contribute an initial indicator to qualitative evaluations (cf. Wastl). In their current form, all experts interviewed advised against using altmetrics in research evaluation.

In terms of drawing conclusions from this hypothesis, however, opinions differ greatly over what role politics should play and in what way altmetrics can be used for research policy: in four of the five interviews, politics was accorded an active – if varying – role in shaping this process. Jürgen Wastl attributed the most active role to politics: He says the essential objective is that politics “fix demands and articulate research issues”, i.e. to create an overarching and binding framework for application. Subsequently, Wastl sees implementation as the responsibility of the science organizations, which would have a corresponding mandate through political requirements. Due to the exploratory developmental state, however, he views politics as being responsible for showing an openness and sensitivity in terms of reacting to the insights that can be gained through altmetrics.

From a sociological point of view, Martina Franzen stressed that this would be an experimental system and that learning through trial and error would be important. She thinks that actively dealing with the topic would lead to a gradual opening of the “black box”. Similarly, all interviewees agreed that scientific reflection, theory development, and in-depth analyses are an integral and indispensable part of the process of generating insights. This particularly includes openness to results which may indicate that altmetrics are not, in fact, usable for research evaluation in the long term.

When examining the application options, a major aspect was to actively shape the process, for example by establishing data concerning relevant issues. These data are “established according to users’ priorities” and represent “an important push factor” (Franzen). This was also confirmed during the workshop with Altmetric.com: customer requests and availability are a key orientation for developments, but also particularly for the resource-intensive expansion of sources such as policy documents and news items. Isabella Peters also stressed resulting requirements from a systemic point of view: “Politics and funding play a major role because science tends to maintain long-established traditions” and no system change is possible without such stimuli.

At the other end of the spectrum, Bornmann promoted a comparatively technocratic approach. He said that politics should refrain from application as long as the scientific knowledge gained is yet to reach a sufficiently advanced stage. Science has the responsibility to first investigate whether altmetrics can be used as a quantitative indicator in research evaluation, and if so, to what extent.

Conclusions

To what extent altmetrics will establish themselves in research policy depends fundamentally on empirical values from practical application in the sense of a learning experimental system. Therefore, potential fields of application are briefly outlined in the following paragraphs.

Science evaluation, performance assessment, and measurement of social impact

Due to the explorative development stage of altmetrics (as described above), they must be used carefully with regard to their application in the performance assessment of institutions and single scientists, for example within the scope of scientific evaluation. In particular, there is a lack of studies investigating how valid and reliable the evaluation of science based on altmetrics is. In the scientific discourse, a deeper understanding of the heterogeneity and the significance of the data must be achieved. In addition, useful indicators must be developed and benchmarking studies have to be conducted. According to current opinion, altmetrics will in the near future be more of a complementary component rather than an independent indicator for the assessment of scientific performance.

In addition, some research topics are more in the focus of society than others without necessarily displaying a larger social impact. In this context, attention should be drawn to the news values theory: it describes factors why some topics are reasonably sure to be reported and some are unlikely to become objects of journalistic reports in mass media [Warren et al., 2017]. Against this backdrop, altmetrics can be viewed as an incomplete indicator for social visibility. To what extent this circumstance will change over time cannot currently be predicted and depends more on the social discourse on science and the opening of the science system than on further methodological developments.

Public relations, visibility, and advertising of activities

A part of communication on science and its visibility in the public sphere is represented by altmetrics. In any case, it should be noted that there is a rising trend in social media activity measured by the frequency of contributions and the number of people involved. Thus, it is becoming increasingly important to use social media platforms in order to proactively draw attention to research, that is, advertise it.

As an example in this context, institutional efforts such as those undertaken by universities or the European Commission, can be observed, which strategically position their own publications and activities. Against the backdrop of the explorative state of these efforts, altmetrics could serve as feedback, for example, to test various approaches aimed at new target groups in society. With regard to research policy, particularly activities with a strong social relevance and their visibility could represent an interesting field of application complementing current evaluation approaches for analyzing media feedback. Initial network analyses are already delivering promising results and their application to research policy issues could be examined. Using specific issues associated with communication propagation, attention could be focused, for example, on the identification of relevant multipliers—for example, science journalists and representatives from politics, industry, and interest groups—in the dissemination of information. Identifying such mechanisms and transmission channels in pilot studies would be promising research priorities in this respect in addition to medial feedback already addressed through established investigation designs. Publishers already use the altmetric score as feedback on articles, albeit in a strongly aggregated and simplified form. Similar efforts are also apparent at universities and research institutions, which are testing the implementation of the Altmetric Donut both with and without the score, although the added value of these efforts has yet to be clarified. As part of a pilot measure, the OECD is currently investigating to what extent the altmetric explorer and the implementation of the altmetric score are suited to determine the social range of policy documents.

Science institutions can also use altmetrics within the scope of science marketing: it is conceivable that altmetrics could be used to focus attention on those publications by an institution that is widely discussed, shared, tweeted, or used in news pieces. This would permit the interface between science and society to be better addressed.

Whether there is any benefit from altmetrics in economics or politics beyond science has not yet been verified. From our viewpoint, there would be benefits if more sources of economic or policy-relevant sources were covered by the altmetrics databases. In this case, it would be possible to regard or measure the contribution of science in economy or policy. With bibliometric instruments, such as publication or citation analyses, it is not possible to measure this contribution since the economic or political world does not publish articles in scientific outlets. With altmetrics one would be able to have a look at, for example, mentions of scientific publications in documents, which influence politics or discussions on the application of scientific research in economics or companies. Generally, it would be worthwhile to identify the impact of scientific contributions on individual groups more easily, if one could associate contributions on social media platforms to particular fields of application.

Reporting reputation

For scientists, the visibility of their publications is essential. The reputation resulting from the use by others of their scientific output in the form of ideas, statements, calculations, and findings is an essential part of the science system. Only the use of the generated output creates sustainable value for an individual scientist, be it in other scientific publications or in web-based communication, social media, or news pieces. Bibliometrics and altmetrics help scientists document the visibility of their work. Thus, the majority of the almost 700 scientists who participated in a survey on the RG platform stated that it is important to them to have a high RG score.

Altmetrics permit scientists to record, regulate, and document their own visibility to a greater extent than was previously possible. Particularly for early-career scientists, there is thus a great opportunity to increase attention and reputation independently from the traditional publication system. In the longer term, altmetrics could assume the function of documenting the mediation of science to society and of making it more transparent.

Support from libraries

Academic libraries are usually where contacts can be found within a scientific institution for issues related to publication data and bibliometric processes/indicators. Librarians' clean data, compile publication profiles, and collect data within the scope of evaluations. They are thus specialists for handling data, particularly data related to publications, user statistics, and stock management.

This is where altmetrics represent a connecting element as they illuminate the use of publications in social media. Thus it is plausible for libraries to be directly involved whenever the issue of altmetrics is addressed at an institution. This makes sense because librarians are in contact with many areas of a scientific institution and offer advice on using information products. Roemer and Borchardt (Roemer & Borchardt, 2015) identified this central role of libraries and summarize:
"[…] librarians serve as natural leaders when it comes to altmetrics […]" (Roemer & Borchardt, 2015). They argue that this is due to the resources and data knowledge of libraries as well as their central position as contact partners for various target groups (Gimpl, 2017).

Outlook

Altmetrics is still under development and testing. They are still far from making a regular contribution to quantitative science evaluations in the near future. But: Altmetrics represent communication, which is very important in science and which increasingly goes beyond scientific journals. This is where we should start and think about incentives for how new forms of communication can be used profitably for science. This is all the more true if the incentives to bring science into society through social media are increased and integrated into the scientific reward system.

References

Butler, J.S., Kaye, I. D., Sebastian, A.S., Wagner, S. C., Morrissey, P. B., Schroeder, G. D., Kepler, C. K. and Vaccaro, A. R. (2017). The Evolution of Current Research Impact Metrics: From Bibliometrics to Altmetrics? Clinical Spine Surgery, 30(5).

European Commission (2017a). Next-generation metrics: Responsible metrics and evaluation for open science. doi:10.2777/337729

European Commission (2017b). Mutual Learning Exercise: Open Science – Altmetrics and Rewards.

Franck, G.: "Aufmerksamkeit - Die neue Währung.", 1996. Available from: http://www.aesthetischepraxis.de/Seminar/Franck_Aufmerksamkeit.pdf [Accessed: February 08, 2019]

Franzen, M. (2017). Digitale Resonanz. Neue Bewertungskulturen fordern die Wissenschaft heraus. WZB Mitteilungen 155, pp. 30–33.

Gimpl K.: Evaluation von ausgewählten Altmetrics-Diensten für den Einsatz an wissenschaftlichen Bibliotheken. Available from: https://publiscologne.th-koeln.de/frontdoor/deliver/index/docId/1034/file/MAT_Gimpl_Kerstin.pdf [Accessed: February 08, 2019]

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. Scientometrics, 101(2), pp. 1419–1430

Haustein, S. and Tunger, D. (2013). Szientо- und bibliometrische Verfahren. Grundlagen der praktischen Information und Dokumentation, 6. Auflage, Chapter: C 10. De Gruyter, pp. 479–492

Haustein, S. (2016a). Vier Tage für fünf Jahre Altmetrics. Bericht über die Konferenz 2AM und den Workshop altmetrics15. b.i.t. online, 19(1): pp. 110–112.

Haustein, S. (2016b). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. Scientometrics. doi: 10.1007/s11192-016-1910-9

Meier, A. and Tunger, D. (2017a). Investigating the Transparency and Influenceability of Altmetrics Using the Example of the RG Score and the ResearchGate Platform. Submitted for publication.

Meier, A. and Tunger, D. (2017b). Survey on Opinions and Usage Patterns for the ResearchGate Platform. Submitted for publication.

Mittermaier, B., Holzke, C., Tunger, D., Meier, A., Glänzel, W., Thijs, B., and Chi, P.-S. (2017). Erfassung und Analyse bibliometrischer Indikatoren für den PFI-Monitoringbericht 2018. http://hdl.handle.net/2128/16265

Roemer RC, Borchardt R. Altmetrics and the role of librarians. Library Technology Reports. 2015;51:31-38

Tunger D, Meier A, Hartmann D. Feasibility study Altmetrics for the German Ministry of Science and Research (BMBF), 2017. Available from: http://juser.fz-juelich.de/record/851696/files/Altmetrics%20Machbarkeitsstudie%20EN.pdf [Accessed: February 08, 2019]

Tunger, D., Clermont, M., Meier, A.: Altmetrics: State of the Art and a Look into the Future; 2018; Available from: https://www.intechopen.com/books/scientometrics/altmetrics-state-of-the-art-and-a-look-into-the-future [Accessed: February 08, 2019]

Warren, H.R., Raison, N., and Dasgupta, P. (2016). The Rise of Altmetrics. Journal of the American Medical Association, 317(2), pp. 131–132.

Wilsdon JR, Bar-Ilan J, Frodeman R, Lex E, Peters I, Wouters P. Next-generation metrics: responsible metrics and evaluation for open science [Internet]. 2017. Available from: http://eprints.whiterose.ac.uk/113919 [Accessed: February 08, 2019]

Wouters P, Thelwall M, Kousha K, Waltman L, de Rijcke S, Rushforth A, Franssen T. The metric tide: Literature review (Supplementary report I to the independent review of the role of metrics in research assessment and management) [Internet]. Available from: http://www.dcscience.net/2015_metrictideS1.pdf [Accessed: February 08, 2019]

The corporate identity of Italian Universities on the Web: a webometrics approach

G. Bianchi[2], R. Bruni[1], A. Laureti Palma[2], G. Perani[2], F. Scalfati[2]

[1] *bruni@diag.uniroma1.it*
Dept. of Computer Control and Management Engineering, Sapienza University of Rome, Rome (Italy)

[2.] *perani@istat.it*
ISTAT – Italian National Institute of Statistics, Via Cesare Balbo 16, 00184, Rome (Italy)

Abstract

In parallel with the increasing marketisation and globalisation of higher education, Universities' corporate websites have become institutional virtual storefronts largely contributing to reinforcing the organisations' brand, to disseminate information on their main achievements and to communicate with both enrolled students and potential "customers" worldwide. Thus, the effectiveness of Universities' websites to deliver value in terms of information on the organisations' activities and to interact with actual and potential students - as well as partner institutions in education and research projects - is to be regarded as a key objective of all Universities. The level of accomplishment of this task, measured so far mostly on a case-study basis, can be more extensively surveyed by adopting a webometric approach combining the use of web analytics as indicators of efficiency with selected indicators of contents collected through web scraping techniques. This approach has been tested on the websites of Italian Universities with the aim of classifying them in terms of quality and impact of their institutional websites, as well as to develop a permanent monitoring framework.

Introduction

The ability of academic institutions to effectively play the multiple roles of educational agencies, research hubs and drivers of innovation processes, in close connection with business enterprises and other organisations (Göransson and Brundenius, 2010), has become a key topic of research and policy action. The 'open innovation' paradigm – increasingly diffused in developed countries – assumes that knowledge can be freely transferred across economic sectors thus making attractive for businesses to give up large internal research facilities and replacing them with a network of potential partners – universities, research centres, start-ups, SMEs, customers, etc. – providing the technical and managerial knowledge needed to feed the innovation processes (Chesbrough, 2003). Universities are a privileged source of knowledge and innovations, as well as of educational services, and research has been long focusing on the measurement of the level of interaction between universities and external actors (West *et al.*, 2014). Leading universities are becoming more open (Lepori *et al*, 2015; Dennis *et al*, 2016; Pharr, 2016; Foroudi *et al*, 2019) as a condition not just for success but for survival in a context of increasing marketisation of higher education. Absolutely crucial is to develop their own identity (Steiner *et al*, 2013) and brand (Delmestri *et al*, 2015) in order to compete for attracting the most talented people (and, as a consequence, an increasing amount of funding). These efforts need to be pursued at a global scale, by adopting standard methods of communication and knowledge sharing and, most important, by extensively use digital technologies as key enablers. The digital transformation and the global competition are thus forcing the universities to foster their ability to communicate on the Web about their activities, capabilities and achievements. These two phenomena are intertwined, as a high-impact Web communication is a powerful driver to improve reputation, developing the brand, connecting with potential partners, and attracting funders and customers (including students).

Objectives of the study

Two methods are commonly used for evaluating the quality of universities, as well as their ability to meet the institutional objectives (higher education, research and the so-called third-mission, i.e. knowledge transfer): a) *institutional evaluation exercises* focusing either on academic programs or research outputs (long, complex, detailed and expensive exercises carried out at national level with a multi-year frequency); b) *rankings* (relying on informal data collections based on a range of available sources, quite often not very detailed, with annual frequency) (Shin *et al*, 2011). University rankings are increasingly using data freely available on the Web but often with a poor ability to properly check them for data quality.

Under the assumption that a corporate website is going to become the main communication channel between universities and external actors by delivering services (Bernier *et al*, 2002), attracting new students (Arslan *et al*, 2018) or interacting with potential or actual partners (Chu, 2005; Seeber et al, 2012 with reference to university-to-university links), a webometric approach can be adopted in order to draw a "university profile" as a result of the analysis of its corporate website. The webometric approach is extensively adopted to produce Web-based statistics (Thelwall, 2009, Thelwall *et al*, 2005; Björneborn and Ingwersen, 2004) and, more specifically, to collect information on higher education institutions from their websites. Several international universities' rankings use, at least partially, data extracted from the Web but no one of them relies exclusively on information available from universities' websites.

Universities' corporate websites have thus become key sources of information being: a) the main gateways to access both general information and specific contents universities are disseminating; b) delivery points of educational services and c) contact points to develop research and co-operation networks. In this respect, to improve the quality of the universities' websites and their effectiveness as information hubs is a key corporate task (Kaur *et al*, 2016).

The exercise described below is aimed at collecting information on Italian universities through their corporate websites, similarly to existing rankings, but with two rather distinctive features:

- *to be exclusively based on information available on universities' websites, thus to be potentially updated with a frequency higher than once per year;*
- *to collect a set of indicators about the efficiency of universities' websites and their effectiveness in disseminating key contents in order to produce a "profiling" of Italian universities, rather than a ranking.*

Methodology

In order to achieve the mentioned objectives, a data collection and data processing exercise has been developed. Raw data have been collected from two main sources:

- A leading provider of Web analytics indicators (http://www.similarweb.com) has been used to draw a set of indicators about the quality/efficiency of Italian university websites.
- A web-mining task has extracted selected contents from the same websites.

Web analytics are regularly produced by highly specialised web companies that monitor the performance of commercial websites, in comparison to their direct competitors, in order to increase their attractiveness for customers and profitability. When using analytics of public or non-profit institutions' websites, the aim is usually that of assessing their effectiveness in communicating with the public or in delivering online services. In this study, still at a pilot stage, the analytics covering the last six months of activity of Italian universities' websites have been collected. The frequency of data collection and the related coverage of time (for instance, extending it to one year) will be increased in the next stage of the study. In parallel, by considering that the choice of the data provider could affect the quality of the data, several commercial data providers will be compared to select the most reliable source.

The study has also included an advanced application of the webometric approach. Shortly, webometrics uses three categories of web mining: web content mining; web structure mining; web usage mining[i]. In the first stage of this study the *web content mining* has been mostly adopted. It involves the analysis of unstructured text data in webpages in order to translate it into structured information (e.g. to find connections between academic web portals and external organisations, as in this study). The web scraping activity has been limited to the ‘second level’ of the websites’ link structure in order to reduce the volume of scraped data.

The scraping activities on universities’ websites has three main steps: a) acquisition of the universities’ web addresses; b) validation of the websites; c) data extraction. By using official sources of information on the Italian tertiary institutions, the official names of universities have been used as *search strings* by search engines in order to keep web addresses with a high matching probability assessed via a machine learning approach. In addition to the identification of universities Web portals, it has also been needed – for a few large universities - to extend the analysis to University Departments’ websites (increasing both data processing time and volume of downloaded data).

Data collection has started by downloading contents from the targeted websites (texts, hyperlinks, HTML tags, meta-keywords, pdf files, etc.) by using a web scraping procedure. This has allowed to explore the websites’ structure and collect all available information by using text mining techniques. The scraped information has been stored in a semi-structured format to allow for efficient information retrieval (Bianchi *et al*, 2018a; Bianchi *et al*, 2018b). These require the integration of natural language processing (to extract meaning from free text) with advanced machine learning (Bruni and Bianchi 2019).

Table 1. Selected indicators to be used for profiling Italian universities’ websites.

No.	*Indicators*	*Area*	*Description*	*Rationale*
1.	Relevance	Analytics	(1/national ranking by visitors) / log(number of students)	Websites’ popularity at the national level is the key indicator of effectiveness for universities mainly enrolling Italian students
2.	Usability	Analytics	Percentage of contacts from mobile devices	Level of use by the mobile-oriented audience (largely including students).
3.	Identifiability	Analytics	100- bounce rate	A higher level of visitors leaving the website after the visualisation of the main page is an indicator of a low ability of the website (or the university) to be identifiable
4.	Intensity of use	Analytics	Number of pages visited * average time spent on the website	A key indicator of website effectiveness: the more time is spent on the website, the more relevant will be available contents for users
5.	International orientation	Analytics	Percentage of foreign contacts	Popularity abroad as a condition to attract customers (incl. students) and partners
6.	Visibility	Analytics	Percentage of direct accesses	Percentage of non-casual visitors as an indicator of popularity and ability to connect to a population of regular users
7.	Use of social media	Analytics	Percentage of accesses from social media	Degree of orientation to the use of social media
8.	Access to information on teaching	Contents	Number of e-mail address / number of professors	Measures the ability of students to easily get in touch with professors
9.	Access to data and outcomes	Contents	Number of pdf documents / log(number of students)	Measures the ability of users to have access to relevant documents (including learning materials and research outcomes) produced by the university
10.	Orientation to external collaborations	Contents	Number of firms+research institutions (IT+EU) mentioned in the website	Measures the ability of the website to provide a comprehensive description of the extent of on-going research (or Third-mission) collaborations
11.	Link impact studies (URL-degree)	Contents	Number of hyperlinks pointing to each University website	Measures of the numbers of hyperlinks pointing to each website

This methodology has been used to produce a set of eleven indicators (Table 1) combining Web analytics and website contents data. Seven variables (analytics) focus on the intensity of use of universities’ websites, as well as highlighting some key features of users and their access modes (whether direct or indirect access, users from Italy or from abroad, Web traffic from social media, etc.). Four indicators have been drawn from scraped data: percentage of professors' e-mail available; number of pdf documents (weighted by university size) i.e. volume of information available to users; number of EU firms or research institutions mentioned in the website; number of hyperlinks pointing to each University website, as a link impact metrics. In particular, tenth indicator referring to develop international partnerships.

In order to identify the key features of the data collection stage of this study, it can be pointed out that it has been designed to be: a) Totally Web-based; b) Fully transparent/reproducible; c) Based on data from a leader company in Web monitoring and from systematic web-scraping to implement a web-mining approach on universities' websites. On the other hand, data processing was assumed to be: a) suitable for replication on a regular basis; b) effective in minimising the influence of university size on websites' effectiveness; c) based on state-of-the-art text mining and advanced machine learning techniques.

Moreover, at this stage, also considering that official data on the third mission of Italian universities do not exist, it is possible to confirm a positive linear dependence (Figure 1a) between key indicators of analytics (relevance) and scraped contents (collaboration, i.e. number of EU firms and research centres cited in the websites).

A critical issue of the study is the potential inconsistency between data analytics set and scraped data set. The eleventh indicator should allow for validating the joined use of this two sets of data. Figure 1b shows a remarkable concordance between the URL-degree (indicator 11) and Relevance (indicator 1) that supports the combined use of the two sets of data.

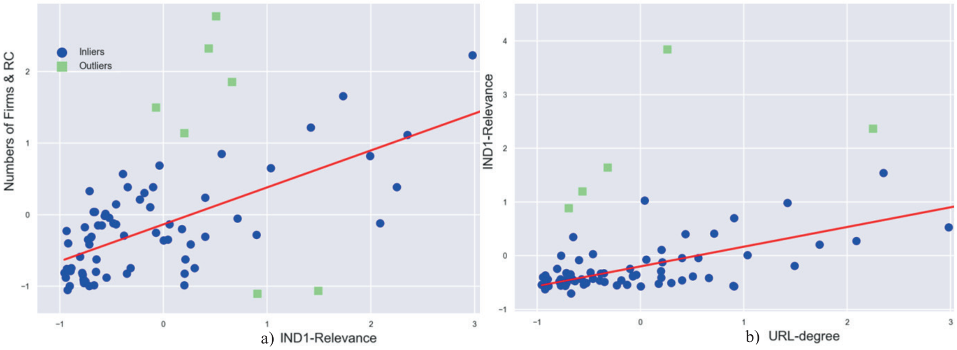


Figure 1. Linear regressions based on random sample consensus fits model. Scatter plot of: a) Relevance (Indicator-1) vs. Orientation to external collaborations (Indicators 10); b) URL-degree (Indicators 11) vs. Relevance (Indicators 1).

Results

The aim of this study was more that of profiling Italian universities according to their Web activity, rather than comparing them and their performances. The potential of the indicators in Table 1 has been tested by running a cluster analysis (FASTCLUS^{ii} in SAS) which allowed for the identification of three main profiles. The analysis focused on 79 Italian universities (two universities for foreign students and all Italian online universities have been excluded).

In Figure 2, such profiles are described with reference to two canonical variables, respectively describing (X) the websites' impact on users (mostly based on indicators 1 'Relevance' and 8 'Information on teaching') and (Y) the level of websites' quality (indicators 2, 3, 4, 6 and 7). As a result, three clusters have been identified. Cluster 1 (red) includes websites with a high number of visitors (which are those of medium-large universities although the access rates were weighted by university size) and providing extensive information about how to get in touch with the teaching staff (i.e., indirectly, to get information on teaching in general).

Cluster 2 (blue) is influenced by the same indicators but rather with a negative sign: low access rates and poor information delivered to users. As compensation, the quality level of these websites is, on average, higher than that of the other clusters. Finally, Cluster 3 (yellow), including several small and highly specialised universities, can be described as poorly performing in terms of Web quality while featuring non-irrelevant access rates. This exercise has been designed to deliver most of its potential by comparing the website performance over time, thus allowing for spotting any progress in the ability of universities to make their websites increasingly attractive and effective. The description which can be given of the current profiling results may be neither relevant per se, nor totally new compared to existing rankings based on structural and economic indicators (Shin, 2011; Aguillo, 2010).

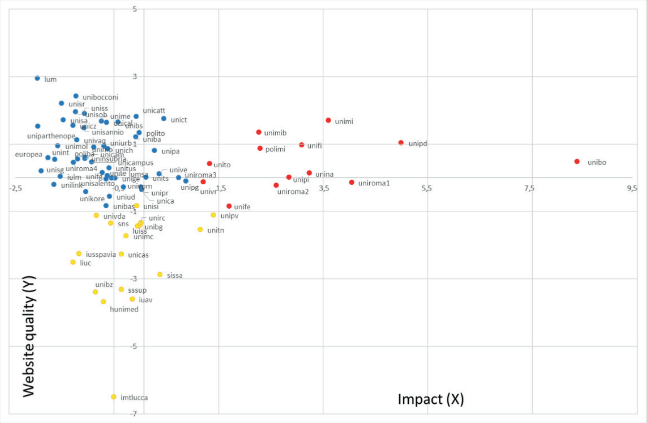


Figure 2. Websites of Italian universities grouped in clusters by quality and impact

Next steps

This project of profiling Italian universities by adopting a webometric approach aims at filling the need for a timely and neutral assessment of the ability to improve their competitiveness in a global context. The first stage is, necessarily, that of profiling them by using data available on the Web (i.e. the information available, in principle, to anyone would be interested to get in touch with them). A second stage will be that, of course, of extending the analysis to several time periods in order to assess the dynamics of an effort to develop Web strategies – including branding – over time in parallel with potential competitors/partners abroad. But again, the stage of profiling, focusing on how they communicate, is a preliminary stage for any meaningful measurement of effective performance in a digitalised environment. On the methodological side, the next stages of the study will improve the framework defined so far with respect to several actions of collecting and processing data. Areas where most of the development efforts will be focused will include: improving the quality of Web analytics; testing a web-scraping activity reading more than two layers of a Web portal structure (i.e. addressing the issue of website quality); developing more accurate machine learning routines to process scraper data.

References

Aguillo, F.I., Bar-Ilan, J., Levene, M & Ortega, J.L. (2010). Comparing university rankings. Scientometrics 85.1 p.243-256.

Arslan, Y., Evren, S. & P. Soner. (2018). The Relationship between International Students Attitude toward the University Website and University Attractiveness. *Journal of the Faculty of Education* 19.1 p. 200-211.

Bernier, J.L., Barchéin, M., Cañas, A., Gómez-valenzuela, C. & Merelo, J. (2002). The services a university website should offer. *Information Society and Education: Monitoring a Revolution. Serie Sociedad de la Educación* 9 p.1746-1750.

Bianchi, G., Laureti Palma, A. & Quaresma, S. (2018a). Prepare your data warehouse for a Big Future, by including Big Data. *European Conference on Quality in Official Statistics 2018, Krakow*.

Bianchi, G, Bruni, R & Scalfati, F. (2018b). Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms, *Mathematical Problems in Engineering*, vol. 2018.

Björneborn, L. & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), p. 1216-1227.

Bruni, R., Bianchi, G. (2019). Robustness Analysis of Classifiers for Website Categorization: the Case of E-commerce Detection. *Expert Systems With Applications*, to appear.

Chesbrough, H. (2003). Open Innovation. The New Imperative for Creating and Profiting from technology. *Harvard Business School Press, Boston*.

Chu, H. (2005). Taxonomy of inlinked Web entities: What does it imply for webometric research? *Library & Information Science Research* 27.1 p. 8-27.

Delmestri, G., Oberg, A. & Drori, G. S. (2015). The unbearable lightness of university branding: Cross-national patterns. *International Studies of Management & Organization* 45.2 p. 121-136.

Dennis, C., Papagiannidis, S., Alamanos, E. & Bourlakis, M. (2016). The role of brand attachment strength in higher education. *Journal of Business Research* 69.8 p. 3049-3057.

Pantea, F., Qionglei, Y., Suraksha, G. & Mohammad M. F. (2019). Enhancing university brand image and reputation through customer value co-creation behaviour. *Technological Forecasting and Social Change* 138 p. 218-227.

Göransson, B. & Brundenius, C. (Eds). (2010). *Universities in transition: The changing role and challenges for academic institutions*. Springer Science & Business Media.

Huang, M. (2012). Opening the black box of QS World University Rankings. *Research Evaluation* 21.1 p. 71-78.

Kaur, S., Kaur, K. & P. Kaur. (2016). An empirical performance evaluation of universities website. *International Journal of Computer Applications* 146.15 p. 10-16.

Lepori, B., Seeber, M. & Bonaccorsi A. (2015). Competition for talent. Country and organizational-level effects in the internationalization of European higher education institutions. *Research policy* 44.3 p. 789-802.

Montazer, G. (2018). University Website Quality Improvement Using Intuitionistic Fuzzy Preference Ranking Model. *Quarterly Journal of Iranian Distance Education* 1.2 p. 9-30.

Pharr, J M. (2016).University Branding 2.0 Harnessing the Power of Social Media for Open-Source Branding and Brand Co-Creation of Colleges and Universities. *Kennesaw State University*, paper.

Rauschnabel, P.A., Krey, N., Babin B.J. & Ivens B.S. (2016). Brand management in higher education: the university brand personality scale. *Journal of Business Research* 69.8 p. 3077-3086.

Seeber, M., Lepori,B., Lomi,A., Agiullo,I. & Barberio,V. (2012). Factors affecting web links between European higher education institutions. *Journal of informetrics* 6.3 p. 435-447.

Shin, J. C., Toutkoushian, R. K. & Teichler, U. (Eds). (2011). University rankings: Theoretical basis, methodology and impacts on global higher education. *Vol. 3. Springer Science & Business Media*.

Steiner, L., Sundström, A. C. & Sammalisto, K. (2013). An analytical model for university identity and reputation strategy work. *Higher Education* 65.4 p. 401-415.

Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. *Synthesis lectures on information concepts, retrieval, and services* 1.1 p. 1-116.

Thelwall, M, Vaughan, L. & Björneborn, L. (2005). Webometrics. *Annual review of information science and technology* 39.1 p. 81-135.

Vaughan, L. & Yang, R. (2013). Web traffic and organization performance measures: Relationships and data sources examined. *Journal of informetrics* 7.3 p. 699-711.

West, J. Salter, J.A., Vanhaverbeke, W. & Chesbrough H.W. (2014). Open innovation: The next decade. *Research Policy*, Volume 43, Issue 5, p. 805-811.

[i] Web mining techniques have been used to extract the Web analytics used in variables 1 to 7 (Table 1).

[ii] Additional details on the analysis carried out and regression results are available on request.

The Impact of Research Funding Agencies on the Research Performance of five European Countries – A Funding Acknowledgements Analysis

Torger Möller[1]

[1] moeller@dzhw.eu
German Centre for Higher Education Research and Science Studies (DZHW),
Schützenstraße 6a, D-10117 Berlin (Germany)

Abstract

This study analyses the funding background of publications in five European countries (Germany, Denmark, Great Britain, France, and the Netherlands) with the aim to identify the major funding agencies and their contribution to the nationwide performance. The national research systems differ in the amount and proportion of domestic and foreign funders. Denmark and Great Britain are more diversified than the Netherlands, France and Germany. In general, foreign funding agencies have a higher impact than domestic ones, even by applying different indicators for an appropriate organizational comparison. The general impact of a country determines the impact a funding agency can have in the respective country.

Introduction

With the emergence of funding acknowledgments in the Web of Science in 2008, research on the funding sources of publications was accelerated. Initially, the focus was on fundamental questions of data cleaning, coverage analyses and distribution of funding information between document types, disciplines and countries (Álvarez-Bornstein, Morillo, & Bordons, 2017; Costas & Leeuwen, 2012; Morillo & Álvarez-Bornstein, 2018; Sirtes & Riechert, 2014). In recent years, some studies have focused on questions of science policy or innovation theory. Mejia and Kajikawa examined the role of funding agencies in breakthroughs in the fields of robotics (Mejia & Kajikawa, 2018). Wang and Shapira showed by citation measurement that in the field of nanotechnology sponsored research exhibits higher impact (J. Wang & Shapira, 2015). Möller, Schmidt and Hornbostel investigated the effects of the German Excellence Initiative on the German research system by using raw grant texts and grant numbers as a data source. They concluded that although an effect can be observed, it has a relatively small influence on the entire German research system (Möller, Schmidt, & Hornbostel, 2016).

From the perspective of science systems, funding agencies have an important role for the overall development of science (Braun, 1998). They provide additional financial resources beyond the block funding of universities and public research organizations. Through the competitive selection and peer review process, funding agencies decide to whom and for what purpose additional funds are granted. Project-related research funding can be understood as a governance instrument that affects not only individual or organisational performance, but also the whole national research system.

Wang et al. investigated ten countries to demonstrate that research funding systems are differentiated to various degrees. China has been described as a country dominated by a single agency (Natural Science Foundation of China, NSFC), while Great Britain has diversified funding sources (X. Wang, Liu, Ding, & Wang, 2012). Huang & Huang emphasized that research funding in the individual countries differs with regard to the topics and perspectives of the funders (Huang & Huang, 2018).

The aim of this study is to compare different countries and their research funding agencies on the basis of their publication outputs and impacts. What are the major domestic and foreign funders in each country? Do funded publications lead to an increase in national research performance? And what are the differences between the countries and domestic and foreign organisations?

Five European countries were selected for this study: Germany (DEU), Denmark (DNK), France (FRA), Great Britain (GBR) and the Netherlands (NLD). The selection of countries was based on a broader research project on governance and performance of research. In the context of this paper, it should be emphasized that the individual countries have different funding structures. Germany with a large science-based and self-organized funding agency (German Research Foundation, DFG). France with a relatively young research funder (French National Research Agency, ANR, established 2005). Great Britain with various discipline-oriented research councils and more influential private research funders in relation to Germany and France. Compared to the three larger countries we include with Denmark and the Netherlands two smaller but powerful science systems in the study.

Methodology

The analysis is based on the Web of Science (WoS) that contains a greater coverage than other data sources as Scopus or Pubmed (Kokol & Vošner, 2018). In addition, own case studies stated that the coverage of the WoS is also greater than Crossref or Dimensions. For this study we use all journal publications (article or review) published in the WoS in 2016 (Science Citation Index Expanded, the Social Sciences Citation Index and the Arts and Humanities Citation Index).

The WoS contains three fields with funding acknowledgement (FA) data: (i) the raw grant text field, (ii) the funding organisation field and (iii) the grant number field. Our analysis is based on the "dirty" (Sirtes, 2013) funding organisation field (FO field). More than four-fifths of the FO entries are linked to only one publication, but there are several entries that refer to thousands of papers. The distribution of publications linked to the FO field is strongly skewed and there exists a large number of different spellings for each funding agency.

In a semi-automatic procedure with subsequent human quality control, the spelling variants were assigned to the respective major funding agency. The term "major" is defined in proportion to the size of the analysed country. The goal of the data cleaning was to identify all funding agencies of one country that contribute to at least 1% of the nationwide publication output.

To measure the impact we apply the percentile-based indicator of the top 10% highly cited papers (HC or also known as PP top 10%). The indicator gives the percentage of worldwide highly cited publications in a given corpus, e.g. of a funding agency or a country. If the value is greater than 10% the performance of the unit is above the worldwide average. Our adjusted calculation (for details see Waltman & Schreiber, 2013) is publication type (journal), document type (article, review) and field (subject categories) normalized. We use a full counting method.

Results

In 2016 there are 328,735 distinct publications in the five countries (DEU, DNK, FRA, GBR, NLD; without multiple counting of papers in collaboration). The publications correspond to 256,187 different entries in the funding organization field (FO field). The semi-automatic cleaning and validation process assigned 63,847 spelling variants (24.9%) to 177 distinct research funding agencies. As a rule, only one funding agency is stored in the FO field. However, we have identified 212 cases (0.3% of validated FO-field entries) in which two to four funders are mentioned. We observed two reasons for multiple entries: (i) several research agencies have been accidentally extracted into a single FO-field. (ii) A joint research programme has been identified, for example DFG/ANR or NWO/NSFC. Of the total of 328,735 publications from the five countries, 215,280 (65.5%) have a funding acknowledgement. Denmark (74.1%) has the highest share of funded publications, followed by Great Britain (67.7%), Germany (67.4%) and the Netherlands (67.4%). Only 64.8% of French papers acknowledge a funding. With the cleaning rate of 24.9% of the spellings, 65.8% of all publications (141,683) could be assigned to at least one funding organisation.

Table 1: Funding agencies with more than 5,000 publications (P) in 2016

	EU -EU	DFG -DEU	EPSRC -GBR	ANR -FRA	NSFC -CHN	NSF -USA	NIH -USA	BMBF- DEU	WT -GBR	NIHR -GBR	NOW -NLD	MRC -GBR	CNRS -FRA
P	33776	26001	9942	9448	9362	8748	7677	7623	6010	5832	5663	5385	5269
FO-fields	12924	6442	2216	4561	999	1771	2941	2730	892	3118	2729	799	1764
P per FO-Field	2.61	4.04	4.49	2.07	9.37	4.94	2.61	2.79	6.74	1.87	2.08	6.74	2.99

Table 1 gives an overview of the most important funding agencies (more than 5,000 publications in 2016). The two by far largest organisations in terms of their publication output (P) are the European Union (including the Research Framework Programme and the European Research Council, abbreviated with EU-EU) and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG-DEU). Table 1 shows how many spelling variants were found for each funding agency and the average number of publications per entry. For instance, the National Science Foundation of China has an average of 9.37 publications, whereas the National Institute for Health Research (NIHR-GBR) has only 1.87 publications. Many publications are linked to a few main spellings of respective funders. 5,693 publications are assigned to a single entry of the National Natural Science Foundation of China (NSFC) and 4,750 to one of the EU. The DFG has five main variants, each linked to more than 1,000 publications (5,691; 3,940; 3,330; 2796 and 1,568).

We have defined a uniform threshold value for the following analyses. We concentrate on the 55 major funders, which contribute to at least 1% of a country's publication output. As some organisations assume the role of major funders in more than one country, the multiple counting gives a total of 106 funding agencies in the five countries.

Table 2: Number of major funding agencies (contribute to at least 1% of the publications in each country)

	DEU	DNK	FRA	GBR	NLD
funding agencies	11	37	16	22	20
thereof domestic	4	12	4	12	5
thereof domestic in %	36.4%	32.4%	25.0%	54.5%	25.0%

Table 2 shows that the number of major funding agencies varies between countries. This is partly due to the different diversification of the funding landscape and partly to the size of the respective country. In larger countries, a funding agency must also be larger and financially more powerful if it is to exceed the 1% threshold. In smaller countries, the funding landscape is more diverse, as smaller funding agencies can more easily exceed the 1% threshold. Great Britain is an exception. Although there is an umbrella organisation for the research councils (until 2018 the Research Councils UK), the seven research funders are independent and differentiated according to their research areas. The five countries also differ in terms of the number and the proportion of domestic funders. These vary from 25% (FRA, NLD) to 54.5% (GBR).

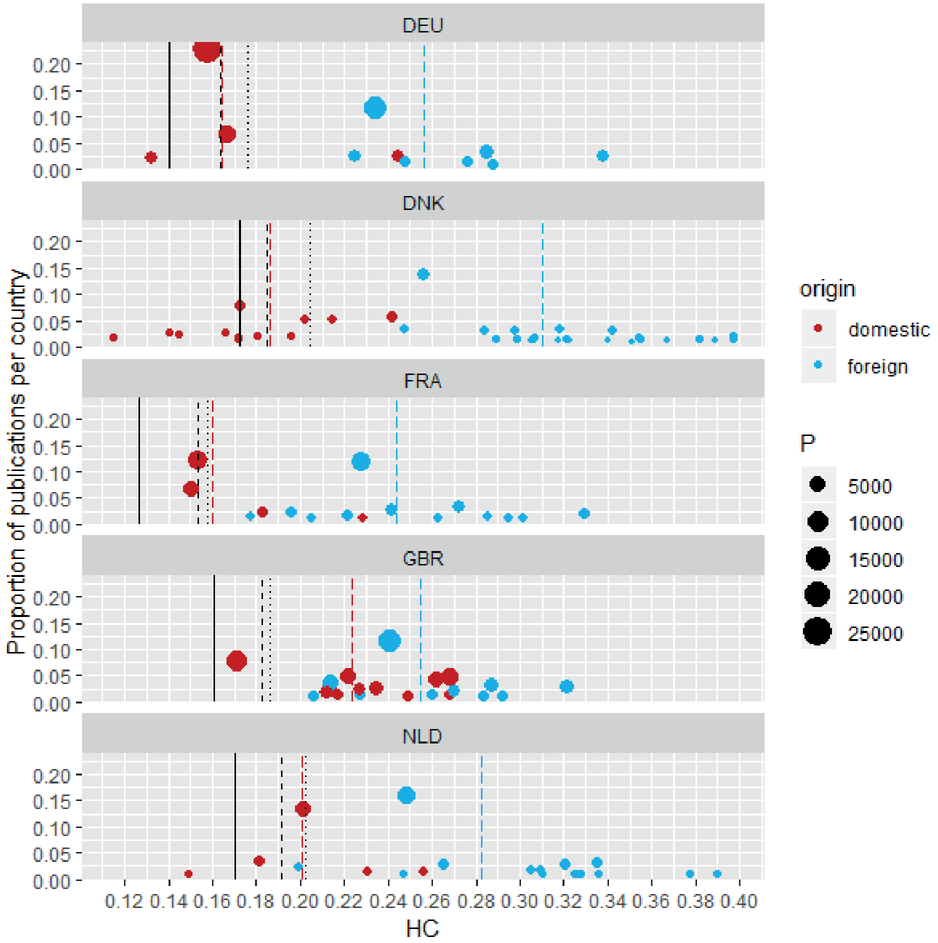


Figure 1: Major funding agencies, their proportion of publications and highly cited impact (HC) in each country

Figure 1 shows all major funding agencies contributing to at least 1% of a country's publication output. The y-axis indicates the proportion of publications in each country, the x-axis the proportion of highly cited publications (HC). The size of the points varies depending on the absolute number of publications. Domestic (dark red) and foreign (light blue) funders are highlighted. Vertical lines reflect the proportion of highly cited papers in each country: (i) black solid line for all publications, (ii) black dashed line for publications with a funding acknowledgement and (iii) black dotted line for international collaborative publications. The coloured, long dashed lines display the share of highly cited papers of all major domestic (dark red) and foreign (light blue) funders.

First of all, the five countries differ in their overall HC rates (black solid line): Denmark has the highest value (17.3%), followed by the Netherlands (17.0%), Great Britain (16.0%), Germany (14.0%) and France (12.7%). In all countries, the HC rate of publications with a funding acknowledgement (black dashed line) is above the national HC value. The impact of publications produced in international cooperation is once again higher in all countries (black dotted line).

There are few and only domestic funding agencies that are below the national HC rates. In Germany, for example, this is the German Academic Exchange Service (DAAD-DEU, 13.2%) with funding programmes that cover the entire personnel spectrum from students, graduates, postdocs to professors. Several smaller funding agencies in Denmark are also below the Danish HC rate. However, it should be noted that Denmark has a significantly higher HC value than the other four countries.

If one takes the HC rate of publications with a funding acknowledgement as a benchmark, the largest domestic funding agency by far, the DFG (HC 15.8%), is below this benchmark (black dashed line, 16.4%). This also applies to the National Centre for Scientific Research (CNRS-FRA, 15.1%) in France. With 15.3% the ANR-FRA meets exactly the HC benchmark of publications with a funding acknowledgement. In Great Britain the Engineering and Physical Sciences Research Council (EPSRC-GBR, 17.1%), in the Netherlands the Dutch Ministry of Economic Affairs and Climate Policy (EZK-NLD, 14.9%) and Netherlands Organization for Health Research and Development (ZonMw-NLD, 18.1%) are below the respective domestic benchmarks for funded publications.

A comparison of major funding agencies shows that the foreign funders (light blue long dashed line) achieve a considerably higher impact than the domestic ones (dark red long dashed line). What is the cause of this difference? Do foreign funders an overall better job? Or are these results influenced by a hidden factor?

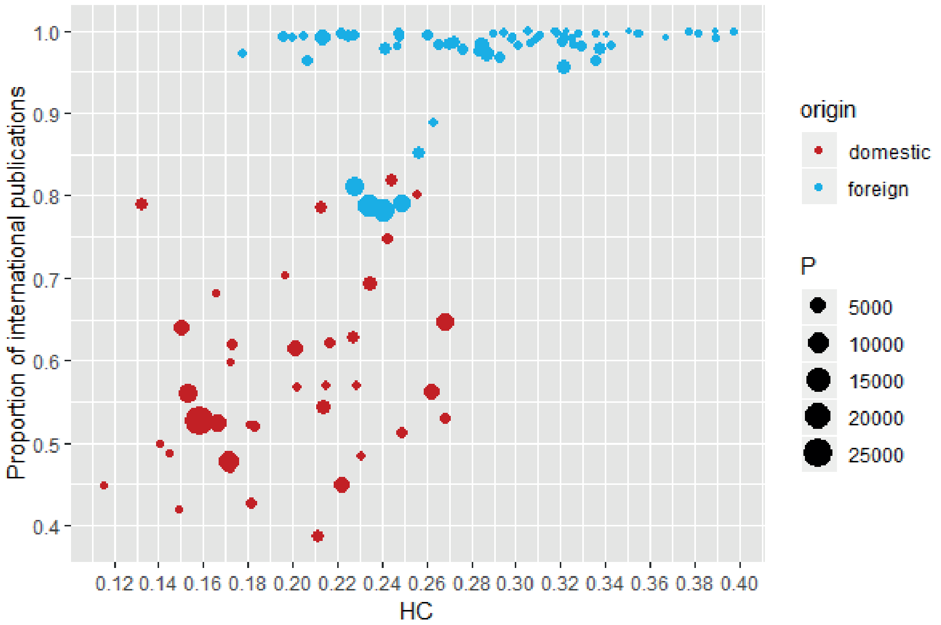


Figure 2: Major funding agencies, their proportion of international publications and highly cited impact (HC)

Figure 2 plots the same 106 funding agencies as Figure 1, but the y-axis represents the proportion of international cooperative publications. In general, foreign funders have a higher share of international papers (generally above 95%) compared to domestic organisations. The supranational EU (five dots, one per country) has lower percentages than foreign funders, but higher rates than most of the domestic ones (EU from 78.2% for GBR up to 85.2% for DNK).

The high proportion of international publications among foreign funding agencies raises the question whether funders actually promote research abroad. It is also possible that each researcher of an international publication mentions his own national grants, without any international financial transfer having occurred. Other reasons for the phenomenon could be international mobility or multiple international affiliations of academics.

In general, international collaborative publications have a positive influence on the HC rate (see the black dotted lines for each country in Figure 1). For example, the HC value for all publications of the German Research Foundation (DFG) is 15.8% and rises to 19.3% if only international publications of the DFG are considered. The comparison of domestic and foreign funding agencies in a given country is therefore distorted by the underlying factor of international cooperation. If we want to measure the influence of domestic funders on nationwide impact, the results shown in Figure 1 are useful. However, if we want to compare research funders within a country, then the comparison should only take into account publications that have been produced in international cooperation.

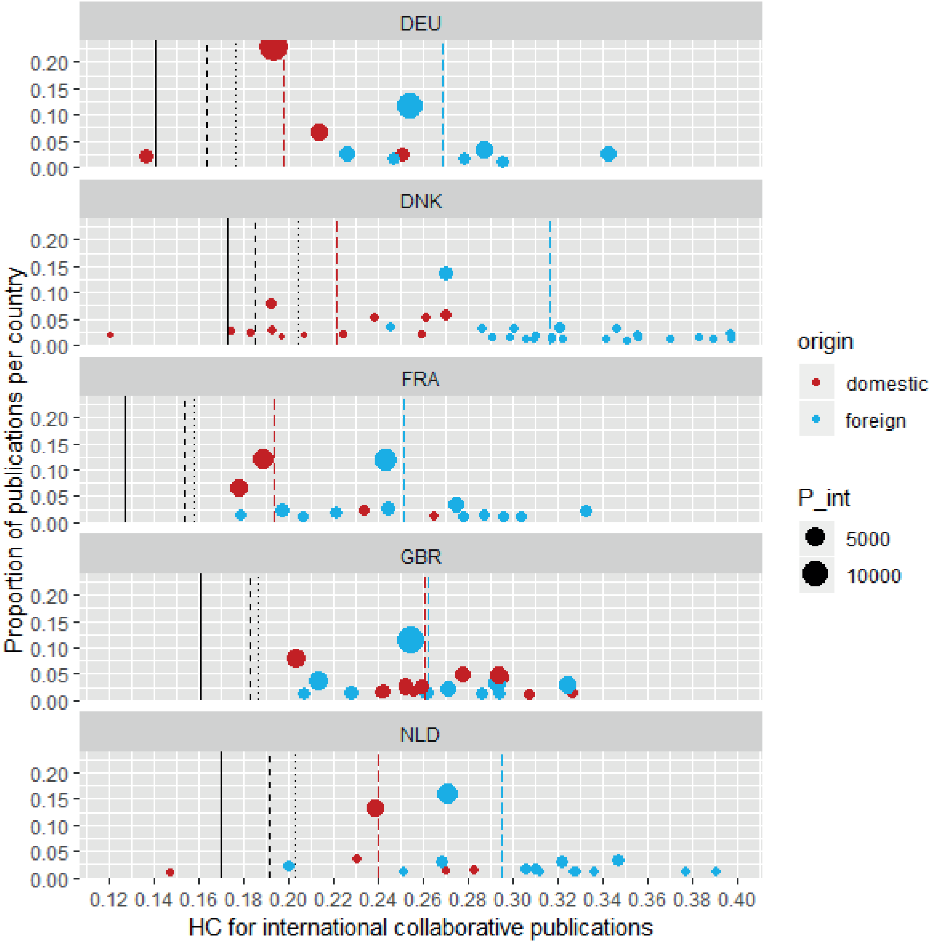


Figure 3: Major Funding agencies, their proportion of publications and highly cited impact (HC) for international collaborative publications (P_int) in each country

In Figure 3, the HC was calculated solely for the international collaborative publications of the domestic and foreign research funding agencies. In particular, domestic funders (dark red dots) have shifted from left to right due to an increased HC rate. The mean impact differences (dark red long dashed line versus light blue long dashed line) have narrowed, but the pattern has not changed considerably. Domestic funders are mainly left, while foreign funders are mainly right and the gap between them is significant. Great Britain represents an exception. On average, the highly cited rate of foreign funders (26.2%) in Great Britain is only 0.1 percentage points higher than the domestic rate. The funding agency with the highest impact is even a domestic organisation (Cancer Research UK; HC 32.6%).

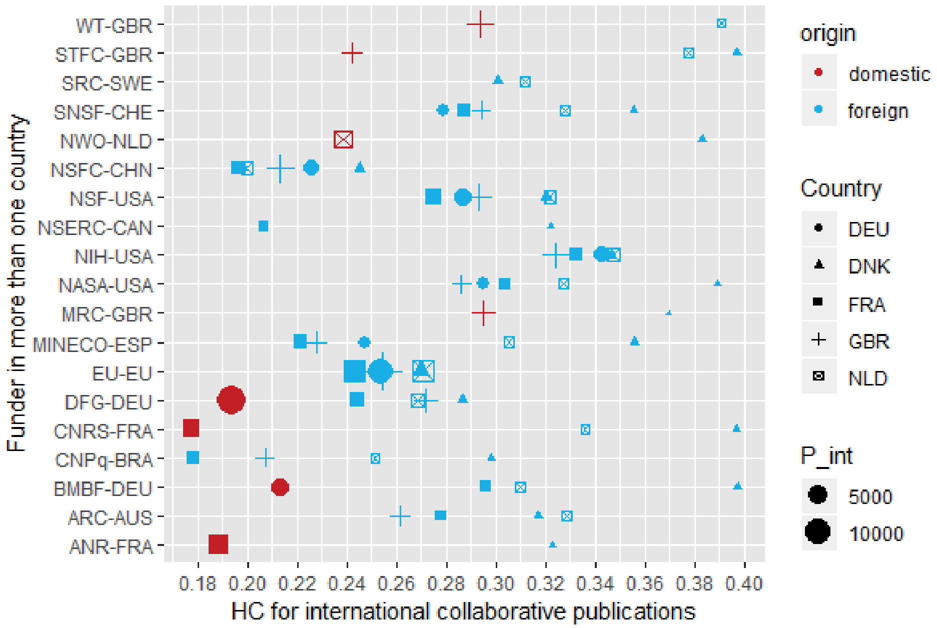


Figure 4: Major funding agencies, their contribution in different countries and highly cited impact (HC) for international collaborative publications (P_int) in each country

Figure 4 presents all major funding agencies that contribute to the publication output in more than one country. For the EU and the DFG, this is valid for all five countries. In order to ensure comparability between domestic and foreign funding agencies, the HC rate was calculated solely for the international collaborative publications. In each case presented, the HC rate is the lowest in the country where the funding agency is located. For example, DFG publications from Germany have a lower performance than publications that are related to one of the other countries, even though all publications were produced in international collaboration. Figure 4 indicates that country patterns influence the performance of sponsored publications. The general country ranking is reproduced with minor differences: Denmark leads, followed by the Netherlands, Great Britain, Germany and France. The general country patterns are remarkably stable at the organisational level.

Discussion and Conclusion

National research funding systems differ in the total number and the proportion of domestic and foreign funding agencies (Table 2). Great Britain and Denmark have the most diversified system of domestic funders (12 organizations), while the Netherlands (5), Germany (4) and France (4) are less diversified. The high proportion of collaborative publications by foreign funding agencies casts doubts as to whether foreign funders really finance research abroad (Figure 2).

Funded publications lead to a general increase in national research performance (Figure 1) and the impact of papers funded by foreign organisations is on average higher than those of domestic ones. Even after a more appropriate indicator was applied for the organisational comparison (Figure 3), the impact of foreign funders – with the exception of Great Britain – is higher. This result raises the question of whether foreign funding agencies are doing a better job. Do they have more suitable funding instruments than domestic funders in France, Germany, Denmark or the Netherlands? Or a more qualified peer review and selection procedure? These questions are hardly measurable with bibliometric methods. However, the bibliometric analysis shows that there are differences that need to be explained.

The major funding agencies considered in this study have on average a higher HC value than all financed publications of a country (coloured long dashed lines versus black dashed lines). The selection of the major funders is therefore not representative for the entire funding system of a country. It can be assumed that smaller organisations achieve a lower impact. This assumption is also confirmed by the results of the Danish system. Especially in a small country, smaller funders are more likely to cross the threshold of 1% national publications. The Danish case illustrates that many smaller domestic funders have a lower HC rate.

It is remarkable that the impact of a funding agency is lowest in its own country (Figure 4). In addition, the general country patterns appear to be extremely powerful. The possibilities of a funder to make the best use of resources are limited by the country-specific impact or in other words by the researchers located in the country.

These are preliminary results of an ongoing investigation. Further steps need to be taken to gain more insights on how domestic and foreign funders contribute to the research performance of different countries.

Acknowledgments

This research was supported by the German Federal Ministry of Education and Research (BMBF, grant numbers 01PQ16001 and 01PQ17001).

References

Álvarez-Bornstein, B., Morillo, F., & Bordons, M. (2017). Funding acknowledgments in the Web of Science: completeness and accuracy of collected data. *Scientometrics*, *112*(3), 1793–1812.

Braun, D. (1998). The role of funding agencies in the cognitive development of science. *Research Policy*, *27*, 807–821.

Costas, R., & Leeuwen, T. N. (2012). Approaching the "reward triangle": General analysis of the presence of funding acknowledgments and "peer interactive communication" in scientific publications. *Journal of the American Society for Information Science and Technology*, *63*(8), 1647–1661.

Huang, M.-H., & Huang, M.-J. (2018). An analysis of global research funding from subject field and funding agencies perspectives in the G9 countries. *Scientometrics*, *115*(2), 833–847.

Kokol, P., & Vošner, H. B. (2018). Discrepancies among Scopus, Web of Science, and PubMed coverage of funding information in medical journal articles. *Journal of the Medical Library Association*, *106*(1), 81–86.

Mejia, C., & Kajikawa, Y. (2018). Using acknowledgement data to characterize funding organizations by the types of research sponsored: the case of robotics research. *Scientometrics*, *114*(3), 883–904. https://doi.org/10.1007/s11192-017-2617-2

Möller, T., Schmidt, M., & Hornbostel, S. (2016). Assessing the effects of the German Excellence Initiative with bibliometric methods. *Scientometrics*, *109*(3), 2217–2239.

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, *117*(3), 1755–1770.

Sirtes, D. (2013). Funding acknowledgements for the German Research Foundation (DFG). The dirty data of the web of science database and how to clean it up. In *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference* (Vol. 1, pp. 784–795).

Sirtes, D., & Riechert, M. (2014). A Fully Automated Method for the Unification of Funding Organizations in the Web of Knowledge. *STI 2014 Leiden*, 594-597.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *64*(2), 372–379.

Wang, J., & Shapira, P. (2015). Is there a relationship between research sponsorship and publication impact? An analysis of funding acknowledgments in nanotechnology papers. *PloS One*, *10*(2), e0117727.

Wang, X., Liu, D., Ding, K., & Wang, X. (2012). Science funding and research output: a study on 10 countries. *Scientometrics*, *91*(2), 591–599.

The role of geographic proximity on citation preferences: the case of Artificial Intelligence

Isabella Cingolani, PhD[1] and Eleonora Palmaro[2]

[1]i.cingolani@elsevier.com
Elsevier B.V., Radarweg 29, 1043 NX Amsterdam (The Netherlands)

[2]e.palmaro@elsevier.com
Elsevier B.V., Radarweg 29, 1043 NX Amsterdam (The Netherlands)

Abstract

This paper presents a methodology to study geography-based assortative mixing patterns of citations preference, under the assumption that geographic proximity plays a role in knowledge diffusion. We modelled citation relationship as a complex network where scientific publications are represented as nodes and each directed link from a publication to another one models the event that authors of a scientific publication have cited another publication. We then studied assortative mixing pattern defined as nodes' preference to be connected to nodes that have similar characteristics. We explore individual variation in assortative mixing patterns using data from a publication set on Artificial Intelligence. In this study, geography proximity has been codified as a country-based classification of publications. We focused the scope of the analysis to those citation patterns interplaying between publications by authors affiliated uniquely to a single country. Results suggest that geographic proximity plays a significant role in determining knowledge diffusion preferences, but also individual variation exists within each geography-based group and that preference is either directed toward the same country or systematically toward other few countries. The latter empirical evidence helps to unveil what are the country-level research bases where scientific credit is mostly directed.

Background

It is a well-known empirical observation that the strength of most interactions decreases with distance between entities. This is true for physical, economic, social and information systems (Barthélemy, 2011; Liben-Nowell, Novak, Kumar, Raghavan, & Tomkins, 2005; Onnela, Arbesman, González, Barabási, & Christakis, 2011). Similarly, scientific interactions are more likely to occur between scholars localised in nearby areas (Ponds, van Oort, & Frenken, 2007). The impact of geographic distance on creation and diffusion of knowledge has been widely discussed in the field of economic geography and knowledge spill overs (Jaffe, Trajtenberg, & Henderson, 1993; Maurseth, & Verspagen, 2002). In this context, tacit knowledge is usually disseminated through interpersonal exchanges and therefore it is tied to the physical and social space (Breschi, Lissoni, & Montobbio, 2005; Ponds et. al, 2007). In addition to this evidence, studies observed a diminishing role of distance on the diffusion of codified knowledge as international collaboration happens to be a more frequent way of conducting research (Barjak, & Robinson, 2008; Hoekman, Frenken, & Tijssen, 2010; Wagner, & Leydesdorff, 2005).
Nevertheless, in the case of codified knowledge diffusion, geographic closeness still affects the process of receiving or giving credits for someone’s contribution as expressed by citations (Frenken, Hardeman, & Hoekman, 2009). For most scientific contributions we can expect to find a decreasing probability of citation with distance, as new findings are typically more visible in the area where the author works. Studies have investigated the role of geography on the recognition of scientific production using publication citations (Börner, Penumarthy, Meiss, & Ke, 2006; Glänzel, & Schubert, 2005; Schubert, & Glänzel, 2006; Onodera, & Yoshikane, 2015; Pan, Kaski, & Fortunato, 2012; Wang & Zhang, 2018).

In the context of complex systems, networks are extensively used to represent and study patterns of connections among agents constituting these systems. When knowledge creation and diffusion systems are considered, then collaboration and citation networks have been used to model and study properties of codified knowledge flows spanning across various scientific disciplines and time scales (Ding, 2011; Wang, Song, & Barabási, 2013; Zeng et al., 2017).
A feature of many networks is called assortative mixing, the tendency of network nodes to be connected to others that are like themselves according to one or more characteristics. Methods have been developed to capture the average mixing behaviour of nodes, i.e. the average preference for members of one group to create connections with another (Newman, 2003). Beyond the assessment of average preference of one group toward itself or others, it is important also to consider and study individual preferences to provide a more complete picture of preference distribution among and within groups (Cantwell & Newman, 2018). To our best knowledge, networks assortative mixing property has never been considered for the study of spatially constrained citation patterns. This work makes a first attempt in this direction, extending and adapting a network-based approach to the modelling and measurement of geographic-based assortativity for the study of domesticity and internationality of scientific credit and recognition.

Data

The database used in this paper is Scopus (© Elsevier). Scopus, with its 72 million documents at the time of writing, represents the global research landscape well. It includes content coming from all over the world for most subject areas. The analysis of this paper is focused on the field of Artificial Intelligence (AI), which has been defined through a 3-steps bottom-up methodology. The first phase was the extraction of key phrases from representative samples of AI publications, patents, text books, and media outlets, using Natural Language Processing techniques. The second phase was the extraction of the publications using a keyword search based on these key phrases in Scopus, followed by an optimisation of the dataset through the use of a trained classifier to remove false positives. The third phase was a semi-automated machine learning co-occurrence clustering of key phrases in the corpus, which revealed seven sub-fields of AI (Elsevier, AI Resource Center, 2018a; Elsevier, AI Resource Center, 2018b). In our study, only publications with one unique country in the authors' affiliations have been included in the analysis. This means that international publications have been excluded. In addition to this, publications without a country information have been excluded. In this set, 150 countries were represented. The dataset analysed includes 396,587 national publications published between 1996 and 2019. The main countries represented are China, the United States, and India with respectively around 87,000, 61,000 and 27,000 AI publications.

Using this publication set, we constructed the network of citations among publications comprising 396,587 nodes and 1,444,969 links, where nodes represent publications and links are representative of citations between publications. The citations analysed here are those connecting exclusively within the AI publication set. We are excluding those citations going to or coming from publications outside the AI publication set as defined in this study.

Methodology

This study adopts methods proposed by Cantwell & Newman (2018) that can account for and quantify preferences at the individual level. We compute individual citation preferences based on the geographical location of each publication's authors.

In computing quantities of our interests, we opted for naïve preference estimates from data taking care of not including publications that could lead to unreliable estimates of preference. We assume to have *C* groups corresponding to the affiliation country of the non-international publication, n nodes and m links. We then define k_i the out-degree of the node *i,* that corresponds to the sum of all out-going links (i.e. count of out-going citations). We also define the k_{ir} as the number of out-going links from node *i* to group *r.* With n_r we denote the number of nodes within group *r*. The groups are mutually exclusive therefore each node belongs to a single group. We finally define the preference of *i* for group *r* as $\widehat{x_{\iota r}} = k_{ir}/k_i$ representing the naïve preference estimate of the node *i* for group *r*.
In the context of this application, where groups correspond to countries, whenever the group for which we are evaluating the preference by node *i* corresponds to the group where the node *i* belongs to, then we will classify that preference as *domestic*. We will instead classify it as *foreign* when the group *r* is different from the group to which the node *i* belongs to.
We can finally compute the network *average in-group preference* as it follows

$$a = \frac{1}{n}\sum_{i=1}^{N} \widehat{x_{\iota g_\iota}}$$

where *a* = 1 in a perfectly assortative network and *a* = 0 in a perfectly disassortative network. This measure is the average fraction of connections that fall within groups. For real networks, we expect a to lie between 0 and 1, with higher values indicating more assortativity. For assortativity we mean the preference to connect with similar nodes within same group. To know what constitute and high value for *a* we need to calculate the expected $\bar{a}$ within an appropriate null model. The null model here is represented by a network with same number of groups, same distribution of nodes across groups and same distribution of in-coming links across groups. The model does not impose any constraint on what should be the group of origin for an in-coming citation into a group *r*. The *average in-group preference* within the null model can therefore be computed as it follows:

$$\bar{a} = \sum\nolimits_{r=1}^{C} \frac{n_r}{n} \frac{K_r}{m}$$

where K_r is the total number of edges incoming to group *r* and *m* the total number of links in the network. When the difference between the estimated value and the expected value within the null model is greater than 0 this means that the preferences are more assortative than we would expect by chance.
In addition to this global network measure we can compute for each group *r* the *average in-group* and *across-groups* naïve preference estimates for those citing nodes within each group as it follows:

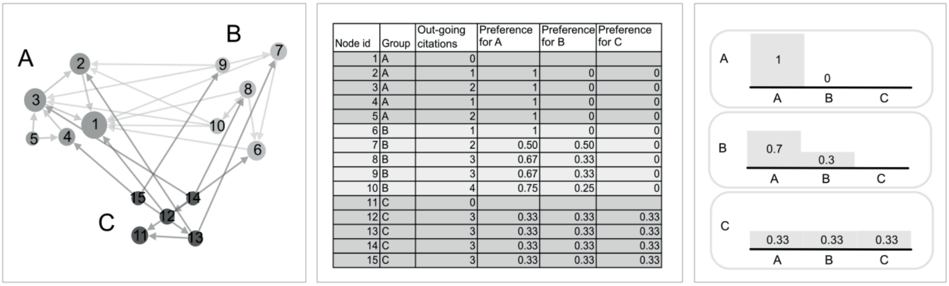
$$a_{rr} = \frac{1}{n_r} \sum\nolimits_{i=1}^{n_r} \widehat{x_{\iota g_\iota}}$$

$$a_{rg} = \frac{1}{n_r} \sum\nolimits_{i=1}^{n_r} \widehat{x_{\iota g_{\iota \notin g}}}$$

If all nodes n_r belonging to the group *r* have at least one out-going citation then it holds that $a_{rr} + \sum_{g \notin r} a_{rg} = 1$. If instead, within a group *r,* there are nodes with no out-going citations n_{r0} , then $a_{rr} + \sum_{g \notin r} a_{rg} = 1 - n_{r0}/n_r$. In this study, we use a_{rr} and a_{rg} , for all $g \notin r$, to assess the distribution of citations among groups distinguishing *domestic* and *foreign* components of citations preference.

In Figure 1 we depict a stylised example of network to illustrate a limited set of ideal cases illustrating how individual and group level preferences can be distributed across groups. Group A's individual members preference seems to be exclusively directed to members of the same group with no variation of individual preferences within the group. We then expect for this group an average in-group preference equal to 1 and a null average across-groups preference. Group B's individual members preference is split between members of the same group and members of group A, with a tendency to be more concentrated on the latter. This group shows a preference for group A higher than the preference for its own group (0.7 vs 0.3). It is also characterised by an average across-groups preference higher than its average in-group preference (0.35 vs 0.3). Finally, group C's members preference is equally distributed between members of the same group and other groups, with an average in-group preference equal to the across-groups preference (0.33).

In this analysis, to avoid having unreliable citation preference estimates in correspondence of low counts of citations, we exclude those citing publications showing less than three citations within the AI network, with three being the 25th percentiles of the distribution of citations per publication.



Node id	Group	Out-going citations	Preference for A	Preference for B	Preference for C
1	A	0			
2	A	1	1	0	0
3	A	2	1	0	0
4	A	1	1	0	0
5	A	2	1	0	0
6	B	1	1	0	0
7	B	2	0.50	0.50	0
8	B	3	0.67	0.33	0
9	B	3	0.67	0.33	0
10	B	4	0.75	0.25	0
11	C	0			
12	C	3	0.33	0.33	0.33
13	C	3	0.33	0.33	0.33
14	C	3	0.33	0.33	0.33
15	C	3	0.33	0.33	0.33

Figure 1 (left panel) a toy citation network with 15 nodes and 25 positive directed links; (central panel) table of computed individual preferences toward each group; (right panel) average preference by couples of groups. Direction of preference goes from each group in x-axis (A, B, C) to group in y-axis, respectively A, B and C, from top to bottom. For nodes with at least one out-going link, the absence of a link connecting them to the others is interpreted as a null preference and therefore is considered in the computation of average values by couple of groups. Nodes with no out-going links have been instead excluded as points of observation of preference behaviour.

Findings

The histogram in Figure 2 plots 20 countries which account for 13% of the overall number of countries. The 20 countries have been ranked by the descending order of the share of publications per country (light grey). The share of in-coming citations (dark grey) and the share of out-going citations (black) per country has been plotted as well.

The 20 countries account for 85% of the publications in our dataset, 88% of the overall in-coming citations, and 86% of the overall out-going citations. The high percentages suggest that AI publications and citations are concentrated in a few countries. The share of in-coming citations has been compared to the share of publications to study if there is a comparable production and dissemination of the knowledge.

The United States ranks first in term of share of the overall in-coming citations and it ranks second based on the share of the overall publications. China shows the complementary pattern. On the other hand, the UK shows the behaviour of a country such as the United States, where the excellence of its research base is well established, being characterised by a relatively higher share of in-coming citations than of its scientific production in the field.

After filtering according to the minimum number of out-going citations per publication, the network comprises 252,990 nodes (65% of initial number of nodes) and 1,082,075 (75% of initial number of links) across 133 countries. For each citing publication, we studied the distribution of citations across countries to study citation preference patterns. We considered first the distribution of publications, of out-going citations and in-coming citations by country. We found that the average global assortativity of AI citation network (0.307) is higher than the assortativity of a corresponding null model (0.087). This means that globally, the citation preferences tend to be directed to similar nodes (same country) more than it would happen by chance.

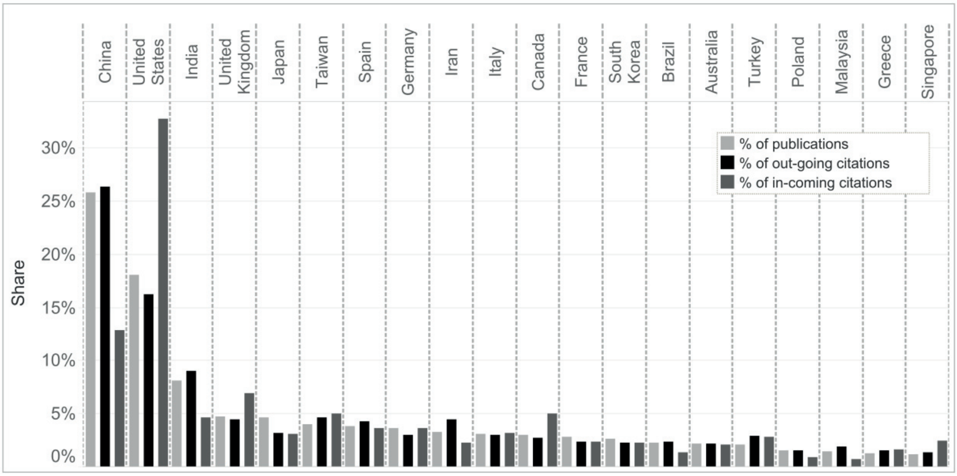


Figure 2 Top 20 countries in terms of publications where all authors are from the same country and their share of publications, in-coming and out-going citations.

In the Figure 3 the average domestic preference and foreign preference is shown. The domestic preference quantitively measures the phenomenon of authors from a country giving scientific credit to authors from the same country. The foreign preference quantitively measures the complementary phenomenon, i.e. authors from a country giving scientific credit to authors from other countries. Overall, countries' domestic preference (0.307 average) is higher compared to the preference given on average to other countries (0.005), meaning that on average, for any given publication in our set, the estimated probability that the authors' of this publication will cite a publication by authors affiliated to institutions located in the same country is around 31%, higher than the estimated probability that the citation will be directed to publications by authors affiliated to another country (0.5%).

For those countries with at least 700 out-going citations (around 0.05% of the overall number of out-going citations), the domestic preference ranges from 0.06 (Bangladesh) and to 0.64 (United States). This can be interpreted that the estimated probability that a publication authored by authors affiliated to United States' institutions will cite a national publication is around 64%, compared to Bangladesh, where the same probability is estimated to 6%. China cites itself with an estimated probability around 31%. The United States and China are big contributor to the average global assortativity of AI citation network, considered their relatively high estimated probabilities to cite themselves and their high number of scholarly outputs. The result shows how much the United States relies on the research developed in its own country to develop new research. Poland (0.35) and Taiwan (0.33) have also a domestic preference higher than the overall average of domestic preferences, respectively. The African countries show low domestic preferences and a number of out-going citations much lower than other countries.

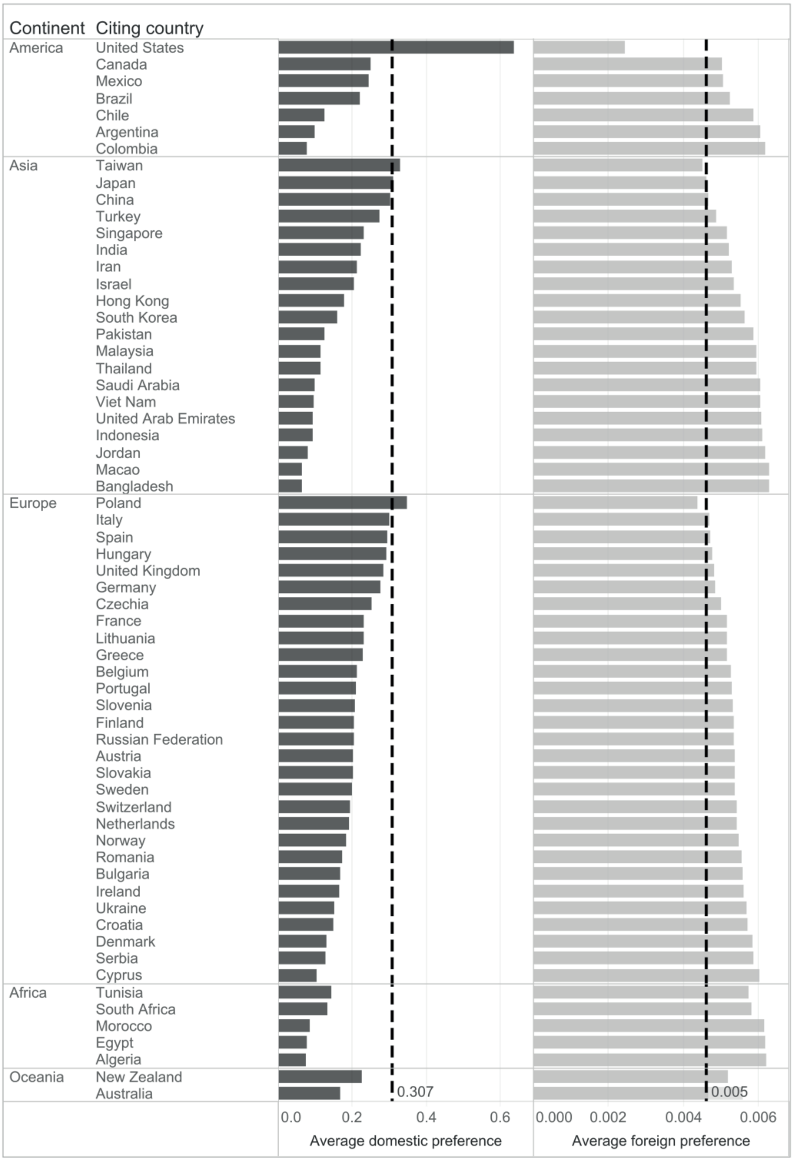


Figure 3 Rank of countries by average domestic preference within each region. Dotted line represents the overall average of domestic (left panel) and foreign (right panel) preferences. Countries with at least 700 out-going citations (0.05% of overall number of citations) are here depicted.

Figure 4 depicts individual preferences variation of selected and ordered countries and regions (Americas, Asia and Europe) according to their varying levels of overall domestic preference. The observed preferences for each citing publication in each country have been classified as domestic and foreign. For each of these two groups, the distribution of preferences has been considered and they have been confronted to each other over 6 preference levels over the range going from 0 to 1, here represented by 5 bins of 0.2 size, with an additional one representing the bin of zero preferences (wherever a publication does not cite any publication from a country).

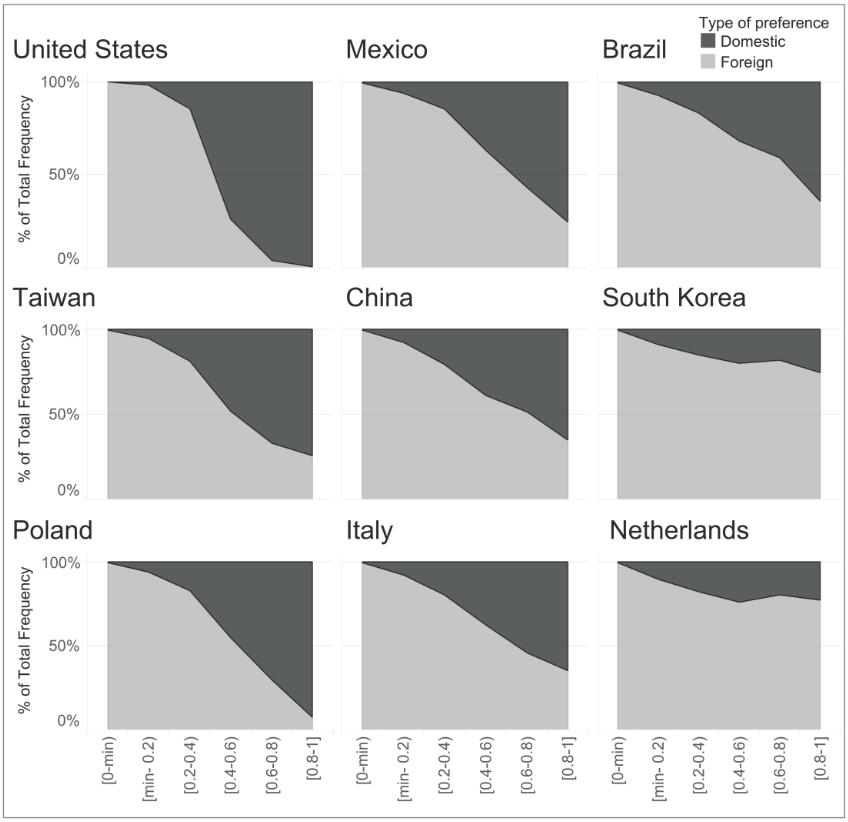


Figure 4 Individual preferences footprints. Citation preferences have been classified in two classes: domestic (dark grey) and foreign (light grey). Domestic preference indicates the tendency of a country to cite itself. Foreign preference instead indicates the complementary behaviour. Countries from three regions (Americas in the top, Asia in the middle and Europe in the bottom) have been ordered from left to right from higher to lower average domestic preference. Countries here depicted are those showing a significant number of observations and have been selected to represent the full spectrum of preference variation within the same region. If the dark grey area increases from left to right (from lower to higher values of preference) the country shows a stronger behaviour to cite itself.

What is evident to the observer is how moving from a low level of individual preference to a high level of individual preference, each country shows a different ‘footprint’ of how its own set of authors tend to distribute their preference toward authors from the same country as opposed to the case they cite authors from other countries. Authors affiliated to United States’ institutions for example show a relatively higher concentration of their citation preference toward national authors (including self-citations) instead of foreign authors. Indeed, darker grey area dominates in correspondence of higher preference levels (darker right side of the normalised frequency area chart). As opposed to the Netherlands authors' that show a relatively higher preference to concentrate the recognition to foreign research fellows from other countries (lighter right side of the normalised frequency area chart).

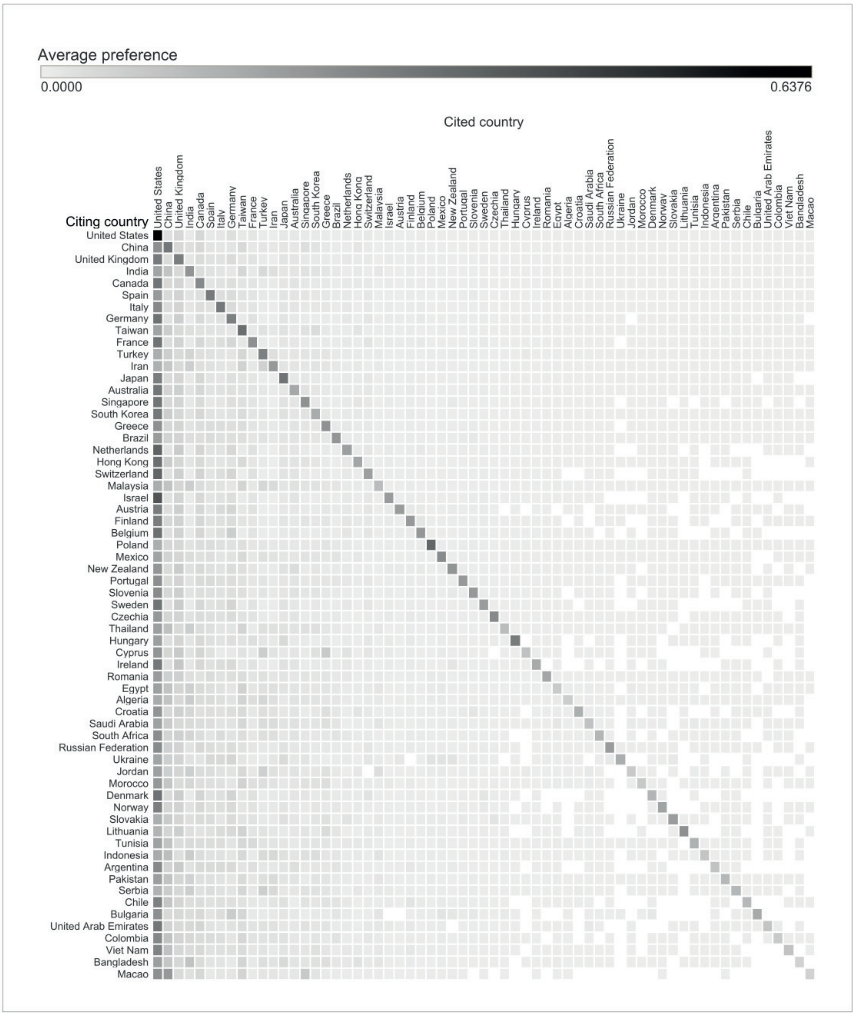


Figure 5 Preference matrix where both citing (rows) and cited (columns) countries have been sorted by their average received preference in a descending order. Each cell is grey-scale-shaded from light grey to black according to the average preference within each couple of countries. Countries for which we observed at least 700 out-going citations (0.05% of all network citations) are here depicted.

Figure 5 shows the structure of citation preferences across countries. The darker shaded diagonal elements of the matrix are indicating that authors within a country tend to cite with a higher probability authors by the same country. The descending rank of countries representing rows and columns of the matrix is a function of their associated average citation preference from all the publications in the dataset, both domestic and foreign. The highest average citation preference is directed toward the United States (0.235 on average), followed by China and United Kingdom (0.062), India (0.041) and Canada (0.040), just to cite the top 5 countries. The almost systematically darker shaded area of the left-hand side matrix in correspondence of the mostly preferred countries, is indicative of the systematic scientific credit they receive from others as opposed to the more puzzled pattern than can be detected elsewhere.

China cites United States with an estimated probability of 24%, while United States' authors cite Chinese peers with an estimated probability equal to 3%, this despite the fact that China owns more than 25% of scientific production in AI. This evidence to stress the asymmetrical nature of scientific credit as a response to real and perceived scientific excellence of the research base of each country.

Discussion

Using a complex network approach, we uncovered mixed assortativity patterns of citation preferences based on the geography of authors. Our analysis of the citation network in the AI research area supports the significant role played by geographical proximity as a driving factor of scientific credit. We assessed that the empirical geography-based assortativity of the network is significantly higher than the assortativity we would obtain by chance. We are also able to assess what are the countries that, beyond national borders, offer the most credited research base in a specific field. This has been possible observing existence of cross-country systematic preferences toward one or a limited number of countries. In the case of AI, we noted that United States is the research base that is systematically preferred by authors from other countries, in some cases more than they would direct to their own research base, notably Germany (32% vs 28%), United Kingdom (30% vs 28%), Netherlands (36% vs 19%) and Singapore (31% vs 23%).

The results show a strong evidence of domestic preference which may be led by the definition of our set of analysis, which includes only publications with authors from the same country. For this set of publications, the tendency to cite the same country could be a more common behaviour that we would expect if international collaborations were included. The choice of including only single country publications was determined by the methodological assumption of classifying publications into distinct geographical groups. Despite the restriction applied to the publications to satisfy the model assumptions, the analysed dataset still counts 65% of the entire set of AI observations. Therefore, it well approximates the overall behaviour of the sample.

The significant evidence of domestic preference could be also influenced by author-level self-citations. A further study could clarify which role the author-level self-citations play in driving these results.

In addition, we focused on the study of citation pattern of AI publications revealing a significant global domestic assortativity and the leading role of United States as the most credited research base beyond domestic borders. The same analysis can be extended to other research areas could give further indications on what are the country-level research bases that attract most of the citations at global level and therefore host the most recognised authors in each field.

Further insights could be gained if a more granular AI classification into more specific topics was adopted. AI is indeed a broad subject area which includes several sub-fields such as

Computer Vision, Neural Networks, Machine Learning and Probabilistic Reasoning, Search Optimisation, Fuzzy Systems and Planning and Decision Making (Elsevier, AI Resource Center, 2018a; Elsevier, AI Resource Center, 2018b).

Conclusions

We proposed a methodology to study citation preference patterns controlling for the geographical location of citing and cited publications, assessing mixing assortativity of individual preferences. This type of analysis helps breaking down the scientific impact accrued by each country by geographic source from which the recognition has been given, from national to international research fellows.

In the context of AI, we observed a significant global assortativity among authors affiliated to the same country but also a tendency to concentrate their citation preference into a limited number of other countries, with the United States playing the leading role as the most scientifically credited country.

The study of spatial citation patters according to the methodology proposed here could be extended to more granular geographical definitions such as cities or sub-regions within countries. A temporal analysis of preferences by country of origin or destination would instead help to determine changes in scientific credit preferences over time. Moreover, it would be relevant to assess how international collaboration conducted by national authors would affect the spatial distribution of the credit received by them.

References

Barjak, F., & Robinson, S. (2008). International collaboration, mobility and team diversity in the life sciences: impact on research performance. *Social Geography*, 3(1), 23-36.

Barthélemy, M. (2011). Spatial networks*. Physics Reports*, 499(1-3), 1-101.

Börner, K., Penumarthy, S., Meiss, M., & Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major US research institutions. *Scientometrics*, 68(3), 415-426.

Breschi, S., Lissoni, F., & Montobbio, F. (2005). The geography of knowledge spillovers: conceptual issues and measurement problems. *Clusters, networks and innovation*, 343-378.

Cantwell, G. T., & Newman, M. E. J. (2018). Mixing patterns and individual differences in networks. *Physical Review E*, 99(4), 042306.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5(1), 187-203.

Elsevier, AI Resource Center (2018a). Technical Background and Methodology for the Elsevier's Artificial Intelligence Report. Retrieved from: https://www.elsevier.com/?a=829143

Elsevier, AI Resource Center (2018b). Artificial Intelligence: How knowledge is created, transferred, and used. Retrieved from: https://www.elsevier.com/research-intelligence/resource-library/ai-report

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3(3), 222-232.

Glänzel, W., & Schubert, A. (2005). Domesticity and internationality in co-authorship, references and citations. *Scientometrics*, *65*(3), 323–342.

Hoekman, J., Frenken, K., & Tijssen, R. J. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, 39(5), 662-673.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly journal of Economics*, 108(3), 577-598.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33), 11623-11628.

Maurseth, P. B., & Verspagen, B. (2002). Knowledge spillovers in Europe: a patent citations analysis. *Scandinavian Journal of Economics*, 104(4), 531-545.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*. 67(2), 026126.

Onnela, J. P., Arbesman, S., González, M. C., Barabási, A. L., & Christakis, N. A. (2011). Geographic constraints on social network groups. *PLoS one*, 6(4), e16939.

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4), 739-764.

Pan, R. K., Kaski, K., & Fortunato, S. (2012). World citation and collaboration networks: uncovering the role of geography in science. *Scientific reports*, 2, 902.

Ponds, R., van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science*, *86*(3), 423–443.

Schubert, A., & Glänzel, W. (2006). Cross-national preference in co-authorship, references and citations. *Scientometrics*, 69(2), 409-428.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research policy*, 34(10), 1608-1618.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, 342(6154), 127-132.

Wang, J., & Zhang, L. (2018). Proximal advantage in knowledge diffusion: The time dimension. *Journal of Informetrics*, 12(3), 858-867.

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, 714, 1-73.

Investigating scientific collaboration through the sequence of authors in the publication bylines and the diversity of collaborators

Yi Bu[1], Chenwei Zhang[2], Yong Huang[3], Cassidy R. Sugimoto[4], and Zaida Chinchilla-Rodríguez[5]

[1] *buyi@iu.edu*
Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.

[2] *zhang334@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.
Qiqihar Institute of Engineering, Qiqihar, Heilongjiang, China

[3] *yonghuang1991@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan, Hubei, China

[4] *sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.

[5] *zaida.chinchilla@csic.es*
Instituto de Políticas y Bienes Públicos (IPP), Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain

Abstract

In scientometrics, it is critical to investigate the patterns of scientific collaboration and how these patterns result in different impacts. In this research, we investigate the relationship between the sequence of authors in the publication bylines and the diversity of their collaborators. The diversity of collaborators is quantified with two dimensions, namely topic and impact diversities. Using the ArnetMiner dataset containing ACM-indexed publications in computer science, we find that the following two patterns tend to lead higher-impact scientific publications: (1) greater topic diversity of collaborators plus more tendency to work as leading authors (including first and/or corresponding authors); and (2) less topic diversity of collaborators plus less tendency to work as leading authors. Meanwhile, from the perspective of impact diversity, the results of our empirical study show that authors who work as more leading authors and collaborate with less impact diversity researchers have tendencies to receive more citations than those with collaborators with greater impact diversity. We also detect different patterns of authors' sequence and diversity of their collaborators before and after their Ph.D. graduation.

Background and Research Objectives

Scientific collaboration is prevalent in various disciplines (Wu *et al.*, 2019). Scientometricians have made great efforts to understanding scientific collaborations from different perspectives, such as scale-free networks (Newman, 2001), homophily and transitivity (Zhang *et al.*, 2018), dependency vs. autonomy (Chinchilla *et al.,* 2018a), geographical proximity (Katz, 1994), science of team science (Stokols *et al.*, 2008), temporal aspects (Bu *et al.*, 2018c), and labour of contribution (Lu *et al.*, 2018).

In bibliometrics and scientometrics, co-authorships are often applied as an important measurement for scientific collaboration. The sequence of co-authors identifies details on "who is accountable for the integrity of the reported study and who deserves what amount of credit for the work" (p. 359), as well as their contributions (He *et al.*, 2012). The patterns of authors' sequence reveal practical implications for scientists, funding providers, and research evaluators; thus, it is crucial to paint a more nuanced picture on the sequence of authors, their collaborators, and the impacts of their co-authored publications (Chinchilla *et al.*, 2019). Another branch of study in scientific collaboration focuses on the diversity of collaborators, a.k.a., members in a research team. For instance, Bu *et al.* (2018b) studied the relationship between an author's impact and his/her collaborators' diversities, namely research topic and impact diversities. They found that high-impact authors tend to have more diverse collaborators in these two dimensions. Likewise, Zhang *et al.* (2019) concluded that the diversity on team members' productivity and scientific ages will increase the team performance. Similarly, a temporal-based analysis demonstrated that co-authors with diverse scientific impact or scientific ages benefit from persistent collaboration more than homogeneous compositions, a.k.a., less diversity (Bu *et al.*, 2018a).

Similarly, the viability and productivity of diversity is impacted by the support it receives from institutional leaders and research funding agencies (Stokols *et al.,* 2019). Many universities, governments, and funding agencies encourage and require cross-disciplinary applicant teams to submit collaboration plans as part of their research proposals (Wang & Shapira 2015; Zhang *et al.*, 2018). However, there are risks especially related with publishing and the allocation of credit in the peer review and academic reward system, institutional barriers, and funding requirement (Bromhan *et al.,* 2016).

Concerning the academic reward system, there exists a lack of credit given to interdisciplinary research in the context of promotion and tenure and limits to career advancement and publishing (Roy *et al.* 2013). Among obstacles are negative perceptions of interdisciplinary research by traditional disciplinary specialists and consequently, troubles publishing because research does not adhere to or fit neatly within traditional disciplinary frameworks (Rafols *et al.,* 2012), and in general, problems related to the peer review system (Wagner *et al.*, 2019). Evidences also suggests that it takes longer for scholars doing interdisciplinary research to establish themselves in their careers (Rhoten & Parker, 2004), and that scholars can be less productive, possible due to cognitive and collaborative challenges associated with such research, which is counterproductive especially in early career stages (Leahey *et al.,* 2017). Institutional review processes may be deeply rooted in disciplinary approaches to evaluation and only senior researchers, who face less-rigid performance evaluations, are better equipped for the complexity associated to with leading and publishing interdisciplinary research projects and publications. Goring *et al.* (2014) coincide in how the current reward structure in academia and other institutions may be misaligned with the current practice of interdisciplinary collaborative science, especially for early career researchers. They advocate for developing strategies behind team building and the requirements for understanding philosophical underpinning to promote interdisciplinary collaborative success. In this research-in-progress paper, we investigate the relationship between the sequence of authors in the publication bylines and the diversity of their collaborators. The diversity of collaborators is measured in two aspects, research topic and impact diversities.

Methodology

Similar to our previous work (Bu *et al.*, 2018a, 2018b), in this paper, we employ the ArnetMiner dataset (Tang *et al.*, 2018b) containing ~2M ACM-indexed computer science publications, as well as ~1.2M authors of these publications and ~8M citation relations between these publications. The authors' names were disambiguated according to the algorithm proposed by Tang *et al.* (2012). Some descriptive statistics can be found in some previous work (e.g., Amjad *et al.*, 2017; Bu *et al.*, 2018b). We follow Bu *et al.* (2018b, 2018c) to focus on articles published between 2001 and 2010, which results in ~450K publications, ~885K distinct authors, nearly 4M different collaboration pairs, and ~606K local citation relations. Note that the ignorance of global citation (e.g., citations from publications outside the current dataset) relations is one of the limitations of the current study, partly because transdisciplinary citations will be missing. We follow Bu *et al.* (2018b) to quantify two dimensions of diversity for an author's collaborators, namely research topic and impact diversities. In terms of research topic diversity of an author's collaborators, we run Author-Conference-Topic model (Tang *et al.*, 2008a), an extended Latent Dirichlet allocation (LDA) model, on our dataset and calculate the cosine similarity between the topic vectors of an authors' collaborators. As for impact diversity of an author's collaborators, we use the normalized standard deviation (NSD) to indicate the degree of impact diversity among the collaborators an author works with, where the *h*-index (Hirsch, 2005) is applied to indicate the impact of the collaborators. We are also interested in the sequence of an author in his/her publication's byline. In computer science, last authors tend to be corresponding authors of a certain publication. Hence, first authors and last authors of the computer science publications are regarded as leading authors (Chinchilla-Rodríguez *et al.*, 2019) in the current paper.

Preliminary Results

Figure 1 contains two sub-figures. In the left sub-figure, the horizontal axis represents the percentage of an author's working as leading authors (i.e., first or last authors), while the vertical axis indicates the cosine similarity of collaborators' research topic—the greater two collaborators' research topic cosine similarity is, the less diversity they are. The color is proportional to the average number of citations received by the corresponding publications. In the left sub-figure, one can find that the top left and the bottom right corners of the heat map feature the most darkness, which demonstrates two patterns that tend to lead higher-impact scientific publications: (1) A greater topic diversity of collaborators plus more tendency to work as leading authors (including first and/or corresponding authors); and (2) a less topic diversity of collaborators plus less tendency to work as leading authors.

The right heat map of Figure 1 reveals the relationship between the percentage of working as leading authors (first and last authors of publications) of an author and his/her collaborators' impact diversities. The only difference between this heat map and its left one is the vertical axis—the current sub-figure shows the NSD of collaborators' *h*-indices while the left cosine similarity of their research topic. We find that authors who work as more leading authors and collaborate with less impact diversity researchers have tendencies to receive more citations than those with collaborators with greater impact diversity.

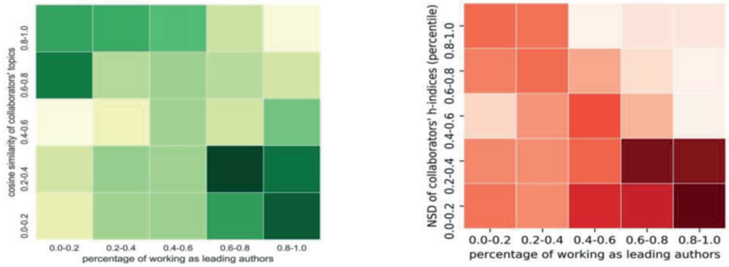


Figure 1. The relationship between the percentage of working as leading authors (first and last authors of publications) of an author and his/her collaborators' research topic (left) and impact (right) diversities. The darkness of cells shows the average number of citations received by the corresponding publications. Note that NSD does not range from zero to one, thus we represent its percentile instead in the vertical axis of the right sub-figure.

Following Amjad *et al.* (2017) and Bu *et al.* (2018d), we also investigate the difference of distributions of authors' sequence before and after their Ph.D. graduation, an important milestone in their scientific career. Due to the limitation of our dataset, we only employ a small sub-set in the dataset (~1K authors), in which we can find the authors' Ph.D. graduation year online. Table 1 shows the basic descriptions, where one can see that before Ph.D.

graduation, more than 70% of an author's publications are first-authored, but the number decreases to ~20% after he/she receives the doctoral degree. Reversely, the percentage of their last-authored publications increases from 8.2% to 39.5% after an author's graduation. The finding makes sense. Before Ph.D. graduation, students tend to work under their supervisor—researchers who tend to lead a study and work as the corresponding authors (more often than not, last authors)—and students themselves tend to write manuscripts, conduct empirical studies, and implement ideas, as pointed out by DeCastro *et al.* (2013) as well as Pachalen and Bhattacharya (2015). Yet, after Ph.D. graduation, authors might have their own students/postdocs, at which stage they might start to lead a certain study and work as corresponding authors (Gingras *et al.*, 2008; Rowlands & Nicholas, 2006). We also quantify research topic and impact diversity of authors' collaborators before and after they receive their doctoral degrees, as shown in the right part of Table 1, where the research topic diversity is equivalent to one minus the cosine similarity of collaborators' research topics and the impact diversity equals to the NSD of collaborators' *h*-indices. Specifically, one can find that the values for both of the two dimensions of diversities increase after an author got his/her Ph.D., though not quite obvious. The dual increasing found in the right part of Table 1 echoes our previous findings in Bu *et al.* (2018b).

Table 1. The distribution of first-, last-, and other-authored publications' percentage before and after authors' Ph.D. graduation, as well as their diversities.

	First-authored publication	*Last-authored publications*	*Middle-authored publications*	*Research topic diversity*	*Impact diversity*
Before graduation	73.8%	8.2%	18.0%	0.48	2.45
After graduation	21.4%	39.5%	39.1%	0.60	2.86

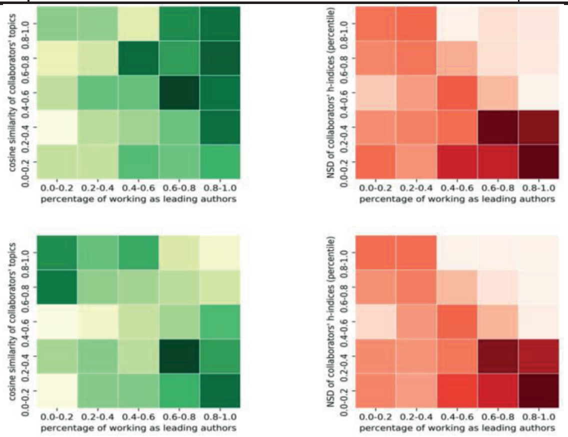


Figure 2. The relationship between the percentage of working as leading authors (first and last authors of publications) of an author and his/her collaborators' research topic (left) and impact (right) diversities. The above row indicates those before Ph.D. graduation, whilst the bottom row shows those after Ph.D. graduation.

We duplicate our experiments shown in Figure 1 on authors that we know their Ph.D. graduation years, and separately consider different patterns in terms of percentage of leading authors and their diversities (research topic and impact diversities) before and after their Ph.D. graduation. The above row in Figure 2 indicates those before Ph.D. graduation, whilst the bottom row shows those after Ph.D. graduation. As shown in the top left sub-figure, one can observe that Ph.D. student who have tendency to work as leading authors with lower research topic diversity of collaborators tend to have higher-impact work, which is different from the pattern revealed in the left sub-figure in Figure 1. Yet, the left bottom sub-figure of Figure 2 looks quite similar to the left sub-figure in Figure 1, partly because of the dominant number of publications authored by post-doctoral researchers among our dataset. As for the impact diversity of collaborators, we surprisingly find that the right two sub-figures in Figure 2 look similar to the right sub-figure in Figure 1, indicating a uniform pattern between authors' sequence and their collaborators' impact diversity, regardless of before or after the authors' Ph.D. graduation.

Conclusion Remarks and Future Work

In this research-in-progress paper, we investigate the relationship between the sequence of authors in the publication bylines and the diversity of their collaborators. The diversity of collaborators is quantified with two dimensions, namely topic and impact diversities. There are many potential implications and applications regarding the finding and the approach of this study. As pointed out by Klein and Falk-Krzensinski (2017), for instance, the Computing Research Association has been grounding generic recommendations in the information science, computing, and engineering fields in a project named *Promotion and Tenure of Interdisciplinary Faculty*. In the project, they not only highlight the interdisciplinarity in job interviews but also emphasize their proposed collaboration-based center/institute to "seek advice on how to balance participation on large team projects with work that establishes a strong individual reputation" (p. 1056).

Using the ArnetMiner dataset containing ACM-indexed publications in computer science, we find that the following two patterns tend to lead higher-impact scientific publications: (1) greater topic diversity of collaborators plus more tendency to work as leading authors (including first and/or corresponding authors); and (2) less topic diversity of collaborators plus less tendency to work as leading authors. Meanwhile, from the perspective of impact diversity, the results of our empirical study show that authors who work as more leading authors and collaborate with less impact diversity researchers have tendencies to receive more citations than those with collaborators with greater impact diversity. We also detect different patterns of authors' first- and last-authored publications before and after their Ph.D. graduation.

There are some future works following this paper. Firstly, we are going to distinguish first and corresponding authors more in detail, as well as other impact related indicators to ensure the robustness of the findings. Secondly, we will follow Bu *et al.* (2018d) to set up several milestones (e.g., Ph.D. graduation, 5 years after Ph.D. graduation) and will consider authors' scientific collaborations with "giants." Thirdly, we will conduct similar empirical studies in various disciplines and implement more comparisons among disciplines with computer science. Moreover, many other issues should be considered in our following-up study, such as collaborators' contribution (Lu *et al.*, 2018) and joint effect of mobility and scientific collaboration (Chinchilla-Rodríguez *et al.*, 2018b). Furthermore, the empirical results shown in Table 1 might be biased. The small sub-dataset used is derived from their "Group A" in Amjad *et al.* (2017), those who have ever collaborated with "giants" at least once in their career. The potential effects of collaborating with "giants" on their author sequence patterns might exist, and this issue will also be researched and discussed in our future work by increasing the sample size for those authors that their Ph.Ds. are known, as well as adding an extra category to distinguish between postdocs and senior researchers are crucial for the reliability of the results.

References

Amjad, T., Ding, Y., Xu, J., Zhang, C., Daud, A., Tang, J., & Song, M. (2017). Standing on the shoulders of giants. *Journal of Informetrics, 11*(1), 307-323.

Bu, Y., Ding, Y., Liang, X., & Murray, D.S. (2018a). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology, 69*(3), 438-448.

Bu, Y., Ding, Y., Xu, J., Liang, X., Gao, G., & Zhao, Y. (2018b). Understanding success through the diversity of collaborators and the milestone of career. *Journal of the Association for Information Science and Technology, 69*(1), 87-97.

Bu, Y., Murray, D. S., Ding, Y., Huang, Y., & Zhao, Y. (2018c). Measuring the stability of scientific collaboration. *Scientometrics, 114*(2), 463-479.

Bu, Y., Murray, D. S., Xu, J., Ding, Y., Ai, P., Shen, J., & Yang, F. (2018). Analyzing scientific collaboration with "giants" based on the milestones of career. *Proceedings of the Association for Information Science and Technology, 5*5(1), 29-38.

Chinchilla-Rodríguez, Z., Miguel, S., Perianes-Rodríguez, A., Sugimoto, C.R. (2018a). Dependencies and autonomy in research performance: examining nanoscience and nanotechnology in emerging countries. *Scientometrics, 115*(3), 1485–1504.

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R., & Sugimoto, C.R. (2018b). Travel bans and scientific mobility: Utility of asymmetry and affinity indexes to inform science policy. *Scientometrics, 116*(1), 569-590.

Chinchilla-Rodríguez, Z., Sugimoto, C.R., & Larivière, V. (2019). Follow the leader: On the relationship between leadership and scholarly impact in international collaborations. *PLoS One*.

DeCastro, R., Sambuco, D., Ubel, P.A., Stewart, A., & Jaqsi, R. (2013). Mentor networks in academic medicine: Moving beyond a dyadic conception of mentoring for junior faculty researchers. *Journal of the Association of American Medical Colleges, 88*(4), 488-496.

Gingras, Y., Larivière, V., Macaluso, B., & Robitaille, J.-P. (2008). The effects of aging on researchers' publication and citation patterns. *PLoS One, 3*(12), e0004048.

Goring, S.J., Weathers, K.C., Dodds, W.K., Soranno, P.A., Sweet, L.C., Cheruvelil, K.S., ... & Utz, R.M. (2014). Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. *Frontiers in Ecology and the Environment, 12*(1), 39-47.

He, B., Ding, Y., & Yan, E. (2012). Mining patterns of author orders in scientific publications. *Journal of Informetrics, 6*(3), 359-367

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academic Science of United States of America, 102*(46), 16569-16572.

Katz, J.S. (1994). Geographical proximity and scientific collaboration. *Scientometrics, 31*(1), 31-43.

Klein, J.T., & Falk-Krzensinski, H.J. (2017). Interdisciplinary and collaborative work: Framing promotion and tenure practices and policies. *Research Policy, 46*(6), 1055-1061.

Leahey, E., Beckman, C.M., & Stanko, T.L. (2017). Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly*, 62(1), 105-139

Lu, C., Ding, Y., Zhang, Y., Bu, Y., & Zhang, C. (2018). Types of scientific collaborators: A perspective of author contribution network. In *Proceedings of iConference 2018*, March 25-28, 2018, Sheffield, U.K.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 404-409.

Packalen, M., & Bhattacharya, J. (2015). Age and the trying out of new ideas. National Bureau of Economic Research. DOI: 10.3386/w20920.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*(7), 1262-1282

Rowlands, I., & Nicholas, D. (2006). The changing scholarly communication landscape: An international survey of senior researchers. *Learned Publishing, 19*(1), 31-55.

Roy, E.D, Morzillo, A.T, Seijo, F, Reddy, S.M.W, Rhemtulla, J.M, Milder, J.C, …, & Martin S.L. (2013). The elusive pursuit of interdisciplinarity at the Human-Environment Interface, *BioScience,* 63: 745-753.

Stokols, D., Hall, K.L., Taylor, B.K., & Moser, R.P. (2008). The science of team science: Overview of the field and introduction to the supplement. *American Journal of Preventive Medicine, 35*(2), S77-S89.

Tang, J., Fong, A.C.M., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transaction on Knowledge and Data Engineering, 24*(6), 975-987.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008a). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.990-998), August 24-27, 2008, Las Vegas, NV., U.S.A.

Tang, J., Jin, R., & Zhang, J. (2008b). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceeding of the Eighth IEEE International Conference on Data Mining* (pp. 1055-1060), December 15-19, 2008, Pisa, Italy.

Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy, 48*(5), 1260-1270.

Wu, L., Wang, D., & Evans, J.A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378.

Zhang, C., Bu, Y., & Ding, Y. (2019). Does diversity of team members affect scientific success of a team? A preliminary study. In *Proceedings of iConference 2019*, March 31-April 3, 2019, Washington D.C., U.S.A.

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology, 69*(1), 72-86.

How a Single Paper Affects the Impact Factor: Implications for Scholarly Publishing

Manolis Antonoyiannakis[1,2]

[1] ma2529@columbia.edu, Department of Applied Physics & Applied Mathematics, Columbia University, 500 W. 120th St., Mudd 200, New York, NY 10027 (USA)

[2] American Physical Society, Editorial Office, 1 Research Road, Ridge, NY (USA)

Abstract

Because the Impact Factor (IF) is an average quantity *and* most journals are small, IFs are volatile. We study how a single paper affects the IF using data from 11639 journals in the 2017 Journal Citation Reports. We define as volatility the IF gain (or loss) caused by a single paper, and this is inversely proportional to journal size. We find high volatilities for hundreds of journals annually due to their top-cited paper—whether it is a highly-cited paper in a small journal, or a moderately (or even low) cited paper in a small and low-cited journal. For example, 1218 journals had their most cited paper boost their IF by more than 20%, while for 231 journals the boost exceeded 50%. We find that small journals are rewarded much more than large journals for publishing a highly-cited paper, and are also penalized more for publishing a low-cited paper, especially if they have a high IF. This produces a strong incentive for prestigious, high-IF journals to stay small, to remain competitive in IF rankings. We discuss the implications for breakthrough papers to appear in prestigious journals. We also question the practice of ranking journals by IF given this uneven reward mechanism.

Keywords:

Science of Science – Impact Factor – Volatility – Indicators – Scientific Impact – Citation Distributions

Introduction

For a performance indicator of a population of papers to be reliable, it needs to be relatively stable and not highly sensitive to fluctuations or outliers—otherwise, the indicator becomes more of a measure of the few outliers than the general population. So, how volatile are Impact Factors, and other citation averages in general? A single research article can tip the balance in university rankings when citation averages are used (Waltman *et al*., 2011; Bornmann and Marx, 2013), due to the skewed nature of citation distributions. It is also known that in extreme situations, a single paper can strongly boost a journal's IF (Dimitrov, Kaveri, and Bayry, 2010; Moed *et al*., 2012). More recently, Liu *et al*. (2018) studied the effect of a highly-cited paper on the IF of four different-sized journals in particle physics and found that "the IFs of low IF and small-sized journals can be boosted greatly from both the absolute and relative perspectives."

The effect of size of a journal or a university department on its citation average cannot be overstated. Previously (Antonoyiannakis, 2018), we discussed the *overall* influence of journal size on IFs, in the context of the Central Limit Theorem. The Theorem tips the balance in IF rankings, because only small journals can score high IFs, while the IFs of large journals asymptotically approach the global citation average in their field via regression to the mean.

In this paper, first, we introduce the IF volatility index as the change, $\Delta f(c)$—or relative change, $\Delta f_r(c)$—when a *single* paper cited c times is published by a journal of Impact Factor f

and size *N*. We study theoretically how $\Delta f(c)$ depends on *c*, *f*, and *N*, and discuss the implications for editorial decisions from the perspective of improving a journal's position in IF rankings. Then, we analyze data from the 11639 journals in the 2017 Journal Citation Reports (JCR) of Clarivate Analytics. We provide summary statistics for the journals' IF volatility to their own top-cited paper. Overall, large values of IF volatility occur for small journal sizes, especially for journals publishing annually fewer than 250 articles or reviews. We discuss the implications for publishing breakthrough papers in high-profile journals.

How a single paper affects the IF: The general case. Introducing the IF volatility index.

Here, we consider what happens when a paper that "brings" *c* citations is published in a journal. The initial IF of the journal is

$$f_1 = \frac{C_1}{N_1}, \qquad (1)$$

where C_1 is the number of citations received in a year and N_1 is the biennial publication count, i.e., the number of published citable items in the previous 2-year period. With the new paper published by the journal, the new IF becomes

$$f_2 = \frac{C_1 + c}{N_1 + 1}. \qquad (2)$$

The change (volatility) in the IF induced by this one paper is then

$$\Delta f(c) = f_2 - f_1 = \frac{C_1 + c}{N_1 + 1} - \frac{C_1}{N_1} = \frac{c - f_1}{N_1 + 1} \approx \frac{c - f_1}{N_1}, \qquad (3)$$

where the approximation is justified for $N_1 \gg 1$, which applies for all but a few journals that publish only a few items per year. So, the IF change $\Delta f(c)$ depends both on the new paper (i.e., on c) and on the journal (size N_1, and citation average f_1) where it is published.

We can also consider the *relative change* in the citation average caused by a single paper, which is arguably a more pertinent measure of volatility. That is,

$$\Delta f_r(c) = \frac{f_2 - f_1}{f_1} = \frac{c - f_1}{f_1(N_1 + 1)} \approx \frac{c - f_1}{C_1}, \qquad (4)$$

where, again, the approximation is justified for $N_1 \gg 1$. The above equation can be further simplified for highly cited papers ($c \gg f_1$) as

$$\Delta f_r(c) \approx \frac{c}{C_1}, \qquad \text{when } c \gg f_1. \qquad (5)$$

Let us now return to $\Delta f(c)$ and make a few remarks.

(a) For $c > f_1$, the additional paper is above-average with respect to the journal, and there is a *benefit* to publication: $\Delta f(c) > 0$ and the IF increases (i.e., $f_2 > f_1$).
(b) For $c < f_1$, the new paper is below-average with respect to the journal, and publishing it invokes a *penalty*: $\Delta f(c) < 0$ and the IF drops (i.e., $f_2 < f_1$).

(c) For $c = f_1$, the new paper is average, and publishing it makes no difference in the IF.
(d) Most important: The presence of N_1 in the denominator means that the benefit or penalty of publishing an additional paper decays rapidly with journal size. This has dramatic consequences, as we will see.

Let us now consider two special cases of interest:

- *Case 1*. The new paper is well above average relative to the journal, i.e., $c \gg f_1$. Here,

$$\Delta f(c) = \frac{c - f_1}{N_1 + 1} \approx \frac{c}{N_1 + 1} \approx \frac{c}{N_1}, \qquad (6)$$

where the last step is justified since in realistic cases we have $N_1 \gg 1$. The benefit $\Delta f(c)$ depends on the paper itself and on the journal size. As mentioned above, the presence of N_1 in the denominator means that publishing an above-average paper is far more beneficial to small journals than to large journals. For example, a journal *A* that is ten times smaller than a journal *B* will have a ten times higher benefit upon publishing the *same* highly cited paper, even if both journals had the same IF to begin with! The *editorial implication* here is that it pays for editors of small journals to be particularly watchful for high-performing papers. From the perspective of competing in IF rankings, small journals have two conflicting incentives: Be open to publishing risky and potentially breakthrough papers on the one hand, but not publish too many papers lest they lose their competitive advantage due to their small size.

For $c \ll N_1$, we have $\Delta f(c) \approx 0$, even when c is large. This means that large journals, even when they publish highly cited papers, have a tiny benefit in their IF. For example, when a journal with $N_1 = 2000$ publishes a highly-cited paper of $c = 100$, the benefit is a mere $\Delta f(100) = 0.05$. For a very large journal of $N_1 = 20{,}000$, even an extremely highly cited paper of $c = 1000$ will produce a small gain of $\Delta f(1000) = 0.05$.

- *Case 2*. The new paper is well below average, i.e., $c \ll f_1$. Again, by "average" we mean with respect to the journal, not the global population of papers. (For journals of low IF, say, $f_1 \leq 2$, the condition $c \ll f_1$ implies $c = 0$.) Here,

$$\Delta f(c) = \frac{c - f_1}{N_1 + 1} \approx -\frac{f_1}{N_1 + 1} \approx -\frac{f_1}{N_1}, \qquad (7)$$

since in realistic cases we have $N_1 \gg 1$. The penalty $\Delta f(c)$ depends now only on the journal parameters (N_1,f_1), and is greater for small-sized, high-IF journals. The *editorial implication* is that editors of small journals—and especially editors of small *and* high-IF journals—need to be more vigilant in pruning low-performing papers than editors of large journals. Two kinds of papers are low-cited, at least in the IF citation window: (a) archival, incremental papers, and (b) some truly ground-breaking papers that may appear too speculative at the time and take more than a couple years to be recognized.

For $f_1 \ll N_1$, we have $\Delta f(c) \approx 0$. So, very large journals have little to lose by publishing low-cited papers.

The take-home message from the above analysis is two-fold. First, with respect to increasing their IF, it pays for all journals take risks. Because the maximum penalty for publishing below-average papers ($\approx f_1/N_1$) is smaller in magnitude than the maximum benefit for publishing above-average papers ($\approx c/N_1$), it is better for a journal's IF that its editors publish a paper they are on the fence about, if what is at stake is the possibility of a highly influential paper that, if proven to be correct, may be ground-breaking. Some of these papers may also reap high citations to be worth the risk: recall that c can lie in the hundreds or even thousands.

However, the reward for publishing breakthrough papers is much higher for small journals. For a journal's IF to seriously benefit from ground-breaking papers, the journal must above all remain small, otherwise the benefit is much reduced due to its inverse dependence with size. To the extent that editors of elite journals are influenced by IF considerations, they have an incentive to keep a tight lid on their risk-taking decisions and perhaps reject some potentially breakthrough research they might otherwise have published. We wonder whether the abundance of prestigious high-IF journals with biennial sizes smaller than $N_{2Y} < 400$ bears any connection to this realization.

On a related note, Wang, Veugelers and Stephan (2017) have reported on the increased difficulty of transformative papers to appear in prestigious journals. They found that "novel papers are less likely to be top cited when using short time-windows," and "are published in journals with Impact Factors lower than their non-novel counterparts, *ceteris paribus.*" They argue that the increased pressure on journals to boost their IF "suggests that journals may strategically choose to not publish novel papers which are less likely to be highly cited in the short run." Our analysis may suggest an additional explanation for their findings that "novel papers encounter obstacles in being accepted by journals holding central positions in science" namely, the punishing effect of journal size on the IF.

Systematic study of the volatility index $\Delta f(c)$, using data from 11,639 journals.

Now let us look at some actual IF data. We ask the question: *How did the IF (citation average) of each journal change by incorporation of its most cited paper, which was cited c^* times in the IF 2-year time-window*? We thus calculate the quantity $\Delta f(c^*)$, where c^* is no longer constant and set equal to some theoretical value, but varies across journals.

First, some slight change in terminology to avoid confusion. We wish to study the effect of a journal's top-cited paper on its IF. Suppose the journal has a citation average f and a biennial publication count N_{2Y}. So, our journal's "initial" state has size $N_1 = N_{2Y} - 1$ and citation average f_1, which we denote as f^*. Our journal's "final" state has $N_2 = N_{2Y}$ and $f_2 = f$, and was produced by incorporation of the top-cited paper that was cited c^* times. We study how $\Delta f(c^*)$ and $\Delta f_r(c^*)$ behave using data from journals listed in the 2017 JCR.

Among the 12,266 journals initially listed in the 2017 JCR, we removed the several hundred duplicate entries, as well as the few journals whose IF was listed as zero or not available. We thus ended up with a master list of 11639 unique journal titles that received a 2017 IF as of December 2018. For each journal in this master list we obtained its Journal Citation Report, which contained the 2017 citations to each of its citable papers (i.e., articles and reviews) published in 2015–2016. We were thus able to calculate the citation average f for each journal, namely, the ratio of 2017 citations to 2015–2016 citable papers. The citation average f approximates the IF and becomes identical to it provided there are no "free" citations in the numerator—that is, citations to *front-matter* items such as editorials, letters to the editor, commentaries, etc., or just "stray" citations to the journal without specific reference of volume and page or article number. We will thus use the terms "IF" and "citation average" interchangeably, for simplicity. Together, the 11639 journals in our master list published 3,088,511 papers in 2015–2016, which received 9,031,575 citations in 2017 according to the JCR. This is our data set.

In Fig. 1 we plot the volatility $\Delta f(c^*)$ vs. N_{2Y} for each journal in our data set. In Table 1 we identify the top-10 journals in terms of $\Delta f(c^*)$, while in Table 2 we show the frequency distribution of $\Delta f(c^*)$ values. Finally, Tables 3 and 4 pertain to the relative volatility $\Delta f_r(c^*)$.

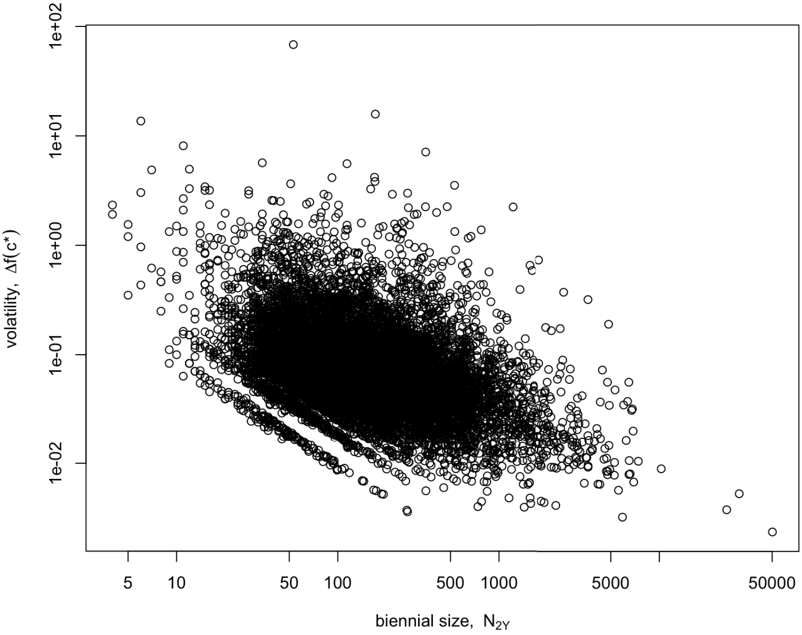


Figure 1. IF volatility, $\Delta f(c^*)$, vs. journal biennial size, N_{2Y}, for all 11639 journals that received an IF in the 2017 JCR.

Our key findings are as follows. A more detailed analysis will be presented in a forthcoming publication (Antonoyiannakis (2019, in preparation)).

1. *Large values of IF volatility occur for small journal sizes, namely, for $N_{2Y} \leq 2000$ and especially for $N_{2Y} \leq 500$.* (That is, for journals publishing *annually* less than 1000 and 250 citable items, respectively.) By large values of volatility, we mean $\Delta f(c^*) \approx 0.5$ and $\Delta f_r(c^*) \approx 25\%$, say.

2. *Many journals experience a large boost in their IF due to their most cited paper.* For instance (see Table 2), there are 381 journals in our data set where $\Delta f(c^*) > 0.5$, i.e., a single paper raises a journal's citation average by at least half a point. For 140 journals we have $\Delta f(c^*) > 1$, while for 41 journals we have $\Delta f(c^*) > 2$, and so on.

3. *For some journals, an extremely highly cited paper causes a large* $\Delta f(c^*)$ *value.* Consider the top 2 journals in Table 1. The journal *CA-A Cancer Journal for Clinicians* published in 2016 a research article that was cited 3790 times in 2017, which accounted for almost 30% of the total citations that entered in its IF calculation that year, with a corresponding $\Delta f(c^*) = 68.3$. Without this paper, the journal's citation average would have dropped from $f = 240.1$ to a "meager" $f^* = 171.8$. Similarly, the *Journal of Statistical Software* published in 2015 a research article that gathered 2708 citations in 2017 and captured 73% of the total citations to the journal that year. Although such extreme levels of volatility are rare, they do occur every year, because of papers cited thousands of times and published in small journals.

4. *A paper needs not be exceptionally cited to produce a large IF boost provided the journal is sufficiently small.* Consider the journals in positions #3 and #4 in Table 1, namely, *Living Reviews in Relativity* and *Psychological Inquiry*. These journals' IFs were strongly boosted by their top-cited paper, even though the latter was much less cited (c^*=87 and c^*=97, respectively) than for the top 2 journals. This happened because journal sizes were smaller also (N_{2Y} = 6 and 11, respectively). Such occurrences are not uncommon, because papers cited dozens of times are much more abundant than papers cited thousands of times, while there are also plenty of very small journals. Indeed, within the top-40 journals (not shown here) in terms of decreasing volatility there are 3 journals whose top-cited paper received 32, 42, and 13 citations respectively, causing a significant $\Delta f(c^*)$ that ranged from 2.3 to 2.6. High values of *relative* volatility $\Delta f_r(c^*)$ due to low-cited or moderately-cited papers are much more common—see Table 3 and journals in positions #2, #3, #4, #9, and #10.

Table 1. **Top-10 journals in volatility $\Delta f(c^*)$, i.e., *absolute* change in IF due to their top-cited paper.**

	Journal	$\Delta f(c^*)$	c^*	$\Delta f_r(c^*)$	f	f^*	N_{2Y}
1	CA-CANCER J CLIN	**68.27**	3790	40 %	240.09	171.83	53
2	J STAT SOFTW	**15.80**	2708	271%	21.63	5.82	171
3	LIVING REV RELATIV	**13.67**	87	273%	18.67	5.00	6
4	PSYCHOL INQ	**8.12**	97	105%	15.82	7.70	11
5	ACTA CRYSTALLOGR C	**7.12**	2499	474%	8.62	1.50	351
6	ANNU REV CONDEN MA P	**5.67**	209	35%	21.82	16.15	34
7	ACTA CRYSTALLOGR A	**5.57**	637	271%	7.62	2.05	114
8	ADV PHYS	**4.96**	85	19%	30.42	25.45	12
9	PSYCHOL SCI PUBL INT	**4.88**	49	33%	19.71	14.83	7
10	ACTA CRYSTALLOGR B	**4.19**	710	199%	6.30	2.11	169

Table 2. Number of journals whose volatility $\Delta f(c^*)$ was *greater than* the threshold value listed in the 1st column.

$\Delta f(c^*)$	No. journals above threshold	% all journals
0.1	3881	33.3%
0.25	1061	9.1%
0.5	381	3.3%
0.75	221	1.9%

1	140	1.2%
1.5	73	0.63%
2	41	0.35%
3	21	0.18%
4	11	0.09%
5	7	0.06%
10	3	0.03%
50	1	0.01%

Table 3. Top-10 journals in *relative* volatility $\Delta f_r(c^*)$, i.e., relative change in IF due to their top-cited paper.

	Journal	$\Delta f(c^*)$	c^*	$\Delta f_r(c^*)$	f	f^*	N_{2Y}
1	ACTA CRYSTALLOGR C	7.12	2499	**474%**	8.62	1.50	351
2	COMPUT AIDED SURG	0.88	9	**395%**	1.10	0.22	10
3	ETIKK PRAKSIS	0.15	4	**381%**	0.19	0.04	26
4	SOLID STATE PHYS	3.03	19	**379%**	3.83	0.80	6
5	CHINESE PHYS C	2.25	1075	**350%**	2.90	0.64	477
6	LIVING REV RELATIV	13.67	87	**273%**	18.67	5.00	6
7	J STAT SOFTW	15.80	2708	**271%**	21.63	5.82	171
8	ACTA CRYSTALLOGR A	5.57	637	**271%**	7.62	2.05	114
9	AFR LINGUIST	0.26	3	**264%**	0.36	0.10	11
10	AM LAB	0.04	5	**247%**	0.05	0.01	136

Table 4. Number of journals whose *relative* volatility $\Delta f_r(c^*)$ was *greater than* the threshold value listed in the 1st column.

$\Delta f_r(c^*)$	No. journals above threshold	% all journals
10%	3403	29.2%
20%	1218	10.5%
25%	818	7%
30%	592	5.1%
40%	387	3.3%
50%	231	2.0%
60%	174	1.5%
70%	140	1.2%
80%	124	1.07%
90%	114	0.87%
100%	50	0.43%
300%	5	0.04%

Conclusions

The above findings corroborate our earlier conclusion (Antonoyiannakis, 2018) that IFs are scale dependent and particularly volatile to small journal sizes, as explained by the Central Limit Theorem. This point is pertinent for real journals because 90% of all journals publish no more than 250 citable items annually (Antonoyiannakis, 2018).

Compared to large journals, small journals have (a) much more to gain by publishing a highly-cited paper, and (b) more to lose by publishing a little-cited paper—that is, more to gain by rejecting a little-cited paper. Therefore, in terms of IF, it pays for small journals to be selective.

The fact that there are more than a hundred journals annually whose highest cited paper suffices to raise their citation average by 1 point demonstrates that the effect we study here is not of academic but of practical interest. If so many journals are that much affected by a single paper, then the usefulness of the IF as a *journal* defining quantity is questioned, as is the practice of IF rankings of journals. This point becomes even more pertinent when we consider the *relative* volatility. Evidently (see Table 4), for 1 out of 50 journals (231 journals) a single paper boosts the IF by 50%. Roughly 1 out of 10 journals (1218 journals) had their IF boosted by more than 20% by a single paper. And for more than a quarter of all journals (3403 journals) the IF increased more than 10% by a single paper.

So, the IF volatility affects thousands of journals. It is not an exclusive feature of a few journals or a statistical anomaly that we can casually brush off, but an everyday feature that is inherent in citation averages (IFs) and affects many journals, every year.

The high volatility of IF values from real-journal data demonstrates that ranking journals by IFs constitutes a non-level playing field, since the IF gain of publishing an equally cited paper scales as the inverse journal size and can therefore span up to 4 orders of magnitude across journals. It is therefore critical to consider novel ways of comparing journals based on more solid statistical grounds. The implications of such a decision may reach much further than producing ranked journal lists aimed at librarians (the original motivation for the IF) and affect research assessment and the careers or scientists.

Disclaimer: The author is an Associate Editor at the American Physical Society. These opinions are his own.

References

Antonoyiannakis, M. (2019, in preparation). Impact Factor volatility to single papers: A comprehensive analysis of 11639 journals.

Antonoyiannakis, M. (2018). Impact Factors and the Central Limit Theorem: Why citation averages are scale dependent, *Journal of Informetrics, 12*, 1072–1088.

Bornmann, L., and Marx, W. (2013). How good is research really? - Measuring the citation impact of publications with percentiles increases correct assessments and fair comparisons, *EMBO REPORTS, 14*, 226–230.

Dimitrov, J.D., Kaveri, S.V., and Bayry, J. (2010). Metrics: journal's impact factor skewed by a single paper, *Nature, 466*, 179.

Liu, W.S., Liu, F., Zuo, C., and Zhu, J.W. (2018). The effect of publishing a highly cited paper on a journal's impact factor: A case study of the Review of Particle Physics, *Learned Publishing, 31*, 261–266.

Moed, HF, Colledge, L, Reedijk, J, Moya-Anegon, F, Guerrero-Bote, V, Plume, A, Amin, M. (2012). Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way. *Scientometrics, 92*, 367–376.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., and van Raan, A.F.J. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics, 87*, 467–481

Wang, J., Veugelers, R., and Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*, 1416–1436.

Matching Education and Scientific Specialization of European Universities: a Micro-based Country Level Analysis

Giuseppe Catalano[1], Cinzia Daraio[1], and Giammarco Quaglia[1]

[1] giuseppe.catalano@uniroma1.it, daraio@diag.uniroma1.it, giammarco.q@gmail.com

Sapienza University of Rome Department of Computer, Control and Management Engineering (DIAG), Via Ariosto 25, 00185, Rome (Italy)

Abstract

The diversity of Higher Education Institutions is well studied in the literature while their disciplinary specialization has received less attention. Even if some authors suggest taking into account the disciplinary specialization of universities to carry out meaningful comparisons, the results obtained thus far about the connection between disciplinary specialization and performance seem still anecdotic and do not find definitive evidence. In this paper, we attempt to contribute to the advancement of this field of research proposing some empirical evidence about the existing matching or mismatching between the educational and the scientific specialization of European universities. To the best of our knowledge, this is the first attempt done thus far. We adopt a multi-level perspective by providing analysis on different levels of aggregation, starting from institution (micro) level data, which are further aggregated at country level. We calculate Balassa indices of specialization on teaching and research and show the existence of *matching* and *mis-matching* in different fields across Europe.

Introduction

The European organizational model of universities, called Humboldtian model (Schimank & Winnes, 2000), is characterized by the coexistence of teaching and research, and by a generalist orientation, based on the coexistence of many disciplines within a single institution. Bonaccorsi and Daraio (2007) provide empirical evidence of the generalist model of European universities. Although the differentiation and diversity of European higher education institutions is a topic well studied in the literature (e.g. Huisman, 1995, 1997, Huisman et al. 2007), the specialization of universities, that is, the orientation to do research and teaching in a few areas as opposed to the traditional model of broad coverage is a relatively unexplored issue. Some authors (e.g. López-Illescas et al. 2011, Bornmann et al. 2013, 2014, Daraio et al. 2015) suggest taking into account the disciplinary specialization of universities to carry out meaningful comparisons, but the results obtained thus far about the connection between disciplinary specialization and performance seem still anecdotic and do not find definitive evidence (Moed et al. 2011). In this paper, we attempt to contribute to the advancement of this field of research proposing some empirical evidence about the existing *matching* or *mis-matching* between the educational and the scientific specialization of European universities. To the best of our knowledge, this is the first attempt done thus far. Previous related studies include the following.

Halffman and Leydesdorff (2010) apply Gini coefficients to university rankings, in order to assess whether universities are becoming more unequal, at the level of both the world and individual nations. They show that universities are not becoming more unequal but present a globalization trend. They observe an increasing output in those countries that are steered by specific policies targeted on this which lead to a global conformation to performance standards. Indeed, Moed et al. (2011) find that concentration and performance in university research are positively related, although the underlying causal relationships are complex. They found no evidence that more concentration of research among a country's universities or among an institution's main fields is associated with better overall performance. They observe a tendency that the research in a particular subject field conducted in universities specializing in other fields outperforms the work in that field in institutions specializing in that field. In their interpretation, it is multi-disciplinary research that is the most promising and visible at the international research front, and that this type of research tends to develop better in universities specializing in a particular domain and expanding their capabilities in that domain towards other fields. The disciplinary specialization of universities on new firms' creation has been investigated by Bonaccorsi et al. (2013). They find that universities specialized in applied sciences and engineering have a broad positive effect on new firm creation while university specialization in basic sciences has an impact on new firm creation in science-based manufacturing industries. Universities specialized in social sciences and humanities have no effect on new firm creation. Robinson-García and Calero-Medina (2014) analyse the problems related to the subject classification of institutions based on bibliometric research data. They conclude by highlighting that "rankings by fields should clearly state the methodology for the construction of such fields".

Data

The data analysed come from the following two databases at disciplinary level: i) The European Tertiary Education Register (ETER) database (https://www.eter-project.com), for information about teaching, and the ii) CWTS Leiden Bibliometric database, for information about academic research. The final dataset contains 1194 institutions that are in both databases. A description of the analysed variables is reported in Table 1.

Table 1. Definition, sources and classification of variables

Source	***Variables***	***Definition***	***Classification***
Teaching (ETER)	Enrolled_students ISCED 5-7	Total students enrolled at ISCED 5-7 (2011-2014)	Fields of Education (FOE) ISCED F 2013
	Grads ISCED 5-7	Total graduates ISCED 5-7 (2011-2014)	Fields of Education (FOE) ISCED F 2013
	ACAD staff_FTE	Total academic staff (expressed in FTE) (2012-2015)	Fields of Education (FOE) ISCED F 2013, but lower coverage (40%)
Research (CWTS Bibliometric database)	Pub_fract	Number of publications (fractional counting) (2012-2015)	Fields of Science (FOS) - 2007
	Pub_top10%	Number of papers in top 10% (2012-2015)	Fields of Science (FOS) - 2007
	PPub_in_top10%	Percentage of papers in top 10% (2012-2015)	Fields of Science (FOS) - 2007
	Pub_int_coll	Percentage of papers with international collaborations (2012-2015)	Fields of Science (FOS) - 2007
	MNCS	Mean normalized citation score (2012-2015)	Fields of Science (FOS) - 2007

Table 2 shows the correspondence applied between the classification of teaching data (based on the UNESCO International Standard Classification of Education: Fields of Education (FOE) and Training (ISCED-F) 2013, draft, May 2013. http://www.uis.unesco.org/Education/Pages/international-standard-classification-ofeducation.aspx) and the classification by fields of science and technology (FOS) that has been introduced in the Frascati manual in the 1960s. OECD conducted the last revision of the FOS classification in 2007 (FOS-2007). The FOS classification is based on principles different than those of FOE. However, at the level of broad fields, as is the case here, the correspondence table reported in Table 2 may be used.

Table 2. Correspondence table FOE – FOS (Source: ETER Handbook, p. 26)

ISCED-F 2013	**Fields of Science FOS - 2007**	**Code of Field**
00 General programmes and qualifications	-	-
01 Education	5.3 Educational sciences	FOES1
02 Humanities and Arts	6. Humanities	FOES2
03 Social sciences	5. Social sciences without 5.2, 5.3 and 5.5	FOES3
04 Business and law	5.2 Economics and Business 5.5 Law	FOES4
05 Natural Science, mathematics and statistics	1. Natural sciences without 1.2	FOES5
06 Information and communication technologies	1.2 Computer and information sciences	FOES6
07 Engineering, manufacturing and construction	2. Engineering and technology	FOES7
08 Agriculture, forestry, fisheries and veterinary	4. Agricultural sciences	FOES8
09 Health and welfare	3. Medical sciences	FOES9
10 Services	-	-

Method

The analyses carried out are based on the calculation of Balassa (1965) specialization indices and their illustrations. We calculate three Balassa indices: i) BInst_inCountry a specialization index of the institution j with respect to the other institutions of the same country; ii) BInst_inEU a specialization index of the institution j with respect to all the other institutions in EU 28+3 (28 EU members + Iceland, Norway and Switzerland); iii) BCountry_inEU a specialization index of a country j with respect to all the other countries in EU 28+3. To save space, in the next section we report selected results on the third Balassa index calculated on research (BR) and on teaching (BT).

The Balassa Research (BR) specialisation index (in the field of Education and Science -FOES *i* of country *j)*, is calculated as: $\boldsymbol{BR}_{i,j} = \frac{\boldsymbol{P_{ij}}/\sum_i P_{ij}}{\sum_j P_{ij}/\sum_{ij} P_{ij}}$, where P_{ij} is the number of publications in the FOES i in country j. The Balassa Teaching (BT) specialization index, *mutatis mutandis*, is calculated using the number of ISCED 5-7 graduates in FOES i in country j instead of the number of publications. The interpretation of the Balassa index is very simple. A value higher than one indicates a specialized unit, a value lower than one identifies an unspecialized unit and a value equal to one points out to an average specialization in that field.

Results

In this section, we illustrate some results obtained thus far. Table 3 reports the specialization indices of teaching and research calculated for five big European countries: France, Germany, Italy, Spain and UK. We find evidence of *matching* between Education and Scientific Specialization of Universities in France and in Germany in Natural Sciences (FOES 5), France in ICT (FOES6), in Italy and in Spain in FOES 7 (Engineering and Technology) and only in Italy in FOES 9 (Heath). UK, departing by the other European countries show a matching in FOES 1 and 2 (respectively Education and Humanities).

The observed mismatch in social sciences and humanities may show the strong bias of the WoS database with respect to national language publications.

Table 3. Balassa indices calculated on research and on teaching dimensions. BR is Balassa Research index (calculated on number of publications, fractional count), BT is Balassa Teaching index (calculated on number of graduate students).

Field code	***France***		***Germany***		***Italy***		***Spain***		***UK***	
	BR	BT	BR	BT	BR	BT	BR	BT	BR	BT
FOES1	0.2577	0.4151	0.4789	**1.053**	0.2996	0.5628	0.9691	**2.147**	**2.016**	**1.064**
FOES2	0.4822	**1.1982**	0.7043	**1.305**	0.4938	**1.115**	0.9368	0.6979	**1.866**	**1.246**
FOES3	0.4038	0.8560	0.7741	0.6312	0.5593	0.9299	0.8205	0.6318	**1.792**	0.9866
FOES4	0.6403	**1.375**	0.8366	0.8445	0.7912	0.9125	**1.165**	0.935	**1.518**	0.9514
FOES5	**1.309**	**1.017**	**1.120**	**1.859**	0.8920	0.9424	**1.049**	0.7576	0.8925	0.9763
FOES6	**1.190**	**1.121**	0.7353	**1.376**	0.9889	0.3205	**1.929**	0.8893	0.8757	**1.103**
FOES7	**1.287**	0.9060	0.8189	**1.307**	**1.162**	**1.346**	**1.275**	**1.195**	0.8446	0.695
FOES8	0.5922	0.1840	0.7351	0.6485	**1.204**	0.8375	**1.709**	0.6455	0.5885	0.3795
FOES9	0.6116	0.9578	**1.070**	0.5333	**1.226**	**1.335**	0.6359	**1.100**	0.9915	**1.238**

Note: the values in bold (higher than 1) identify countries specialized in that field. The cells highlighted in grey show matching between specialization of teaching and specialization of research.

Figure 1 illustrates the existing matches or mis-matches for other 8 European countries. In ICT (FOES6) we observe a matching in the specialization of teaching and research (in Austria, Greece and Finland). In Engineering and Technology (FOES 7) there are Greece, Finland, Ireland and Portugal that show the existence of a matching between teaching and research specialization.

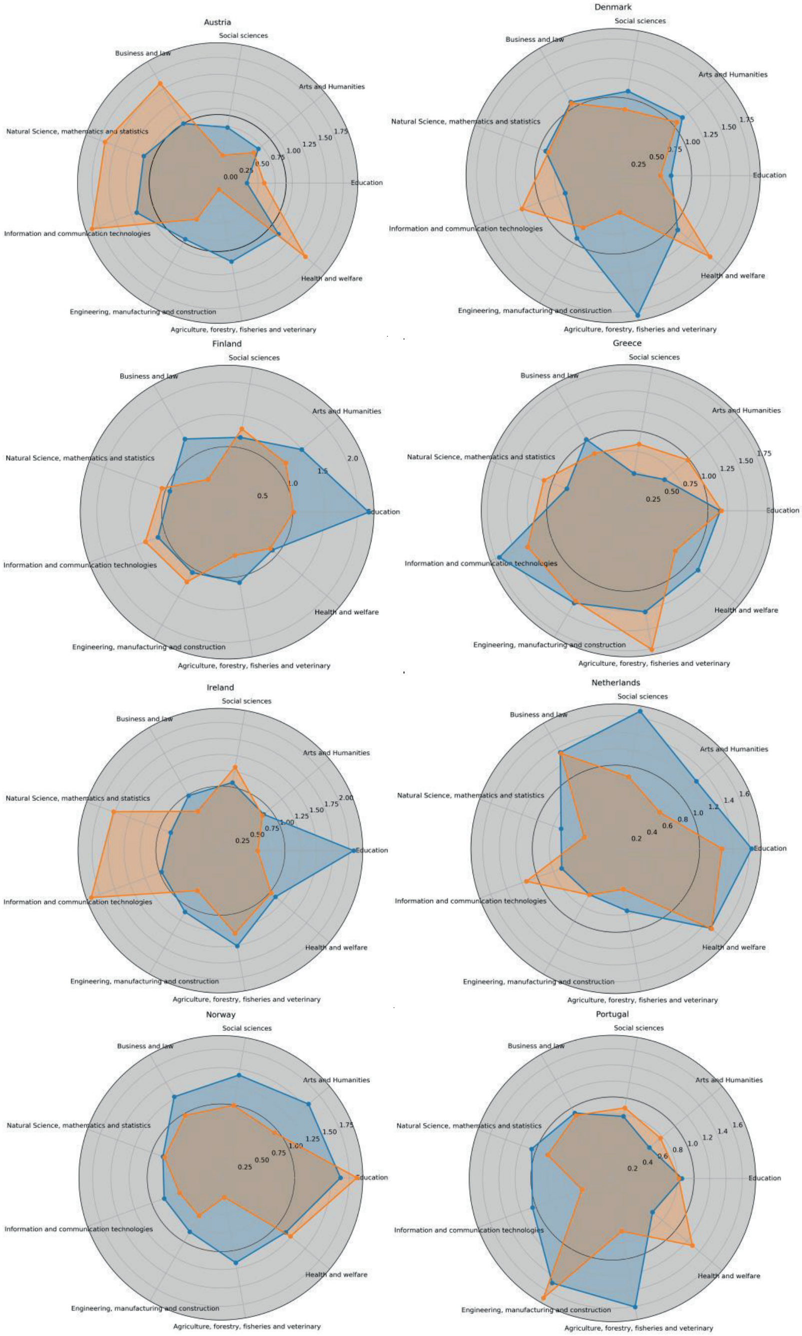


Figure 1. Radar plots of the Balassa indices calculated on research and teaching dimensions, by country. The Balassa Research (BR) specialization index is the blue line, while the Balassa Teaching (BT) is the orange line. The circle corresponding to the value 1 is reported in bold.

Conclusions

Building on a concordance table between Field of Education and Field of Science, we calculated Balassa specialization indices at institution (micro) level and at a country (macro) level. The latter is illustrated in previous section. After that, we investigate the existence of matching or mis-matching between the research and teaching specialization in a given field. The obtained results may be very helpful to inform policy makers about fields in which efforts (resources or measures) should be put.

Extensions of the analysis include calculating Balassa Research (BR) specialization indices using the other bibliometric indicators (see Table 1) about research; extending the analysis including data regarding the patenting activities from the PATSTAT database.

Acknowledgments

The financial support of the Italian Ministry of Education and Research (through the PRIN Project N. 2015RJARX7), of Sapienza University of Rome (through the Sapienza Awards no. PH11715C8239C105) and of the EU Horizon 2020 RISIS2 Project (grant agreement N° 824091) is gratefully acknowledged. We warmly thank Ed Noyons for providing us the CWTS bibliometric data analysed in this paper.

References

Balassa, B. (1965). Trade liberalisation and ''Revealed'' comparative advantage. The Manchester School, 33(2), 99–123.

Bonaccorsi, A., Colombo, M. G., Guerini, M., & Rossi-Lamastra, C. (2013). University specialization and new firm creation across industries. Small Business Economics, 41(4), 837-863.

Bornmann, L., Moya-Anegón, F., & Mutz, R. (2013). Do universities or research institutions with a specific profile have an advantage or a disadvantage in institutional rankings? A latent class analysis with data from the SCImago Ranking. Journal of the American Society for Information Science and Technology.

Bornmann, L., Stefaner, M., de Moya Anegón, F., & Mutz, R. (2014). Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level models. Online Information Review, 38(1), 43-58.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. European Journal of Operational Research, 244(3), 918-930.

Huisman, J. (1995). Differentiation, diversity and dependency in higher education. Utrecht: Lemma.

Huisman, J. (1997). Institutional and programmatic diversity. A comparative analysis of national higher education systems in nine Western European countries. University of Twente: Netherlands. CHEPS (Centre for Higher Education Policies Studies)—Thematic report II.

Huisman, J., Meek, L., & Wood, F. (2007). Institutional diversity in higher education: A cross-national and longitudinal analysis. Higher Education Quarterly, 61, 563–577.

López-Illescas, C., de Moya-Anegón, F. & Moed, H. F. (2011). A ranking of universities should account for differences in their disciplinary specialization. Scientometrics, vol. 88, n.2, pp. 563–574.

Moed, H. F., de Moya-Anegón, F., López-Illescas, C., & Visser, M. (2011). Is concentration of university research associated with better research performance?. Journal of Informetrics, 5(4), 649-658.

Robinson-García, N., & Calero-Medina, C. (2014). What do university rankings by fields rank? Exploring discrepancies between the organizational structure of universities and bibliometric classifications. Scientometrics, 98(3), 1955-1970.

Schimank, U., & Winnes, M. (2000). Beyond Humboldt? The relationship between teaching and research in European university systems. Science and public policy, 27(6), 397-408.

Coping with Altmetrics' Heterogeneity – A Survey on Social Media Platforms' Usage Purposes and Target Groups for Researchers

Steffen Lemke[1] and Isabella Peters[2]

[1] s.lemke@zbw.eu, [2] i.peters@zbw.eu
ZBW – Leibniz Information Centre for Economics, Düsternbrooker Weg 120, 24105 Kiel (Germany)

Abstract

As the online platforms used as sources for altmetrics are highly heterogeneous regarding the usage purposes they fulfil, aggregating altmetrics from different platforms to conflating scores means to amalgamate the results of actions with possibly completely different meanings and intentions. This impedes the informative value and complicates the interpretation of altmetric scores. To arrive at a more differentiated understanding of the motivations under which interactions with research products on platforms that are potential sources for altmetrics take place, we surveyed 1,018 researchers about the usage purposes that 18 popular social media platforms serve for them, as well as about the target groups they aim to reach by being active on these platforms. By performing hierarchical clustering on basis of the response data, we reveal similarities and differences between the examined platforms regarding the goals they help to fulfil and the communication partners they are used to address. Our findings contribute to a better differentiation between altmetrics derived from different sources and thus aim to increase the informative value of different web-based metrics for research evaluation.

Introduction

The utilization of altmetrics for the evaluation of research is still impeded by severe gaps of understanding regarding what they truly mean. One central challenge in this regard is posed by the heterogeneity of the various sources that are used to acquire altmetric data, i.e. the online platforms on which interactions with scientific outputs are observed and counted (Haustein, 2016). Sources of altmetric data include a variety of platform 'classes', e.g., social networks, microblogging platforms, literature management services, news outlets, blogs, and more. And even if one was to compare two instances from the same class of platforms, for example *Facebook* and *LinkedIn* as representatives of the class 'social network', anyone somewhat familiar with both examples could most likely quickly point out substantial differences in the goals they usually help their users to fulfil.

The motivations for which stakeholders of science – especially researchers – use a certain online platform affect the meaning of their interactions with research products on it: a mention of a scientific article on a platform that is predominantly used by most of its users to promote the own research projects for instance will likely have a different meaning than a mention on a platform that is first and foremost used for entertainment purposes. In altmetrics, which are often reported as conflated scores that comprise indicators derived from a variety of different platforms, these nuances are usual indiscernible. Such aggregations of regarding their underlying motivations possibly deeply heterogeneous indicators reduce the altmetrics' informative value and make their appropriate interpretation more difficult.

One way of arriving at less ambivalent altmetrics would be to group altmetric sources regarding their similarity and only perform aggregations for sources that do not reach a certain threshold of dissimilarities. There are various possible ways how one could define such similarity: one could for example strive for similarity of sources regarding their technical affordances, their data volume, or their user demographics (see also Lemke, Mehrazar, Mazarakis, & Peters, 2018). In this article however – for the reasons stated above – we suggest to compare potential sources for altmetric data on basis of the purposes they usually fulfil for their users. As we are interested specifically in the usage of online platforms in relation to scholarly publications, we focus on the group of researchers, which we assume to be the user group most commonly interacting with research products online (see also Tsou, Bowman, Ghazinejad, & Sugimoto, 2015; Vainio & Holmberg, 2017).

To better understand the motivations with which researchers use various online platforms during their work, we aim to answer the following research question using response data from an online survey: *Which purposes do different social media platforms (that are potential sources for altmetrics) serve for researchers?* To additionally identify how researchers use different platforms to communicate with different stakeholders of science, we also aim to answer a second research question: *Which groups do researchers try to reach by being active on different social media platforms?*

Several studies have been investigating on researchers' usage of social media, although rarely specifically focussing on platforms which are potential sources for altmetrics. Van Noorden (2014) surveyed researchers who reported to regularly visit social media sites in detail about the purposes for which they use six popular social media platforms, revealing that *Mendeley, Facebook, Twitter* and *LinkedIn* all differ considerably from each other regarding the ways they are used by researchers, while *ResearchGate* and *Academia.edu* fulfil widely similar sets of purposes. In a follow-up survey, Harseim (2017) asked over 3,000 researchers about the tasks they do on social media in relation to their work, finding the discovering/reading of content to be the overall most prevalent of seven examined tasks. Focussing on science-specific social networking services, Nentwich & König (2014) identified *communication and cooperation, public relations and self-marketing, e-teaching* and *job exchange* to be main usage practices for researchers. LaPoe, Carter Olson, & Eckert (2017) interviewed 45 media scholars about their usage of social media in their professional lives, identifying the promotion of academic work and the use of social media as a communication and mentorship tool to be common motives for scholars. In a survey with over 20,000 responses, Kramer & Bosman (2016) suggested a set of 17 research activities from six areas in which online platforms and tools provide support to the work of researchers.

Methods & Data

To collect data to help us answer our research questions we designed an online survey questionnaire of 14 questions about the use cases 18 popular social media platforms fulfil for researchers. The platforms to include were largely based on the most popular online tools among researchers according to the **metrics*-project's survey from 2017 (see also Lemke et al., 2018). From the ranking established in that survey we removed all platforms that were either used by less than 100 survey respondents or did not match the social media definition by Kaplan & Haenlein (2010), which describes social media as "Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content", leaving us with 16 platforms. To this set we added *Xing* and *Quora*, which both had been mentioned frequently as free text answers to an "Other" field in 2017's survey. The full list of the 18 platforms included in this study can be seen in Figure 1 and Figure 2.

Due to the **metrics*-project's disciplinary focus, a primary goal during dissemination was the recruitment of a representative number of researchers from the social sciences and economics. The call for participation was sent to about 27,000 mail addresses of researchers, ~6,000 of those addresses stemmed from a mailing list administered by the *ZBW Leibniz Information Centre for Economics,* which primarily contained mail addresses of economists from the German-speaking parts of Europe. The remaining ~21,000 mail addresses belonged to authors of recent publications from the fields of social sciences or economics found on *RePEc* and *Web of Science*. The dissemination of the survey took place from June 25th to July 14th 2018, a wave of reminders was sent out in the second week of August 2018. The survey was then kept open till August 27th 2018. As an additional incentive, participants could enter a drawing of 25 10€-*Amazon.com* vouchers at the end of the questionnaire.

To investigate on the usage purposes that individual platforms fulfil for our respondents, we first asked them which of the 18 platforms included in our survey they had used at least once for work till now. All the platforms selected this way were then presented to the respective respondent on the y-axis of the subsequent matrix question *"Why do you use the following services for your work?"*. The x-axis of that question included 25 usage purposes for social media in the research workflow. To come up with this list of purposes we reviewed existing literature on scholarly use of social media (see above) to identify as many purposes as possible, which we then merged and reduced to the aforementioned 25 cases to keep the questionnaire reasonably short. The full list of purposes can be seen in Figure 1.

In a similar fashion we later in the questionnaire asked our respondents in a matrix question *"Whom do you want to reach by being active on the following services?"* to answer our second research question. The y-axis of that question again included the platforms the respondent previously had ticked as used, the x-axis showed 10 groups of stakeholders in scholarly communication, as can be seen in Figure 2.

The responses to both questions were transformed into heat maps by calculating for every pairing of platform and purpose (or platform and target group) the share among the users of the respective platform that had ticked that purpose (or target group) as complying with their usage behavior. To more easily identify similarities between platforms regarding both research questions a hierarchical clustering was performed on the data using *R's* package *gplots*. Clusters were built with the complete-linkage method using Euclidean distances.

Results & Discussion

In the following section we will present survey demographics and the findings from our cluster analyses of the survey responses.

Survey Demographics

A total of 1,018 researchers responded to the survey, meaning a response rate of ~4%. Of all respondents stating a gender 69% identified as male, 31% as female, and <1% as another gender. The majority of respondents stated Germany (28%) as their current country of affiliation, followed by the USA (14%), the UK (6%), Italy (6%), and France (4%). In total respondents stated 70 different countries of affiliation.

Discipline-wise, the vast majority of respondents reported to primarily work in economics (70%), followed by social sciences (18%), other disciplines (7%), engineering/technology (2%), arts/humanities (1%), life sciences (1%), medicine (1%), physical sciences (<1%), and law (<1%). Regarding their current career stage most respondents stated to be professors (28%), followed by associate professors (17%), research assistants/PhD students (17%), postdocs/senior researchers (15%), assistant professors (12%), and other career stages (11%).

Research Question 1 – Researchers' Usage Purposes of Social Media

Figure 1 shows to which degrees individual platforms fulfil various purposes for their respective users. Each cell reflects the percentage of users of that specific platform among our survey respondents who stated that they would use it for that respective purpose – the darker the cell, the higher the share. The area on the left of the heat map shows the dendogram resulting from a hierarchical clustering of the data, the lines of the heat map are ordered respectively. The column on the right of the heat map shows the share of survey respondents who reported to have used the respective platform for work purposes before.

Through a first visual examination of the platforms' distribution over the heat map, a concentration of social networking services in its lower half becomes apparent. This group of six platforms is united by prevalently serving the purposes of facilitating *networking* as well as *maintaining a personal profile*. Moreover, all six platforms commonly help with *receiving updates/news from the scientific community*, *self-promotion*, *discovering/announcing job opportunities*, and *personal communication,* although it can be seen that more specialized networks like *Academia.edu* or *Xing* seem to be slightly less versatile regarding their use cases than the more general platforms *Facebook* or *Twitter*.

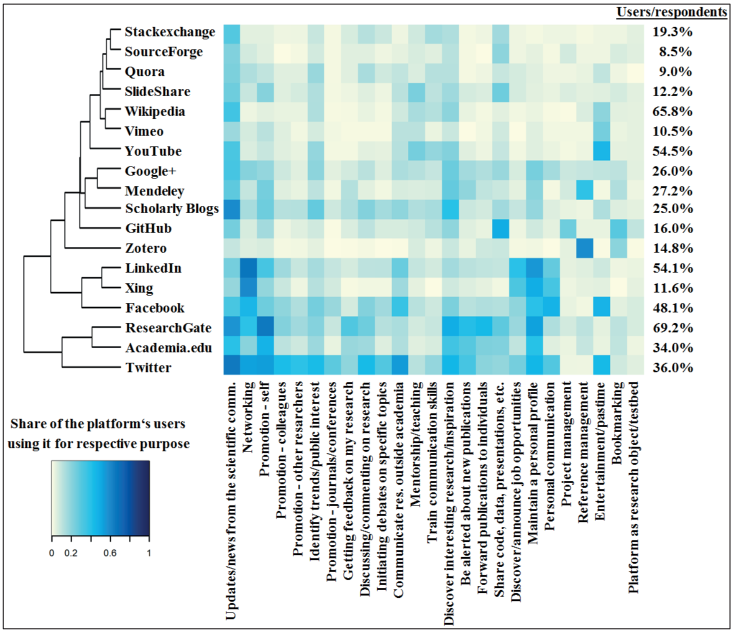


Figure 1: Usage purposes fulfilled by social media platforms for researchers.

The remaining 12 platforms in the upper area of the heat map mostly appear to be more specialized in that each of them tends to serve only few usage purposes for large shares of its users. For many of them the reported fulfilment of usage purposes is very low in general, which could indicate that the response options in our survey did not cover the true usage purposes these platforms fulfil. The groupings resulting from the cluster analysis for these platforms should therefore be interpreted with caution.

Looking at the clusters on the lowest levels, we see some expectable pairings regarding served purposes in the lower half of the diagram: to probably little surprise *LinkedIn* behaves similar to its German counterpart *Xing*, both being used much for *networking*, *discovering or announcing job opportunities*, *maintaining a personal profile* and *personal communication*. In a similar fashion the academic social networks *ResearchGate* and *Academia.edu* form one cluster. Interestingly, one level higher *Facebook* forms a cluster with the two employment-oriented platforms, as all three services share particular emphases on the purposes of *personal communication* and *networking*. The two academic social networks on the other hand form a cluster with *Twitter*, shared focuses lying on *updates/news, self-promotion*, *discovering interesting research* and *alerts about new publications*.

Examining individual lines of the heat map, especially Twitter's outstanding role as a platform with high versatility sticks out. Almost every purpose is for a considerable share of users fulfilled by Twitter, primarily except the most specific purposes that only highly specialized platforms cater to, e.g., project management or reference management.

Research Question 2 – Audiences Targeted by Researchers on Social Media

The heat map in Figure 2 shows the shares of users of respective platforms among our survey respondents that stated that they would aim to reach the respective target group on this platform. Dendogram and user shares are arranged in a way analogous to Figure 1.

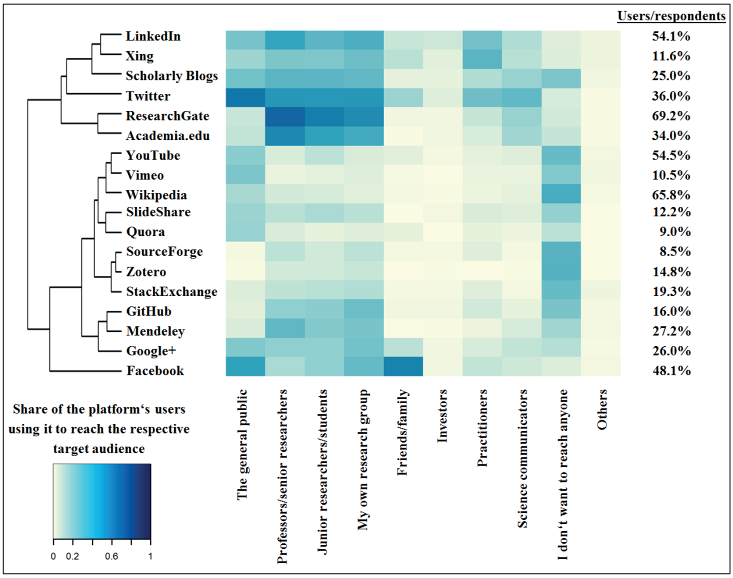


Figure 2: Audiences targeted by researchers on social media platforms.

Examinations of the platforms' order in the dendogram as well as the highest level of clustering suggest a rough subdivision of the platforms into two groups: first, platforms that are prevalently used to reach out to other researchers (*LinkedIn, Xing, Scholarly Blogs, Twitter, ResearchGate, Academia.edu*) and second, platforms on which this is not the case. Among the latter are platforms which are at least fairly commonly used to reach the general public (e.g., *YouTube, Vimeo, Facebook*) but also those on which most users do not try to actively reach anyone at all (e.g., *Wikipedia, SourceForge, Zotero*). A particular use case is fulfilled by *Facebook*, which is prevalently used to communicate with friends and family.

Conclusion

We conducted an online survey to get to a better understanding of which usage purposes several popular social media platforms serve for researchers, in particular regarding the task of scholarly communication. Our analysis revealed distinct clusters of platforms that behave similar regarding the communication goals they fulfil for researchers. These insights contribute to the aim of achieving a more evidence-based foundation for the reasonable interpretation and aggregation of altmetrics measured on these platforms by indicating which services cater to similar communication needs and might therefore be aggregated with comparatively little loss of information. The findings also help to characterize the scholarly information we can expect to find on different platforms regarding its amount and complexity. On *Facebook* and *Twitter* for example it seems to be more common for researchers to address non-academic audiences than on other social media, suggesting that mentions of research products on *Facebook* and *Twitter* will to a higher degree reflect efforts of disseminating research to the general public, compared to mentions on for instance scholarly blogs, academic-, or business-oriented networks. These hypotheses will have to be backed up by further research though.

Moreover, regarding its usage purposes especially *Twitter* stood out as a particularly versatile platform for researchers. This suggests that on *Twitter* an especially varied and complex interplay of user motivations might affect the scholarly communication taking place, underlining the need of its thorough exploration to fully enable *Twitter* as a source for informative altmetric data.

A limitation of this study lies in its sample's bias towards social scientists and economists, which impedes its validity to other disciplines. And – as is typical for online surveys – our sample will be subject to self-selection bias, meaning a likely overrepresentation of researchers with a comparatively high interest in the topics of social media or research metrics. Also, although it might be reasonable to assume that researchers are the user group most actively citing research online, there are other stakeholders interacting with scientific products on the Web – and therefore affecting altmetrics – whose user behaviour is not captured by our survey. Future work should go into addressing these limitations as well as into the discussion of this study's implications for the construction of indicators.

Acknowledgments

This work is part of the DFG-funded research project **metrics* (project number: 314727790). We wish to thank the researchers who helped us by participating in our survey.

References

Harseim, T. (2017, July 15). How do researchers use social media and scholarly collaboration networks (SCNs)? : Of Schemes and Memes Blog. *of schemes and memes*. Retrieved February 8, 2019, from http://blogs.nature.com/ofschemesandmemes/2017/06/15/how-do-researchers-use-social-media-and-scholarly-collaboration-networks-scns

Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics*, *108*(1), 413–423. doi:10.1007/s11192-016-1910-9

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. doi:10.1016/j.bushor.2009.09.003

Kramer, B., & Bosman, J. (2016). Innovations in scholarly communication - global survey on research tool usage. *F1000Research*, *5*, 692.

LaPoe, V. L., Carter Olson, C., & Eckert, S. (2017). "Linkedin Is My Office; Facebook My Living Room, Twitter the Neighborhood Bar": Media Scholars' Liminal Use of Social Media for Peer and Public Communication. *Journal of Communication Inquiry*, *41(3)*. doi: 10.1177/0196859917707741

Lemke, S., Mehrazar, M., Mazarakis, A., & Peters, I. (2018). Are There Different Types of Online Research Impact? *Proceedings of the 81st ASIS&T Annual Meeting: Building an Ethical & Sustainable Information Future with Emerging Technology* (pp. 282–289). Presented at the 81st ASIS&T Annual Meeting in Vancouver, Silver Springs, MD, USA: American Society for Information Science.

Nentwich, M., & König, R. (2014). Academia Goes Facebook? The Potential of Social Network Sites in the Scholarly Realm. In S. Bartling & S. Friesike (Eds.), *Opening Science* (pp. 107–124). Cham: Springer International Publishing.

Tsou, A., Bowman, T. D., Ghazinejad, A., & Sugimoto, C. R. (2015). Who tweets about science? In *Proceedings of the 2015 International Society for Scientometrics and Informetrics* (pp. 95–100). Istanbul, Turkey.

Vainio, J., & Holmberg, K. (2017). Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions. *Scientometrics*, *112*(1), 345–366. doi:10.1007/s11192-017-2368-0

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature*, *512*(7513). Retrieved from http://www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711

The P-model:
An Indicator that Accounts for Field Adjusted Production as well as Field Normalized Citation Impact.

Erik Sandström[1], Ulf Sandström[2] and Peter van den Besselaar[3]

[1]erik.sandstrom@zoho.com Stockholm (Sweden)
[2] ulf.sandstrom@indek.kth.se KTH Royal Inst Technol, Stockholm (Sweden)
[3]p.a.a.vanden.besselaar@vu.nl Vrije Universiteit, Amsterdam (the Netherlands)

Abstract

Any type of scientific study or evaluation of research quality and impact enters into two types of problems if there is more than one topic area involved in the study: (1) How to account for differences in (paper) production? (2) How to account for differences in citation impact, i.e. influence over subsequent literature? This paper aims to show that these questions can be answered with the help of two methods; the Field Adjusted Production (FAP) indicator and a percentile indicator which is designed to include the FAP. Consequently, they are used in combination in order to express a score that includes both paper production an impact into one figure. Thereby is constructed a score that can be used for ranking of universities, departments, individuals. The paper first explains the background of the method, and then how to calculate the indicators belonging to the P-Model. Then the paper indicates some examples and will discuss methods for validation of the proposed indicator.

Introduction - a long discussion

Performance indicators seem to be subject of continued interest for the bibliometric community. After a period of 10-15 years doing good work with quite some interest from the professional society of research administrators, the fall of the crown indicator has stimulated renewed interest and a critical discussion of bibliometric indicators. Suddenly it was (re)-discovered that mean values was not the best way to handle bibliometric data. Re-discovered because it was already indicated a long time ago by Paul R McAllister, Francis Narin and James G Corrigan in a paper with the title, "Programmatic Evaluation and Comparison Based on Standardized Citation Scores" (IEEE Transactions on Engineering Management 1983)[1].
Their indicator uses a transformation to the logarithm of the number of citations (plus one-half to include the zero-cited papers) and measures this in standard deviations from the mean. This means that many of the ingredients of the indicator discussion that has taken place since Lundberg (2007) already have been available a good time before the Leiden and Leuven indicators were created and launched in Europe (Moed & van Raan 1988; Schubert, Glänzel & Braun 1988). This cultural divide seems even more idiosyncratic as the National Science Indicators in the US have used percentiles since long time ago, but in Europe, they didn't gain interest as the so-called *crown indicator* had such a strong market position. It took a situation where the trust or confidence in indicators had reached its bottom (Wilsdon et al. 2015, Hicks et al 2015) before the percentiles came into the European discussion. After that first wave of the European discussion pioneered by Leydesdorff and Bornmann (in several papers) it is now time to start building stronger and more sustainable indicators.

This paper presents a composite indicator called the P-model which combines production *and* impact into one score and is size-dependent to its nature. The act of combining papers and citations has been done before e.g by the Leiden group (P time MNSC) but with important problems on both sides of the multiplication. Here we suggest solutions to these problems.
We will evaluate the P-model using criteria suggested by Yves Gingras in the edited book *Beyond Bibliometrics* (ed. Sugimoto 2014): 1) Adequacy; 2) Sensitivity and 3) Homogeneity.

[1] The paper is accessed at <https://www.forskningspolitik.se/files/dokument/programmatic-evaluation-and-comparison-based-on-standardized-citation-scores.pdf>

Field Adjusted Production – Waring distributions

Field differences in production are well known; medical researchers tend to produce more, often shorter papers where methodology and prior knowledge is codified in citations. Engineering scientists are known to produce less frequently and have fewer cross-references (Narin and Hamilton, 1996; Glänzel, 1996). These field differences affect both citation rates and the number of papers per author, differences that are to some extent explained by the shifting coverage of publication activity in the WoS database.

Let us say that we want to stay with the WoS database due to its good features: selection of sources, prudence, etc. How do solve the problems? In order to compute a field adjusted factor, we have to get rid of certain obstacles: publication databases give information on the authors that are active during a given period, not all the potential authors. As the non-contributors (non-publishing authors) are unknown it is difficult to calculate an average publication rate per author taking all potential authors into account. But, there is a proposed mathematical solution to this problem: bibliometric data are characteristical "Waring distributions" (Schubert and Glänzel, 1984). Using information on the distribution of author publication frequencies an estimate of the average publication rate per researchers (contributors and non-contributors) in a given field and country can be computed (Telcs, Glänzel & Schubert, 1985).

The approach is based on mathematical statistics and a theoretical discussion can be found in papers by Braun, Glänzel, Schubert & Telcs during the second half of the 1980s. Inspired by Irwin (1963) they showed that bibliometric material had the properties of "Waring distributions". A straight line should be obtained by plotting the truncated sample mean of these distributions (Telcs, Glänzel & Schubert, 1985). By extrapolating this series to Origo, the numbers of non-contributors are included. The intercept of this line is the average productivity of all potential authors during a given period of time (Braun, Glänzel & Schubert, 1990). In our model, this value is used as a reference value and is computed per field for Nordic data. Several successful empirical tests using the Field Adjusted Production (FAP) model have been implemented (e.g. Schubert and Glänzel 1984; Schubert and Telcs, 1986; Buxenbaum, Pivinski & Ruberg, 1987; Schubert and Telcs, 1989; Sandström and Sandström, 2008b). A more complete article on this method was published by Koski, Sandström & Sandström (2016). Here we follow that latter source for the explication of the method.
The Field Adjusted Production is calculated as follows:

$$\sum_{i=1}^{n} \frac{P_i}{r_i}$$

where P_i is the number of papers in field i and r_i is the (estimated) average number of papers per researcher in field $_i$. The estimation of the reference values is performed for each field by first calculating the s-truncated sample mean of each field as follows:

$$\frac{\sum_{i=s}^{\infty} i n_i}{\sum_{i=s}^{\infty} n_i}$$

Where n_i is the number of authors having exactly i papers. The truncated sample means are plotted versus s and the intercept of the fitted line, using weighted least squares linear regression, is used as an estimate the number of papers per author for the entire population The regression is weighted using proposed method for that by Telcs et al. (1985).

When applying this model, authors with an address at Nordic universities are used as data. Homonyms and similar problems are taken care of by automatic procedures in combination with manual procedures. This was done for all Nordic universities (Sweden, Finland, Denmark, and Norway) and the operation yielded almost 400,000 unique authors for the period 2008–2011.

Field delineation is maybe the most important issue here. The Thomson/ISI subject categories are used for citations, but these some 260 categories create too small samples when Nordic authors are used to constructing the productivity data. There are several alternative ways of producing macro classes (e.g. the Clarivate ESI 22 field categories). We have been using journal inter-citations as proximity values (Boyack and Klavans, 2006), and with the least frequent relation as decisive in order to distinguish, as far as possible, between basic and applied sciences. It has been shown by Rinia, van Leeuwen, Bruins, van Vuren and van Raan (2002) that applied sciences tend to cite back to more basic sciences, not the other way around. But the clustering procedures that were tried didn't really work as good as we wanted and therefore we decided, after some reiterations, that the suggested macro fields in the Science Metrix classification would fulfill the requirements we knew where needed, e.g. to distinguish a category of applied science fields. So, we had in the final round five different clusters (fields); humanities, social science and economics, applied sciences, health sciences, and natural sciences.

The methodology described was used to establish a reference value based on disambiguated researchers from all Nordic universities. By using the count of paper fractions per author and relate that to the reference value (the field factor) we obtain the relative quantity of production performed by the person or the unit (the indicator is further explained below, see Table 2). This indicator is called the "Field Adjusted Production (FAP)" and can be explained as the expected production per area over a period of time (in this case a four year period) and for a "normal" researcher with all other assignments at the same time. One can say that the indicator expresses how many persons the actual production score accounts for, if the value is ten for a group of people, then that can be related to the actual number of people in the group. So, if they are five and they publish in the range of ten persons then the production is 100 % higher than what would be expected.

Citation impact – percentile distributions

The literature on citation impact is wider and more diverse than the one on productivity but much of it is somewhat dated and irrelevant (Waltman 2016; Abramo 2018). There are three major questions that we will touch upon before we present the methods that were applied in this project. 1) What do we mean with the term "citation impact"; 2) Percentiles instead of averages; and 3) Size-dependent vs. size independent indicators.

The discussion on citation impact from research has intensified and has been widened over that last ten years (Bastow et al. 2014). Opening the concept of impact to all types of influence on society has many advantages in the dialogue with politics and funding agencies but at the same time the concept a bit vague. Therefore, it should be possible to talk about two different concepts of impact, the first one is the restricted impact and the second one is the wider and looser concept of impact. Depending on the fact that this exercise is a quantitative study of the relation between gender diversity and research performance we lean towards the first version of the concept of impact which is neatly laid out by Abramo (2018).

In the understanding of Abramo (2018) a paper might have an impact on the subsequent literature and for this, he reserves the concept "citation impact". It follows that we can use a very precise measure based on how articles are cited even if it should be considered as a proxy as the reference behavior is an non-harmonized process, many different types of behavior are detectable, but in the long run and with large stocks of papers there should be possible to use statistical methods that do not suffer from the noise in the data.

Calculation methods built on averages are of less interest as we can easily understand that there are drawbacks with methods that measure impact as a mean of all papers over a period of time. When the same author publishes another paper, the first papers' impact does not disappear or diminish. On the contrary, it can be made stronger by new evidence. Therefore, the overall impact of the two papers cannot be measured by an average, and instead, an additive method is required. In order to proceed with that method, we apply the FAP score introduced in the former section and illustrated in detail below.

Instead of averages, the method for performance analysis is partly based on a percentile approach. All articles in each group of articles are ranked based on citations. The field is defined according to the subject categories specified in the Web of Science database, and the articles are divided into percentile classes, the top 1% (99th percentile), 5 %, 10 %, 25 %, 50 % and below 50 %. Measures based on percentiles have the advantage of not being affected by causes of bias in citation distributions (Rousseau 2005). In certain disciplinary areas, a few publications with very numerous citations otherwise boost the mean, which can result in 70% of the articles in the area being below this mean (c.f. Campbell 2017 [STI 2017 paper], c.f. Thelwall 2019).

With this, we turn to the Percentile Model and how it allocates points for each article. The points are based on probability. An article that is among the most highly cited 1% of articles is assigned 100 points; one in the top 5 % is given 20 points and so forth (see Table 1). An article that is among the 50 % least cited is given 1 point, which means that a researcher can never lose from getting an article published. The points thus received by each article are then corrected by the field-adjusted production (FAP) method to compensate for differences between research areas in the rate of scholarly production. Such an approach provides a lot of information and should be useful to summarise performance in a single value. The method is preliminary called P-model or the Influence Factor.

Table 1. Points allocated per percentile group in the P-model

Percentile group	Points
TOP1 % (99th percentile)	100
TOP5 %	20
TOP10 %	10
TOP25 %	4
TOP50 %	2
TOP100 %	1

Note: Based on Sandström & Wold (2015)

The idea to allocate points is inspired by Leydesdorff (2012) and Leydesdorff & Bornmann, (2011). They suggested the following: *"(...a method to) calculate a mean of the ranks weighted by the proportion of papers in each. The minimum is 1, if all papers are in the lowest rank; the maximum is 6 if they are all in the top percentile."* Although of interest this method is dubious as the groups are not of the same size.

One major problem with the points in the P-model (see Table 1) is that there is quite a large difference between top1% and top2% which has been pointed out as reactions to evaluations based on this model (Henreksson, personal communication). Pragmatic reasons are to a large extent behind the model as it is dependent on the programming of the indicators and the calculations would be too extensive for a normal personal computer of today.

The model ingredients in detail

Now we have the two components of the model and in the following, we will go through it in detail so that if the reader would have the two different basic calculations done (FAP and P-model) he/she should be able to finalize it to a score. Here is the method for calculation and the text refers to Table 2 below:

(1) Each publication is from a source (SO), the periodical (journal) name.

(2) REF stands for reference value based on Nordic values; the first line is about 0.86 which means that an article by one author alone would account for more than what would be expected from one author in the Nordic countries, 1.0 is the expected value.

(3) Then Frac P showing the fraction of a paper or how many authors were involved in the production, in this case, four authors.

(4) This is transformed to a FAP value of 0,29 (i.e. Frac P/REF). This indicates that over a four-year period an author is expected to publish about four such papers.

(5) Then follows six different columns giving information on the percentile fraction, a fractionalization based on the fact that we are working with integers and therefore there can be many with a same number of citations at the border of a percentile group. An elegant solution to this is fractional counting suggested by Waltman and Schreiber (2013).

(6) Based on that there are six FTOP columns which give the product of the calculation [FAP*(Fraction) Percentile Group]*P-Model points(100; 20; 10; 4; 2; 1).

(7) Total sum in the column to the right gives the total per article. When totals are calculated they can be summarised per person or per research team or any unit of interest.

Table 2. Example showing the calculation of the Percentile Model (P-Model) indicator

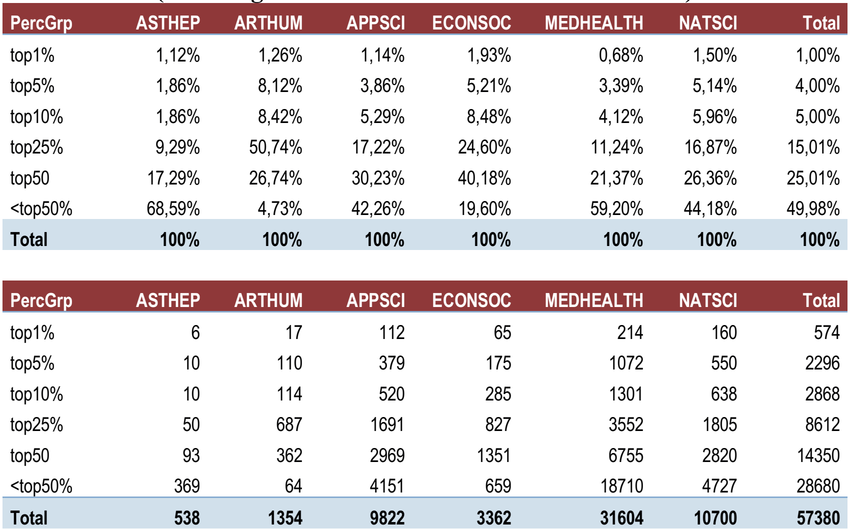
SO	REF	Frac P	FAP	TOP1	TOP5	TOP10	TOP25	TOP50	TOP100	FTOP1	FTOP5	FTOP10	FTOP25	FTOP50	FTOP100	P-Model points
LIGHTING RESEARCH & TECHNOLOGY	0,862958833	0,25	0,289700957	0	0,8	0	0	0	0	0	4,6352	0	0	0	0	4,635215314
JOURNAL OF CRYSTAL GROWTH	1,255448818	0,25	0,199131973	0	0	0,6667	0	0	0	0	0	1,32755	0	0	0	1,327547149
JOURNAL OF BIOSOCIAL SCIENCE	0,697673	0,333333333	0,477778749	0	0	0	1	0	0	0	0	0	1,91111	0	0	1,911114998
JOURNAL OF PROSTHODONTICS	1,171782937	0,333333333	0,284466792	0	0	0	0	0	1	0	0	0	0	0	0,284467	0,284466792

It should be remembered that we adhere to the principles of normalization to the field as has been practiced by the Leuven group (Glänzel et al. 1988) and implemented by the Leiden group (Moed & van Raan 1988). This is a central feature of bibliometric work: all types of performance, e.g. citation performance, are relative to the field where the object of evaluation has its publications. By publishing in a specific type of journals, authors tell the community of scientist that they want to be evaluated and measured by the standards in each subcategory of the fields available. There is some criticism towards the subject categories developed by Web of Science (see Leydesdorff 2008) or Scopus for that matter. But, the critique often fails to understand that a subject category are far from one-dimensional, instead, they include multi-assignations of each journal and it is, therefore, correct to say that there are thousands of categories in the WoS due to the multi-assignation methodology.

All calculations used here are based on three databases (SCI-E, SSCI, A&HCI) and four document categories only: Articles, Letters, Proceeding Papers, and Reviews. No other document categories are involved in the calculation of citation scores or the calculation of percentile groups. Author-based self-citations are deleted when the citation scores are calculated (based on the first author name).

Examples from the Swedish database showing how different areas are represented at every level of performance and that the indicator fulfils the criteria of equality between areas. But, there are obvious problems due to the differences between areas, e.g. Medical science are heavy in the bottom and social science is top heavy due to many low fraction authors in the former and full fraction authors in the latter domain.

Table 3. Fields distribution over Percentile Groups (disambiguated Swedish researchers - 2012-2015)



PercGrp	ASTHEP	ARTHUM	APPSCI	ECONSOC	MEDHEALTH	NATSCI	Total
top1%	1,12%	1,26%	1,14%	1,93%	0,68%	1,50%	1,00%
top5%	1,86%	8,12%	3,86%	5,21%	3,39%	5,14%	4,00%
top10%	1,86%	8,42%	5,29%	8,48%	4,12%	5,96%	5,00%
top25%	9,29%	50,74%	17,22%	24,60%	11,24%	16,87%	15,01%
top50	17,29%	26,74%	30,23%	40,18%	21,37%	26,36%	25,01%
<top50%	68,59%	4,73%	42,26%	19,60%	59,20%	44,18%	49,98%
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

PercGrp	ASTHEP	ARTHUM	APPSCI	ECONSOC	MEDHEALTH	NATSCI	Total
top1%	6	17	112	65	214	160	574
top5%	10	110	379	175	1072	550	2296
top10%	10	114	520	285	1301	638	2868
top25%	50	687	1691	827	3552	1805	8612
top50	93	362	2969	1351	6755	2820	14350
<top50%	369	64	4151	659	18710	4727	28680
Total	**538**	**1354**	**9822**	**3362**	**31604**	**10700**	**57380**

Note: ASTHEP is Astronomy & High Energy Physics; ARTHUM is Humanities; APPSCI is Applied Sciences; ECONSOC is Social Sciences; MEDHEALTH is Medical Sciences, and NATSCI is Natural Sciences. Upper table shows relative frequencies and lower table show raw numbers.

Further work

An important issue is how to validate the approach. Relevant validation criteria are:

1) Adequacy
2) Sensitivity
3) Homogeneity

These will be discussed in the presentation.

Inventor Turnover and Knowledge Transfer: The Case of Wind Power Industry

Chun-Chieh Wang[1,3] and Dar-Zen Chen[2,3]

[1] *wangcc@ntu.edu.tw*
Dept. of Bio-Industry Communication and Development, National Taiwan University, Taipei, Taiwan R.O.C.

[2] *Corresponding Author: dzchen@ntu.edu.tw*
Dept. of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan R.O.C.

[3] Center for Research in Econometric Theory and Applications, National Taiwan University, Taipei, Taiwan R.O.C.

Abstract

Despite various studies regarding both patentometrics and innovation performance of companies, there are still a lack of discussion for inventor turnover and knowledge transfer among companies. The purpose of this study is to quantitatively analysis the knowledge transferred through inventor turnover. Main companies in the wind power industry were observed to measure inventors transferring among them. A total sample of 32 companies with 3,242 patents and 2,497 inventors were collected to measure the patterns of turnover-inventor. Results show that turnover-inventors invent relatively more patents per inventor and their patents are also with higher average patent cited count. It means that turnover-inventors have more technical knowledge than those remain-inventors, and their turnover activity would leads to the technical knowledge transfer. Turnover patterns in major companies show that many companies with lower transfer-in rate no matter their transfer-out rate higher or not, and they prefer R&D in exploitative innovation. Patents invented by transfer-out inventors disperse company's patent portfolio, thus after these inventor transfer out, the company would become more concentrated on their patent portfolio. This is a prime study that tries to connect the link inventor turnover and knowledge transfer.

Introduction

Turnover is an important topic in the business management due to the employee turnover could affect a company's performance (Hinkin & Tracey, 2000). Prior Research found that employee turnover negatively impacts organizations due to the loss of production-oriented knowledge (Shaw, Duffy, Johnson, & Lockhart, 2005). Companies facing such problems and studies how to reduce turnover are received much attention (Griffeth, Hom, & Gaertner, 2000). To realize what reason causes individuals to turnover could help companies to response this challenge and reduce the negative effect (Felps, Mitchell, Hekman, Lee, Holtom, & Harman, 2009; Mossholder, Settoon, & Henagan, 2005).

Patent document is an informational resource for studies about research and development (R&D) in company. Based on the patent document, the details of individual inventor turnover among companies could be observed easily. The main purpose of this study is to investigate the relationship between inventor turnover and knowledge transfer. Firstly, the patent performance of turnover and remain inventors will be observed to clarify the different performance between them. Secondly, turnover patterns in major companies will be analysed to classify various turnover types of them. Finally, the effect of turnover-inventor in companies' patent portfolio will be measured to illustrate the important actors of turnover-inventor for R&D management.

Turnover and Knowledge Transfer

Knowledge could be separated as explicit and tacit components, where the former refers to knowledge that can be codified and is represented in processes, procedures, writings, and drawings (Nonaka and Von Krogh, 2009), and the latter refers to knowledge that is difficult to articulate and is embedded within skilful actions, routines, values, and beliefs (Nonaka and Von

Krogh, 2009). The tacit elements of production knowledge are particular concern with employee turnover because such knowledge is vital to the performance of production-oriented tasks (Michele Kacmar, Andrews, Van Rooy, Chris Steilberg, & Cerrone, 2006), and transferring tacit knowledge to existing or new employees prior to an employee's departure is costly or, in many cases, infeasible (Van Wijk, Jansen, & Lyles, 2008). While tacit knowledge can be effectively transferred through apprentice and other mentoring approaches (Nonaka, 1994), such mechanisms are time-consuming and quickly become impractical at higher levels of employee turnover. Furthermore, employees may be unwilling or unable to appropriate such knowledge prior to departure. As a result, one would expect that substantial amounts of critical knowledge could be lost when organizations experience high levels of turnover and that the loss of such knowledge is likely to have substantial negative implications for the efficiency and subsequent performance of the company (Eckardt, Skaggs, & Youndt, 2014).

Turnover Impacts Organizational Innovation

Referring the model proposed by Guidice, Thompson Heames and Wang (2009), the research model in this study is drafted as Figure 1. The organizational innovation would be effected by inventor turnover and the innovative performance of turnover-inventors as the moderator. The inventor turnover could be classified into functional and Dysfunctional turnover. Abelson and Baysinger (1984) stated that companies have an "optimal" level of turnover which "minimizes the sum of the costs of turnover plus the costs associated with reducing it." Dess and Shaw (2001) argued that turnover rates are dysfunctional (a condition of malfunction) when they detract from organizational effectiveness. Dysfunctional turnover as occurring when an organization fails to retain a significant portion of its critical knowledge workers or when there is little, if any, churn in an organization's pool of knowledge workers. In contrast, functional turnover exists when an organization achieves a moderate rate of knowledge worker turnover (Guidice, Thompson Heames and Wang, 2009).

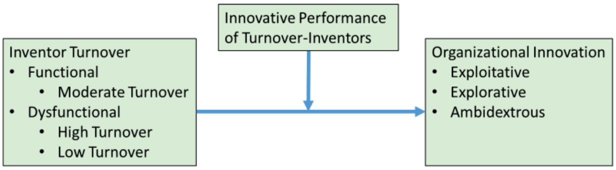


Figure 1. Research Model in This Study

Organizational innovation are separated into three types: exploitative, explorative, and ambidextrous. Exploitative innovation is sequential and incremental, and thus does not fall outside traditional thought processes or policies of the organization (Slater and Narver, 1995). With focus resting on making core competencies and processes more efficient, exploitative innovation results in doing existing things better (Hayes and Allinson, 1998). In contrast, explorative innovation requires searching for new possibilities rather than exploiting old certainties (March, 1991). Explorative innovation results in doing things differently or doing different things (Hayes and Allinson, 1998).

The moderator, innovative performance of turnover-inventors, includes not only the quantity and quality patent performance of turnover-inventors but the betweenness centrality of turnover-inventors in the inventor citation network. Betweenness centrality moderates the relationship between functional turnover and ambidextrous innovation such that the positive relationship is strongest when a greater proportion are positioned within network clusters as structural holes (Guidice, Thompson Heames and Wang, 2009).

Methodology

Patent data was collected from United States Patent and Trademark Office (USPTO). Patent bibliometrics was utilized to analysis the inventor turnover patterns. Betweenness centrality from Social Network Analysis was also utilized to measure the importance of inventors in the citation network. These research methodologies are described as follows.

Data Collection

In this study, patents issued by the USPTO between 2009 and 2018 were gathered for analysis. The patent data was collected based on the Cooperative Patent Classification (CPC) system. The CPC is the result of a partnership between the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO) in their joint effort to develop a common, internationally compatible classification system for technical documents, in particular patent publications, which are used by both offices in the patent granting process (European Patent Office and United States Patent and Trademark Office, 2019). Patents related to wind power technologies are categorised under the CPC code: Y02E 10/70 (European Patent Office and United States Patent and Trademark Office, 2019b). Artificial check for the precision of wind power technology was processed after patent database quarried.

Patent Bibliometrics

Patent bibliometrics (or patentometrics) is a theoretical method using mathematics, statistics and logic. It studies and analyses the quantity, quality and application of patent literature, e.g. patent counts and patent citations (Narin, 1994). Patent bibliometrics can be used to understand the development of patented technologies (including that of individuals, organisations and countries). What is more, it can be used to study the links between researchers, organisations and countries through the relationship of patent citations.

Patent analysis calculates the patent counts and the frequency distribution of patents based on the selected units of analysis (e.g. country, company/organisation, inventor and technology field), and can be used to identify the major activities of the selected units. Patent citation analysis focuses on the references (including patent and non-patent references) cited in the patent specification. As a result, potential links can be explored through patent citation counts and citation relations. This study mainly focuses on the turnover and knowledge transfer in the wind power industry using patent bibliometrics.

Betweenness Centrality

Social network analysis (SNA) is a quantitative technique based on graph theories in mathematics. The network constitutes nodes and lines that connect the nodes. Nodes can be individual actors, groups of people, events or organisations. Lines between nodes can be used to indicate the existence of the relationships as well as the direction, strength, content and formats of the relationships. Quantitative indicators can be used to analyse the relationship, length and density of the lines between nodes (Freeman, 1979, 1991).

Betweenness centrality is to measure the total number of links between an actor and other actors. When the centrality of an actor is high, it means that the actor plays a vital role as an information channel in the overall network, has advantages of dominating the network, and shows more innovation in outputs and a higher speed of product development when compared with other actors (Ahuja, 2000; Deeds & Hill, 1996). Brandes (2001) proposed compactly betweenness centrality as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where σ_{st} is total number of shortest paths from node s to node t and σ_{st} (v) is the number of those paths that pass through v. The betweenness may be normalised by dividing through the number of pairs of vertices not including v, which for directed graphs is $(n\text{-}1)(n\text{-}2)$ and for undirected graphs is $(n\text{-}1)(n\text{-}2)/2$.

Results

Thirty two companies granted 3,242 patents, shared 50.4% in total 6,434 patent, and 2497 of 40.3% in 6194 inventors, are selected as the analysis samples in this study.

Patents Invented by Turnover and Remain Inventors

The patent granted in wind power industry are categorized into two types: patent invented by turnover inventors and patent invented by remain inventors. The statistics are shown in Table 1. Based on this table, there are only 20.7% patents invented by turnover inventors, and these turnover inventors only share 5.6% of total inventors. Although there are only 5.6% inventors ever transfer among companies, but they invent more patents with higher average patent cited count than those patents invented by remain inventors. Besides, these turnover inventors also get higher average normalized betweenness in the inventors' citation network. Thus, turnover inventors not only invent more patents but their patents are with higher quality. It means that the turnover inventors are with more technical knowledge and their turnover among companies also accompany with technical knowledge transfer.

Table 1. Patents Categorized into Turnover and Remain Inventors

	Turnover Inventors	*Remain Inventors*	***Total***
Patent Count	672 (20.7%)	2,570(79.3%)	**3,242**
Inventor Count	140(5.6%)	2,357(94.4%)	**2,497**
Average Patent Count per Inventor	4.80	1.09	**1.30**
Average Patent Cited Count	3.87	3.56	**3.63**
Average nBetweenness in Citation Network	1.039	.819	**.874**

Turnover Patterns in Major Companies

Thirty two companies granted more than ten patents are selected as major companies in this industry to analysis the turnover patterns, and the result are shown in Table 2. Transfer-out inventors are those inventors appear in latter patents filed by other assignee, and transfer-in inventors are those inventors appeared in prior patents filed by other assignee. In Table 2, we can observe some company are with moderate and balance turnover rate: Vestas, LM Wind Power, and Robert Bosch, and these companies could be classified as with functional innovation. Companies also with functional innovation are those companies with higher transfer-in rate than transfer-out rate, for example: Siemens, Senvion, Alstom, Wobben Properties, Envision Energy, Hamilton Sundstrand, XEMC Darwind, Blade Dynamics, and Lockheed Martin.

Some companies are with dysfunctional innovation, their net turnover rate (transfer-out rate minus transfer-in rate) higher than 10%, for example: Wilic s.ar.l., Northern Power Systems, SSB Wind Systems, Frontier Wind, Genedics Clean Energy, FloDesign Wind Turbine, and Modular Wind Energy. Other companies with dysfunctional innovation are those with none transfer-in inventors or lower turnover rate, fewer knowledge interchange with others, for example: GE, Mitsubishi Electric, Nordex Energy, Hitachi, Google, Acciona Energy, ABB Group, Delta Electronics, Boeing, Airbus, Sony, AMSC Windtec, and Moog.

The R&D styles of company could be classified as exploitative, explorative, and ambidextrous innovation based on their turnover patterns. Companies with exploitative innovation are those ones with lower turnover rate or only higher transfer-out rate: GE, Vestas, Mitsubishi Electric, Nordex Energy, Hitachi, Google, Acciona Energy, ABB Group, Delta Electronics, Northern Power Systems, Boeing, SSB Wind Systems, Genedics Clean Energy, Airbus, FloDesign Wind Turbine, Sony, Modular Wind Energy, AMSC Windtec, and Moog; Companies with explorative innovation are those ones with higher transfer-in and transfer-out rates: LM Wind Power, Robert Bosch, Hamilton Sundstrand, Wilic s.ar.l., Frontier Wind, and Blade Dynamics; Companies with ambidextrous innovation are those ones with higher transfer-in rate but lower transfer-out rate: Siemens, Senvion, Alstom, Wobben Properties, Envision Energy, XEMC Darwind, and Lockheed Martin.

Table 2. Turnover patterns in Major Companies

	Total		*Transfer-out Inventor*		*Transfer-in Inventor*	
Assignee	*Patent*	*Inventor*	*Rel_Patent (%)*	*Inventor (%)*	*Rel_Patent (%)*	*Inventor (%)*
GE	984	838	49(5.0)	16(1.9)	7(0.7)	10(1.2)
Vestas	573	387	48(8.4)	19(4.9)	32(5.6)	12(3.1)
Siemens	480	310	6(1.3)	1(0.3)	112(23.3)	27(8.7)
Mitsubishi Electric	203	131	82(40.4)	9(6.9)	8(3.9)	3(2.3)
Senvion	126	79	0(0.0)	0(0.0)	12(9.5)	6(7.6)
LM Wind Power	103	75	20(19.4)	4(5.3)	19(18.4)	3(4.0)
Alstom	88	62	0(0.0)	0(0.0)	9(10.2)	4(6.5)
Nordex Energy	87	66	2(2.3)	2(3.0)	0(0.0)	0(0.0)
Wobben Properties	69	58	3(4.3)	1(1.7)	32(46.4)	8(13.8)
Hitachi	64	93	0(0.0)	0(0.0)	1(1.6)	3(3.2)
Google	58	25	0(0.0)	0(0.0)	3(5.2)	1(4.0)
Robert Bosch	45	60	3(6.7)	2(3.3)	3(6.7)	2(3.3)
Acciona Energy	39	44	0(0.0)	0(0.0)	0(0.0)	0(0.0)
Envision Energy	33	14	0(0.0)	0(0.0)	18(54.5)	7(50.0)
Hamilton Sundstrand	32	22	10(31.3)	3(13.6)	10(31.3)	3(13.6)
ABB Group	30	47	1(3.3)	1(2.1)	0(0.0)	0(0.0)
Delta Electronics	25	39	0(0.0)	0(0.0)	3(12.0)	1(2.6)
Wilic s.ar.l.	20	17	10(50.0)	7(41.2)	2(10.0)	2(11.8)
Northern Power Systems	20	18	6(30.0)	3(16.7)	0(0.0)	0(0.0)
Boeing	20	31	0(0.0)	0(0.0)	1(5.0)	1(3.2)
SSB Wind Systems	19	10	2(10.5)	1(10.0)	0(0.0)	0(0.0)
Frontier Wind	18	16	9(50.0)	3(18.8)	3(16.7)	1(6.3)
XEMC Darwind	15	11	0(0.0)	0(0.0)	7(46.7)	1(9.1)
Blade Dynamics	13	5	6(46.2)	2(40.0)	9(69.2)	2(40.0)
Genedics Clean Energy	13	2	13(100)	2(100.0)	0(0.0)	0(0.0)
Airbus	13	29	0(0.0)	0(0.0)	1(7.7)	1(3.4)
FloDesign Wind Turbine	12	12	12(100)	3(25.0)	0(0.0)	0(0.0)
Sony	12	15	1(8.3)	1(6.7)	0(0.0)	0(0.0)
Modular Wind Energy	11	5	10(90.9)	4(80.0)	0(0.0)	0(0.0)
Lockheed Martin	11	16	0(0.0)	0(0.0)	1(9.1)	1(6.3)
AMSC Windtec	10	6	0(0.0)	0(0.0)	0(0.0)	0(0.0)
Moog	10	16	1(10.0)	1(6.3)	0(0.0)	0(0.0)
Sum	**3,242**	**2,497**	**294(9.1)**	**85(3.4)**	**293(9.0)**	**99(4.0)**

The inventor turnover among major companies are illustrated as Figure2. GE, Vestas, and Siemens are the top three companies with most patents, and they also are the main hubs in which many inventors transfer to. Siemens receives most turnover-inventors transfer from other companies, but GE and Vestas get more inventors transfer out from themselves.

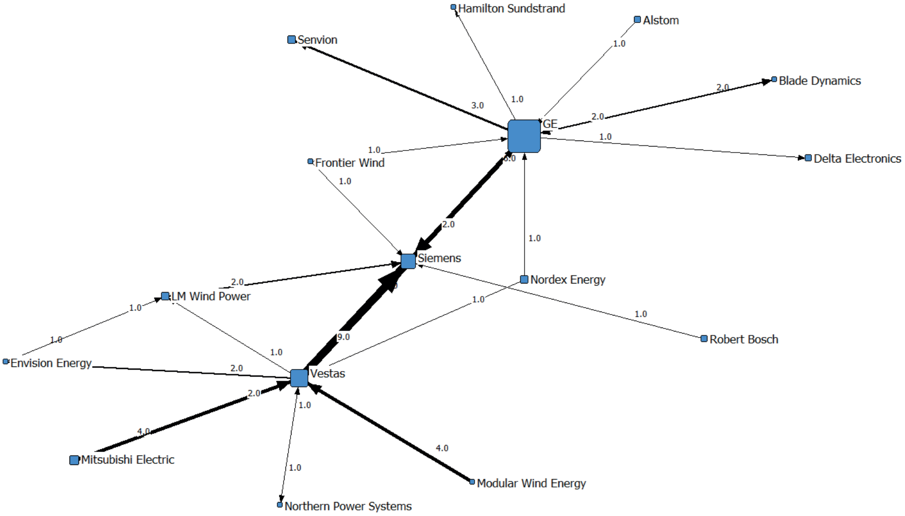


Figure 2. Inventors Turnover Network in Major Companies

The Effect of Turnover-Inventor in Company's Patent Portfolio

Based on Table 1, we can observed that turnover-inventors invent more patents with higher quality. Thus, the turnover-inventor would be the important human capital for companies, and their effect in the innovation performance could be analysed by the change of company's patent portfolio. The Herfindahl-Hirschman index (HHI) is a measure of market concentration. It is calculated by squaring the market share of each firm competing in a market and then summing the resulting numbers. In this study, the HHI is utilized to analysis the technology concentration of companies. The primary International Patent Classification (IPC) code is selected to identify the technology field of each patent. Hamilton Sundstrand, Blade Dynamics, and Wilic s.ar.l. are the three companies selected as the samples to observe their patent portfolio effected by the turnover-inventor, and the results are shown in Table 3.

Hamilton Sundstrand gets 0.24 in the HHI of total patents. While the patents invented by transfer-out inventors excluded, the HHI is increasing 22.13%. But the HHI is decreasing 7.12% while the patents invented by transfer-in inventors excluded. It means that the transfer-out inventors disperse company's patent portfolio but those transfer-in inventors concentrate company's patent portfolio. These turnover patterns would make Hamilton Sundstrand becoming more focus on the developments of B64D, F01D, and F03D technology fields.

Blade Dynamics gets 0.349 in the HHI of total patents. While the patents invented by transfer-out inventors excluded, the HHI is increasing 116.29%. And the HHI is also increasing 7.42% while the patents invented by transfer-in inventors excluded. It means that the transfer-out and transfer-in inventors both disperse company's patent portfolio. These turnover patterns would make Blade Dynamics becoming more focus on the developments of the F03D technology field.

Wilic s.ar.l. gets 0.455 in the HHI of total patents. While the patents invented by transfer-out inventors excluded, the HHI is increasing 45.05%. But the HHI is decreasing 9.1% while the patents invented by transfer-in inventors excluded. It means that the transfer-out inventors disperse company's patent portfolio but those transfer-in inventors concentrate company's patent portfolio. These turnover patterns would make Wilic s.ar.l. becoming more focus on the developments of the F03D technology field.

Table 3. The Effect of Turnover-Inventor in Companies' Patent Portfolio

		Patents Excluded	
	Total Patent	*Invented by Transfer-out Inventors*	*Invented by Transfer-in Inventors*
Hamilton Sundstrand			
B63H	1	0	1
B64D	11	9	7
F01D	7	4	4
F03D	8	6	6
F04B	1	0	1
H02K	3	3	2
H02P	1	0	1
HHI (Change Rate)	**.240**	**.293(22.13%)**	**.223(-7.12%)**
Blade Dynamics			
B63H	2	0	1
E04C	2	0	0
E04H	1	1	0
F01D	1	0	1
F03D	7	6	2
HHI (Change Rate)	**.349**	**.755(116.29%)**	**.375(7.42)**
Wilic s.ar.l.			
B63H	1	1	1
F03B	1	0	1
F03D	13	8	11
F16J	1	0	1
F24H	1	0	1
H02K	3	1	3
HHI (Change Rate)	**.455**	**.66(45.05%)**	**.414(-9.1%)**

Conclusions

Employee turnover has been an important issue for business management. This study tries to explore the patterns of inventor turnover among wind power companies. Totally 3,242 patents invented 2,497 inventors in 32 major companies are analysed to primer realize the effect of turnover-inventor to the patent portfolio of companies. According to the results, we can conclude three arguments as follows:

Turnover Inventors invent more patents with higher quality

Turnover-inventors invent relatively more patents per inventor and their patents are also with higher average patent cited count. This result shows that turnover-inventors have more technical knowledge than those remain-inventors, and their turnover activity would leads to the technical knowledge transfer. Besides, turnover-inventors also perform higher betweenness centrality in the inventor citation network. It means that turnover-inventors also perform key gates in the knowledge spillover path. Thus, obtaining external knowledge through transfer-in inventor has become an important issue in R&D management.

Most Major Companies prefer exploitative innovation

Turnover patterns in major companies show that 19 companies all with lower transfer-in rate no matter their transfer-out rate higher or not, and they prefer R&D in exploitative innovation. These companies prefer developing isolated technology, thus they need not external technical knowledge import. Although some companies are with lower transfer-in rate but higher transfer-out rate: Vestas, Mitsubishi Electric, Northern Power Systems, SSB Wind Systems, Genedics Clean Energy, FloDesign Wind Turbine, Sony, Modular Wind Energy, and Moog. These exploitatively innovative companies need not external technical knowledge import, but their technical knowledge are needed by others. It reflects to their higher transfer-out rate.

Patents from Transfer-out Inventors Disperse Company's Patent Portfolio

Based on the effect of turnover-inventor in three companies' patent portfolio, their HHI all increase while the patent invented by transfer-out inventors excluded. These patents invented by transfer-out inventors disperse company's patent portfolio. In other words, after these inventor transfer out, the company would become more concentrated on their patent portfolio. It implicates that while the company growing as a mature enterprise with clear R&D direction, other surrounding technical knowledge would be not essential. Those inventors have surrounding technical knowledge will become more easily to transfer out to other company.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 108L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. MOST 107-3017-F-002-004- and Grant No. MOST 107-2634-F-002-001-.

References

Abelson, M. A., & Baysinger, B. D. (1984). Optimal and dysfunctional turnover: Toward an organizational level model. *Academy of management Review, 9*(2), 331-341.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology, 25*(2), 163-177.

Dess, G. G., & Shaw, J. D. (2001). Voluntary turnover, social capital, and organizational performance. *Academy of management review, 26*(3), 446-456.

Eckardt, R., Skaggs, B. C., & Youndt, M. (2014). Turnover and knowledge loss: An examination of the differential impact of production manager and worker turnover in service and manufacturing firms. Journal of Management Studies, 51(7), 1025-1057.

European Patent Office and United States Patent and Trademark Office (2019a). *About CPC.* Retrieved January 20, 2019 from: https://www.cooperativepatentclassification.org/about.html.

European Patent Office and United States Patent and Trademark Office (2019b). *CPC Scheme and Definitions.* Retrieved January 20, 2019 from: https://www.cooperativepatentclassification.org/cpcSchemeAndDefinitions.html.

Felps, W., Mitchell, T. R., Hekman, D. R., Lee, T. W., Holtom, B. C., & Harman, W. S. (2009). Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting. *Academy of management journal, 52*(3), 545-561.

Freeman, L.C. (1979). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Freeman, L.C. (1991). Networks of innovators: A synthesis of research issues. *Research Policy, 20*(5), 499-514.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of management, 26*(3), 463-488.

Guidice, R. M., Thompson Heames, J., & Wang, S. (2009). The indirect relationship between organizational-level knowledge worker turnover and innovation: An integrated application of related literature. *The Learning Organization, 16*(2), 143-167.

Hayes, J., & Allinson, C. W. (1998). Cognitive style and the theory and practice of individual and collective learning in organizations. *Human relations, 51*(7), 847-871.

Hinkin, T. R., & Tracey, J. B. (2000). The cost of turnover. *Cornell Hospitality Quarterly, 41*(3), 14.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization science, 2*(1), 71-87.

Michele Kacmar, K., Andrews, M. C., Van Rooy, D. L., Chris Steilberg, R., & Cerrone, S. (2006). Sure everyone can be replaced… but at what cost? Turnover as a predictor of unit-level performance. Academy of Management journal, 49(1), 133-144.

Mossholder, K. W., Settoon, R. P., & Henagan, S. C. (2005). A relational perspective on turnover: Examining structural, attitudinal, and behavioral predictors. *Academy of management journal, 48*(4), 607-618.

Narin, F. (1994). Patent bibliometrics. *Scientometrics, 30*(1), 147-155.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization science, 5*(1), 14-37.

Nonaka, I. and Von Krogh, G. (2009). 'Tacit knowledge and knowledge conversion: controversy and advancement in organizational knowledge creation theory'. *Organization Science, 20*, 635–52.

Shaw, J. D., Duffy, M. K., Johnson, J. L., & Lockhart, D. E. (2005). Turnover, social capital losses, and performance. *Academy of management journal, 48*(4), 594-606.

Slater, S. F., & Narver, J. C. (1995). Market orientation and the learning organization. *Journal of marketing, 59*(3), 63-74.

Van Wijk, R., Jansen, J. J., & Lyles, M. A. (2008). Inter-and intra-organizational knowledge transfer: a meta-analytic review and assessment of its antecedents and consequences. Journal of management studies, 45(4), 830-853.

Why Sociologists Should Not Bother with Theory: The Effect of Topics on Citations

Radim Hladík[1, 2]

radim.hladik@fulbrightmail.org
[1] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, 100-0003 Tokyo (Japan)
[2] Czech Academy of Sciences, Institute of Philosophy, Jilska 1, 110 00 Praha (Czechia)

Abstract

This paper utilizes odds ratios of topical words to model cumulative citation counts in a corpus of Czech sociological articles. It builds atop a novel method for topic modeling based on hierarchical stochastic block models and notes why network based topic models may be a more productive approach if the output of topic models is expected to serve as an input to further statistical models. The demonstrated technique shows that topics of articles have significant effects on citations and possibly interact with other important variables, such as the gender or number of authors.

Introduction

Despite their many limitations in that role, citations remain an important indicator of research impact. One of the challenges when working with citation data is the need to account for different rates and patterns of referencing previous work in individual disciplines. Particular scientific fields vary not only by objective measures, such as the number of researchers working in a scientific field, but also by the less tractable, but possibly more important, effects of multiple citation practices and cultures. Scholarly work on citations that transcends disciplinary boundaries therefore needs to account for such differences. Luckily, disciplines are considerably stable social institutions (Abbott 2001) that can be usually identified with relative ease. Things get trickier when it comes to more subtle distinctions, such as research specialties or topics. Researchers have approximated these fine-grained intellectual structures and trends e.g. by linking specialty journals to subfields (Moody 2004) or by mapping co-authorship (Abbasi, Altmann, & Hossain 2011), co-citation, or co-words networks of people and documents (Guan, Yan, & Zhang 2017; Leydesdroff 1989). This kind of research typically has descriptive goals.

Automated topic modeling has been growing in popularity as another way to infer semantic structure of documents and its results have been applied to citations data (Mann, Mimno, & McCallum 2006) with the motivation to develop a model predictive of trending topics (Fu & Aliferis 2010). In this paper, we reverse the perspective and ask if the topical structure of a discipline can have effect on citation counts. If research topics would arise solely due to the cognitive demands, we should expect the scientific merit (or whichever quality citations actually reflect) to be distributed randomly within each and across all topics. However, it is also conceivable that topics can receive citation bonuses for various reasons. For example, some authors can be influential across the entire field and thus boost the importance of their topic relative to others. Other topics may become fashionable due to internal disciplinary discourse or through external intervention, whereby some topics receive more funding or public recognition and, in turn, raise the scholarly profile of the topic. Limited evidence suggests that topics can indeed be relevant factor for citations (Klavans & Boyack 2017); (Tahamtan, Afshar, & Ahamdzadeh 2016)

Sociology as a research object

Hicks (2004) dubbed sociology as "paradigmatic social science", because it encapsulates many divisive and centrifugal tendencies that reign in the social science domain, such as quantitative and qualitative approaches, insiders and outsiders communication of knowledge, or explanatory and interpretative paradigms. Moody (2004; Moody & Light 2006) examined the influence of sociological "specialty" areas on the odds of collaborative work in sociology. He found that, overall, specialties are important factor in determining co-authorship levels, although the effect differs across specialties. Ultimately, Moody, argues, we may be witnessing a divisive trend with two main group of specialties, which differ in their intensity of collaboration. The main driver of this trend seems to be the presence or lack of thereof of quantitative work. The complex and fragmented structure of sociology makes it a good candidate for investigating its topical structure.

Data

To examine the effect of sociological topics on citations, we use the available data on Czech sociology. A dataset from a small country has the advantage of serving as a comprehensive but manageable example of larger trends. The full-text data consists of sociological articles published between 1993–2016 in the Czech Sociological Review (CSR), which is a "core" and generalist journal of Czech sociology. Original scientific papers were identified in the sociological corpus using metadata on the articles' categories and manual inspection of attributes such as an existing abstract or a list of references. Ultimately, 522 articles were used for topic discovery. Citation data were retrieved from the Web of Science database. Of the 522 articles, 499 were successfully matched with their corresponding citation records based on the last names of first authors, year of publication, and the starting page of the article. The fulltexts were morphologically tagged and only nouns appearing more than 2 times were retained. Such reduction not only leads to more efficient computations, but also to better semantic coherence in topic modeling (Martin & Johnson 2015).

Methods

Computational social scientists and digital humanists have been adopting the method of automated topic analysis known as Latent Dirichlet Allocation (LDA) from the field of natural language processing and information retrieval where it was introduced by Blei et al. (Blei, Ng, & Jordan 2003). The advantage of the method is not only its capacity to process large number of texts, but also its ability to detect topics - defined as specific distribution of words in a corpus - which the researchers would not necessarily expect to find if they were only working with predefined topics as in the traditional content analysis. Additionally, while the method can be used for classification of documents (a document is assigned to the topic with the highest probability), it also provides descriptions of documents as mixtures of topics, which is a more realistic representation of any longer text.

In the current applications of topic modeling research, descriptive approaches prevail. Structural topic models (STM) allow for topics to correlate with document-level metadata (Roberts et al. 2014). This is a significant contribution to exploitation of topic models, but because STMs enter metadata into the topic modeling process, their results cannot be used for inferential purposes and should be treated only descriptively. STMs are designed for detecting topic based on metadata and not vice-versa.

Regardless of the particular variant of the LDA approach, their strength – representing documents as mixtures of multiple topics – is also their fundamental weakness in further applications. Because the topic loadings come from mixtures, a value assigned to any one topic is fully determined by other topics in the model. The normalized probabilistic distribution output of LDA-based model therefore makes them problematic for uses in linear modeling. Discarding some topics based on an arbitrary threshold as done by (Antons et al. Antons, Joshi, & Salge 2018) may allow a model to work computationally, because the introduced zeroes break the constraint, but such a hack does not conceptually transcend the essential problem of dealing with mixtures. Sound regression models for compositional data exist, but rely on transformations that do not make the already challenging interpretation any easier.

Leydesdorff and Nerghes (2017) showed that probabilistic topic modeling has superseded semantic networks as the method of choice for mapping the content of large corpora. They argue that semantic networks are a viable alternative to LDA. The attractive feature of semantic networks for our purposes is that they can be used to generate communities of words whose presence in a document can be simply counted. Semantic networks, however, are cumbersome when it comes to determining the number of communities (which requires manipulation of cut-off values) and the ranking of words in each community.

A recently published method TopSBM model based on hierarchical stochastic block models in bi-partite networks (Gerlach, Peixoto, & Altmann 2018) resolves the main conundrums of both LDA and co-word networks. For example, it does not require setting of hyperparameters. It shares mathematical properties with probabilistic topic modeling, but the mixtures it yields for both documents and topics are not based on Dirichlet priors. Instead, they are calculated from the number half-edges incident on the nodes of a bi-partite network made up separately of words (i.e. word-types) and documents. The only intervention required from the researcher is to choose a level of hierarchy from several options available in the output of the model.

We take advantage of this novel method, but modify it slightly to remove the per document constraint. We adapt the output of the published TopSBM model to obtain a list of words in topics and their counts in each document. This then enables us to calculate topic prevalence on the corpus level. We then construct a contingency table for each topic and document as two partially overlapping categories of the corpus. This allows us to express the keyness of topics in each documents as log odds ratios. Consequently, we are able to treat each topic as a fully independent covariate in further analytical steps. Conceptually, we treat topics as observed variables on the corpus level and each document is a particular instantiation of the each topic. This contrasts with LDA that perceives topics as latent variables in each document.

By eschewing a probabilistic representation of topical structure of documents, we are able to treat each topic as a feature independent of others at the document level. In TopSBM, the composition of topics is only constrained at the corpus level. Once we obtain our measurement of the log odds ratios for each topic in each document, we have a range of possibilities for employing topics as features in for statistical modeling. Since our task is to examine the effects of topics on citations count, which is an over-dispersed variable in our dataset, we choose a model for negative binomial regression. In the lack of consensual alternatives, this is a common choice for modeling citation data (Thelwall & Wilson 2014), (Ajiferuke & Famoye 2015). Control variables include binary author-level data: single vs. team authorship, man vs. woman lead author, the age of publication, and lexical variety. Document length is not included, because variance inflation factor suggested that the information about document length was sufficiently captured by the lexical variety variable.

Table 1. The Effects of Topic Odds Ratios on the Citation Counts of Articles in the Czech Sociological Review Estimated by Negative Binomial Regression

Variable	*Model 1 & 2*	*Model 3*
(Intercept)	2.141 (0.258) ***	1.47 (0.509) **
First author sex	0.273 (0.124) *	0.037 (0.13)
Publication age	0.01 (0.009)	0.02 (0.01) *
Collaborative authorship	0.26 (0.133) *	0.117 (0.129)
Lexical variety	-4.565 (0.72) ***	-1.976 (1.005) *
1 school phase transition	0.013 (0.049)	-0.057 (0.057)
2 party system country	-0.017 (0.058)	0.047 (0.067)
3 family woman child	-0.036 (0.037)	0 (0.056)
4 help household income	0.092 (0.044) *	0.006 (0.055)
5 level attitude status	0.204 (0.083) *	0.011 (0.101)
6 city resident municipality	0.233 (0.032) ***	0.124 (0.049) *
7 question society case	0.225 (0.222)	0.165 (0.243)
8 election government democracy	-0.018 (0.04)	0.088 (0.051) +
9 table survey respondent	0.056 (0.042)	-0.03 (0.065)
10 parent reproduction mother	-0.094 (0.046) *	0.019 (0.059)
11 relationship life form	-0.108 (0.153)	0.172 (0.201)
12 conclusion method comparison	-0.141 (0.087)	0.044 (0.108)
13 sociology sociologist image	-0.128 (0.057) *	0.168 (0.084) *
14 place time space	0.112 (0.066) +	0.09 (0.076)
15 equality class risk	-0.06 (0.045)	-0.065 (0.049)
16 population age market	0.119 (0.043) **	0.011 (0.06)
17 theory principle action	-0.36 (0.06) ***	-0.169 (0.084) *
18 affect danger threat	-0.101 (0.045) *	-0.004 (0.049)
19 number data information	0.36 (0.073) ***	0.198 (0.103) +
20 education opportunity effect	-0.035 (0.048)	0.039 (0.06)
21 bond identity generation	0.032 (0.067)	-0.048 (0.078)
22 network interaction community	0.218 (0.043) ***	0.15 (0.047) ***
23 center facility transportation	0.269 (0.044) ***	0.064 (0.065)
24 position demand claim	-0.229 (0.065) ***	0.12 (0.08)
25 world thing science	-0.306 (0.071) ***	-0.244 (0.108) *
26 sign communication intention	-0.174 (0.055) ***	-0.113 (0.066) +
27 price populace unit	0.262 (0.041) ***	0.05 (0.061)
28 care partner discourse	-0.024 (0.041)	-0.023 (0.05)
29 practice interview event	-0.133 (0.056) *	-0.036 (0.07)
30 majority member support	-0.018 (0.086)	-0.261 (0.106) *
31 citizen functioning worker	-0.079 (0.055)	-0.106 (0.063) +
32 history knowledge religion	-0.112 (0.046) *	0.014 (0.052)
33 sex old-age ageing	-0.042 (0.055)	-0.092 (0.057)

Note:

Model 1: Control variables only. First author sex: female = 0, male = 1; Publication age: years since publication; Collaborative authorship: single = 0, team = 1; Lexical variety: types/tokens ratio. Pseudo-R^2 = 0.1.

Model 2: Each topic entered the regression separately. Estimates and significance levels for control variables are not reported.

Model 3: Topics entered the regression simultaneously. Pseudo-R^2 = 0.28.

Significance levels: + 0.05–0.1; * 0.01–0.05; ** 0.001–0.01; *** 0–0.001

Standard errors reported in parentheses.

Results and Discussion

Table 1 conveys the results of the regression model for cumulative citations. Model 1 consists of control variables only. Model 2 is actually a combined report for multiple models, whereby each topic was added to control variables separately. Following Moody (2004), the results reported as model 2 attempt to account for the effect of a topic against all other topics. In contrast, model 3 considers all topics simultaneously and the effects are therefore at mean values for all other topics. Such model is unlikely to represent any real-world constellation in a document, but it shows that some topic effects hold significance even under such scenario. In addition, it shows that when topics are considered, the effects of author-level variables lose significance. This suggests that topics can moderate author attributes, but post-hoc tests are needed to examine this eventuality more closely. We also note that dispersed lexical focus does not favor high citation counts. Topics themselves are reported through their three highest ranking words.

Thanks to building a regression model from a feature space made up of topics, we can detect that quantitative topics and social geography yield considerable advantage in citations. Perhaps a bit surprisingly, topics with a focus on theory have negative effect on citations. The opposite effects for theoretical and quantitative topics can point to an important disconnect between theory and empirical research in Czech sociology. The majority of topics do not seem to affect citations significantly, but those that do can have non-negligible positive or negative effect on citation counts. Overall, the results suggest that topics are a useful feature to consider in the models of citations – and possibly other bibliometric indicators. Further research should focus on refining topic measurements and investigate interactions of topics among themselves as well as with other variables.

Acknowledgments

The author conducted the work as the overseas researcher under Postdoctoral Fellowship of Japan Society for the Promotion of Science and with the support by Grant-in-Aid for JSPS Fellows no. 17F17769.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, *5*(4), 594–607.

Abbott, A. D. (2001). *Chaos of disciplines*. Chicago: University of Chicago Press.

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics*, *9*(3), 499–513.

Antons, D., Joshi, A. M., & Salge, T. O. (2018). Content, Contribution, and Knowledge Consumption: Uncovering Hidden Topic Structure and Rhetorical Signals in Scientific Texts. *Journal of Management*, 0149206318774619.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257–270.

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, *4*(7), eaaq1360.

Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics*, *11*(2), 407–422.

Hicks, D. (2004). The Four Literatures of Social Science. In *Handbook of Quantitative Science and Technology Research* (pp. 473–496). Springer, Dordrecht.

Klavans, R., & Boyack, K. W. (2017). Research portfolio analysis and topic prominence. *Journal of Informetrics*, *11*(4), 1158–1174.

Leydesdorff, L., & Nerghes, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N < 1,000). *Journal of the Association for Information Science and Technology*, *68*(4), 1024–1035.

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, *18*(4), 209–223.

Mann, G. S., Mimno, D., & McCallum, A. (2006). Bibliometric Impact Measures Leveraging Topic Analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 65–74). New York, NY, USA: ACM.

Martin, F., & Johnson, M. (2015). More Efficient Topic Modelling Through a Noun Only Approach. *Proceedings of Australasian Language Technology Association Workshop*, 111–115.

Moody, J. (2004). The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213–238.

Moody, J., & Light, R. (2006). A view from above: The evolving sociological landscape. *The American Sociologist*, *37*(2), 67–86.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225.

Thelwall, M., & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, *8*(4), 963–971.

Understanding Multiple References Citation

Gege Lin[1], Haiyan Hou[2] and Zhigang Hu[3]

[1] *lingegenl@mail.dlut.edu.cn*
WISE Lab, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian City,
Liaoning Province (China)

[2] *htieshan@dlut.edu.cn*
WISE Lab, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian City,
Liaoning Province (China)

[3] *huzhigang@dlut.edu.cn*
WISE Lab, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian City,
Liaoning Province (China)

Abstract

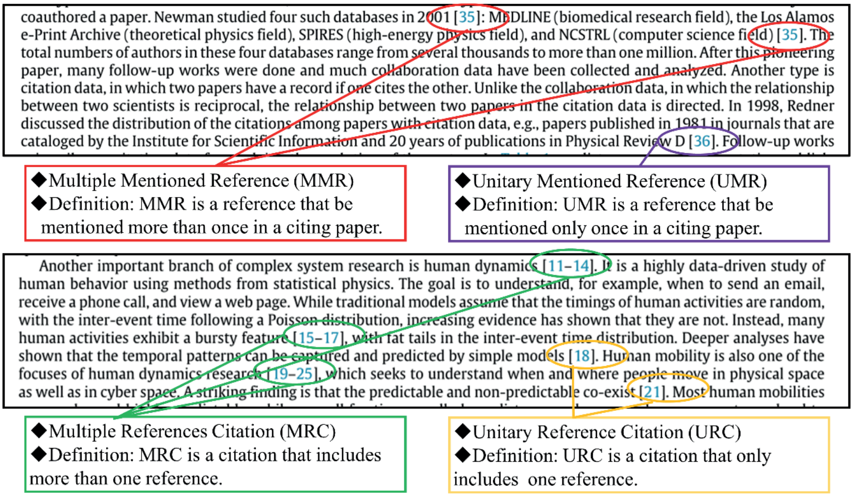
The purpose of this study is to make a comprehensive comparison between multiple references citation (MRC) and unitary reference citation (URC) in several aspects, including their location in full-text, their shares of self-citing citation, their citation age, etc. We chose all the 797 articles published in the JOI from 2007 to 2019 as the sample. Their full-text in XML format were crawled by employing Elsevier ScienceDirect API, and then parsed to extracted all the in-text citations using our developed python programs. As the results show, the percentage of MRC are approximately 25%. There are more MRC in the beginning and the ending of the paper. Scholars prefer to cite their own papers in MRC. So do journal self-citing citation. The average references age of MRC tends to be larger than that of URC. There is no significant correlation between the number of reference and the percentage of MRC in articles.

Introduction

Citations are essential components for scientific articles. In academic writing, the literature used in an article usually is referred to twice as follows (Hu, Lin, Sun, & Hou, 2017): first, in the body of the text with the name of the author and the publication year of the work, enclosed in parentheses, which is called a citation or in-text citation; and second, in a bibliography list at the end of the document with the full details of the publication, such as author(s), title, source, volume and page numbers, which is called a reference.

The correspondence between reference and in-text citation usually is not a one-to-one match. A reference could have several in-text citations (Liu, Guo, & Cronin, 2013; Shahid, Afzal, & Qadir, 2015; Zhao & Strotmann, 2016), while an in-text citation could include more than one reference. The latter citation we called Multiple References Citation (APA, 2009; Chicago, 2010), and the Unitary Reference Citation refers to an in-text citation only contains one reference. There is lots of research about the former (Ying, Liu, Guo, & Cronin, 2013;Cano, 1989; Zhao & Strotmann, 2014;Boyack, van Eck, Colavizza, & Waltman, 2018; Tahamtan & Bornmann, 2018), named multiply mentioned references. Ding et al. (2013) have used “CountX mentions” to weigh references by the number of times they are mentioned within citing papers. Zhao and Strotmann (2016) used “re-citation” to express a similar meaning.

Study of in-text citations and related text from scientific documents using full text sources has a long history. Although both positional (the location of references) and semantic (the meaning of references) studies have been pursued, here we focus primarily on the positional aspect. The terminology used in previous studies of in-text citations is not consistent. Thus, to avoid confusion, we define our terminology here. According to a reference mentioned times, we divided reference into two types: Unitary Mentioned Reference (UMR) and Multiple Mentioned Reference (MMR). We also divided in-text citation into two types: Unitary Reference Citation (URC) and Multiple References Citation (MRC). Their definition and example are shown in figure 1. In previous work, we characterized properties of in-text citations using 350 papers from Journal of Informetrics, and focusing on MMR(Hu, Lin, Sun, & Hou, 2017). In this study, we focus on MRC and want to find its characteristic and distribution in citing paper.



coauthored a paper. Newman studied four such databases in 2001 [35]: MEDLINE (biomedical research field), the Los Alamos e-Print Archive (theoretical physics field), SPIRES (high-energy physics field), and NCSTRL (computer science field) [35]. The total numbers of authors in these four databases range from several thousands to more than one million. After this pioneering paper, many follow-up works were done and much collaboration data have been collected and analyzed. Another type is citation data, in which two papers have a record if one cites the other. Unlike the collaboration data, in which the relationship between two scientists is reciprocal, the relationship between two papers in the citation data is directed. In 1998, Redner discussed the distribution of the citations among papers with citation data, e.g., papers published in 1981 in journals that are cataloged by the Institute for Scientific Information and 20 years of publications in Physical Review D [36]. Follow-up works

◆Multiple Mentioned Reference (MMR)
◆Definition: MMR is a reference that be mentioned more than once in a citing paper.

◆Unitary Mentioned Reference (UMR)
◆Definition: UMR is a reference that be mentioned only once in a citing paper.

Another important branch of complex system research is human dynamics [11–14]. It is a highly data-driven study of human behavior using methods from statistical physics. The goal is to understand, for example, when to send an email, receive a phone call, and view a web page. While traditional models assume that the timings of human activities are random, with the inter-event time following a Poisson distribution, increasing evidence has shown that they are not. Instead, many human activities exhibit a bursty feature [15–17], with fat tails in the inter-event time distribution. Deeper analyses have shown that the temporal patterns can be captured and predicted by simple models [18]. Human mobility is also one of the focuses of human dynamics research [19–25], which seeks to understand when and where people move in physical space as well as in cyber space. A striking finding is that the predictable and non-predictable co-exist [21]. Most human mobilities

◆Multiple References Citation (MRC)
◆Definition: MRC is a citation that includes more than one reference.

◆Unitary Reference Citation (URC)
◆Definition: URC is a citation that only includes one reference.

Figure 1. The definition and example of UMR, MMR, URC, MRC

Almost relevant works have focused on MMR, while the research about MRC is very few. With the increasing availability of full digital texts, in-text citation analysis is becoming more feasible and opening. In this study, we make a comprehensive comparison between MRC and URC. The remainder of the paper is structured as follows. Section 2 shows data acquisition, data processing and research questions. Section 3 shows the comparison between MRC and URC in several aspects. Finally, Section 4 discusses the results and concludes the paper.

Research Questions and Data

Research questions

The aim of this study is to give insights into Multiple References Ctitation (MRC) rather than to provide comprehensive information. As a young research topic in full-text citation analysis, it may evolve soon and attractive more researchers' attention. The following exploratory research questions drive the study.

1. How many references are there in each citation? What is the percentage of MRC and URC?

2. What is the distribution of MRC and URC? For example, Which type is more at the beginning of the citing paper?

3. Which do include more self-citing, MRC or URC? We have analyzed two kinds self-citing, author self-citing citation and journal self-citing citation.

4. Which do include more new or recent reference, MRC or URC? In other words, whose citation age is smaller, MRC or URC?

5. If a citing paper has more references, will increase the percentage of MRC or URC?

Data acquisition

In traditional citation analyses, Web of Science, Scopus and Google Scholar are the mostly frequently used data sources. However, research of MRC has to be based on full texts, because in-text citation existing in the body of citing papers. Scopus Search API (https://dev.elsevier.com/documentation/ScopusSearchAPI.wadl) is provided by Elsevier. We can use it to download some full-text in XML format. In this research, we selected the paper that published in Journal of informetrics from 2007 to 2019 as a case data. Finally, 797 research articles were retrieved and downloaded by Scopus Search API in January 2019.

Data processing

Using developed python programs, these XML-format full-text are parsed and all citations are extracted. The details of each in-text citation, including its location, references, etc., were recorded and imported into database tables. By querying these database tables, the number of reference in each citation could be determined.

Among the 28155 in-text citations within the 797 articles of JOI, 21095 citations (accounting for 75%) are URC and 7060 (about25%) citations are MRC. The more analysis and details are as follows.

Results and Discussion

The number of reference in each in-text citation

Statistics were carried out to figure out the number of reference in each in-text citation that located in citing paper published in JOI from 2007 to 2019. As shown in Table 1, about 75% in-text citation is URC that there is only one reference in this citation anchor, while approximately 25% in-text citation is MRC containing at least two references in the citation location. In MRC, two-reference citation in the same citation anchor is highest proportion among them, following by three-reference citation (5.3%) and four-reference citation (2.2%).

Table 1. The number of reference in each in-text citation (2007-2019)

Year	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*≥10*	*Total*
2007	507	121	38	14	7	3	1	0	2	1	694
2008	614	107	47	17	8	5	6	4	0	1	809
2009	801	172	51	29	17	9	3	2	2	1	1087
2010	1366	284	98	47	17	16	5	8	2	5	1848
2011	1405	297	78	46	20	11	3	3	3	8	1874
2012	1445	288	113	39	17	14	5	5	1	2	1929
2013	2311	474	134	60	26	22	17	9	3	5	3061
2014	2192	414	124	54	27	15	6	3	3	4	2842
2015	2209	441	200	85	38	16	15	6	4	7	3021
2016	2461	483	127	67	37	20	7	4	0	7	3213
2017	2499	521	214	71	39	27	7	6	5	9	3398
2018	3043	570	260	96	46	22	14	7	5	4	4067
2019	243	46	12	6	4	1	0	0	0	0	312
Total	21095	4218	1496	631	303	181	89	57	30	54	28155
Share(%)	74.92	14.98	5.31	2.24	1.08	0.64	0.32	0.20	0.11	0.19	100

Table 2. The share of URC and MRC in full-text (2007-2019)

Year	*Paper*	*URC*	*MRC*	*Total Citation*
2007	31	507 (73%)	187 (27%)	694
2008	33	614 (76%)	195 (24%)	809
2009	32	801 (74%)	286 (26%)	1087
2010	64	1366 (74%)	482 (26%)	1848
2011	61	1405 (75%)	469 (25%)	1874
2012	69	1445 (75%)	484 (25%)	1929
2013	93	2311 (75%)	750 (25%)	3061
2014	82	2192 (77%)	650 (23%)	2842
2015	80	2209 (73%)	812 (27%)	3021
2016	80	2461 (77%)	752 (23%)	3213
2017	81	2499 (74%)	899 (26%)	3398
2018	82	3043 (75%)	1024 (25%)	4067
2019	9	243 (78%)	69 (22%)	312
Total	797	21095 (75%)	7060 (25%)	28155

As shown in Table 2, the percentage of MRC stabilizes around 25%, while the proportion of URC fluctuates about 75%. Although the number of in-text citation dramatically increases in the past decade, the percentage of URC and MRC in full-text change little.

The distribution of URC and MRC

Within the body of an article, in-text citations or mentions are located in sections. The density of in-text citations in different sections varies greatly (Hu et al., 2017). Besides, different paper has different sections. As shown in Figure 2, 5-section structure paper is the highest among all kinds of structures, accounting for 35.1%. The figure for 4-section structure is the second highest (30%), following by 6-section structure (roughly 20%). These three types structure contain most paper and total percentage of them reach up to about 85%.

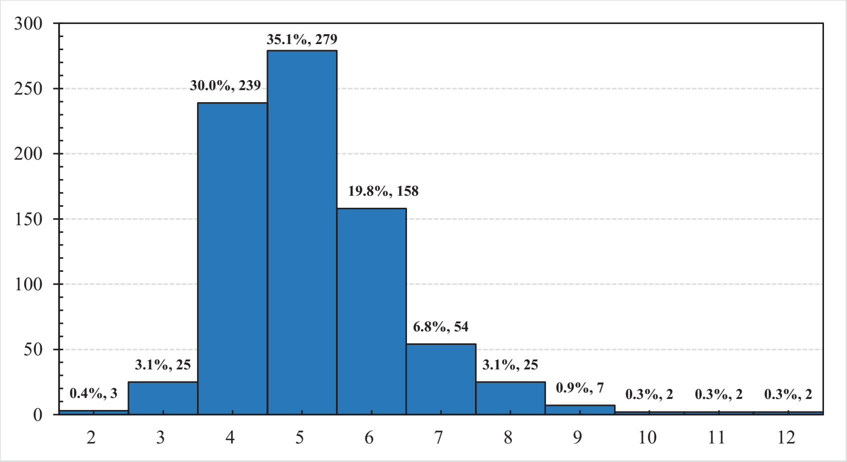


Figure 2. The number of section in citing paper

A typical scientific article follows the IMRaD structure or its variants (Agarwal & Yu, 2009). IMRaD usually includes Introduction, Methods, Results, and Discussion. Generally, 4-section structure paper is IMRaD structure or its variants. Figure 3 shows that in the start the paper, usually introduction and method separately, have more MRC than that in the finish of the paper, while URC in the same way.

As shown in Figure 4, the figure for second section has biggest proportion of MRC in 5-section structure paper, following by first section and third section. However, the first section in 5-section structure paper make up the largest portion of URC, with the second section following behind.

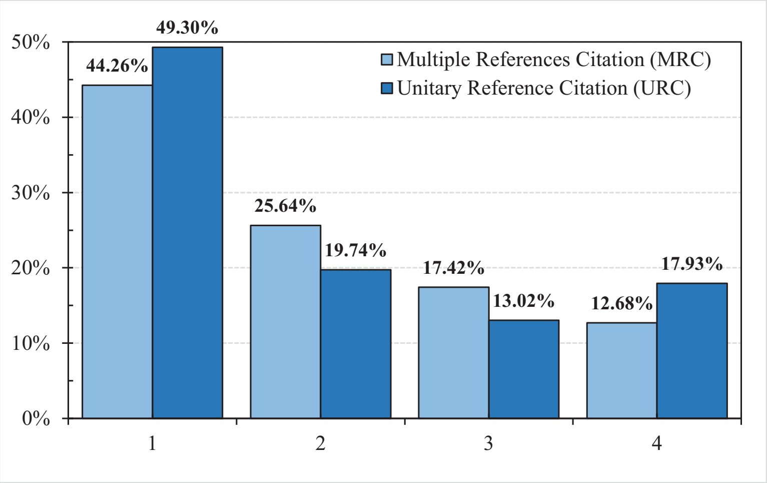


Figure 3. The distribution of URC and MRC in 4-section structure paper

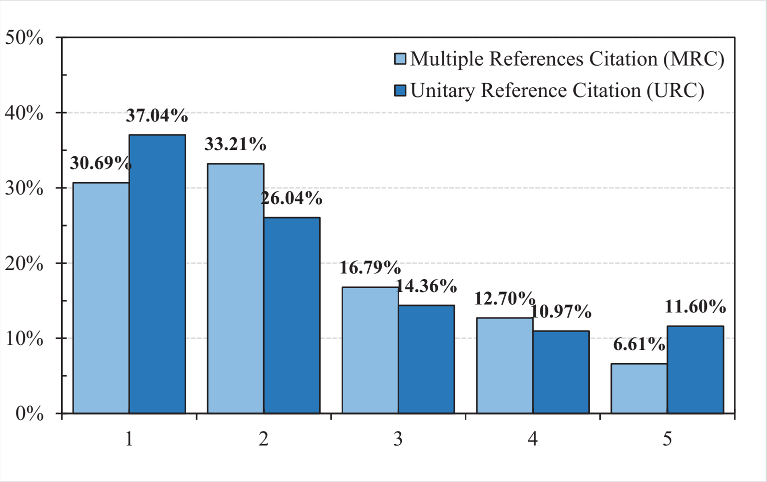


Figure 4. The distribution of URC and MRC in 5-section structure paper

Figure 5 shows that the percentage of MRC in first and second section are significantly highest than the other sections, up to about 60%, the others all below 15%. Maybe there are lots of papers need to be included in introduction and literature review, so the Multiple Reference Citation (MRC) would be the better choice because of the limitations of coverage. Similarly, the figure for URC in first and second section is highest than the other seconds, and the other sections are all lower than15%.

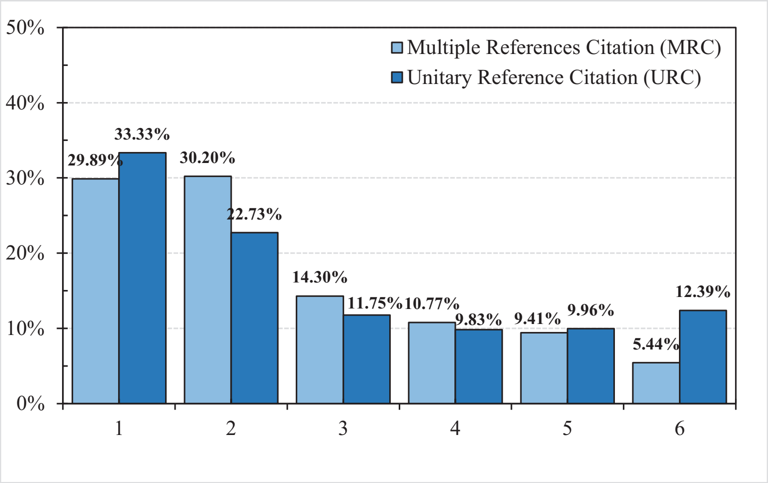


Figure 5. The distribution of URC and MRC in 6-section structure paper

The self-citing citation in URC and MRC

Which kinds of in-text citation are most likely to be author self-citing citation? The hypothesis is that the more relevant a reference is, the more likely it will be cited. Based on this hypothesis, author self-citing citation will be examined first and the journal self-citing citation will be second. Author self-citing citations are expected to be cited easily because self-citing articles are the continuation of their previous work and are thus extremely relevant to the citing articles.

a. Author self-citing citation

To examine this hypothesis, the proportion of author self-citing citation is compared between URC and MRC. An author self-citation is defined as there is at least one same author name between the citing and the cited publication that be contained in an in-text citation. Note that as long as one publication in MRC location has one author name in common with the citing publication, the in-text citation would be considered as an author self-citation.

As shown in Table 3, among the 28155 in-text citations from JOI, there are 21095 (75%) URC and 7060 (25%) MRC. For URC, about 14% is author self-citing citation, while for MRC, the figure is over 21%. Therefore, it is clear that scholars prefer to cite their own papers in MRC rather than URC.

Table 3. The author self-citing citations of two different type citations

In-Text Citation	*Count*	*Author Self-Citing Citation*	*Share*
URC	21095	2926	13.87%
MRC	7060	1514	21.44%
All	28155	4440	15.77%

b. Journal self-citing citation

Similar to the author self-citation hypotheses, it is also reasonable to assume that journal self-citing citation tend to be more in MRC rather than URC. In the same way, a journal self-citing citation is defined as the citing and the cited paper that be contained in an in-text citation publish in same journal. Note that as long as the journal of any cited paper in MRC location is same with that of citing publication, the in-text citation would be considered as a journal self-citing citation.

As Table 4 shows, 19.05% of journal self-citing citation is in MRC, while only 9.6% of that is in URC. On average, a journal self-citing citation is about 12%. If an author decides to cite more than one reference in one citation anchor, there is one in five chance that he will choose to cite a paper that published in the journal which he also wants to contribute to.

Table 4. The journal self-citing citations of two different type citations

In-Text Citation	*Count*	*Journal Self-Citing Citation*	*Share*
URC	21095	2030	9.62%
MRC	7060	1345	19.05%
All	28155	3375	11.99%

The reference/citation age in URC and MRC

The time gap between a reference and its citing paper is called the citation age (Burrell, 2002). Based on that definition, reference age has the same meaning with the citation age of Burrell. Due to MRC with more than one references, the citation age in this study refers to average reference age in the same citation anchor.

Table 5 shows that almost references have collected their publication year. The negligence of authors or databases may result in the publication year missing of reference. As shown in Figure 6, For URC, the figure for reference age under five is about 44%, ranking in the highest percentage; And reference age 1 accounts for biggest (about 12%). While for MRC, the figure for reference age under five is also highest (39%); and reference age 2 has the most proportion among them.

Table 5. The number of reference that has publication year in URC and MRC

In-Text Citation	*Citation Count*	*Reference*	*Reference that has publication year*
URC	21095	21095	20222 (95.86%)
MRC	7060	20161	18792 (93.21%)
All	28155	41256	39014 (94.57%)

Likewise, the vast majority of in-text citation can calculate their citation age (or called average citation age). Table 6 shows that about 97% citation has publication year overall. As shown in Figure 7, citation age (average reference age) 4 in MRC has the highest percentage, about 9%, while citation age 1 in URC ranks first (about 12%). When the citation age over 3, the percentage of MRC no less than that of URC.

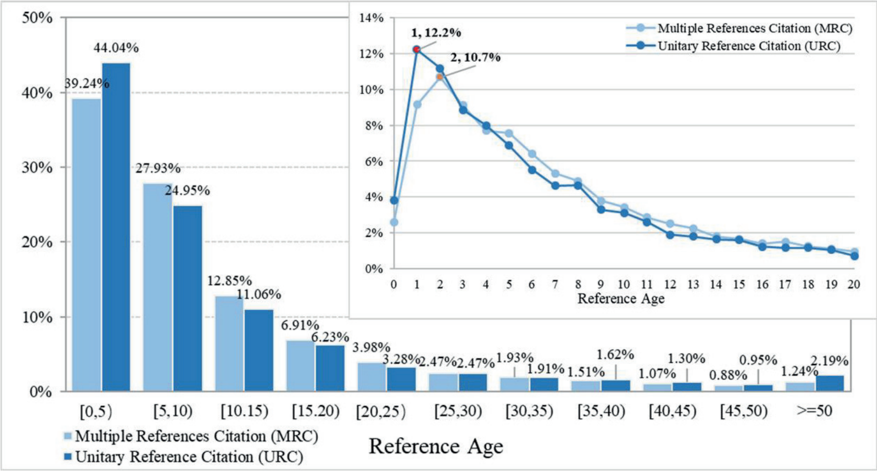


Figure 6. The reference age in URC and MRC

Table 6. The number of citation that has publication year in two different type citations

In-Text Citation	*Citation Count*	*Citation that has publication year*
URC	21095	20222 (95.86%)
MRC	7060	6961 (98.60%)
All	28155	27183 (96.55%)

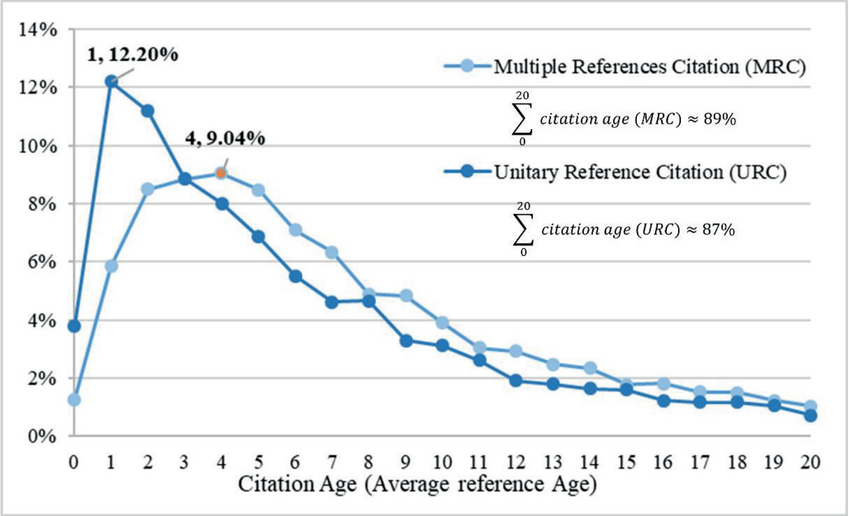


Figure 7. The citation age (average reference age) in URC and MRC

The relationship between the number of reference and the percentage of URC and MRC

Average numbers of references has increased over time at a higher rate (Boyack et al., 2018). So the hypothesis is that the more references in a paper, the more MRC will have. Figure 8 shows that negative correlation was found between the number of reference and the percentage of MRC or URC. Maybe because of the limitations of space and increasing reference in each paper, the MRC would be more and more popular.

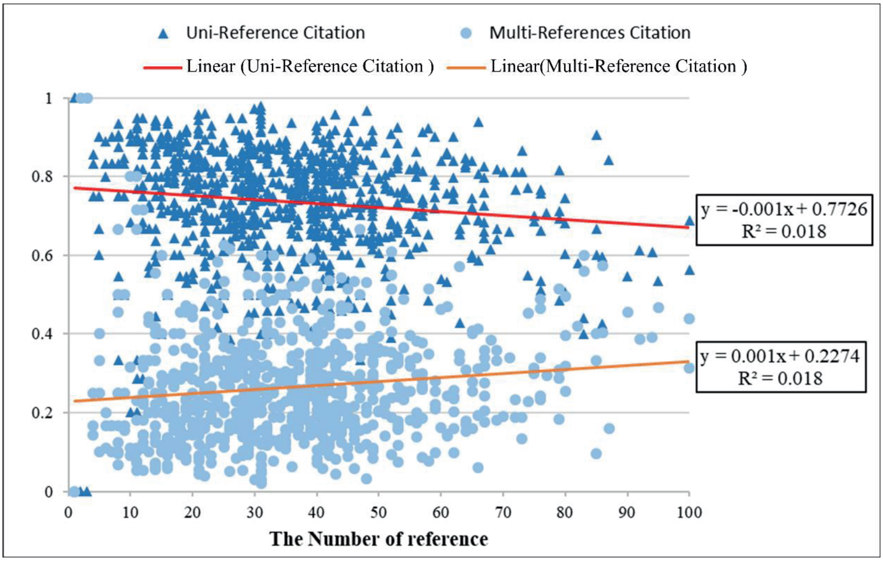


Figure 8. The relationship between the number of reference and the percentage of URC and MRC

Conclusions

In scientific writing, Multiple References Citation (MRC) should not be ignored both in proportion and their significance to citing papers. In the Journal of Informetrics, approximately 25% of all citations are multiple references citation (MRC). On average, a MRC includes 2.86 references. In contrast, roughly 75% of all citations are Unitary Reference Citation (URC). Meanwhile, the percentage of MRC and URC both changed little over the past decade.

In this study, we make a comprehensive comparison between MRC and URC in several aspects, including their location in full-text, their shares of self-citing citation, their citation age, etc. First, in the start of the citing paper, usually introduction and method, have more MRC than that in the middle and the finish of the paper, so dose URC. Second, scholars prefer to cite their own papers in MRC rather than URC. So does journal self-citing citation. Third, When the citation/reference age over 3, the percentage of MRC no less than that of URC. Finally, negative correlation was found between the number of references and the percentage of MRC or URC.

There are several limitations to this study that should be noted. First, while the study is large, it still covers only a relatively modest share of the articles published in recent years. Additional data from other sources could show different results. Second, only two citation types were considered, and it is possible that MRC could be studied in more detail depending on the number of reference. Despite these limitations, we consider the results to be robust and reliable. Such studies have the potential to influence our understanding of citation theory and behavior, and to have practical influence on applications such as information search and retrieval and accurate modeling of the structure and dynamics of science(Hu, Lin, Sun, & Hou, 2017; Boyack, van Eck, Colavizza, & Waltman, 2018).

Acknowledgments

The research was supported by the Natural Science Foundation of China (NSFC) under Grant 71503031.

Reference

Agarwal, S., Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics, 25*(23), 3174-3180.

APA. (2009). *Publication manual of the American Psychological Association (6th ed.)*. Washington, DC: American Psychological Association Washington.

Boyack, K. W., van Eck, N. J., Colavizza, G., Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*(1), 59-73.

Burrell, Q. L. (2002). Modelling citation age data: Simple graphical methods from reliability theory. *Scientometrics, 55*(2), 273-285.

Cano, V. (1989). Citation Behavior: Classification, Utility, and Location*.*, *40*(4), 284-290.

Chicago, U. O. (2010). *The Chicago manual of style (sixteenth ed.)*. Chicago: University of Chicago Press.

Ding, Y., Liu, X., Guo, C., Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Hu, Z., Lin, G., Sun, T., Hou, H. (2017). Understanding multiply mentioned references. *Journal of Informetrics, 11*(4), 948-958.

Shahid, A., Afzal, M. T., Qadir, M. A. (2015). Lessons Learned: The Complexity of Accurate Identification of in-Text Citations. *The International Arab Journal of Information Technology, 5*(12), 481-488.

Tahamtan, I., Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics, 12*(1), 203-216.

Ying, D., Liu, X., Guo, C., Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Zhao, D., Strotmann, A. (2014). In-text author citation analysis: Feasibility, benefits, and limitations. *Journal of the Association for Information Science and Technology, 65*(11), 2348-2358.

Zhao, D., Strotmann, A. (2016). Dimensions and uncertainties of author citation rankings: Lessons learned from frequency-weighted in-text citation counting. *Journal of the Association for Information Science and Technology, 67*(3), 671-682.

Large-scale comparison of bibliographic data sources: Web of Science, Scopus, Dimensions, and Crossref

Martijn Visser, Nees Jan van Eck and Ludo Waltman

{visser,waltmanlr,ecknjpvan}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University (the Netherlands)

Abstract

We present a large-scale comparison of four multidisciplinary bibliographic data sources: Web of Science, Scopus, Dimensions, and Crossref. Scopus is compared in a pairwise manner with each of the other data sources. We first analyze differences between the data sources in the coverage of documents. We then study differences in the completeness and accuracy of citation links. Based on our analysis, we discuss strengths and weaknesses of the different data sources.

Introduction

Over the past 15 years, Web of Science (WoS; Schnell, 2017), Scopus (Schotten, el Aisati, Meester, Steiginga, & Ross, 2017), and Google Scholar have been the three most important multidisciplinary bibliographic data sources, providing metadata on scientific documents and on citation links between these documents. It is very challenging to perform large-scale analyses using Google Scholar. WoS and Scopus have therefore long been the only options for large-scale bibliometric studies. This has changed in recent years with the introduction of two new multidisciplinary bibliographic data sources: Microsoft Academic (Sinha et al., 2015) and Dimensions (Hook, Porter, & Herzog, 2018). At the same time, Crossref has become an increasingly interesting data source (Van Eck, Waltman, Larivière, & Sugimoto, 2018). Thanks to the Initiative for Open Citations (I4OC; https://i4oc.org/), hundreds of millions of citation links between documents have been made openly available in Crossref.

Both for bibliometric research and for bibliometric practice, it is important to understand the strengths and weaknesses of different bibliographic data sources. Because most researchers do not have large-scale access to data sources such as WoS and Scopus, bibliographic data sources are typically compared in small-scale case studies, focusing for instance on documents in a specific research field or on a small number of researchers and the documents they have authored (e.g., Harzing, in press). Alternatively, bibliographic data sources have been compared at a large scale, but at the level of journals rather than individual documents (Mongeon & Paul-Hus, 2016).

In this paper, we present a large-scale document-level comparison of four bibliographic data sources: WoS, Scopus, Dimensions, and Crossref. Google Scholar is not included because we do not have large-scale access to this data source. Studies of Google Scholar typically focus on relatively small numbers of documents (e.g., Martín-Martín, Orduna-Malea, & López-Cózar, 2018; Martín-Martín, Orduna-Malea, Thelwall, & López-Cózar, 2018). Microsoft Academic is not included because we did not have sufficient time to include this data source. However, we are currently working on an extended version of this paper in which Microsoft Academic will be included as well.

The comparison that we present in this paper focuses on differences between the data sources in the coverage of documents. In addition, differences in the completeness and accuracy of citation links are also studied. To keep the analysis manageable, the focus is on pairwise comparisons of Scopus with each of the other three data sources.

Data sources

In our comparison, we consider the following four bibliographic data sources:

- *WoS*. WoS consists of multiple citation indices. We consider the Science Citation Index Expanded, the Social Sciences Citation Index, the Arts & Humanities Citation Index, and the Conference Proceedings Citation Index. Our center has full access to these citation indices for documents starting from 1980. The Emerging Sources Citation Index and the Book Citation Index, which are also part of the so-called WoS Core Collection, are not considered, because our center does not have access to them. We use WoS data updated until week 26 in 2018. The data was delivered to our center in XML format.
- *Scopus*. Our center has full access to Scopus for documents starting from 1996. We use Scopus data delivered to our center in April 2018.
- *Dimensions*. Our center has full access to Dimensions. We use Dimensions data delivered to our center in December 2018. In March 2019, we received an update of the disciplinary classification of documents in Dimensions. In our analysis, we make use of this updated classification. In addition to scientific documents, Dimensions also covers clinical trials, grants, patents, and policy documents. We do not consider this content in our analysis.
- *Crossref*. We use Crossref data downloaded in August 2018 through the public REST API of Crossref. We downloaded the data in JSON format.

Our comparison focuses on documents from the period 1996–2017.

The different data sources have different content selection policies. WoS has an internal Editorial Development team for content selection. WoS emphasizes the selectivity of its content selection policy for the WoS Core Collection.[1] Scopus works together with an international group of researchers, referred to as the Content Selection and Advisory Board, to perform content selection.[2] Scopus claims to be "the largest abstract and citation database of peer-reviewed literature".[3] Compared with the WoS Core Collection, Scopus therefore appears to focus more on comprehensiveness and less on selectivity. Dimensions has an even stronger focus on comprehensiveness: "The database should not be selective but rather should be open to encompassing all scholarly content that is available for inclusion … The community should then be able to choose the filter that they wish to apply to explore the data according to their use case." (Hook et al., 2018).

Crossref is a special case. It is a registration agency for Digital Object Identifiers (DOIs). If a scientific publisher works with Crossref to register a DOI for a document, Crossref obtains basic metadata for this document. Crossref then makes this metadata openly available (with the possible exception of the reference list, for which the publisher determines whether it can be made openly available or not). In this way, Crossref has become a bibliographic data source that is of significant interest for bibliometric analyses. The completeness and the quality of the data available in Crossref depend on what publishers provide to Crossref. Crossref itself does not actively collect and enrich data.

Matching of data sources

We matched documents in Scopus with documents in WoS, Dimensions, and Crossref. To match documents in Scopus with documents in WoS and Dimensions, we developed a matching procedure. Documents were matched by comparing the following attributes: (1) DOI, (2) first author (i.e., last name and first initial), (3) title, (4) source (i.e., ISBN, ISSN, or

[1] See https://clarivate.com/essays/evaluation-criteria-web-science-core-collection-journals/. See also Schnell (2017).

[2] See https://www.elsevier.com/solutions/scopus/how-scopus-works/content/scopus-content-selection-and-advisory-board. See also Schotten et al. (2017).

[3] https://www.elsevier.com/solutions/scopus. See also Schotten et al. (2017).

source title), (5) publication year, (6) volume and issue number, and (7) article number and begin and end page.

For a pair of two documents, one in Scopus and one in WoS or Dimensions, a score was assigned for each of the above attributes if the attribute had the same value for both documents. In the case of the first author, title, and source attributes, we also allowed for partial matches. To do so, we used a fuzzy matching approach based on the Levenshtein distance. The smaller the Levenshtein distance, the higher the score that was assigned. A match between two documents was established when the sum of the scores of all attributes exceeded a certain threshold. The threshold was set in such a way that precision of the matching procedure was favored over recall.

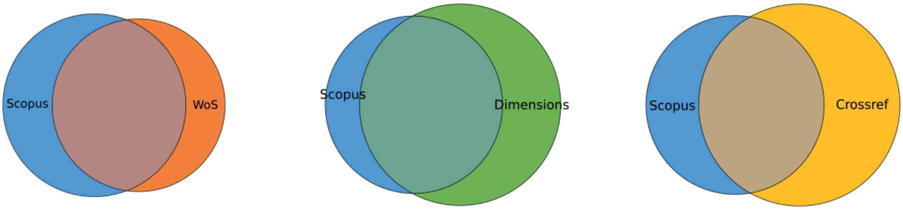
Documents in Scopus were matched with documents in Crossref simply based on DOI. All documents in Crossref have a DOI, but a substantial share of the documents in Scopus do not have a DOI (Gorraiz, Melero-Fuentes, Gumpenberger, & Valderrama-Zurián, 2016). Documents without a DOI in Scopus could not be matched with documents in Crossref. In the exceptional case in which multiple documents with the same DOI were found in Scopus, no match was established. In this way, we also excluded duplicate documents. Duplicate documents have been reported to be a significant problem in Scopus (Valderrama-Zurián, Aguilar-Moya, Melero-Fuentes, & Aleixandre-Benavent, 2015; Van Eck & Waltman, 2017).

Comparison of coverage of documents

As already mentioned, in our comparison of WoS, Scopus, Dimensions, and Crossref, we use Scopus as a baseline. Figure 1 shows the differences in coverage of documents between Scopus on the one hand and WoS, Dimensions, and Crossref on the other hand. Scopus covers almost 45 million documents. With 40 million documents, WoS is smaller than Scopus. Dimensions and Crossref are of similar size. They both cover between 57 and 58 million documents, which is substantially more than Scopus and WoS. Since Dimensions relies strongly on data from Crossref (Hook et al., 2018), these two data sources largely cover the same documents. However, certain types of content that are covered by Crossref, such as data sets, are not covered by Dimensions. The other way around, Dimensions covers documents that are not covered by Crossref. A substantial share of these documents seem to originate from PubMed.

As can be seen in Figure 1, WoS has an overlap of 29 million documents with Scopus. For Dimensions, an overlap of 35 million documents with Scopus is found. Based on the simple DOI-based matching of documents in Scopus and Crossref, there is an overlap of 29 million documents between Scopus and Crossref. However, this is likely to be a considerable underestimation of the true overlap between these two data sources.

The high-level statistics presented in Figure 1 of course hide many important differences between the various data sources. We analyze these differences in the next subsections.



Scopus: 44.9M
WoS: 40.1M
Overlap: 29.1M

Scopus: 44.9M
Dimensions: 57.9M
Overlap: 35.1M

Scopus: 44.9M
Crossref: 57.3M
Overlap: 29.2M

Figure 1. Overlap of documents between Scopus and the other data sources.

Differences in coverage by publication year

Figure 2 shows the time trend in the number of documents covered by the different data sources and the overlap of documents between Scopus and the other data sources. The yearly number of documents in Dimensions and Crossref is quite similar. This illustrates the strong reliance of Dimensions on data from Crossref.

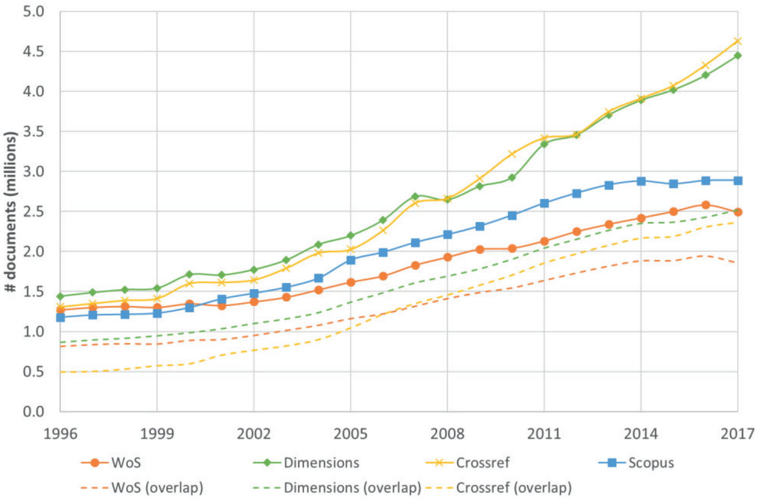


Figure 2. Breakdown by publication year for all documents in each data source and for the overlap with documents in Scopus.

Differences in coverage by document type

The top-left plot in Figure 3 provides a breakdown by document type for all documents in Scopus and for the overlap with documents in the other data sources. The document type classification of Scopus is used. The plot shows that for many articles and conference papers in Scopus there are no matching documents in the other data sources.

The other plots in Figure 3 provide the opposite perspective. Using the document type classifications of WoS, Dimensions, and Crossref, these plots offer a breakdown by document type for all documents in WoS, Dimensions, and Crossref and for the overlap with documents in Scopus. The top-right plot shows that meeting abstracts and book reviews are missing in Scopus. Also, for many proceedings papers in WoS, there are no matching documents in Scopus. On the other hand, almost all articles in WoS can also be found in Scopus.

Unfortunately, the document type classifications in Dimensions and Crossref are less detailed. The bottom plots therefore offer less information. They show that for many articles and book chapters in Dimensions and Crossref there are no matching documents in Scopus. Importantly, however, any document published in a journal is classified as an article in Dimensions and Crossref. This even includes content such as the list of editorial board members of a journal or the cover of a journal issue.

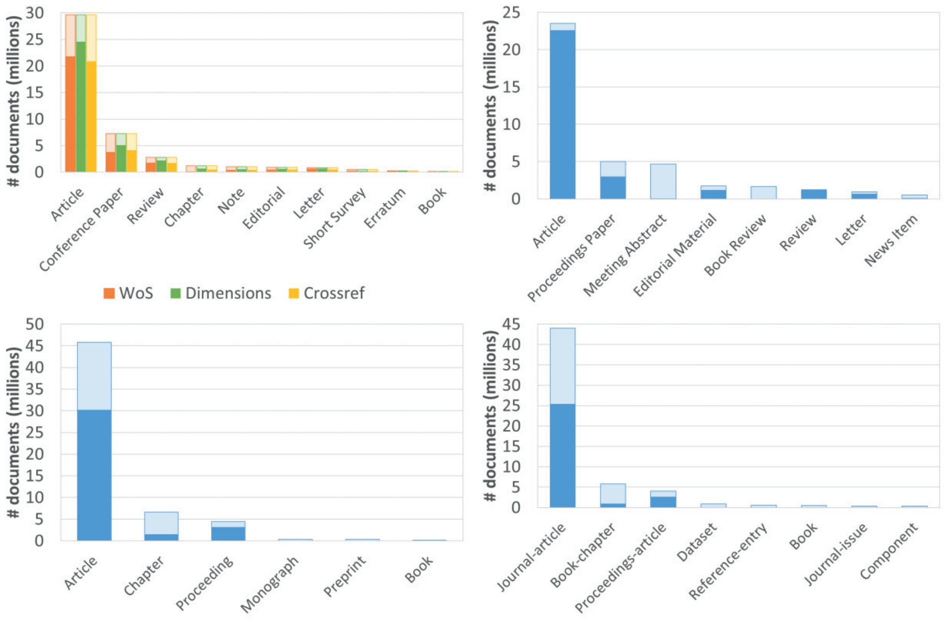


Figure 3. Top-left plot: Breakdown by document type for all documents in Scopus and for the overlap with documents in the other data sources. Other plots: Breakdown by document type for all documents in WoS (top right), Dimensions (bottom left), and Crossref (bottom right) and for the overlap (in dark blue) with documents in Scopus.

Differences in coverage by discipline

We now compare the coverage of documents by broad discipline. In Scopus, documents are assigned to four broad disciplines: *Health Sciences*, *Life Sciences*, *Physical Sciences*, and *Social Sciences & Humanities*. In WoS, we make use of an assignment of documents to five broad disciplines: *Arts & Humanities*, *Life Sciences & Biomedicine*, *Physical Sciences*, *Social Sciences*, and *Technology*. In Dimensions, we rely on a classification of documents into 22 fields, which we further aggregate into four broad disciplines: *Arts & Humanities*, *Biomedical Sciences*, *Physical Sciences*, and *Social Sciences*. Crossref also provides a classification of documents into broad disciplines, but most documents are not included in this classification. We therefore do not use this classification.

In the disciplinary classifications of Scopus and WoS, documents are assigned to disciplines based on the source in which they have appeared. In Scopus, documents in multidisciplinary sources (e.g., *Nature*, *PLOS ONE*, *PNAS*, *Science*, and *Scientific Reports*) are assigned to the *Health Sciences* discipline. In WoS, these documents do not have an assignment to a discipline. Some documents belong to multiple disciplines in the classifications of Scopus and WoS. We use a fractional counting approach to handle these documents. We note that in an earlier study significant inaccuracies were identified in the disciplinary classification of Scopus (Wang & Waltman, 2016).

In the disciplinary classification of Dimensions, documents are assigned to disciplines independently of the source in which they have appeared. The accuracy of the disciplinary classification of Dimensions has been questioned (Bornmann, 2018; Herzog & Lunn, 2018; Orduña-Malea & Delgado-López-Cózar, 2018). The classification also has the limitation of being incomplete. Many documents in Dimensions do not have an assignment to a discipline.

The top-left plot in Figure 4 provides a breakdown by discipline for all documents in Scopus and for the overlap with documents in the other data sources. The disciplinary classification of Scopus is used. This for instance means that a document that is covered both by Scopus and by WoS is assigned to the discipline to which it belongs in the disciplinary classification of Scopus. The disciplinary classification of WoS plays no role. The plot shows that the overlap between Scopus and the other data sources is largest in the *Life Sciences* discipline. In the *Social Sciences & Humanities* discipline, the overlap between Scopus on the one hand and Dimensions, Crossref, and especially WoS on the other hand is quite small.

The other plots in Figure 4 provide the opposite perspective. Using the disciplinary classifications of WoS and Dimensions, these plots offer a breakdown by discipline for all documents in WoS and Dimensions and for the overlap with documents in Scopus. As can be seen in the top-right plot, in the *Life Sciences & Biomedicine* discipline, a large number of documents in WoS do not have matching documents in Scopus. Many of these documents are meeting abstracts, which are not covered by Scopus. From a relative point of view, the large share of the documents in the *Arts & Humanities* discipline in WoS that do not have matching documents in Scopus is noteworthy. There are two main explanations for this. First, there are various types of documents that play a prominent role in the *Arts & Humanities* discipline in WoS and that Scopus does not seem to cover at all. The most important one is the WoS document type *Book Review*. Other examples are the WoS document types *Film Review*, *Theater Review*, *Poetry*, and *Fiction, Creative Prose*. Second, Scopus has a rather low coverage of documents in the arts and humanities in the earlier years of our analysis, while it has a much higher coverage in recent years. Hence, the small overlap between Scopus and WoS for documents in the arts and humanities is not entirely representative for the situation in recent years.

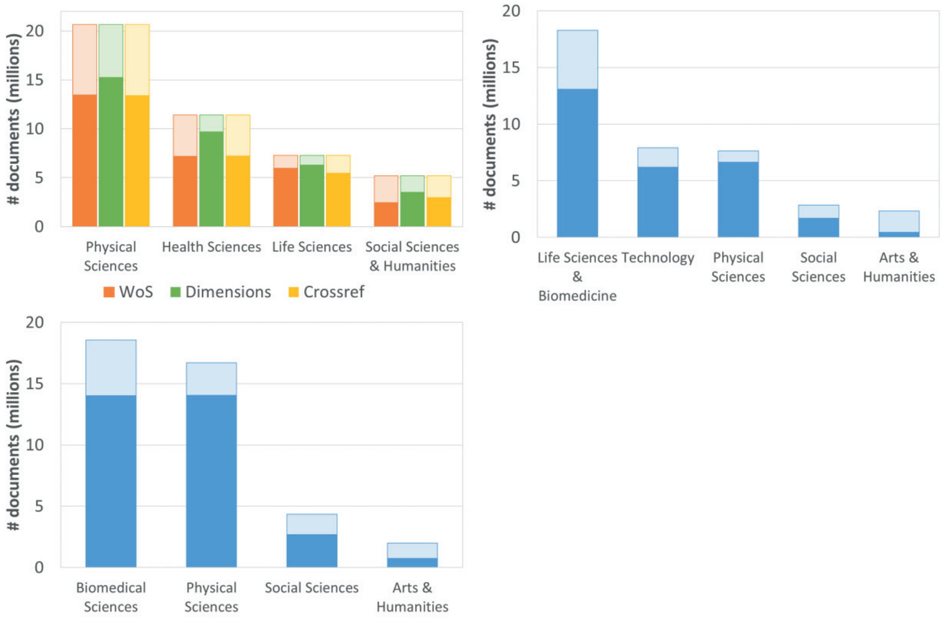


Figure 4. Top-left plot: Breakdown by discipline for all documents in Scopus and for the overlap with documents in the other data sources. Other plots: Breakdown by discipline for all documents in WoS (top right) and Dimensions (bottom left) and for the overlap (in dark blue) with documents in Scopus.

The patterns observed for Dimensions, presented in the bottom-left plot in Figure 4, are similar to those observed for WoS. However, more than 16 million documents in Dimensions do not have an assignment to a discipline. These documents are not included in the bottom-left plot in Figure 4.

Differences in coverage by language

Scopus, WoS, and Dimensions are strongly dominated by documents written in English (see also Mongeon & Paul-Hus, 2016), although they also cover documents written in Chinese, French, German, Portugese, Spanish, and other languages. For Crossref, we do not have language information. For most of the documents in Scopus that are not in English, no matching documents were found in the other data sources. Likewise, most of the documents in WoS and Dimensions that are not in English do not have matching documents in Scopus. Hence, the overlap between Scopus and the other data sources is biased toward English language documents.

Differences in coverage by number of references

The number of references of a document may be used as a rough proxy of the scientific importance of the document. Although there are all kinds of exceptions, a document with many references (e.g., a full research article) may often be considered to have a higher scientific importance than a document with only a few references or no references at all (e.g., an editorial, a letter, or a meeting abstract). For this reason, we look at a breakdown by number of references of the overlap between the different data sources.

The left plot in Figure 5 provides a breakdown by number of references for all documents in Scopus and for the overlap with documents in the other data sources. Documents with a large number of references are overrepresented in the overlap between Scopus and the other data sources. However, even among documents in Scopus with more than 50 references, there are a substantial number for which no matching documents were found in the other data sources.

The right plot in Figure 5 provides an opposite perspective. It offers a breakdown by number of references for all documents in WoS and for the overlap with documents in Scopus. There are only a very limited number of documents in WoS that have a large number of references and that do not have a matching document in Scopus.

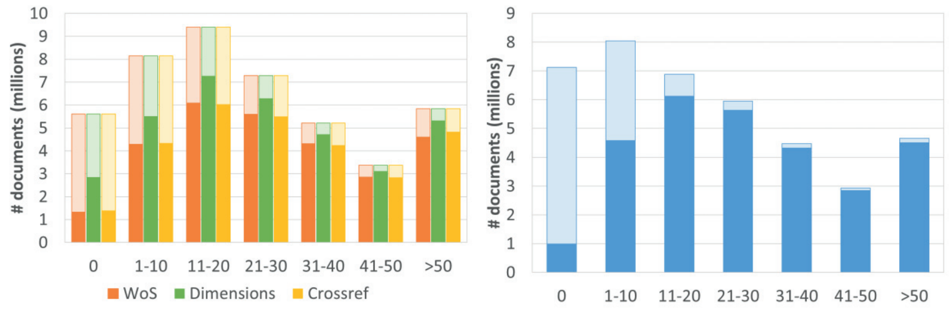


Figure 5. Left plot: Breakdown by number of references for all documents in Scopus and for the overlap with documents in the other data sources. Right plot: Breakdown by number of references for all documents in WoS and for the overlap (in dark blue) with documents in Scopus.

We do not show results from the viewpoint of Dimensions and Crossref. In Dimensions, we do not know the total number of references of a document. We know only the number of references that have been matched with a cited document. In Crossref, we do not know the number of references of documents for which the references have not been deposited in Crossref.

Differences in coverage by number of citations

Like the number of references, the number of citations of a document offers a proxy of the scientific importance of the document. We therefore look at a breakdown by number of citations of the overlap between the different data sources.

The top-left plot in Figure 6 provides a breakdown by number of citations for all documents in Scopus and for the overlap with documents in the other data sources. Documents with a larger number of citations are overrepresented in the overlap between Scopus and the other data sources. However, there are still a substantial number of documents in Scopus with more than five citations for which no matching documents were found in the other data sources.

The other plots in Figure 6 provide the opposite perspective. These plots offer a breakdown by number of citations for all documents in WoS, Dimensions, and Crossref and for the overlap with documents in Scopus. Almost all documents in WoS with a large number of citations have matching documents in Scopus. In Dimensions and Crossref, there are quite some documents with a large number of citations for which no matching documents were found in Scopus.

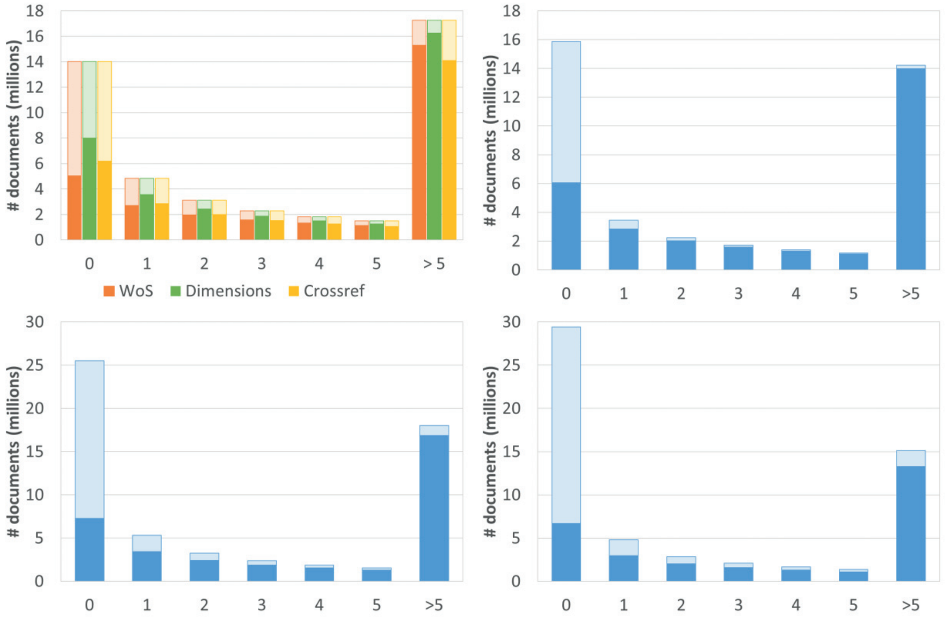


Figure 6. Top-left plot: Breakdown by number of citations for all documents in Scopus and for the overlap with documents in the other data sources. Other plots: Breakdown by number of citations for all documents in WoS (top right), Dimensions (bottom left), and Crossref (bottom right) and for the overlap (in dark blue) with documents in Scopus.

Comparison of completeness and accuracy of citation links

To compare the completeness and accuracy of citation links, we again use Scopus as a baseline. We present pairwise comparisons between Scopus on the one hand and WoS, Dimensions, and Crossref on the other hand. Importantly, in these pairwise comparisons, we consider only citation links between documents that are covered by both data sources. Hence, we compare the completeness and accuracy of citation links after correcting for differences in the coverage of documents.

Figure 7 shows the overlap of citation links between Scopus and the other data sources. Scopus and WoS have the largest overlap. Nevertheless, the discrepancies between the two data sources are quite significant. 3.2% of the citation links in WoS cannot be found in Scopus. Conversely, 5.0% of the citation links in Scopus cannot be found in WoS.

The discrepancies between Scopus and Dimensions are even larger. 5.0% of the citation links in Dimensions cannot be found in Scopus. Moreover, for 13.0% of the citation links in Scopus, there is no corresponding citation link in Dimensions.

Finally, comparing Scopus and Crossref, we find that 63.1% of the citation links in Scopus cannot be obtained from Crossref. There are three main reasons for this. First, some publishers deposit documents in Crossref without depositing their references. Second, there are publishers (in particular ACS, Elsevier, IEEE, IOP Publishing, and Wolters Kluwer Health) that deposit references in Crossref but do not make these references openly available. Third, Crossref has suffered from a technical problem due to which a large number of openly available references incorrectly have not been linked to cited documents (Bilder, 2019).

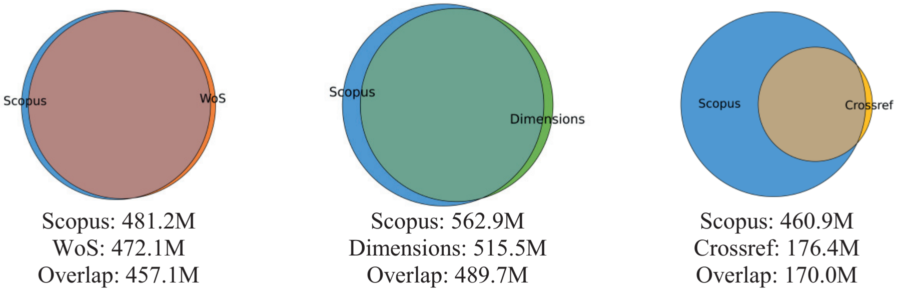


Figure 7. Overlap of citation links between Scopus and the other data sources.

Figure 7 makes clear that Dimensions has an important advantage over Crossref. Our earlier results indicate that Dimensions and Crossref have a similar coverage of documents, but Figure 7 shows that Dimensions provides access to many more citation links than Crossref. Although Dimensions relies strongly on data from Crossref, it enriches this data in various ways, in particular by adding citation links, but also by adding abstracts, affiliation data, and so on.

Analysis of incompleteness or inaccuracy of citation links

An important explanation for the discrepancies in the citation links covered by the various data sources is that for some documents no reference list is available in some of the data sources. Missing reference lists are an important explanation for the 73 million citation links in Scopus for which there is no corresponding citation link in Dimensions. For 52 million of these citation links (70.7%), the citing document does not have a reference list in Dimensions. In Crossref, missing reference lists are a major problem. Of the 291 million citation links in Scopus for which no corresponding citation link can be obtained from Crossref, 257 million

(88.4%) are due to missing reference lists in Crossref. Either these reference lists have not been deposited in Crossref at all or they have been deposited but they have not been made openly available. In WoS, almost all documents have a reference list. Of the 24 million citation links in Scopus for which there is no corresponding citation link in WoS, only 0.4 million (1.5%) are due to missing reference lists in WoS. Finally, in Scopus, the problem of missing reference lists is more significant than in WoS but less serious than in Dimensions and Crossref. About one-quarter of the citation links in WoS, Dimensions, and Crossref for which there is no corresponding citation link in Scopus are due to missing reference lists in Scopus. For instance, due to missing reference lists in Scopus, there are 3.9 million citation links in WoS without a corresponding citation link in Scopus.

In earlier work (Van Eck & Waltman, 2017; see also Olensky, Schmidt, & Van Eck, 2016), we studied inaccuracies of citation links in WoS and Scopus. For WoS, three problems were identified. First, some references are missing in the reference lists of documents in WoS. Second, sometimes there is an error in a reference in WoS, such as an incorrect publication year or volume number. Third, some references in WoS have been incorrectly matched with a cited document, leading to so-called phantom citations (García-Pérez, 2010). For Scopus, the opposite problem was identified. Some references incorrectly have not been matched with a cited document, even though all information needed to make a match seems to be available.

To get more insight into inaccuracies of citation links in Dimensions, we manually examined a number of discrepancies in the citation links covered by Dimensions and Scopus. Like in Scopus, we found some examples in Dimensions in which a citation link incorrectly has not been established. In addition, we found a few cases in which Dimensions provides a citation link to a journal article while the citing document actually refers to a different version of the document, such as a preprint or a paper in a conference proceedings. A citation link in Dimensions may point to a journal article even if the proceedings paper that is actually cited is covered by Dimensions.

Conclusions

We have presented a large-scale comparison of four multidisciplinary bibliographic data sources: WoS, Scopus, Dimensions, and Crossref. Our main conclusions can be summarized as follows:

- Comparing Scopus and WoS, it turns out that Scopus covers a large number of documents that are not covered by WoS, including documents with substantial numbers of references and citations. Almost all journal articles covered by WoS are also covered by Scopus. However, WoS covers meeting abstracts and book reviews, which are not covered by Scopus. A substantial share of the proceedings papers covered by WoS are not covered by Scopus either.
- The results of the comparison of Scopus with Dimensions and Crossref are more difficult to interpret. This is partly due to limitations of the document type classifications of Dimensions and Crossref. These classifications do not distinguish between different types of documents published in journals. In the case of Crossref, this is also partly due to limitations of the DOI-based matching procedure (see below). Dimensions and Crossref turn out to have a similar coverage of documents. This illustrates the strong reliance of Dimensions on data from Crossref. Dimensions and Crossref cover a large number of documents that have been published in journals and that are not covered by Scopus. However, some of these documents are of little scientific significance (e.g., the list of editorial board members of a journal or the cover of a journal issue). Dimensions and Crossref also cover many book chapters that are not covered by Scopus. Some of the documents that are covered by Dimensions and Crossref and not by Scopus have received a substantial number of citations.

Scopus covers quite some proceedings papers that are not covered by Dimensions and Crossref.

- The overlap of documents between the different data sources is smaller in the social sciences and humanities than in other disciplines. The disciplinary classifications of Dimensions and especially Crossref have significant shortcomings.
- All data sources suffer from problems of incompleteness and inaccuracy of citation links. The problem of incompleteness of citation links is more significant in Dimensions than in WoS and Scopus. Dimensions also does not provide data for references that have not been matched with a cited document. This makes it more difficult to analyze the quality of citation links in Dimensions and to correct inaccuracies. In Crossref, incompleteness of citation links is a major problem. To a large extent, this is caused by publishers that deposit references in Crossref but do not make these references openly available. Crossref does take these closed references into account in the citation counts that it provides for documents (Heibi, Peroni, & Shotton, 2019). This for instance explains why Harzing (in press) concludes that Crossref has "a similar or better coverage" of citations than WoS and Scopus.

How the differences between the data sources should be assessed depends on the purpose for which the data sources are used. For many purposes, a broad coverage of documents is valuable, for instance to make sure that locally relevant research is taken into account (e.g., Hicks, Wouters, Waltman, De Rijcke, & Rafols, 2015). However, for other purposes, it may be desirable to work within a more restricted universe of documents (e.g., López-Illescas, de Moya Anegón, & Moed, 2009). For instance, to enable meaningful international comparisons of universities, documents that have not been published in international scientific journals are deliberately excluded from the calculation of the bibliometric statistics reported in the CWTS Leiden Ranking (www.leidenranking.com).

Limitations

Our work has a number of important limitations. First, we have performed pairwise comparisons between Scopus and the other data sources. WoS, Dimensions, and Crossref have not been compared directly with each other. Second, in the case of WoS, the Emerging Sources Citation Index and the Book Citation Index have not been included in the analysis. Third, in the case of Crossref, matching with Scopus has been done based exclusively on DOIs. Because of missing DOIs in Scopus (Gorraiz et al., 2016), there are probably quite a lot of documents in Crossref that incorrectly have not been matched with documents in Scopus. We are currently implementing a more sophisticated procedure for matching documents in Crossref with documents in Scopus.

Finally, there are many features of bibliographic data sources that we have not taken into account. For instance, the completeness and accuracy of abstracts, affiliation data, and funding data has not been analyzed. Also, other aspects of bibliographic data sources, such as the conditions under which a data source can be used, the cost of the use of a data source, and the degree to which a data source provides up-to-date data, have not been considered.

Acknowledgement

We would like to thank Susanne Steiginga from Scopus and Christian Herzog from Dimensions for their helpful comments on an earlier draft of this paper.

References

Bilder, G. (2019, February 5). Underreporting of matched references in Crossref metadata [Blog post]. Retrieved from https://www.crossref.org/blog/underreporting-of-matched-references-in-crossref-metadata/

Bornmann, L. (2018). Field classification of publications in Dimensions: A first case study testing its reliability and validity. *Scientometrics*, *117*(1), 637–640.

García-Pérez, M.A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of *h* indices in Psychology. *JASIST*, *61*(10), 2070–2085.

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics*, *10*(1), 98–109.

Harzing, A.-W. (in press). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*.

Heibi, I., Peroni, S., & Shotton, D. (2019, February 7). Crowdsourcing open citations with CROCI – An analysis of the current status of open citations, and a proposal [Blog post]. Retrieved from https://opencitations.wordpress.com/2019/02/07/crowdsourcing-open-citations-with-croci/

Herzog, C., & Lunn, B.K. (2018). Response to the letter 'Field classification of publications in Dimensions: A first case study testing its reliability and validity'. *Scientometrics*, *117*(1), 641–645.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, *520*, 429–431.

Hook, D., Porter, S., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, *3*, 23.

López-Illescas, C., de Moya Anegón, F., & Moed, H.F. (2009). Comparing bibliometric country-by-country rankings derived from the Web of Science and Scopus: The effect of poorly cited journals in oncology. *Journal of Information Science*, *35*(2), 244–256.

Martín-Martín, A., Orduna-Malea, E., & López-Cózar, E. D. (2018). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: A multidisciplinary comparison. *Scientometrics*, *116*(3), 2175–2188.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228.

Olensky, M., Schmidt, M., & Van Eck, N.J. (2016). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of Science. *JASIST*, *67*(10), 2550–2564.

Orduña-Malea, E., & Delgado-López-Cózar, E. (2018). Dimensions: Re-discovering the ecosystem of scientific information. *El Profesional de la Información*, *27*(2), 420–431.

Schnell, J.D. (2017). Web of Science: The first citation index for data analytics and scientometrics. In F.J. Cantu-Ortiz (Ed.), *Research analytics: Boosting university productivity and competitiveness through scientometrics* (pp. 15–29). Taylor & Francis.

Schotten, M., el Aisati, M., Meester, W.J.N., Steiginga, S., & Ross, C.A. (2017). A brief history of Scopus: The world's largest abstract and citation database of scientific literature. In F.J. Cantu-Ortiz (Ed.), *Research analytics: Boosting university productivity and competitiveness through scientometrics* (pp. 31–58). Taylor & Francis.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proc. of the 24th International World Wide Web Conference* (pp. 243–246).

Valderrama-Zurián, J.C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics*, *9*(3), 570–576.

Van Eck, N.J., & Waltman, L. (2017). Accuracy of citation data in Web of Science and Scopus. In *Proc. of the 16th International Conference of the International Society for Scientometrics and Informetrics* (pp. 1087–1092).

Van Eck, N.J., Waltman, L., Larivière, V., & Sugimoto, C. (2018, January 17). Crossref as a new source of citation data: A comparison with Web of Science and Scopus [Blog post]. Retrieved from https://www.cwts.nl/blog?article=n-r2s234

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364.

Measuring disagreement in science

Dakota Murray[1], Wout Lamers[2], Kevin Boyack[3], Vincent Larivière[4], Cassidy R. Sugimoto[1], Nees Jan van Eck[2], Ludo Waltman[2]

[1] *dakmurra@iu.edu; sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, USA

[2] *w.s.lamers@cwts.leidenuniv.nl; ecknjpvan@cwts.leidenuniv.nl; waltmanlr@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, Leiden, Netherlands

[3] *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, USA

[4] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

Abstract

Dispute in science is central to the production of new knowledge. Such disputes leave traces in scholarly documents, generally through the form that is taken by citations. Based on the full text of scholarly papers from the Elsevier ScienceDirect database published between 1980 and 2016, this paper develops a methodology for investigating disagreement in science. Several signal phrases of disagreement are tested, and two are used ("contradict" and "conflict", with the filter phrases "studies" or "results") to assess the prevalence of disagreement across position within a paper and across disciplines. Results show that disagreement is relatively more common in the introduction and discussion sections of papers, as well as in fields of biomedical sciences, health sciences, social sciences, and humanities.

Introduction

Scientific disputes are central to the creation of new knowledge. More than 350 years ago, Robert Boyle and Thomas Hobbes debated the meaning of experimental results produced using Boyle's newly-created air pump; from this controversy emerged the basis of modern scientific research (Shapin & Schaffer, 1985). Scholars have long been interested in studying controversy as it relates to the production of knowledge at the individual level (Latour, 1988), and the macro-development of science (Kuhn, 1962) and, often, make explicit norms that remain otherwise implicit (Gingras, 2014). More recently, doubts over scientific findings, such as for climate change research, has led scholars to measuring the degree of consensus of specific research areas (e.g., Oreskes, 2004; Shwed & Bearman, 2010). Information scientists have also studied disagreements among scientific literature, leveraging bibliometric tools to understand the development of scientific fields (Evans, 2007), characterize their differences (Fanelli & Glänzel, 2013), predict future scientific impact (Radicchi, 2012), measure uncertainty surrounding scientific claims (Chen et al., 2018), and to classify the function of citations (Catalini, Lacetera, & Oettl, 2015; Moravcsik & Murugesan, 1975).

This paper uses the full-text of scholarly publications to explore the degree of *controversy*, *disagreement*, and *dissonance* (henceforth referred to only as *disagreement*) in scientific literature. We examine sentences containing citations and identify a set of cue phrases that broadly signal disagreement between citing and cited paper, or within the cited literature. We assess the reliability of these cue phrases and use the top performing phrases to identify instances of disagreement in citing papers. This analysis provides a preliminary analysis of the degree of disagreement within fields. We also hope to establish a methodological basis for future analyses of the disciplinary, temporal, and spatial aspects of scientific controversies through the lens of textual analysis.

Operationalizing disagreement

We used a broad operationalization of disagreement between a citing and cited paper, or within two cited papers. Under our definition, we consider such disagreement to include direct contradiction between conclusions, as well as disagreement based on incompatible model assumptions (even if findings are not in conflict). Examples of the types of disagreement we consider are shown in table 1.

Table 1. Examples of our notion of disagreement

Citation sentence	***Type of disagreement***
A: *"Coffee causes cancer"* ***B:*** *"Coffee does not cause cancer"*	Direct disagreement in conclusions
C: *"Based on a model which assumes that coffee increases the probability of cancer by 50%, the predicted life expectancy for the Dutch population equals 80 years."* ***D:*** *"Based on a model which assumes that coffee does not cause cancer, the predicted life expectancy for the Dutch population equals 85 years."*	Disagreement as a result of incompatible model assumptions, not necessarily because of conclusions
E: *"There remains controversy in the scientific literature over whether or not coffee is associated with an increased risk of cancer (A, B, C, D)"*	Disagreement in the broader literature

The main challenge is to obtain accurate signals of disagreement. For this purpose, we focus on sentences that include a citation, and that include a word or sequence of words signaling disagreement. We refer to this sequence of words as a *disagreement signal phrase*. In addition, other words appearing near this phrase may reinforce the likelihood that a disagreement signal phrase represents true disagreement—we call such words *disagreement filter phrase*s

Data

We used data from the Elsevier ScienceDirect database hosted at the *Centre for Science and Technology Studies* at Leiden University. This data contains the full-text of nearly five million English-language research articles, short communications, and review articles published between 1980 and 2016. Sentences containing in-text citations are extracted from the full-text of these articles following the procedure outlined by Boyack et al. (2018).

Reliability

We considered four disagreement signal phrases: *contradict*, *contrast*, *conflict*, and *differ*. Queries include morphological variants of disagreement signal phrases, such that we include terms such as "conflict", "conflicted", and "conflicting". For each term, we used them as a standalone term (with no additional filters applied), and with one of four disagreement filter phrases: "ideas", "methods", "studies", and "results". Disagreement filter phrases must appear within a four-word window of the signal.

For each combination of disagreement signal and filter phrase, we randomly sampled 100 citation sentences from the full-text database that contain the combination of terms. Disagreement filter phrases must occur within a four-word window of the corresponding signal phrase. For each set of 100 sentences, two independent coders assigned a value of *valid*, or *invalid*, where valid means that the sentence represents a true example of our notion of *disagreement*. In some cases, the proportion of *valid* instances was so low that coders did not code all 100 instances. Consider for example the four sentences listed below: the first is invalid, because the signal term, “conflict”, refers to an object of study, and not a scientific dispute; the second sentence is also invalid because the term “conflicting” refers to results within a single study, not between studies; the third and fourth sentence are both examples of sentences that would be marked as valid.

1. **Invalid:** “To facilitate conflict management and analysis in Mcr (…), the Graph Model for Conflict Resolution (GMCR) (…) was used.”
2. **Invalid:** “The 4-year extension study provided ambiguous […] and conflicting post hoc […] results.”
3. **Valid:** “These observations are rather in contradiction with Smith et al.’s […].”
4. **Valid:** “Although there is substantial evidence supporting this idea, there are also recent conflicting reports (…).”

The validity—the proportion of sentences coded by both reviewers, and identified as *valid* by both reviewers—was calculated for each query (Figure 1). We find that the best performing disagreement signal phrases are “contradict” and “conflict”, and that these perform best when they occur alongside the disagreement filter phrases “studies” or “results”.

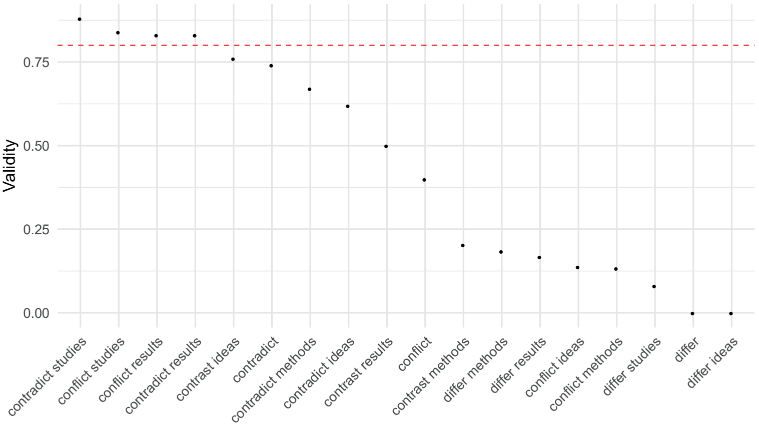


Figure 1: Validity of eighteen combinations of cue and signal words. We plot a threshold horizontal line of 0.80, showing the cut-off for choosing the top performing terms.

Analysis of “conflict” and “contradict” sentences

Due to their high validity, we focus our analysis on citation sentences containing the disagreement signal phrases “conflict” or “contradict”, which occur alongside filter phrases “studies” or “results”. There are, respectively, 62,667 and 63,035 “conflict” and “contradict” sentences in the text of our set of publications, each representing 0.04% of all citing sentences. During the period 1998–2016, the percentage of citing sentences having “conflict” or “contradict” has remained fairly stable over time. Below, we first report an analysis of the location of “conflict” and “contradict” sentences within the full text of publications. We then present a disciplinary comparison in which we examine the distribution of “conflict” and “contradict” sentences across scientific fields.

Location in full text

Figure 2 shows the distribution of “conflict” and “contradict” sentences within the full text of publications. The horizontal axis indicates text progression, expressed relatively to the total length of the full text of a publication. The vertical axis indicates the number of “conflict” or “contradict” sentences in a specific part of the full text of a publication relative to the total number of “conflict” or “contradict” sentences in the entire full text of a publication. The figure also shows the distribution of all citing sentences.

Consistent with earlier work (Bertin et al., 2016; Boyack, Van Eck, Colavizza, & Waltman, 2018), citing sentences are overrepresented in the early and to a lesser extent, the end parts of publications. For “conflict” and “contradict” sentences, this pattern is more pronounced. “contradict” sentences are especially overrepresented at the end of a publication, though less so in the early sections. In biomedical publications, the discussion of related work is often presented in the conclusion, thus leading to a large number of citing sentences at the end of publications (Boyack et al., 2018). As we will see below, “conflict” and “contradict” sentences occur most often in biomedical literature; potentially explaining why these sentences are overrepresented towards the ends of publications.

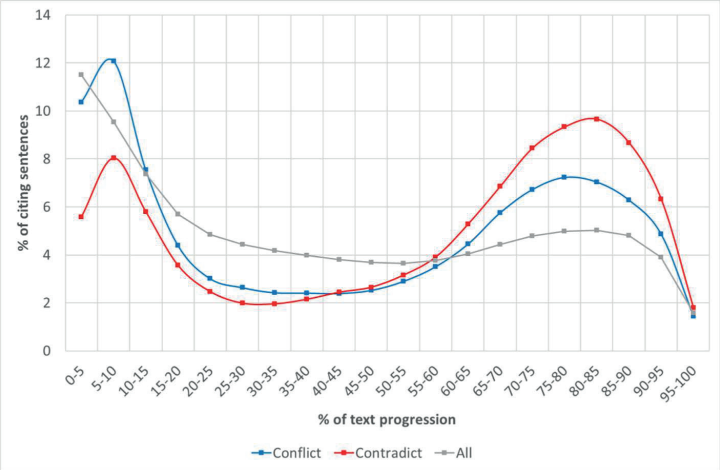


Figure 2. Distribution of “conflict” and “contradict” sentences within the full text of publications.

Disciplinary comparison

Our disciplinary comparison relies on all 2000-2017 publications indexed in the Web of Science, which were clustered into 868 fields through citation links, following the methodology introduced by Waltman and Van Eck (2012). For each field, we queried our Elsevier corpus and counted the total number of citing sentences, as well as the number of “conflict” and “contradict” sentences. Figure 3 presents visualizations of the 868 fields (nodes), produced using the VOSviewer software (Waltman & Van Eck, 2012). The size of a field indicates the total number of citing sentences in the field. The distance between two fields reflects the relatedness of the fields in terms of citation links: the smaller the distance between two fields, the larger the number of citation links between publications in the two fields. Most importantly, the color of a field indicates the relative number of “conflict” or “contradict” sentences in the field, expressed as the binary logarithm of the ratio of the actual and the expected number of “conflict” or “contradict” sentences. A field is colored blue if the number of “conflict” or “contradict” sentences is lower than the expected value, grey if it equals the expected value, and red if it is above the expected level.

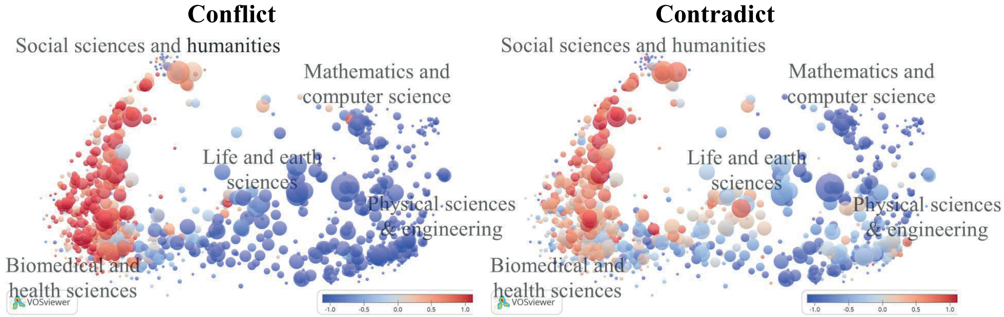


Figure 3. Distribution of "conflict" (left) and "contradict" (right) sentences over fields.

"Conflict" sentences were strongly concentrated in the biomedical and health sciences and in certain fields in the social sciences and humanities (roughly the top left and bottom left of each visualization); this was in strong contrast to the physical sciences, computer science, and mathematics (roughly top right and bottom right). In many of these fields, the number of "conflict" sentences was twice or more below expectation, whereas in the biomedical and health sciences, social sciences, and humanities, the number of conflict sentences was more than twice what was expected.

A similar, though less pronounced, trend was apparent for "contradict" sentences, with smaller disciplinary differences than for "conflict" sentences. In most biomedical and health fields, the number of "contradict" sentences was above expectation, but there were also some fields that fell below the expected number. Conversely, there were some fields in the life, earth, and physical sciences in which the number of "contradict" sentences was above expectation.

Table 2 lists the top 5 fields with the largest relative number of "conflict" and "contradict" sentences. We manually labelled each field by examining the titles of the journals with the largest number of publications in the field. The field labelled *International relations* had the largest relative number of "conflict" sentences. However, this was a methodological artefact. *International relations* studies political conflicts, and the term "conflict" referred mainly to conflicts as an object of study rather than conflicts in the scientific literature. Leaving out this field, all fields listed in table 2 were in the biomedical sciences and in psychology.

Table 2. Top 5 fields with the largest relative number of "conflict" and "contradict" sentences.

Conflict			**Contradict**		
Label	Absolute	Relative	Label	Absolute	Relative
International relations	414	3.11	Bioelectromagnetics	91	2.09
Cancer	67	3.04	Laboratory animals	45	1.71
Sleep medicine	229	2.26	Child psychology	89	1.50
Cardiothoracic surgery	143	2.25	Psychological methods	74	1.49
Cardiology	681	2.22	Cancer	23	1.48

Conclusion

This exploratory study assessed the degree to which *disagreement* between scientific literature exists across scientific fields. We defined a novel indicator of disagreement, and assessed the validity of a set of cue phrases that indicate disagreement between a citing and cited paper, or within the literature cited in a paper. We identified all citation sentences in our dataset that contained one of the two cue phrases ("contradict" and "conflict", with the filter phrases "studies" or "results"). Using these data, we investigated how the incidence of disagreement signal phrases differed based on their position in papers, noting key differences between sentences containing "contradict" or "conflict" signal phrases. We also investigated the incidence of disagreement signal phrases across 868 scientific fields represented by the Web of Science. We observed that the number of citing sentences containing disagreement signal phrases occurred above expected levels in the biomedical sciences, health sciences, social sciences, and humanities, and less than expected in the fields of mathematics, computer sciences, and physical sciences; we also noted that disciplinary differences were more extreme for "conflict" than for "contradict" sentences. Finally, we found that the fields with the largest proportion of conflict and contradict sentences were in the biomedical sciences and psychology.

This study marks the first step in an investigation into how disagreement and controversy function in scientific discourse. In future work, we will refine our notion of disagreement and expand our analysis to include additional disagreement signal and filter phrases. Building on this method, we hope to further investigate the extent to which disagreement and controversy relate to scientific impact; the evolution of the incidence of disagreement over time; how disagreement varies according to the country and institution of affiliation; and the incidence of disagreement as a function of the demographic characteristics of authors.

References

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology, 67(1), 164-177.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73.

Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(45), 13823–13826.

Chen, C., Song, M., & Heo, G. E. (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics*, *12*(1), 158–180.

Evans, J. H. (2007). Consensus and knowledge production in an academic field. *Poetics*, *35*(1), 1–21.

Fanelli, D., & Glänzel, W. (2013). Bibliometric Evidence for a Hierarchy of the Sciences. *PLOS ONE*, *8*(6), e66938.

Gingras, Y. (Editor) (2014) Controversies. Accords et désaccords en sciences humaines et sociales, Paris: CNRS. 278 p

Latour, B. (1988). *Science in Action: How to Follow Scientists and Engineers Through Society* (Reprint edition). Cambridge, Mass: Harvard University Press.

Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*(1), 86–92.

Oreskes, N. (2004). The Scientific Consensus on Climate Change. *Science*, *306*(5702), 1686–1686.

Radicchi, F. (2012). In science "there is no bad publicity": Papers criticized in comments have high scientific impact. *Scientific Reports*, *2*.

Shapin, S., & Schaffer, S. (2011). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (With a New introduction by the authors edition). Princeton, N.J: Princeton University Press.

Shwed, U., & Bearman, P. S. (2010). The Temporal Structure of Scientific Consensus Formation. *American Sociological Review*, *75*(6), 817–840.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Waltman, L., & Eck, N. J. van. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.

An empirical analysis on the relationship between publications and academic genealogy

Rogério Mugnaini[1], Rafael J. P. Damaceno[2] and Jesús P. Mena-Chalco[3]

[1] *mugnaini@usp.br*
University of São Paulo, School of Communication and Arts, Dept. of Information and Culture, Av. Prof. Lúcio Martins Rodrigues 443, São Paulo, SP 05508-020 (Brazil)

[2] *rafael.damaceno@ufabc.edu.br*
Federal University of ABC, Center for Mathematics, Computation and Cognition, Av. dos Estados 5001, Santo Andre, SP 09210-580 (Brazil)

[3] *jesus.mena@ufabc.edu.br*
Federal University of ABC, Center for Mathematics, Computation and Cognition, Av. dos Estados 5001, Santo Andre, SP 09210-580 (Brazil)

Abstract

In the belief that collaboration between advisor and students is a means of following a scientific path for the discovery of new knowledge, this study examines the relationship between bibliometric indicators for publications and academic genealogy. In this study, we analysed the curricular information of more than 40,000 PhD advisors, registered in a huge Curriculum Vitae dataset. This involved displaying different patterns of academic fecundity and publications, concerning several areas and mentoring ages. It was found that productivity in co-authorship with academic sons is an established practice in the hard sciences, while in the soft sciences it is only a reality for researchers until the 25 years (mentoring age). In addition, in the case of the output produced without the participation of the students, there was a constant distribution among the mentoring age groups (with the exception of Agricultural Sciences and Engineering, where there was a gradual decline over the period). Finally, there was a number of advisors that performed best in fecundity but worst in production, which suggests that the involvement in mentoring impairs the advisor's capacity for research. It can be concluded that a separate analysis of the researchers' output is needed, since student participation may be important for an assessment of scholarly performance.

Introduction

Brazilian scholarly output is concentrated in universities, particularly in public universities, where almost all of the Graduate Programs are run. The Coordination of Improvement of Higher Level Personnel (CAPES) has been the national agency responsible for the funding, evaluation, and support of Brazilian graduate education since 1976. The accredited researcher in this system will have his/her academic performance evaluated on the basis of teaching experience, the training of master´s and doctoral students and scientific production. Moreover, although the production of the researcher is the main object of the assessment, the students´ output is also evaluated, as well as their collaboration with a supervisor in producing co-published articles (Oliveira & Amaral, 2017).

Scholarly output can be observed in different ways. In this paper, we focused in both publications and academic mentoring in the training of young researchers. We believe that these terms are compatible when we look at the scholarly output as well as the relations formed by student-advisor pairs over time, which allows us to identify and analyse patterns in the academic genealogy.

Studies based on different approaches have focused on the relationship between publications and academic mentoring, to find out if the students are now publishing articles in greater quantity (Green & Bauer, 1995; Pinheiro, Melkers & Youtie, 2014). They are also concerned with analysing the relationship of co-authored publications between advisor and student (Tuesta et al, 2015) and investigating the involvement and positive influence of advisors in doctoral studies (Horta & Santos, 2016).

In light of this, Larivière (2012) analysed the involvement of more than 27,000 doctoral students in peer review publications. The author states that the grouping of doctoral students in research teams both assists and and encourages students to take part in different research projects. At the same time, Qui et al (2017) showed that collaboration with first-class scientists significantly improves young researchers' careers. They provided evidence that the benefits of working with an outstanding scientist are more noticeable in the early stages of a young student's career. These factors explain why it is important for advisors to give encouragement to students to conduct scientific research very early on.

Apart from the questions of publications, as well as the output that can derive from the mentoring process, many factors can lead to academic success (Reskin, 1977; Bäker, 2015). For this reason, the concern of a supervisor and of scientific policies should be to ensure the continuity of scientific knowledge throughout the generations, which may have more value than the publications or awards that a scholar might receive (Andraos, 2005). In light of this, it is not a question of stimulating the academic fecundity of advisors (or academic fathers), and thus ensuring that their students (or academic sons) are, for example, productive, but of encouraging advisors to foster productive and also fecund sons (Malmgren, Ottino & Amaral, 2010; Heinisch & Buenstorf, 2018). This means that the research output is no longer the target, but rather, the focus is on the widening and transmission of capacity (Bozeman, Dietz & Gaughan, 2001).

This study has sought to describe the relationship between the quantitative indicators of publications and academic genealogy. However, although these factors should be correlated, there is a need for comparing its behaviour, not only among the different areas of knowledge but also among scholars of different age groups. Despite the correlation, this supports the hypothesis that there might be advisors whose performance in publications and academic mentoring performance are antagonistic, because there are those who concentrate their efforts, in one area, to the detriment of another.

We have used a dataset of academic curricula nationwide, which contains the student-advisor pairs, permitting to establish the academic genealogy and respective metrics. Our decision was not to carry out a longitudinal study (Levin & Stephan, 1991), which despite being ideal for measuring the effect of specific factors over time, would be impossible for comparing groups of researchers from different generations – these have been subjected to different scientific policies.

Material and Methods

In this study, we measured both the scientific publications and the academic mentoring relationships of PhD researchers working in Brazil. With regard to scientific publications, we count the total number of publications of each advisor in scientific journals, conferences, book chapters, and books and determine how many of these publications are co-authored with their academic sons. Concerning the academic mentoring, i.e., the training of new researchers, three genealogical metrics were used, namely, academic fecundity, descendants (offspring) and the genealogical index.

Fecundity (F) refers to the number of mentoring relationships that a researcher has already established. Descendants (D) indicate the number of mentoring relationships established with the students, and also the future relationships of these students with their own students. It takes into account all the generations of a researcher, i.e., it also includes the direct academic sons, the indirect relationships (grandsons, great-grandsons, and so on). The genealogical index (GI) of an academic is defined as the largest number of *g* sons of an academic that have at least *g* sons (Rossi et al., 2017).

The procedure shown in Figure 1 is adopted to analyse the academic genealogy combined with scientific publications. It also contains a flowchart divided into five stages: (i) collecting and cleaning, (ii) extraction of scientific publications, (iii) merging, (iv) selection of researchers and (v) analysis (see Figure 1). The following sections describe these stages, which generate a dataset containing researchers with information regarding their academic genealogy and scientific publications.

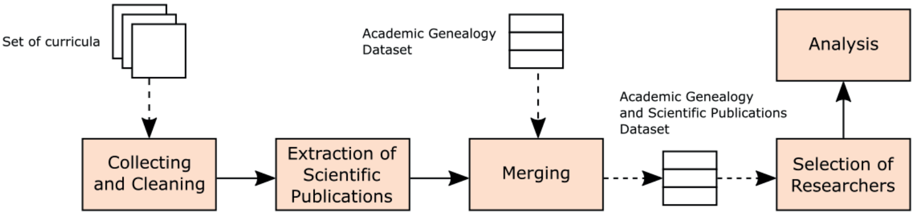

Set of curricula
Academic Genealogy Dataset
Analysis
Academic Genealogy and Scientific Publications Dataset
Collecting and Cleaning
Extraction of Scientific Publications
Merging
Selection of Researchers

Figure 1. A flowchart that shows the five stages of the method applied in this work: collecting and cleaning, extraction of scientific publications, merging, selection of researchers and analysis.

Collecting and Cleaning

The study started from an existing dataset consisting of the PhD researchers working in the Brazilian Graduate Program, which was compiled in our previous study (Damaceno et al. 2019). The process of compiling this dataset involved drawing on information collected from the set of academic curricula of PhD researchers registered in the Lattes Platform (a large Curriculum Vitae dataset). In this dataset, there is information on each researcher's field of study (or areas of knowledge), individual identifier, academic degrees, and mentoring relationships, as well as the genealogical metrics of academic fecundity, the descendants and genealogical index. This dataset also contains the full names and curricular identifiers of the researchers' students (academic sons).

When adding information about the scientific publications in the genealogical dataset, we had to collect the same set of curricula that was to form the academic genealogy. It is essential to use the same set of curricula, i.e., obtained at the same time as the original dataset was formed, since the scientific publications must correspond with the same time as the information about the academic genealogy. After we collected these curricula (dated August 2017), we carried out the data cleaning. All diacritic marks were excluded as well as the characters with accents were transformed to English alphabet. All the characters were transformed to the lowercase. Any articles, books or book chapters without a title were not taken into account.

Extraction of Scientific Publications

The curriculum of each researcher has a section called “Bibliographic Production”, from which we extracted all the information that refers to articles, books or book chapters. In the case of articles, we only took into account the full papers published in journals or in the proceedings of a conferences (expanded abstracts were not included). In the case of books or book chapters, only full texts, encyclopedias, catalogs or collections in both printed or digital versions, were included. The total number of scientific publications of a researcher was calculated as the sum of all the articles, books and book chapters that he/she had published.

Additionally, this study counts the works produced in collaboration with academic sons. Each researcher's publication has a list of its co-authors' names (the initial of the forenames and the complete surname) and its co-authors' identifiers (IDs that identify their curricula in the dataset). A comparison was made between the identifiers and names to check if a researcher's publication was co-authored with some of the researcher's academic sons. With regard to each researcher's publication, the co-authors' identifiers must be identical to some identifier in the list of the researcher son's identifiers (collected from the dataset) to ensure it was co-authored with academic sons. If a co-author does not have an identifier, his/her name must be in the list of the researcher son's names (initial of the first name and the complete surname). The same publication was only counted once since a researcher could have co-authored it with two or more students.

Merging

The process of incorporating data about scientific publications into the genealogical dataset relied on the individual identifier of each researcher in the Lattes Platform (also included in the genealogical dataset). Each scientific publication obtained in the last stage is linked to an individual identifier - the same that is included in the genealogical dataset. Hence, the process of including this information resulted in a dataset consisting of both the academic genealogy and the scientific publications for each researcher in the original dataset.

Selection of Researchers

The academic genealogy dataset, together with the scientific publications added to it, contains information regarding 271,370 PhD researchers. We only analysed a proportion of these that met two requirements: (i) researchers that have at least one mentoring relationship completed in the doctoral studies (ruling as an advisor), and (ii) researchers that have at least one publication since the year they completed their first mentoring relationship.

Analysis

The researchers were separated into eight groups that represent the eight areas of knowledge defined by CAPES, which are as follows: Agricultural Sciences (AGR), Biological Sciences (BIO), Engineering (ENG), Exact and Earth Sciences (EXA), Health Sciences (HEA), Humanities (HUM), Linguistics, Language & Literature and Arts (LIN) and Applied Social Sciences (SOC). We analysed the areas of knowledge globally, and in accordance with the mentoring age defined in this study as the time passed (in years) since a researcher has finished the mentoring of his/her first PhD student. There are ten mentoring age groups, which are as follows: 1 to 5, 6 to 10, 11 to 15, 16 to 20, 21 to 25, 26 to 30, 31 to 35, 36 to 40, 41 to 45 and 46 to 50 years.

Two metrics were used to analyse the scientific publications: Production with Academic Sons (PAS) and the remaining part of the Total Production (TP), calculated by TP - PAS. TP refers to all the work published by a researcher, since he/she completed the first mentoring relationship (ruling as an advisor). TP - PAS refers to the part of these scientific publications that a researcher co-authored with his/her students. For both measurements, we only took into account book chapters or entire books and the full papers published in journals or conferences. Further, we calculated a coefficient, that is the ratio between each scientific publication metrics (PAS and TP - PAS) and the “Fecundity” metric.

Dataset

The dataset obtained as a result of the five stages previously described, contains information about the knowledge area of the researchers, such as their mentoring age, the total number of their scientific publications and the percentage of these publications that was undertaken with students. These data are divided into two groups: N1, which represents all the academics that met all the requirements set out in the “Selection” section, and N2, a subset of N1. This includes meeting another requirement: researchers that have a score higher than zero in the genealogical index (or those that have at least one grandson).

Table 1 shows the total number of researchers and the median and average values for the mentoring age and of the researchers for N1, grouped by their area of knowledge. N1 represents the main dataset containing 40,368 researchers. Information about N2 is also shown, in which there are 10,996 researchers.

Table 1. Number and percentage of academics and their average and median mentoring age for each area of knowledge. The last three columns show the number of academics and include a sub-dataset consisting of all the academics with genealogical index greater than or equal to 1.

	N1				*N2*		
Area	*Academics*		*Mentoring Age*		*Academics*		
	N	*%*	*Avg.*	*Med.*	*N*	*%*	*% N1*
AGR	4,012	9.94	11.79	10	1,065	9.68	26.54
BIO	6,023	14.92	12.44	10	1,703	15.49	28.27
ENG	4,371	10.83	13.11	12	1,224	11.13	28.00
EXA	6,693	16.58	13.16	11	1,804	16.41	26.95
HEA	7,004	17.35	12.40	11	2,032	18.48	29.01
HUM	6,337	15.70	11.09	9	1,698	15.44	26.79
LIN	2,223	5.51	11.40	10	579	5.26	26.04
SOC	3,705	9.18	10.63	9	891	8.10	24.05
All	40,368	100.00	12.13	10	10,996	100.00	27.24

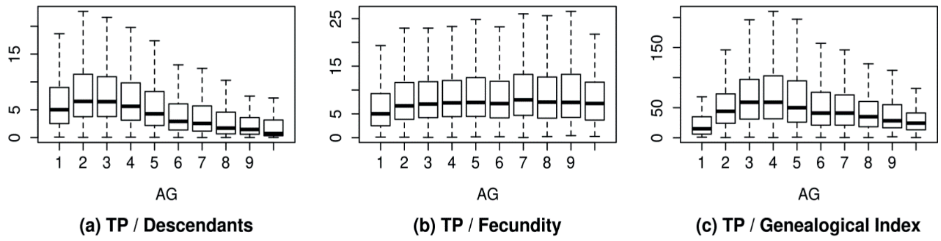
Results and discussion

First of all, we analysed the publication profile of the academics in the eight areas of knowledge (see Table 2). The TP does not reveal notable differences between the areas, while PAS shows a trend of HUM, LIN, and SOC to publish in a smaller quantity with students.

Table 2. TP, FP and percentage of FP for each area of knowledge.

Area	*TP*		*PAS*		*TP - PAS*	
	Avg.	*Med.*	*Avg.*	*Med.*	*Avg.*	*Med.*
AGR	64.45	39	26.46	9	37.99	25
BIO	49.36	29	19.27	6	30.09	20
ENG	78.77	51	32.95	12	45.82	31
EXA	52.65	30	18.53	5	34.12	21
HEA	62.46	38	19.67	6	42.79	27
HUM	38.38	23	6.46	1	31.92	20
LIN	29.75	17	2.27	0	27.49	16
SOC	43.14	25	7.77	1	35.38	22
All	53.49	31	17.41	4	36.08	22

We then proceeded to analyse the behaviour of publication coefficients concerning the genealogical metrics of groups of academics of different mentoring age groups. As can be seen in Figure 2b, the distributions of the different age groups have a very similar profile when account is taken of the total number of publications and fecundity, meaning there is no increase in productivity.



(a) TP / Descendants

(b) TP / Fecundity

(c) TP / Genealogical Index

Figure 2. Distribution of academics by Age Group (AG) and the three coefficients, as follows: (a) TP / Descendants, (b) TP / Fecundity, and (c) TP / Genealogical Index. The X axis represents the mentoring age groups and the Y axis the boxplot of the respective coefficient.

Figures 2a and 2c, consider the coefficients by weighing the publications, in terms of the number of descendants and genealogical index respectively. The profile shown is different, since older researchers accumulate more people in their lineage, and scientific publications do not increase in the same proportion (perhaps because co-publication largely occurs with their academic sons, who are their direct descendants). A decline was noted in both cases, beginning from the fourth age group (16-20 years) when this includes the number of offspring; and from the fifth age group (21 to 25 years), in the case of the genealogical index.

As these are cumulative variables, it should be noted what happens to each one, independently. On the one hand, some features of the distributions, such as the coefficients obtained with the descendants and fecundity metrics (Figures 2a and 2b), show absolute values in a very similar range. On the other hand, there is a difference shown by the descendants, which, in addition to declining for the older age groups, reduces their dispersion (mainly among researchers with a low number of scientific publications, denoting a more marked asymmetry).

This behaviour reveals that in the case of Figure 2a, the cumulative effect of descendants reduces the coefficient. However, in the case of Figure 2b, there was an increase in the advisors' productivity. In other words, if their extra-mentoring publications showed a significant growth, the effect of the coefficients on the age groups would be inversely proportional (in terms of increase and dispersion). In view of this scenario, we thought it would be of value to broaden the analysis of the fecundity variable, separating the scientific production in two parts: Production with Academic Sons (PAS) and the remaining publications of the fathers (TP - PAS), or extra-mentoring publications (those not co-authored by the academic sons).

Figure 3 illustrates that, in general, there is a clear difference between the range of variables, for hard and soft sciences. In the case of soft sciences, co-authorship with students is very low, which may be due both to the low level of collaboration in these areas and to the fact that the academic sons are less involved in the advisor's research (Larivière, 2012). SOC is the area with the largest range of production without academic sons' participation, among all the areas. This is due to the tenth age group (more than 45 years), whose productivity is significant - this profile is usually not observed in most studies, which restrict publications to journal articles. Among the hard sciences, BIO has the highest proportion of production in co-authorship with students, and EXA the lowest.

In light of the distributions of box-plots from hard sciences, it is clear that production without students' participation (BIO, EXA and HEA) is increasing in the first age groups, and remains constant throughout most of the groups (from the fifth age group). On the other hand, AGR fluctuates between the intermediate groups, while ENG decreases from the fourth age group - both show a significant reduction in dispersion among the older researchers. With regard to co-authorship with their academic sons, there is less productivity for the first age groups, which may be due to the fact that the younger advisors have not yet consolidated into groups or formed a research network that makes it easier for the students to be incorporated.

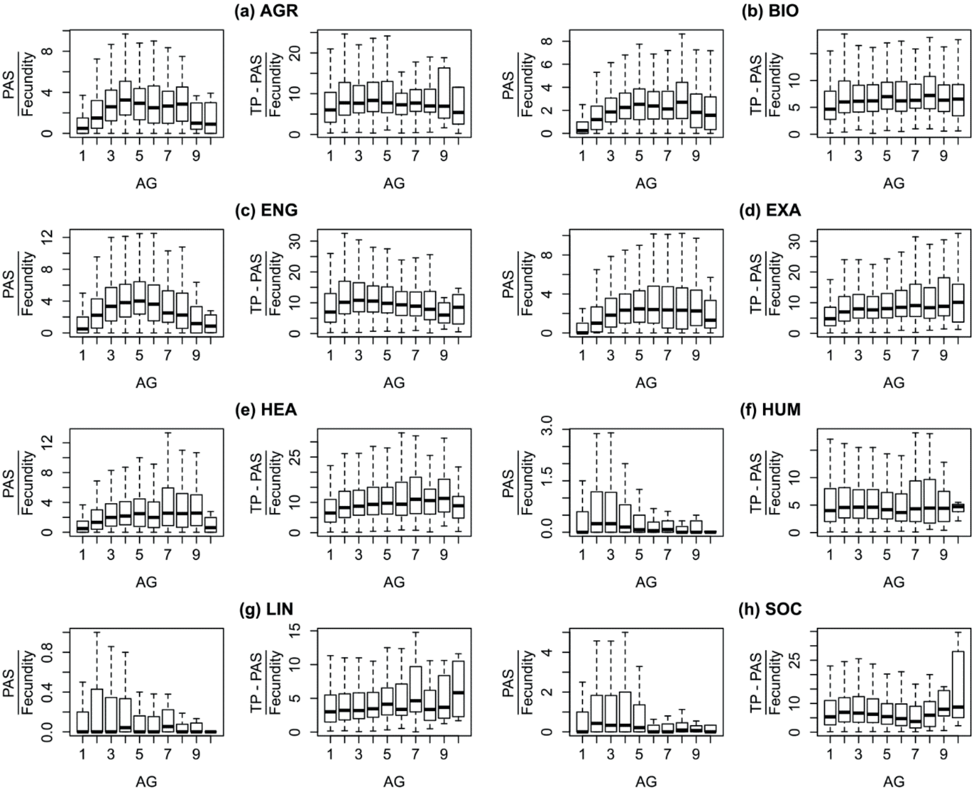


Figure 3. Distribution of academics by Age Group (AG), and coefficients of Production with Academic Sons (PAS) and and the remaining part of the production (TP - PAS). The academics are discretized in major areas of knowledge: (a) AGR, (b) BIO, (c) ENG, (d) EXA, (e) HEA, (f) HUM, (g) LIN and (h) SOC.

The soft sciences have an opposite profile, in which productivity in co-authorship with their academic sons is more pronounced among the younger researchers (up to 20 or 25 years), which can be attributed to the criteria governing the graduate scientific policy, which were laid down in 1998 and encourages co-authorship with students.

Thus, with regard to production without students' participation, it can be seen that, in addition to BIO, EXA and HEA, the soft sciences also showed a more constant distribution among the age groups. The fact that these areas represent about 80% of the total number of researchers explains the behaviour observed in Figure 2b.

Figure 4 shows the performance of the same pair of variables examined above, where the medians are in the scatter plots, with an arrow identifying the oldest age group. The series on the left side (y scale) shows the values for the parents in collaboration with their sons (PAS), while the series on the right side (y scale) relates to the remaining papers of the fathers (TP - PAS). There is an increase in the relationship between fecundity and the production variables that can be observed through almost all the age groups, with regard to the hard sciences. The main difference is that (except for ENG), while the fecundity declines significantly in the older age groups, the production in co-authorship with the students decreases more than the production without them. Among the social sciences: HUM shows a clear linear growth through almost all the age groups, in the case of production without the students, while LIN and SOC show a decline in fecundity among the older age groups. With regard to the production with the students, there is an increase until the third age group to HUM, and the fourth to SOC, while LIN shows than the median of production is zero for almost all the age groups.

It should be noted that, in contrast with the longitudinal approaches, which revealed that during academic life there is a decline in productivity - in terms of scientific publications (Levin & Stephan, 1991), this study covers the entire scientific publications of academics. For this reason, it is not possible to analyse the academics' careers. The comparisons between age groups are made with different groups of academics, which causes an increase in the number of publications originating from the growth in the number of scientific publications produced by the oldest academics.

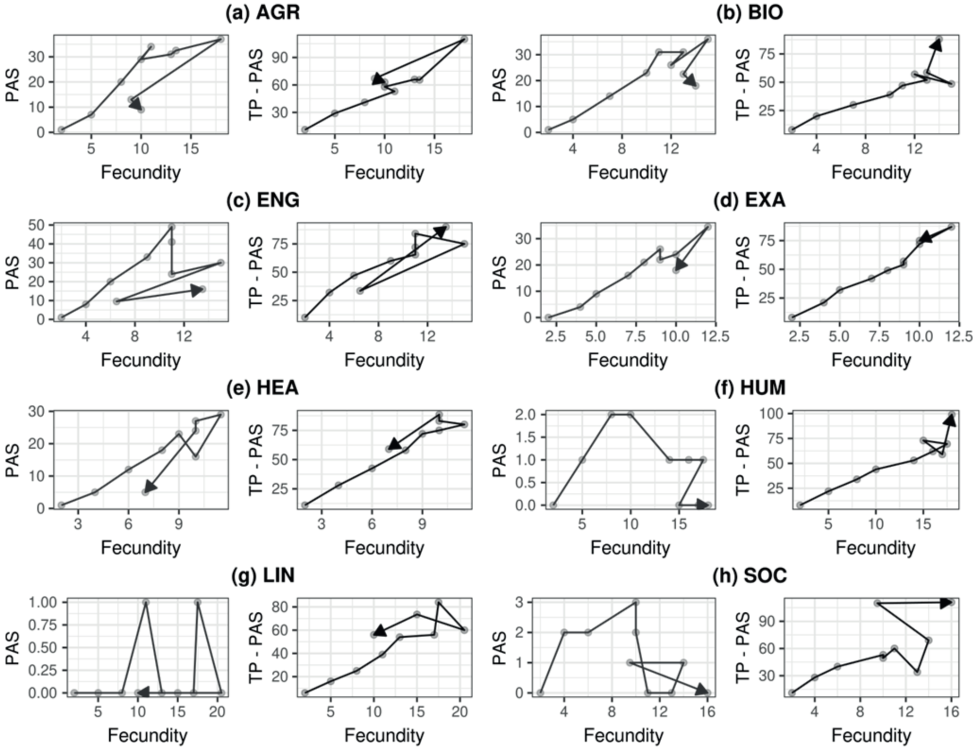


Figure 4. Distribution of academics by Fecundity (X), PAS (Y), and TP - PAS (Y). The academics are discretized in major areas of knowledge: (a) AGR, (b) BIO, (c) ENG, (d) EXA, (e) HEA, (f) HUM, (g) LIN and (h) SOC. Each point corresponds to the median of X and Y for the ten mentoring age groups concerned. The arrow represents age group 10.

Finally, some analytical factors should be noted with regard to one of the objectives of this study, which refers to researchers whose performance in production and fecundity shows a contrast (i.e., Q1 for the former and Q4 for the latter, or opposite). Moreover, the percentages of researchers whose performance in each of the measures is similar (i.e., Q1 or Q4 in both measures) are also displayed, since they represent the expected relationship between the variables (which is the association between them, since the more mentoring relationships there are, the higher the scientific output derived from this relationship).

An examination of Table 3, shows that the respective percentages of production without students' participation (TP - PAS) and fecundity, when performance is better (Q1 for both) and worse (Q4 for both), are between 13.06% and 18.03%. As for production in co-authorship with the academic sons, these percentages are between 16.22% and 23.26% - with bigger percentages in the last quartile of both variables. Additionally, it can be seen that the association between the variables is stronger when the production is co-authored with students.

When the areas are compared, it is evident that ENG, HUM and LIN show the lowest values concerning Q4 (both variables), and production without the participation of academic sons; while BIO, EXA and HEA are the biggest. In the case of Q1 (both variables), there is less variability, with HUM and LIN performing best.

When the production in co-authorship with the students is analysed, the percentages are slightly higher, with BIO and HEA having the biggest percentage in Q4 (both variables) and BIO and ENG in Q1 (both variables).

As noted in Table 2, the medians of PAS for the soft sciences are very low (at most, one), making it impracticable to determine Q4, which explains the presence of empty cells in Table 3 - and the same occurred with EXA, which even had a median of 5.

Table 3. Percentage of academics in the Q1 and Q4 with regard to: (a) (TP - PAS) vs. Fecundity (F) and PAS vs. Fecundity (F).

Area	$(TP - PAS)_{Q1} \wedge F_{Q4}$	$(TP - PAS)_{Q4} \wedge F_{Q1}$	$(TP - PAS)_{Q4} \wedge F_{Q4}$	$(TP - PAS)_{Q1} \wedge F_{Q1}$	$PAS_{Q1} \wedge F_{Q4}$	$PAS_{Q4} \wedge F_{Q1}$	$PAS_{Q4} \wedge F_{Q4}$	$PAS_{Q1} \wedge F_{Q1}$
AGR	0.52	2.04	14.98	13.06	0.17	0.17	20.09	17.07
BIO	0.50	1.61	17.8	13.76	0.33	0.08	23.04	18.28
ENG	0.41	1.49	16.93	13.09	0.41	0.09	22.19	17.25
EXA	0.49	2.29	18.03	13.67	-	-	-	-
HEA	0.50	1.83	17.99	13.34	0.43	0.16	23.26	16.22
HUM	0.38	1.59	16.10	14.69	-	-	-	-
LIN	0.40	1.80	15.29	15.38	-	-	-	-
SOC	0.92	2.97	17.27	13.09	-	-	-	-

Finally, when the opposite kinds of behaviour are analysed in Table 3, it can be seen the first and fifth columns, with production (Q1) and fecundity (Q4), have the smallest values. This suggests that higher productivity is less probable when it is less fecund - and obviously, this situation in more pronounced in the production that is co-authored with students. It is clear that SOC has the highest percentage in the first column, followed by AGR, BIO and EXA. In the production with the participation of the students (fifth column) HEA and ENG are highlighted.

The opposite situation is more pronounced, when there are higher percentages of researchers that perform worse in production, even though they perform best in fecundity. This is more pronounced in the production without students, suggesting that the effort in mentoring disables the advisor's research productivity. It should also be noted that SOC has the highest percentage in the second column, followed by EXA and AGR. Regarding the scientific publication with the participation of the students (sixth column) AGR and HEA are positively highlighted.

Conclusion

On the one hand, scientific productivity has been measured in the past, to a significant extent, by means of bibliometric measures, i.e., those based on the production of papers, books, and scientific citations, among other factors. On the other hand, recent works have measured scientific output also in terms of academic genealogy, i.e., through the formation of human resources. In this study, we conducted an empirical analysis of the relationship between scientific publications and academic genealogy of the PhD researchers who participated in the formation of scholars related to Brazilian science.

The evidence of a relationship between publications and genealogical metrics has made it possible to observe that fecundity is more closely related to publications. This result suggests that the stimulation of the scientific policy may be contributing to the research conducted by scholars. Yet, this may in some way be limited to the research conducted by the students, or else an increase in productivity would be observed. On the other hand, it might be owing to a strong involvement of students in the advisor's research, which is not necessarily the case in some areas, since productivity with academic offspring is declining among the age groups. As was expected, the main differences were found between the hard and soft sciences, and this is worth noting because productivity in co-authorship with sons is only a reality for young researchers in the latter category (i.e. soft sciences). Additionally, some specific features should be highlighted in these areas: Biological Sciences showed the highest proportion of production in co-authorship with academic, which may be the result of the students being more closely involved in their advisor's research; in contrast, the Social Sciences had the largest coefficient for productivity without the participation of the sons, in absolute numbers.

Finally, the analysis of ´antagonism´ in the performance of advisors with regard to publications and academic mentoring, revealed the following: higher productivity is less probable when it is less fecund, and is a factor that is more pronounced in production with students; in sharp contrast, and in a more pronounced way, it was found that there were higher percentages of researchers that perform worse in terms of production, even though they are best in fecundity. It was even more evident in the production without students, which suggests that the involvement in mentoring impairs the advisor's capacity for research.

In subsequent studies, it may be useful to find out if the other genealogical metrics (descendants and genealogical index) are related to the impact measured in citations, as an outcome of the indirect relationship established by the genealogy. Studies of genealogical metrics and their relationships with publications and citation impact may offer a wider perspective and could thus be included in the evaluative processes such as those found in Brazil.

Acknowledgements

This work would not have been possible without the financial support of São Paulo Research Foundation (FAPESP), grants #2012/00255-6; #2019/07118-3. The authors thank anonymous reviewers for constructive comments on improving the presentation of this work.

References

Andraos, J. (2005). Scientific genealogies of physical and mechanistic organic chemists. *Canadian journal of chemistry*, 83(9), 1400-1414. doi:10.1139/v05-158

Bäker, A. (2015). Non-tenured post-doctoral researchers' job mobility and research output: An analysis of the role of research discipline, department size, and coauthors. *Research Policy*, 44(3), 634-650. doi:10.1016/j.respol.2014.12.012

Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: an alternative model for research evaluation. *International Journal of Technology Management*, 22(7-8), 716-740. doi:10.1504/IJTM.2001.002988

Damaceno, R. J. P., Rossi, L., Mugnaini, R., Mena-Chalco, J. P. (2019). The Brazilian academic genealogy: Evidence of advisor-advisee relationships through quantitative analysis. *Scientometrics.* 119(1), 303-333. doi:10.1007/s11192-019-03023-0

Green, S. G., & Bauer, T. N. (1995). Supervisory mentoring by advisers: Relationships with doctoral student potential, productivity, and commitment. *Personnel Psychology*, 48(3), 537-562. doi:10.1111/j.1744-6570.1995.tb01769.x

Heinisch, D. P., & Buenstorf, G. (2018). The next generation (plus one): an analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. *Scientometrics*, 117(1), 351-380. doi:10.1007/s11192-018-2840-5

Horta, H., & Santos, J. M. (2016). The impact of publishing during PhD studies on career research publication, visibility, and collaborations. Research in Higher Education, 57(1), 28-50. doi:10.1007/s11162-015-9380-0

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, 90(2), 463–481. doi:10.1007/s11192-011-0495-6

Levin, S. G., & Stephan, P. E. (1991). Research productivity over the life cycle: Evidence for academic scientists. The American Economic Review, 81(1), 114-132.

Malmgren, R. D., Ottino, J. M., & Amaral, L. A. N. (2010). The role of mentorship in protégé performance. *Nature*, 465(7298), 622. doi:10.1038/nature09040

Oliveira, T. M. de, & Amaral, L. (2017). Public Policies in Science and Technology in Brazil: challenges and proposals for the use of indicators in evaluation. In: Mugnaini, R.; Fujino, A.; Kobashi, N. Y. (Orgs.), *Bibliometrics and scientometrics in Brazil: scientific research assessment infrastructure in the era of Big Data* (pp.189-217), São Paulo: ECA/USP. doi:10.1344/BiD2018.40.19

Pinheiro D., Melkers J., & Youtie J. (2014). Learning to play the game: Student publishing as an indicator of future scholarly success. *Technological Forecasting & Social Change*, 81, 56–66. doi:10.1016/j.techfore.2012.09.008

Qi, M., Zeng, A., Li, M., Fan, Y., & Di, Z. (2017). Standing on the shoulders of giants: The effect of outstanding scientists on young collaborators careers. *Scientometrics*, *111*(3), 1839-1850. doi:10.1007/s11192-017-2328-8

Reskin, B. F. (1977). Scientific productivity and the reward structure of science. American sociological review, 42(3), 491-504. doi:10.2307/2094753

Rossi, L., Freire, I. L., and Mena-Chalco, J. P. (2017). Genealogical index: A metric to analyze advisor-advisee relationships. Journal of Informetrics, 11(2), 564-582. doi:10.1016/j.joi.2017.04.001

Tuesta E. F., Delgado K. V., Mugnaini R., Digiampietri L. A., Mena-Chalco J. P., & Pérez-Alcázar J. J. (2015). Analysis of an Advisor–Advisee Relationship: An Exploratory Study of the Area of Exact and Earth Sciences in Brazil. PLoS ONE, 10(5), e0129065. doi:10.1371/journal.pone.0129065

Research in Progress:
The career of postdocs in Norway

Hebe Gunnes[1] and Paal Boring[2]

[1] *hebe.gunnes@nifu.no*
NIFU Nordic Institute for studies in Innovation, Research and Education, Økernveien 9, N-0608 Oslo, Norway

[2] *paal.boring@nifu.no*
NIFU Nordic Institute for studies in Innovation, Research and Education, Økernveien 9, N-0608 Oslo, Norway

Abstract

What is the probability for a postdoc to obtain a tenured position at a Norwegian higher education institution (HEI)? And who is likely to leave academia? In this study, we follow the academic careers of approx. 3000 postdocs who were affiliated at a Norwegian HEI, health trust or research institute in 2001, 2005 and 2009. We examine how different individual characteristics such as gender, age, field of science, funding and PhD obtained in Norway/abroad, affect the academic careers of the postdocs. We have used logistic regression to calculate the probability of a) obtaining a tenured position and b) leaving academia.

The article is based on a study conducted by NIFU in 2015[1], funded by the Research Council of Norway. The original study had a 6 years span for the logistic regression, but the dataset is now updated, and we follow the academic career of the postdocs for 8 years.

We find that field of science is the key variable when it comes to the postdoc obtaining a tenured position. Whether the PhD is obtained in Norway or abroad (i.e. international mobility) is also of importance, as is the age of the postdoc. Gender is of less importance than expected.

Introduction

This article will map the career of approx. 300 postdocs affiliated at Norwegian higher education institutions, health thrusts and research institutes in 2001, 2005 and 2009. We have a unique empirical dataset, and most analyses are already complete, but the literature review is not yet finished. We intend to see our findings in connection with studies conducted by researchers such as Huisman et. al. (2002), van der Weijden (2015), Cartwell (2011), Xuhong Su (2013) and Åkerlind (2005), and to link our findings with the international discourse on the career of postdocs. There is also an ongoing discussion in Norway on the role of the postdoctor that we want to addres; are postdocs temporary, cheap academic labour, or is the postdoc position a step on the academic career ladder?

Source of the empirical data is NIFU's Register of Research personnel[2], which contains information about former postdocs who are employed at Norwegian higher education institutions (HEIs), research institutes and as researchers as university hospitals and other health trusts. Variables included in the mapping are field of science, gender, age, funding of the postdoc position, region of PhD-awarding institution (i.e. Norway or abroad).

[1] Gunnes & Boring 2015. The results of the study is only availbale in Norwegian.

[2] NIFU's Register of Research personnel is part of the official Norwegian R&D statistics on the Higher education sector and the Institute sector. The register covers researchers/university graduated personnel that participated in R&D at Norwegian higher education institutions, as well as research institutes and health trusts. The register includes information on position, age, gender, educational background and work place. The register does not cover special part time affiliations, with the exception of adjunct professors/Professor II. Personnel data is retrieved from the administration of the R&D-performing institutions per October 1st, and the registry goes back to the 1960s. Mobility, gender balance and career paths of the academic staff in Norway has thus been monitored through several decades through this rather unique database.

The Norwegian context

Postdoc is a rather new position in Norway (Kyvik et al 2003), and the number of postdocs has increased from approximately 500 in 2001 to 2.300 in 2017. A postdoc fellowship lasts 2-4 years and may exceptionally be extended. The majority of the Norwegian postdocs are affiliated within natural sciences or medical and health sciences, and the gender balance is 48-52 in the favour of men. The Research Council of Norway funds close to half of the postdoc fellowships in Norway.

The tasks of a postdoc in Norway

The main task for a postdoc in Norway is research and development (R&D), as described in the national regulations. In order to qualify for tenured positions, normally associate professor, postdocs also need teaching experience. For postdoc positions that last more than 3 years, duty work in the form of teaching and supervision of master and PhD students, takes up some of their time. The postdocs are required to undertake administrative tasks, hereof administration of research projects, as well as department meetings etc.

According to the last Time Use Survey at Higher education institutions in Norway (Gunnes 2018), an average postdoc spends 72 per cent of his or her working hours doing research and development (R&D). A total of 17 per cent of their working hours are spent on teaching or supervision, and 7 per cent on administration. 3 per cent is related to externally directed activities, and 2 per cent to other activities, such as professional practice, museum related activities or artistic activities. There are, however, some variations in the time use, related to field of science, see figure 1.

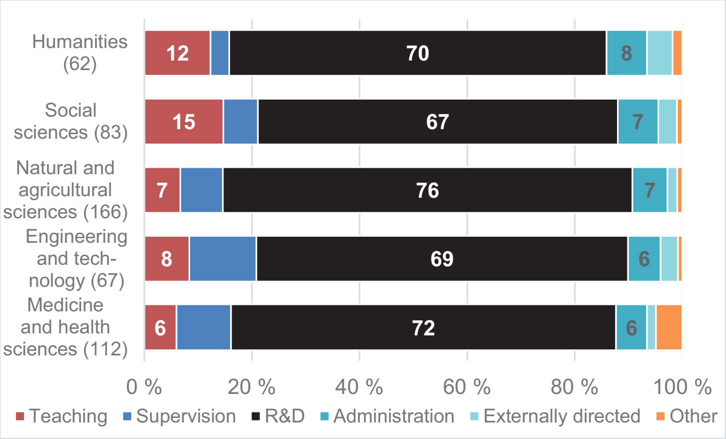


Figure 1. Time use for the average postdoc in Norway by field of science: 2016. (Number of respondents).

Source: NIFU, Register of Research Personnel

Characteristics of the postdocs

The cohort of postdocs that is mapped in this study, consists of 2953 individuals. In 2001, there were 507 postdocs. Four years later, in 2005, there were 1015, and in 2009 the number had increased to 1431. There was a gender balance (i.e. 40-60% representation on each men and women) in both Humanities, Social sciences and Medicine and health sciences, whereas there was 38 per cent women in Natural sciences and 29 per cent women in Engineering and technology, se figure 2.

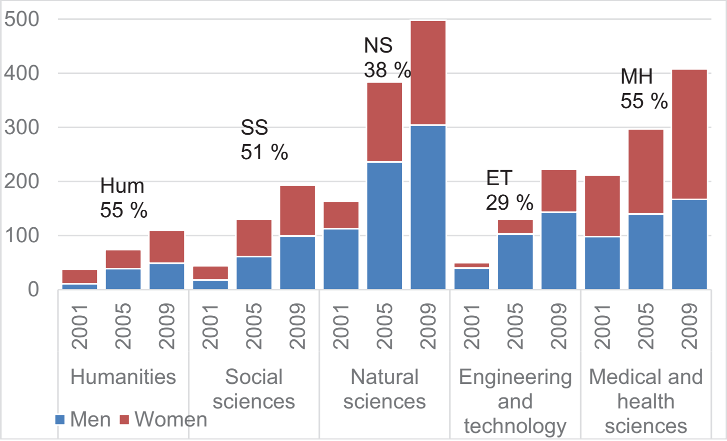


Figure 2. Number of postdocs in Norway in 2001, 2005 and 2009 by field of science and gender. Share of women as percentage.

Source: NIFU, Register of Research Personnel

Altogether, 34 per cent of the postdocs in the study had obtained their PhD outside Norway and are thus considered internationally mobile researchers. The share of postdocs with a PhD from abroad had increased over the years, from 25% in 2001 to 40% in 2009. Approximately half of the postdoc positions in the cohort were funded by the Research Council of Norway (RCN), while 20 per cent were funded by the basic funding. The RCN funded 73 per cent of the postdocs within Natural sciences and 67 per cent within Engineering and technology.

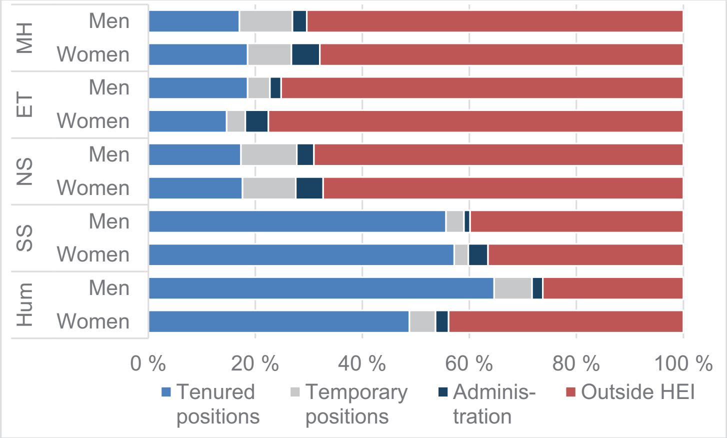


Figure 3. Postdocs in 2001, 2005 and 2009 by affiliation[1] 8 years after reference year.

[1] Tenured position includes full professor, associate professor, assistant professor and lecturers. Temporary positions covers researchers and postdocs. Former postdocs employed outside Norwegian HEIs may be affiliated in the Norwegian institute sector, at health trust, in the Business enterprise sector or in another country.

Source: NIFU, Register of Research Personnel

We have examined to which extent the postdocs have obtained a tenured position 8 years after they were registered as a postdoc. Tenured position comprises full professor, associate professor, assistant professor and lecturers at Norwegian universities, specialized university institutions and university colleges of applied sciences.

We find differences in the academic career patterns after 8 years, and field of science is the key variable, see figure 3. In Humanities and Social sciences, more than half of the postdocs have obtained tenured positions, and fewer have left academia than is the case in other fields. The share of postdocs in temporary positions is somewhat higher within Natural sciences and Medical and health sciences than in the other fields – these are also the fields with the highest number of postdocs.

Logistic regression of the postdoc's careers

Through logistic regression, we have examined the probability of obtaining a tenured position eight years after reference year as postdoc. Reference variables are gender, age group, field of science, funding source and type of PhD. The analysis shows covariation between the probability and each variable, controlled for the other characteristics. The aim of the analysis is to determine how a particular variable influence the probability of obtaining a tenured position at a Norwegian HEI. We have used a "reference person", who is representative for the total population, to test the explanatory variables in the regression. Our reference person is male, between the age of 30 and 34, affiliated within the natural sciences and in a postdoc position which is funded by the Research council of Norway. He has obtained his PhD in Norway and was a postdoc in 2005.

Table 1 shows the probability of obtaining a tenured position after 8 years. The rows marked in gray are the reference categories. In the other rows, we have calculated the same probability for a person who has the same characteristics as our reference person, save one.

The overall probability of obtaining a tenured position after 8 years, is 22 per cent for our reference person. Female postdocs have a 20 per cent probability of obtaining a tenured position, a little less than the male postdocs, but we find no statistical significance in the probability for women and men to obtain a tenured position. The older age groups of postdocs have a higher probability to obtain a tenured position than the younger, and the postdocs funded by basic funds have a higher probability than those funded by the Research Council of Norway or other sources. Postdocs from 2001 have a higher probability of obtaining a tenured position then those from 2005 and 2009. This is not surprising, as the number of postdocs in Norway has increased substantially in this period, whereas the number of open tenured positions has not.

The most striking differences are found when examining field of science and type of PhD. More than half of the postdocs within Social sciences and Humanities (SSH) have obtained tenured positions 8 years after the reference year, while this is the case for approximately 20 per cent of the postdocs within Natural sciences, Technology and engineering and Medicine and health sciences (STEM fields). There are less postdoc positions in SSH fields than in the STEM fields, and the generation shift has been more extensive in the SSH fields in Norway over the last decade, so there have been more open tenured positions in these fields.

Postdocs who have obtained their PhD in Norway, have a significantly higher probability of obtaining a tenured position than those with a PhD from abroad. Postdocs with a PhD from abroad are younger then those with a PhD from Norway and are thus more mobile. Some of them comes to Norway only for the postdoc period, and goes back to their country of origin, or moves on, after the postdoc period is finished.

Table 1. Logistic regression where the outcome variable is whether the post doctor has obtained a tenured position after 8 years (=1) or not (=0), and with individual characteristics as explanatory variables.

	Coefficient	Significance level	Standard error	N	Proba-bility
Constant	-1,252	***	0,130		
Gender					
Male				1620	22%
Female	-0,113		0,097	1331	20%
Age group					
Under 30 years	-0,347		0,257	185	17%
30-34 years				1116	22%
35-39 years	-0,050		0,116	987	21%
40-44 years	0,106		0,151	375	24%
45 years or older	0,048		0,167	288	23%
Field of science					
Humanities	1,496	***	0,170	222	56%
Social sciences	1,467	***	0,143	367	55%
Natural sciences, incl. agricultural sciences				1044	22%
Engineering and technology	-0,176		0,166	402	19%
Medicine and health sciences	-0,209		0,132	916	19%
Funding sources					
General university funds	0,705	***	0,110	687	37%
Other external sources	-0,010		0,133	643	22%
Research Council of Norway				1621	22%
Type of PhD					
PhD obtained in Norway				1957	22%
PhD obtained abroad	-0,899	***	0,116	994	10%
Year of employment					
2001	0,227	*	0,134	505	26%
2005				1015	22%
2009	-0,218	**	0,106	1431	19%
LR chi2(14)					469,560
Prob > chi2					0,000
Pseudo R2					0,143
Log likelihood					-1404,883
N					2951

*Notes: 1) *** Significant at the 1 per cent level, ** significant at the 5 per cent level, * significant at the 10 per cent level. 2) The data set consists of all post doctors in 2001, 2005 and 2009. 3) N is the number of persons.*

We have conducted the same regression analyses related to the probability of leaving academia. This table will not be included here, due to the limited number of pages. We found that the reference person had a 26 per cent probability to leave academia after 8 years. Women left at the somewhat same rate, 25 per cent. Postdocs with a PhD obtained abroad had a probability of 53 per cent of leaving academia 8 years after the reference year. When looking at field of science, the probability of leaving academia was highest within Engineering and technology and Medical and health sciences (approx. 35%), while it was lowest within social sciences (15%).

Conclusion

Of the 3000 postdocs at Norwegian HEIs in 2001, 2005 and 2009, a total of 25 per cent had obtained a tenured position 8 years after the reference year. 63 per cent had left academia, while 8 per cent were employed in temporary positions as researchers, and 4 per cent where in technical or administrative positions. This implies that a postdoc position does not necessarily function as a recruitment position for a tenured track.

We found that the key variable for determining the post doctor's career path is field of science. Within Humanities and Social sciences, more than 50 per cent of the postdocs had achieved tenured positions after 8 years, while this was true for less than 20 per cent of the postdocs within engineering and technology and natural sciences.

Gender had less impact on the post doctor's career paths than we anticipated, at least on an overall level. We see some minor differences when it comes to field of science. A further investigation of the dataset might show larger disciplinary differences, but there are limitations to the number of postdocs in each discipline. The share of men who had left the Norwegian HEIs 8 years after the reference year amounted to 65 per cent, which was somewhat higher than for the women (63 per cent). This difference is not statistically significant.

References

Cantwell, B. (2011): *Academic in-sourcing: International postdoctoral employment and new modes of academic production*. Journal of Higher Education Policy and Management, 33(2), 101–114.

Huisman, J., Weert E. de & Bartelse, J. (2002): *Academic Careers from a European Perspective*. The Journal of Higher Education, 73:1, 141-160

Kyvik, S., Olsen T. B. & A. Vabø (2003): *Postdoktorordningen*. Oslo, NIFU working paper 2003:37.

Gunnes, H. (2018): *Tidsbruksundersøkelse for universiteter og høgskoler*. Oslo, NIFU working paper 2018:2.

Gunnes, H. (2018): *The development of diversity statistics for Norwegian research and higher education*. Poster at the STI2018 conference in Leiden, the Netherlands.

Gunnes, H & P. Boring (2015): *Veien fra postdoktor til akademia: En statistisk analyse av postdoktorenes karriere ved utdannings- og forskningsinstitusjonene*. Oslo, NIFU working paper 2015:15.

Nerad, M. & J. Cerny (1999): *Postdoctoral Patterns, Career Advancement, and Problems.* Science 285, 1533 (1999).

Van der Weijden, I, Teelken, C, de Boer, M & Drost, M. (2016): *Career satisfaction of postdoctoral researchers in relation to their expectations for the future*. Higher Education, Volume 72, Issue 1, pp 25-40.

Xuhong Su (2013): *The Impacts of Postdoctoral Training on Scientists' Academic Employment*. The Journal of Higher Education, Volume 84, Number 2, March/April 2013, pp. 239-265.

Åkerlind, G. S. (2005): *Postdoctoral researchers: roles, functions and career prospects*. Higher Education Research & Development Vol. 24, No. 1, February 2005, pp. 21–40.

Disciplines at the crossroads: scientific re-orientation of economics and chemistry after German reunification

Andreas Rehs[1]

[1]*rehs@incher@uni-kassel.de*
University of Kassel and INCHER Kassel, Mönchebergstr. 17, 34127 (Germany)

Abstract

We empirically approach the identification and measurement of research topic differences by the example of East and West German chemists as well as economists and business economists before and after German reunification. The political transition, which came along with German reunification, is expected to have influenced the research topics of East German scientists in the two disciplines to differing degrees. Our dataset builds on dissertation titles in economics and business administration as well as chemistry from 1980 to 2010. We use university affiliation and year of the dissertation to train a structural topic model and test the model on a set of unseen dissertation titles. Subsequently, we compare the topic distribution of each title pair with cosine similarity and a linear regression framework. Our results on East German economics and business administration suggest substantial differences before the reunification and a rapid assimilation in research topics thereafter. In chemistry we observe minor differences before the reunification and a slightly increasing similarity after the reunification.

Introduction

What is researched by actors in the scientific system naturally differs. Countries, as one entity, vary in their research profiles and so do universities, faculties, individual researchers and scientific journals. Some of these differences are by the nature of the subject, but other, subtle differences, are only obvious to experts in the respective scientific fields. Journal editors, as such kind of experts, are able to classify and distribute a submitted paper to referees of who they know are experts in the specific topic of the submitted paper. The editor can therefore be considered as a topic classification specialist. However, with every step of aggregation this classification task becomes more and more boundary spanning for the editor. Differentiating whether female, East or West German authors in their journal are more engaged in specific topics, would require the editors to process and classify every submitted paper by the desired features again. This is time consuming, subjective and, at a certain point, no longer feasible for the editor as a human expert. Information on differences in research topics at various institutional and socio-demographic levels, are, however, still of interest. Policy makers need a basis upon which they can react to desired and undesired developments in the scientific system. Specialization, convergence and divergence patterns may be such cases and need to be detected early on. In this paper we apply structural topic modelling in combination with cosine similarity and a regression approach to empirically address the identification of research topic differences. Structural topic modelling is firstly able to detect semantic relationships between research topics, and secondly, to model the semantic relationships with document level covariates, e.g. the regional background of the dissertation. This allows potential differences in research topics of scientists to be identified more precisely than other topic modelling algorithms. We investigate the research topics of East and West German economists and business economists as well as chemists before and after German reunification. German reunification is especially suited because it provides an unexpected shock to German researchers and their choice of topic. It went hand in hand with the transition of the political and scientific system in East Germany. German reunification led to the dismantling of a large number of chairs, institutes and whole research organizations as well as a broad institutional restructuring in academia in East Germany. Reasons included political motives, but in several instances also a mismatch between what had been researched under the old system and what was considered interesting in the new one. This affected social sciences more severe than sciences. Especially in the case of economics and business administration, whole faculties were completely rebuilt and often newly founded at East German universities. Therefore, the 1990s saw a comparatively large number of vacancies in this discipline. The new generation of professors, which replaced the free chairs, quite often came from West Germany and brought with them West German topics to East Germany. For the personnel replacement Kolloch (2001) reports that by 1994, 90 % of the overall chairs were replaced with West Germans. A significant or even complete thematic assimilation between East and West Germany is therefore reasonable and should be detected by our approach. In our second discipline, chemistry, the preconditions are different. Ideological involvement at the individual and institutional level was less pronounced than that seen in economics and business administration. The replacement of chairs and phase-out of institutions was therefore probably smaller than that seen in East German economics and business administration. Nevertheless, the discipline was considered to be an economic driver and strictly aligned to the year-plans of the state commanded economy of the German Democratic Republic (in the following GDR). The choice of research problems was therefore to a large extent restricted for chemists. This changed after the reunification. East German chemists could choose their research topic independently, but were faced with other (and fewer) industry demands in a reunited Germany. Albeit less directly, East German chemistry was therefore probably urged to adapt to West German chemistry research. As the switching of topics is associated with costs, the East German chemists most likely re-oriented to topics similar to their former ones. Chemistry therefore serves as an example how a moderate shock affects topic choices by scientists.

Our procedure includes the analysis of PhD-theses titles in economics and business administration as well as chemistry handed in at universities in East and West Germany before and after 1990. The dissertations titles, which constitute a very condensed form of research content, are processed with various text-mining methods, such as stopwords, stemming and n-gram detection. On the basis of 75 % (10,361 in chemistry; 6855 in economics and business administration) of these processed titles, we estimate a structural topic model. The remaining 25 % (2,580 in chemistry; 1767 in economics and business administration) are used to test the model and allow us to estimate the topic distribution of every title. Using these distributions, we calculate the cosine similarity score between every title pair. We test our hypothesis by comparing the similarity of topic-title distributions written in the same part of Germany vs. ones written in different parts with a linear regression framework. In economics and business administration, our findings suggest considerable differences in research topics before reunification. After reunification, we observe a strong and rapid assimilation. In chemistry there are small differences before reunification. Afterwards, we observe a moderate thematic convergence.

Problem choice and scientific change

Why scientists choose to investigate one particular research topic over another is known as the problem of problem choice in the economics of sciences and the sociology of science literature. Problem choices are shaped by myriads of overlapping and sometimes interdependent factors, which we will cluster into economic, disciplinary, societal, scientific, institutional and individual factors. Before reviewing the factors, the term research problem (or research topic) offers a fundamental question regarding its concept and delimitation. The question is: What is a research topic? Gläser, Glänzel & Scharnhorst (2017) argue that the conceptualization of research topics faces two main difficulties. First, knowledge can be structured in different valid ways, i.e. experts may come to different conclusions how to build a taxonomy of research topics in their field. Second, the purpose determines the best structuring of topics as well as the single best scope of a certain topic. This complicates from an inter-temporal perspective because boundaries of disciplines, areas and single problems are constantly in motion. In particular, when it comes to interdisciplinary questions as well as small or emerging subjects, the scope of the subject may change by every single scientific contribution. Research problems can therefore only be ranked ordinarily with other size categories in a scientific taxonomy. Here, a research problem has the smallest scope, whereas research domain and discipline refer to larger categories. In the absence of any other reasonable concept known to us, we will stick to the categorical delimitation of research problems.

Economic factors influence the problem choice in manifold ways. Economic uncertainty is one of the most decisive factors, since the production of knowledge generally brings with uncertainty (Dapgusta & David, 1994). Future conditions and outcomes, such as the availability of funding or political circumstances related to some particular research problems, are to a large extent unpredictable ex ante. In the same manner research collaboration possibilities (Leahy & Reikowsky, 2008), results, complexity of the subject matter as well as payoffs in the scientific reward system (Foster, Rzhetsky & Evans, 2015) and probable commercialization outcomes of scientific findings are characterized by uncertainty. Merton (1957) emphasizes the role of the priority of discovery and uncertainty in science. Here, competition and probable anticipation from peers working at the same problem is unclear to the scientists. In this sense, science is like a contest where the winner takes it all. Even if the researcher achieves initial success with a novel problem, other researchers are attracted to related problems in the same research area, which stimulates subsequent competition (Zuckerman, 1978; Ziman 1987). Borjas & Doran (2012) investigate research topic competition empirically and show the consequences of an unexpected shock in the competition for research problems. They investigate how the immigration of ex-soviet scientists to the US changes the thematic competition after 1990. The influx of ex-Soviet mathematicians led to a negative productivity effect on US mathematicians whose research overlapped with that of the incoming scientists. Uncertainty thus affects numerous economic factors and is generally related to all future outcomes of a scientific problem choice.

Regarding the past and current perspective related to problem choice, prior experience in a topic and associated switching effort to a new one lead to opportunity costs. Also age and career stage affect problem choices and the number of problems picked. Horlings & Gurney (2013) present a scientometric method to longitudinally map the scientific output of scientists in physics. In their investigation of 18,235 publications, they find that scientists constantly add research topics to their agendas throughout the whole academic life. They find an association between career phases and start/end of topics. However, they find no empirical evidence that topics exactly start/end at career moments, such as the appointment to professor or postdoc. Also the selection of new research topics is more related to postdocs than to professor stage. For professors, they find higher numbers of last authorship positions and proportionally more co-authored papers and argue that the work of professors is thus more managerial and collaborative. Problem choices are also affected by academic advisorsships. Professors transfer knowledge, norms and behaviours to their (doctoral) students (Bunestorf & Geisler, 2013). Disciplinary factors such as the disciplinary reward system, feasibility and scientific frontiers of the discipline as well as interdisciplinarity and the disciplinary culture (Abbott, 2001; Knorr-Cetina, 2009) are relevant as well. In new and emerging problem areas or disciplines, for instance, researchers may choose problems differently than in comparison to established disciplines and areas (Debackere & Rappa, 1994). This is because they lack a common paradigm and only rely on recent research (Foster, Rzhetsky & Evans 2015). Similarly, scientists in interdisciplinary areas and disciplines may not be locked into a single disciplinary culture and are in this way unconstrained and can attain acceptance for their chosen problems more easily. Research problems choice of an individual scientist is therefore inextricably linked to the current paradigm and subsequent reward system of the discipline. Problem choice is also affected by the institutional context. The institutional context includes (institutional) culture (Fisher, 2005), geographical, political and societal embeddedness of the institution as well the type and the specialization of the institution. Institutions with profit orientation, directly or indirectly require their scientists to pick research problems that have a higher chance of leading to commercilizable inventions (Cooper, 2009). Specialization profiles of institutions may in the same way obligate individual scientists to pursue appropriate research topics. Here, indirect influences may consist of peer effects - scientific colleagues that affect the choice of one's own research problem through their research problems. The Political and societal context in which the scientist and the superordinate institution are embedded are highly relevant for the research problem choice mechanism. Here, political influence and societal norms can directly or indirectly influence the choices of the scientists. Lastly, individual factors may be amongst the most crucial determinants in research problem choices. In the first place, problem choice is a matter of individual taste, but also of other personal attributes such as beliefs, talents, risk taking attitude and intrinsic motivation. For intrinsic motivation Kuhn (1962) draws the analogy to science as puzzle solving, where researchers are intrinsically motivated to figure out the solution to a problem. The choice of topics more accepted topics by colleagues and the broader scientific community may also be driven by the desire for recognition. Colleagues also drive the problem choices in other respects. The social and geographical proximity with colleagues exposes the respective scientists to the particular problem choices of their colleagues (e.g. through common seminars). This increases the likelihood of choosing each other's problems. Finally, funding partners may also influence problem choices. The previous non-exhaustive overview of the problem of problem choice shows that thematic decisions are driven by diverse factors. Owing to the different disciplinary histories, we argue that these factors take effect to different extents in the problem choices of East and West German chemists and of economists and business economists before and after German reunification.

Hypotheses

Regarding East and West German chemistry before reunification, we hypothesize that problem choices of scientists in both countries should differ. This is primarily because the GDR directly and indirectly interfered with the problem choices of East German scientists. The prime example of direct influence being exerted directly were the official year plans for science and technology, which forced chemistry to meet industry demands of East Germany. The economic and societal restrictions in the GDR also had an influence on problem choices. Collaboration, for instance, was only possible with colleagues from other socialist countries. This prevented thematic spillovers, which could have been resulted from the collaboration with West German colleagues. The different characteristics of uncertainty and problem choices in the GDR may also have had an indirect influence on problem choices. The academic labour market in the GDR was, for instance, in full employment at any point in time, albeit with a considerable hidden unemployment rate, since it was socialist state doctrine to employ everyone. Picking risky research problems was therefore not associated with risky labour market outcomes for East German chemists and scientists in general. All of the previously mentioned circumstances lead us to the conclusion that the problems picked by East and West German chemists before reunification should be different from one another.

H1A: East German chemistry topics before the reunification differ from West German chemistry topics before reunification.

With German reunification East German chemistry was institutionally adapted to West German chemistry. This was associated with personnel replacements and the abandonment of the state-controlled economy. At first sight, this suggests a thematic convergence after reunification. However, some of the chemists in the East who had not been dismissed faced significant switching costs, because they had previously been engaged in typical East German chemistry topics. They are therefore to some extent locked in their old topics. Consequently, their best strategy is not to exactly copy the West German topics as this entails substantial time costs and monetary expenses. Moreover, they faced high competition in that problems by West Germans and those East German chemists, who already researched respective topics in the GDR. The only viable alternative for East German scientists with obsolete topics is therefore to switch to related, but still recognized topics in reunified Germany. Both groups of East German chemists, those with adequate and those with obsolete topics, should consequently account for a thematic assimilation between East and West after the reunification. This convergence is reinforced by the replacement of East German chairs with West German personnel after the reunification. We therefore propose H1B accordingly. After reunification, East German chemistry faced different demand from industry and was no longer forced to research applied topics. East German chemists experienced access to new literature, colleagues, materials and laboratories, which were previously not available due to the Berlin wall. This may have incentivized them to change their topics after reunification. However, they were again confronted with switching costs as well as competition and therefore possibly chose not to abandon their former topics completely. To some extent West Germans also took over free chairs in East Germany, which probably led to an inflow of West topics to East Germany. Following the previous augmentations as well as the argumentation for H1A, the East German topics before reunification should be distinct from those after reunification. We propose H1C accordingly.

H1B: East German chemistry topics after reunification become more similar to West German chemistry topics after the reunification.

H1C: East German chemistry topics before reunification differ from East German chemistry topics after reunification.

In economics and business administration the differences were more pronounced before reunification. The discipline was extremely important in the ideological framework of the GDR. The research of economists and business economists therefore had to be vetted in line with the socialist ideology more than in other disciplines. Capitalistic topics, which were researched in western countries such as West Germany, were de facto banned. This probably had a substantial effect on the differences in problem choices of East and West German economists and business economists before the reunification. We propose H2A accordingly. Massive personnel replacement as well institutional redirection took place in East German economics and business administration after the reunification. The free chairs were predominately filled with West Germans economists and business economists. Consequently, the problems picked by new these scientists should be very different from the topics of the dismissed East German scientists and their predecessors. We propose H2B accordingly. After reunification, the newly appointed West Germans should stick to their existing topics. Therefore, they should be very similar to the West Germans at West German universities. This leads us to hypothesize that East and West German economics and business administration topics become similar after reunification. H2C is proposed accordingly.

H2A: East German economics and business administration topics before reunification differ from West German economics and business administration topics before reunification.

H2B: East German economics and business administration topics before reunification differ from East German economics and business administration topics after the reunification.

H2C: East German economics and business administration topics become more similar to West German West German economics and business administration topics after reunification.

Data and Methods

We probe into the issue described above by relying on PhD theses handed in at universities in East and West Germany after reunification as a formalized representation of scientific work. Our work rests on a number of presumptions: First, in Germany the advisor (often dubbed the "Doktorvater") has a strong influence on the advisee and their choice of research topic. Moreover, the advisor is usually required to have a chair at a university as only they are entitled to award PhDs. The second important assumption is that the title of a thesis represents its content in a very condensed form. Together, both assumptions lead to the conjecture that the research focus of a chair is reflected in the titles of theses handed in at an entity (a university) that he is presumably affiliated with. We utilize the online catalogue of the German National Library (Deutsche Nationalbibliothek, short: DNB) as the basis for our analysis. The catalogue lists the vast majority of PhD theses handed in at German universities including the former GDR. There are entries for approximately one million PhD theses, which are classified by subject. We use this classification to distinguish between economics and business administration and chemistry. Due to the peculiarities of German medical dissertations, we have eliminated dissertations which are cross-listed in chemistry and medicine. Furthermore, we employ information on university location (cities, name of university or a combination of both) to separate East from West dissertations[1]. We assume that re-orientation of research topics after the reunification took place up until 2010. To obtain a picture of the thematic landscape before reunification, we consider the years 1980 to 1989. The years 1990-1995 have been eliminated from our data, since the replacement of East German chairs took several years and the number of observations for East Germans dissertations dropped significantly in those years. In the next step, we paste every title and subtitle into one string and standardize this string. Our pre-processing includes standard methodology of: transformation to lowercase, removal of punctuation, language detection and removal of non-German titles[2], stemming, n-gram detection[3], as well as the removal of very frequent, rare words[4] and short titles[5].

To approach our research question we use structural topic modelling, which is a statistical model for text analysis. Topic modelling aims to automatically discover latent semantic structures (topics) from texts (Blei, 2012). In the past, topic modelling has found various applications in scientometrics. In one of the pioneering works in topic modelling, Blei (2007) investigated topics that construct scientific publications. One fundamental property of topic modelling is to consider every topic as a probability distribution over all the words appearing in the collection of documents. Every document is in turn a probability distribution over topics. It is therefore impossible to label the constructed topics as 'organic chemistry', 'bio chemistry' or other. The topic model can only infer which words are more or less probable for a topic. Here, the various topic model algorithms basically work in a similar way. They repeat the assignment of words to topics as well as topics to documents a large number of times. In every step they update their statistical inference on how words are associated with topics and topics with documents. As mentioned, the number of topics have to be set in advance, since the algorithm needs to draw from a distribution that equals the number of topics. There is, however, no right solution for choosing the optimal number of topics for a given set of documents. Regarding our approach, the number of topics should be more or less equal to the number of research problems existing in the discipline. Due to the size and discipline characteristics, chemistry as well as economics and business administration might naturally arrive at a different number of topics. To determine a number of topics with optimal statistical properties, we rely on spectral initialization (Arora et al. 2013). For our further procedure we will use a structural topic model as described in Roberts et al. (2014, 2016). Structural topic models allow document level covariates to enter the topic modelling process. In this way, information on where a dissertation was written can model how word-topic, and topic-document probabilities are built. Technically, topic prevalence (how much of a document is about a topic) is dependent on covariates in structural topic models. On aggregate level, we can therefore model how university affiliation shifts topic prevalence towards or away from certain topics. The inclusion of covariates allows the topic model to gain statistical quality and improves the reliability of word-topic as well as topic-document probabilities. We apply the resulting topic models in chemistry and economics and business administration to a set of unseen documents. This procedure allows to draw causal inference, since the training and the test set are separate data sets. As we know when and where a dissertation was written, we incorporate dummies for the universities where the dissertation was written and the submission year of the dissertation as covariates into the structural topic models. From the processed titles of 1980 to 2010 (without 1990-1995), we use 75 % of the titles in each discipline to train topic models in chemistry and in economics and business administration. On the remaining 25 %, we apply the topic models. Training and test set sizes in economics and business administration are 6,855 and 1,767 respectively. In chemistry, sizes are 10,361 and 2,580. As one result of the two topic model applications, we obtain a topic distribution for every title. To compare the distributions between two titles, we use the cosine similarity measure, which has found various applications in the comparison of topic model outcomes (see e.g. Ramage, Dumais, & Liebling, 2010). The cosine similarity is a measure for the distance between two vectors and is defined between zero and one; values towards zero indicate similarity. As topic portions per document are vectors of the same length, the cosine similarity allows a comparison of the topic distribution between two documents. In order to approach our hypothesis, our detailed procedure is now as follows: We calculate the cosine similarity between all topic-document distribution pairs. This means, topic distribution of title 1 is compared to title 2, title 3 and so on. We drop duplicate observations (e.g. cosine similarity between 2 and 3 is the same between 3 and 2). For every observation of the cosine similarity, we know when and where both dissertations titles are written and employ this information in creating variables that can test our hypotheses. In order to ease the interpretation, we require both titles to be from the same year. In the next step, different subsets of the data are built to address the specific hypotheses. For hypothesis 1, regarding East and West German chemistry before reunification, we subset to chemistry titles prior to 1990. The dataset addressing H1B uses the opposite subset (titles only after 1990). For the test of H1C only East German dissertations are used. In economics and business administration we proceed accordingly and use titles from before 1990 to address H2A, only East German ones to address H2B, and titles from both periods and parts of the country to address H2C. All titles in economics and business administration are relevant for H2C. In the next step, we create a dummy *diff_part* that describes whether the two underlying dissertations for every similarity score are from different parts of Germany. This allows us to test our hypotheses 1, 2, 4 and 6 in a linear regression framework, where the similarity score for each pair of dissertations is the dependent variable and *diff_part* is the independent variable. Regarding H1C and H2B, the main independent variable is a dummy *post95*, indicating whether a dissertation was written after 1995. We add university dummies to control for differences in similarity scores arising from universities. As the similarity score is calculated between two dissertations that are most often written at different universities, we consequently add dummies for both. The dummy *sameuni* indicates whether both titles in a pair are from the same university.

Results

The spectral initialization approach determines 76 topics in the chemistry topic model as well as 69 in economics and business administration topic model. Fig. 1 represents the top words with the highest probability to topics 11 and 40 in economics as well as their yearly mean probability across all titles. A list of words associated with other topics can be found in appendix A. Top words are measured by the highest beta probability. Since every topic is a probability distribution over words, top words may provide some indication of the underlying subject. However, interpretation should be done very cautious, since the most probable words only represent a small fraction of the probability distribution. Moreover, most probable words are not necessarily the most exclusive words to a topic. Fig. 3. suggests a sharp decline in topic 11 after 1989. This is reasonable, since the depicted words suggest association with socialism. For topic 40 the case is different; the topic sees a significant increase after 1990 and is most likely related to management. Table 1 and 2 address our hypotheses in a linear regression framework. H1A on the difference between East and West chemistry topics before reunification. Both pre models of Table 1 arrive at significantly negative coefficients of the variable *diff_part*. This indicates that the cosine similarity between two dissertations is lower when they were written in different parts of Germany. H1A can therefore be confirmed. H1B concerned the differences after reunification. We proposed that the differences between East and West German chemistry become smaller due to personnel exchange and fluctuation. Full period model 1 suggest that this is empirically the case. The interaction of *diff_part* and *post95* is positive and statistically significant. This indicates increasing similarity between East and West German chemistry after the reunification. However, the effect diminishes after including universities dummies and the variable *sameuni.* H1B can therefore not be confirmed. The last chemistry hypothesis concerned the thematic change within East German chemistry. We suggested that East German chemists change their topics considerably after 1990. The statistically insignificant coefficient of *post95* in the East German models do not support our conjecture. H1C must therefore be rejected.

Table 1. Chemistry OLS regression

				Dependent variable: Cosine similarity				
	Full period (1)	Full period (2)	Pre (3)	Pre (4)	East (5)	East (6)	West (7)	West (8)
diff_part	-0.004**	-0.009***	-0.004**	-0.007**				
	0.0017	(0.002)	(0.002)	(0.004)				
post95	-0.017**	-0.020***			0.009	-0.002	-0.018***	-0.020***
	0.009	(0.001)			(0.008)	(0.010)	(0.001)	(0.001)
*diff_part*post95*	0.013***	0.002						
	(0.002)	(0.002)						
sameuni		0.098***		0.092***		0.119***		0.096***
		(0.002)		(0.004)		(0.011)		(0.003)
Constant	0.217^{***}	0.266^{***}	0.217^{***}	0.182^{***}	0.237^{***}	0.228^{***}	0.217^{***}	0.260^{***}
	(0.001)	(0.012)	(0.001)	(0.030)	(0.007)	(0.024)	(0.001)	(0.013)
Uni dummies	NO	YES	NO	YES	NO	YES	NO	YES
Observations	148,647	148,647	47,329	47,329	2,029	2,029	116,696	116,696
R^{2-}	0.002	0.029	0.0001	0.041	0.001	0.080	0.003	0.029
Adjusted R^2	0.002	0.028	0.0001	0.039	0.0001	0.066	0.003	0.028

Note: *p < 0.1; **p <0.05; ***p <0.01

In economics and business administration we proposed that East and West German economics and business administration differ before reunification. Here, regression results of model 3 and 4 (see Table 2) confirm our conjecture. The cosine similarity between two dissertation title topic distributions decreases on a large scale when both were written in different parts of Germany and before the reunification. Model 5 and 6 of Table 2 represent the regression results addressing H2B. We proposed that East German economics and business administration differs in topics before and after reunification. In both models we reach significance and a substantial effect of -.27 and -.22 respectively. H2B can therefore be accepted. The last hypothesis, H2C, concerned the question towards increasing similarity between East and West after the reunification. The positive interaction term of *diff_part* and *post95* in the full period models suggests that this the case; H2C can therefore be confirmed.

The coefficient sizes of cosine similarity we used indicate that the effects observed in economics and business administration are of relevant size. This was expected, since the discipline underwent a drastic reorientation after German reunification. In chemistry, the statistically significant effects are much smaller. Chemistry served as an example of how a moderate shock to a discipline affects the problem choices of scientists. Accordingly, East German chemists were therefore expected to change their topics only slightly. The descriptive results suggest that the full convergence of both parts of the country in economics and business happened very quickly, whereas in chemistry convergence was already reached before reunification (see Fig. 1).

Table 2. Economics and business administration OLS regression

	Dependent variable:							
	Cosine Similarity							
	Full period	Full period	Pre	Pre	East	East	West	West
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
diff_part	-0.169***	-0.187***	-0.169***	-0.202***				
	(0.002)	(0.003)	(0.003)	(0.003)				
post95	-0.105***	-0.069***			-0.272***	-0.222***	-0.042***	-0.041***
	(0.002)	(0.001)			(0.007)	(0.009)	(0.002)	(0.002)
*diff_part*post95*	0.147***	0.129***						
	(0.003)	(0.004)						
sameuni		0.086***		0.086***		0.105***		0.074***
		(0.004)		(0.008)		(0.007)		(0.005)
Constant	0.3623***	0.358***	0.362***	0.383***	0.520***	0.546***	0.299***	0.323***
	(0.002)	(0.009)	(0.002)	(0.017)	(0.0004)	(0.042)	(0.002)	(0.009)
Uni dummies	NO	YES	NO	YES	NO	YES	NO	YES
Observations	62,586	62,586	12,784	12,784	3,104	3,104	40,875	40,875
R^{2-}	0.069	0.123	0.202	0.416	0.332	0.419	0.008	0.037
Adjusted R^2	0.069	0.121	0.202	0.409	0.332	0.409	0.008	0.034

Note: *p < 0.1; **p <0.05; ***p <0.01

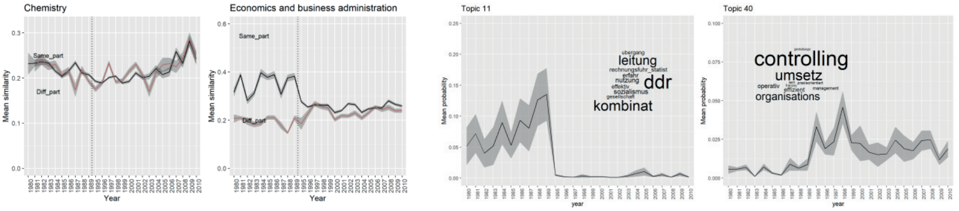


Fig. 1. Yearly mean similarity of dissertation pairs (left) and topical prevalence of two economics and business administration topics (right)

Discussion and Conclusion

In this paper we have shown how research problem choices of scientists change after an unexpected political transition. We investigated dissertation titles in the disciplines of economics and business administration as well as chemistry before and after German reunification in East and West Germany. We find differences between the two parts of Germany in both disciplines before the reunification. These differences decrease partly after the reunification. Moreover, our results suggest that topics predominately composing East German dissertations titles before reunification are significantly different from topics of East German titles after reunification in the field of economics and business administration. The method used in this paper - structural topic modelling and a cosine similarity based approach - aimed to detect differences in research topics of East and West German scientists. As demonstrated, this turned out to be successful; our trained model detects reasonable differences in a set of unseen titles. The inclusion of document level variables, like regional affiliation to universities, into training a topic model can be considered as a decisive advantage of our approach. Research problem choice is dependent on various factors such as regional and temporal background. Therefore, those factors should be considered as non-random in the topic model process. From the visual inspection of the most probable words in economics and business administration, we conclude that our model was able to learn meaningful relationships. The usage of short documents, as in our case dissertation titles, did not turn out to be a problem. In the application to the unseen documents, which provided the basis for our hypothesis testing, our algorithm worked well. Therefore, our methodical approach can be considered as valid. As topic modelling does not aim to label the detected topics, we can only guess what the found differences and their underlying topics most likely refer to. This is a major disadvantage of any sort of topic modelling. The foundation of this problem arises from language as a dynamic, complex and strongly context related semantic system. Topic models can only learn about the relations in this system, but not understand and label them accordingly. It is therefore beyond the scope of our paper to find reasonable labels for topics we detected. The linkage of our data to scientific success and impact measures could provide interesting further research questions. It could be investigated what topical choices are associated with payoffs in the academic reward system. Also the investigation of other types of documents could be of promising. Abstracts and scientific articles may contain document level information which could in the same way shift topic proportions as the variables used in this paper did. Due to the longer documents, the topic model algorithms would exponentially increase calculation time in these cases, but gain statistical properties and topic quality. Therefore, our used method of structural topic modelling in combination with a cosine similarity based regression framework generally offers hug potential for applications in scientometrics and higher education research.

[1] We excluded observations from Uni Berlin, since it is not sure whether the underlying university is in East or West Berlin.

[2] Different languages can distort the outcomes of the algorithm considerably. This problem is a matter of tokens. Although words can have the exact same meaning in two languages, they are statistically considered as different tokens in text machines. Solutions based on translation provoke more problems than they solve. Our approach is therefore to exclude all titles written in English. We might miss some important dissertations, which are addressed to an international auditory, with this procedure. Moreover, dissertations written in German might also differ in quality of the underlying thesis. Nevertheless, as our used language identification algorithm (Ooms, 2018) shows, English titles account for roughly 10% percent of the dissertations. The small number of English titles would therefore distort the statistical inference based on topic modelling. All titles identified as neither German or English are defaulted to German.

[3] Some words are bounded by nature, like "United" and "States". To improve the performance of our topic model, we want the algorithm to treat these words as one token. Bi-grams are two bounded and tri-grams three bounded words. In both corpora we counted to most frequent bi- and tri-grams. We assume that only the top bi- and tri-grams are adding relevant context for the later algorithm. For the both disciplines we set the boundary for relevant n-grams at top 1 $^{\%}$. We proceed by searching these n-grams in every string. If they occur, we add them to the string and remove the words that composed them.

[4] We set the threshold at upper 0.1 % bound of the top words. This is because of complexity reduction and minor relevance for topic modelling.

[5] Since topic modelling infers the topic distribution for every title by drawing words from the title a large number of times, titles consisting of only few words can be problematic, because there is less room for randomness in each title. We therefore exclude titles containing less than five words.

References

Abbott, A. (2001). Chaos of disciplines. Chicago: University of Chicago Press.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M. (2013). A Practical Algorithm for Topic Modelling with Provable Guarantees. In S. Dasgupta, D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning, Volume 28 of Proceedings of Machine Learning Research*, (pp. 280–288). Atlanta: PLMR.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1), 17-35.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Borjas, G. J., & Doran, K. B. (2012). The collapse of the Soviet Union and the productivity of American mathematicians. *The Quarterly Journal of Economics*, 127(3), 1143-1203.

Buenstorf, G., & Geissler, M. (2014). Like Doktorvater, like Son? Tracing Role Model Learning in the Evolution of German Laser Research. *Jahrbücher für Nationalökonomie und Statistik*, 234(2-3), 158-184.

Cooper, M. H. (2009). Commercialization of the university and problem choice by academic biological scientists. Science, *Technology, & Human Values*, 34(5), 629-653.

Debackere, K., & Rappa, M. A. (1994). Institutional variations in problem choice and persistence among scientists in an emerging field. *Research Policy*, 23(4), 425-441.

Dapgusta, P., & David, P. A. (1994). Toward a new economics of science. *Research policy*, 23(5), 487-521.

Fisher, R. L. (2005). The research productivity of scientists: How gender, organization culture, and the problem choice process influence the productivity of scientists. Lanham: University Press of America.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5), 875-908.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2), 981-998.

Horlings, E., Gurney, T. (2013). Search strategies along the academic lifecycle. *Scientometrics*, 94(3), 1137-1160.

Kuhn, T. S. (1962). The structure of scientific revolutions. Chicago: University of Chicago press.

Knorr-Cetina, K. (2009). Epistemic cultures: How the sciences make knowledge. Cambridge: Harvard University Press.

Kolloch (2001). Abwicklung und Neuaufbau der wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin zwischen November 1989 und Dezember 1993. In Theißen, F., Editor, Zwischen Plan und Pleite. Erlebnisberichte aus der Arbeitswelt der DDR. Bühlau Verlag

Leahey, E., Reikowsky, R. C. (2008). Research specialization and collaboration patterns in sociology. *Social Studies of Science*, 38(3), 425-440.

Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American sociological Review*, 22(6), 635-659.

Ooms, J. (2018). cld3: Google's Compact Language Detector 3. Retrieved Feburary 7, 2019 from: https://cran.r-project.org/web/packages/cld3/cld3.pdf, version 1.1.0.

Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing microblogs with topic models. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 130-137). Menlo Park: The AAAI Press

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., Rand, D.G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. Journal of the American Statistical Association, 111(515), 988-1003.

Ziman, J. M. (1987). The problem of "problem choice". *Minerva*, 25(1-2), 92-106.

Zuckerman, H. (1978). Theory choice and problem choice in science. *Sociological Inquiry*, 48(3-4), 65-95.

Appendix: Top 4 words highest β probability

Topic	Economics and business administration	Chemistry
1	‘and,‘the’,‘portfolio’,‘development’,‘model’	‘radikal’,‘selektiv’,‘alk’,‘alkohol’,‘additions’
2	‘integration’,‘kost’,‘internationalisier’,‘beruf’,‘option’	‘kohlenwasserstoff’,‘konzept’,‘oxidativ’,‘methan’,‘methanol’
3	‘prozess’,‘wissenschaft_technisch’,‘technisch_fortschritt’,‘rationalisier’,‘wissenschaft_technisch_fortschritt’	‘funktionalisiert’,‘baustein’,‘verbruckt’,‘chrom’,‘gold’
4	‘schwerpunkt’,‘konzentration’,‘steuerpolit’,‘entwurf’,‘steuerreform’	‘oberflach’,‘adsorption’,‘wasserstoff’,‘wechselwirk’,‘ftir’
5	‘bereich’,‘energie’,‘rationell’,‘brd’,‘konsumgut’	‘the’,‘complex’,‘with’,‘based’,‘catalyst’
6	‘einsatz’,‘erfolgsfaktor’,‘internet’,‘onlin’,‘medi’	‘elektron’,‘clust’,‘zust’,‘fest’,‘spin’
7	‘wandel’,‘osterreich’,‘qualitativ’,‘organisator’,‘inner’	‘olefin’,‘einsatz’,‘stabilisier’,‘homog’,‘ylid’
8	‘mittelstand’,‘intern’,‘berat’,‘modell’,‘unternehmensberat’	‘basis’,‘vorstuf’,‘hinblick’,‘para_phenyl’,‘poly_para’
9	‘industri’,‘effekt’,‘branch’,‘preis’,‘west’	‘platin’,‘komplexbild’,‘cis’,‘stabilitat’,‘phenyl’
10	‘problem’,‘sozialist’,‘beding’,‘volks’,‘aufgab’	‘stereoselektiv’,‘enantioselektiv’,‘enantiomerenrein’,‘aminosaur’,‘diastereoselektiv’
11	‘ddr’,‘kombinat’,‘leitung’,‘sozialismus’,‘nutzung’	‘dynam’,‘synthet’,‘natur’,‘membran’,‘relaxation’
12	‘strategi’,‘ziel’,‘orientiert’,‘unternehmenskris’,‘bewalt’	‘hoh’,‘flussigkristall’,‘niedermolekular’,‘mesog’,‘nemat’
13	‘industriell’,‘aspekt’,‘determinant’,‘organisator’,‘entwicklungsland’	‘ubergangsmetall’,‘phosphan’,‘redox’,‘cyclopentadienyl’,‘fragment’
14	‘extern’,‘prognos’,‘bau’,‘qualitatssicher’,‘steuerungs’	‘for’,‘element’,‘paramet’,‘gallium_indium’,‘aluminium_gallium_indium’
15	‘okonomi’,‘geld’,‘sozial’,‘kritik’,‘okologi’	‘flussigkristallin’,‘amphiphil’,‘monom’,‘phasenverhalt’,‘grenzflach’
16	‘produkt’,‘innovativ’,‘finanz’,‘backed_securiti’,‘rentenversicher’	‘pfeil_recht’,‘eis’,‘typs’,‘eta’,‘mangan’
17	‘polit’,‘institution’,‘quality’,‘histor’,‘islam’	‘bzw’,‘alkaloid’,‘strukturaufklar’,‘pyrrol’,‘cyclisier’
18	‘hintergrund’,‘funktion’,‘okonometr’,‘verander’,‘jung’	‘rhodium’,‘carb’,‘iridium’,‘alkin’,‘zweikern’
19	‘unt’,‘logist’,‘bes_beruck’,‘textil’,‘sektor’	‘elektro’,‘uberbruckt’,‘chromatographi’,‘komplexier’,‘sigma’
20	‘beurteil’,‘anhand’,‘usa’,‘wettbewerbspolit’,‘betracht’	‘unt’,‘phosphor’,‘dimethylamino’,‘phosphoran’,‘nitro’
21	‘forder’,‘mittl’,‘auswahl’,‘massnahm’,‘qualitats’	‘verhalt’,‘cycloaddition’,‘dien’,‘abfang’,‘triazin’
22	‘okolog’,‘nachhalt’,‘sozial’,‘global’,‘umwelt’	‘diel_ald’,‘neutral’,‘hetero_diel’,‘hetero_diel_ald’,‘selektivitat’
23	‘rahm’,‘komplex’,‘nutzung’,‘weiter’,‘beitr’	‘strukturell’,‘kupf’,‘praparativ’,‘oxid’,‘aspekt’
24	‘basis’,‘fuzzy’,‘marktforsch’,‘neuronal_netz’,‘einkaufsstattenwahl’	‘situ’,‘111’,‘adsorption’,‘surfac’,‘non’
25	‘steu’,‘ermittl’,‘kapitalgesellschaft’,‘finanzier’,‘grenzuberschreit’	‘aromat’,‘alkyl’,‘phenol’,‘aliphat’,‘chloriert’
26	‘forschung’,‘betriebs’,‘kost’,‘sicher’,‘kontroll’	‘ausgewahlt’,‘vergleich’,‘ungesattigt’,‘gegenub’,‘substrat’
27	‘bank’,‘roll’,‘kulturell’,‘kund’,‘unternehmenskultur’	‘methyl’,‘total’,‘hydroxy’,‘zugang’,‘est’
28	‘optimier’,‘kommunikation’,‘softwar’,‘mittel’,‘losung’	‘stickstoff’,‘phosphor’,‘schwefel’,‘kohlenstoff’,‘sauerstoff’
29	‘verbesser’,‘qualitat’,‘verwend’,‘neuronal_netz’,‘kunstlich_neuronal’	‘aufbau’,‘messung’,‘druck’,‘temperatur’,‘mpa’
30	‘technisch’,‘informations’,‘rechnergestutzt’,‘darstell’,‘betriebs’	‘iii’,‘oxo’,‘tris’,‘vanadium’,‘chlor’
31	‘struktur’,‘japan’,‘gesellschaft’,‘dimension’,‘alternativ’	‘analyt’,‘modifiziert’,‘hplc’,‘biolog’,‘trennung’
32	‘information’,‘integration’,‘verteilt’,‘heterog’,‘wertorientiert’	‘thermisch’,‘photochem’,‘omega’,‘isomerisier’,‘lamda’
33	‘dynam’,‘optimal’,‘linear’,‘investition’,‘finanzplan’	‘stereo’,‘verwandt’,‘tran’,‘cis’,‘grundlag’
34	‘bezieh’,‘zusammenarbeit’,‘industrieland’,‘nord’,‘kapitalist’	‘modell’,‘einfach’,‘quantenchem’,‘porphyrin’,‘chinon’
35	‘schweiz’,‘wandel’,‘natur’,‘welt’,‘option’	‘metall’,‘modell’,‘chelat’,‘rhenium’,‘haltig’
36	‘uber’,‘zentral’,‘regel’,‘gesetz’,‘plan’	‘dihydro’,‘eta’,‘kenntnis’,‘lambda’,‘sigma’
37	‘sicht’,‘institutionen’,‘betracht’,‘schweizer’,‘wettbewerbsfah’	‘naturstoff’,‘transformation’,‘allyl_substitution’,‘beitr’,‘biolog_aktiv’
38	‘entscheid’,‘computergestutzt’,‘raum’,‘werbung’,‘grenz’	‘verwend’,‘amorph’,‘loslich’,‘kohlenhydrat’,‘materiali’
39	‘risiko’,‘risik’,‘privat’,‘ventur_capital’,‘banking’	‘peptid’,‘konformation’,‘modifizier’,‘zyklisch’,‘racematspalt’
40	‘controlling’,‘umsetz’,‘organisations’,‘operativ’,‘effizient’	‘las’,‘ungewohn’,‘immobilisiert’,‘matrix’,‘studium’
41	‘wirkung’,‘tourismus’,‘mark’,‘stadt’,‘verhaltenswissenschaft’	‘delta’,‘trag’,‘tetra’,‘symmetr’,‘kristallisation’
42	‘markt’,‘industrie’,‘ausgewahlt’,‘fallstudi’,‘transformation’	‘ubergangs’,‘titan’,‘rontgenstrukturanalys’,‘koordination’,‘semiempir’
43	‘hilf’,‘landlich’,‘gebiet’,‘technisch’,‘kennzahl’	‘hilf’,‘infrarot’,‘lichtinduziert’,‘zeitaufgelost’,‘berechn’
44	‘perspektiv’,‘neu’,‘system’,‘regionalpolit’,‘reformvorschlag’	‘gas’,‘massenspektrometr’,‘nachweis’,‘elementar’,‘partiell’
45	‘automobilindustri’,‘netzwerk’,‘kooperation’,‘virtuell’,‘interkulturell’	‘poly’,‘styrol’,‘polystyrol’,‘initiator’,‘copolymerisation’
46	‘servic’,‘financial’,‘engineering’,‘performanc’,‘integration’	‘analoga’,‘festphasen’,‘aufbau’,‘kombinator’,‘strategi’
47	‘arbeit’,‘einflussfaktor’,‘diagnos’,‘grundsatz_ordnungsmass’,‘grundsatz_ordnungsmass_bilanzier’	‘molekul’,‘photo’,‘fluoreszenz’,‘raman’,‘induziert’
48	‘aspekt’,‘ergebnis’,‘land’,‘licht’,‘studi’	‘wassrig’,‘gamma’,‘sio2’,‘al2o3’,‘tio2’
49	‘personal’,‘syst’,‘evaluation’,‘fuhrungskraft’,‘fuhrung’	‘bindung’,‘aktivier’,‘funktionalisier’,‘aktiviert’,‘alkylier’
50	‘staatlich’,‘staat’,‘zusammenhang’,‘land’,‘gesellschaft’	‘optisch’,‘magnet’,‘farbstoff’,‘elektr’,‘schicht’
51	‘makro’,‘fundiert’,‘verhalt’,‘erklar’,‘arbeitsmarkt’	‘amin’,‘amino’,‘ring’,‘aryl’,‘substituent’
52	‘alternativ’,‘losung’,‘geldpolit’,‘finanziell’,‘entscheidungs’	‘molekul’,‘theoret’,‘ion’,‘zeolith’,‘umlager’
53	‘produktion’,‘effektivitat’,‘flexibl’,‘fertig’,‘vorbereit’	‘silicium’,‘kristall’,‘silan’,‘sol_gel’,‘silicat’
54	‘innovation’,‘erfolgreich’,‘fallbeispiel’,‘innovations’,‘organisational’	‘ternar’,‘kristall’,‘lithium’,‘alkali’,‘lanthanoid’
55	‘aufbau’,‘praktisch’,‘rahm’,‘unternehmensfuhr’,‘ansatzpunkt’	‘nickel’,‘koordinations’,‘zink’,‘cobalt’,‘silb’
56	‘marketing’,‘national’,‘einzelhandel’,‘determinant’,‘interaktion’	‘katalyt’,‘mono’,‘aufklar’,‘hydrier’,‘umwandl’
57	‘bestimm’,‘simulation’,‘hilf’,‘system’,‘eignung’	‘cyclisch’,‘umsetz’,‘nucleophil’,‘bzw_beziehungsweis’,‘elektrophil’
58	‘prozess’,‘modellier’,‘unternehmens’,‘dynam’,‘mittel’	‘grupp’,‘element’,‘amid’,‘nebengrupp’,‘moglich’
59	‘regional’,‘studi’,‘rechnungsleg’,‘ifr’,‘bilanzier’	‘oxidation’,‘mechanismus’,‘ruthenium’,‘reduktion’,‘gegenwart’
60	‘gross’,‘markt’,‘operationalisier’,‘bereitstell’,‘erfolgswirk’	‘katalysator’,‘palladium’,‘polymerisation’,‘eth’,‘katalys’
61	‘integriert’,‘unterstutz’,‘technologi’,‘ganzheit’,‘prozessorientiert’	‘optisch_aktiv’,‘pro’,‘baustein’,‘alkohol’,‘katalys’
62	‘einfuhr’,‘business’,‘gruppenarbeit’,‘organisator’,‘produktionsbereich’	‘analys’,‘gebund’,‘optimier’,‘spektr’,‘gaschromatograph’
63	‘dienstleist’,‘relevanz’,‘beschaff’,‘zusammenarbeit’,‘kooperation’	‘wass’,‘syst’,‘thermodynam’,‘mischung’,‘kritisch’
64	‘steuer’,‘konzeptionell’,‘handel’,‘dezentral’,‘handels’	‘addition’,‘versuch’,‘lithium’,‘aldehyd’,‘ungesattigt_ungesattigt’
65	‘rahmenbeding’,‘institutionell’,‘kommunal’,‘bundesland’,‘medizin’	‘wechselwirk’,‘festkorp’,‘hilf’,‘schwach’,‘saur’
66	‘basis’,‘verfahr’,‘entscheidungsorientiert’,‘krankenhaus’,‘energieversorgungs’	‘oligo’,‘sensor’,‘dendrim’,‘kunstlich’,‘potentiell’
67	‘bedeut’,‘entwicklungsland’,‘implikation’,‘wirtschaftspolit’,‘gegenwart’	‘linear’,‘anelliert’,‘thioph’,‘oligom’,‘nichtlinear_optisch’
68	‘region’,‘untersucht’,‘china’,‘strukturwandel’,‘berlin’	‘molekular’,‘anion’,‘experimentell’,‘supramolekular’,‘modellier’
69	‘einfluss’,‘zeitverwend’,‘ausgewahlt’,‘grenz’,‘faktor’	‘protein’,‘dna’,‘wechselwirk’,‘enzymat’,‘inhibitor’
70		‘via’,‘diel_ald’,‘lewis_saur’,‘steroid’,‘intramolekular_diel’
71		‘massenspektrometri’,‘kopplung’,‘icp’,‘prob’,‘direkt’
72		‘struktur’,‘organo’,‘rontgenograph’,‘schwingungs’,‘alkali’
73		‘typ’,‘mechanism’,‘extraktion’,‘chemistry’,‘imidazolin’
74		‘donor’,‘biphenyl’,‘wirt_gast’,‘helical’,‘axial’
75		‘bildung’,‘verfahr’,‘effekt’,‘zerfall’,‘berucksicht’
76		‘verschied’,‘kation’,‘voraussetz’,‘carbonyl’,‘induziert’

Constructing vision-driven indicators to enhance interaction between science and society

Asako Okamura[1] and Keisuke Nishijo[1,2]

[1] *a-okamura@grips.ac.jp*
SciREX Center, National Graduate Institute for Policy Studies (GRIPS), 7-22-1 Roppongi, Minato-ku, Tokyo 106-8677(Japan)
[2] *knw1996knw@gmail.com*
Department of Civil Engineering, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 (Japan)

Abstract

In this fast-paced modern world, science, technology, and innovation (STI) impacts all areas of life, at both the individual and organizational levels. The degree of acceptance and absorption capacity of STI vary across people and organizations both within countries and at a cross-cultural level. We need a deeper understanding of the cultural and social aspects of STI, as the basis of individual and collective values, choices, behaviours, and risk-preferences related to STI and a source of knowledge creation. However, the question is how do we measure these concepts and acquire meaningful indicators to inform society and policymaking. This paper introduces our project's experiment of designing such indicators through 'vision-driven' approaches.

Introduction

In today's fast-paced world, science, technology, and innovation (STI) impacts all areas of life, at both the individual and organizational levels. The degree of acceptance and absorption capacity of STI vary for different people and organizations, both within countries and at a cross-cultural level. Given its impacts on individual and collective values, choices, and behaviours, it is clear that there is need for a deeper understanding of the cultural and social aspects of STI. To do so, we must first define these concepts and acquire meaningful indicators to inform society and policymaking practices. This study in progress proposes a design for creating such indicators. Since 2016, the SciREX Center's 'Measurement of STI and Society' project has attempted to identify indicators that can be used to describe an ideal relationship between STI and society. Such indicators can be used to encourage behavioural changes in individual actors and also assess the expected impacts of STI and society interactions. As a step towards that, we adopt a 'vision-driven' approach by holding several workshops to develop a plan for creating those indicators. The project consisted of researchers, mainly in science policy studies, policymakers, and graduate students as interns. This paper introduces the process of that discussion and updates Okamura (2017).

Rationale

As STI continues to penetrate more deeply into society, the problems related to its risks and governance are becoming more pronounced. There is a need to reconsider the direction that the progress in STI is taking, in mind of the incorporation of individual values and overarching social vision. Moreover, individuals and organizations' unique values, practices, and morals affect their perceptions and acceptance behaviours, and, therefore, have the potential to contribute to the development of STI in different ways. It is important to deal with this matter also on a policy level, as has been done in efforts oriented towards 'Responsible Research and Innovation (RRI)' in Europe. To advance such efforts in practices, it is necessary to identify indicators that can be used to measure the relationship between society and science.

Statistics and analysis for grasping STI activities, such as human resources in STEM fields and expenditure on Research and Development, have advanced mainly on long-standing efforts by the Organization for Economic Cooperation and Development (OECD) and similar organizations, however, the role of individuals and society in the creation and dissemination of STI has received less attention.

We incorporated into our study certain features of international efforts to develop indicators of science and society, particularly those underway in Europe (Peter et al., 2018) as well as Sustainable Development Goal Indicators (SDGIs). The study is also aimed to assist in the monitoring of relevant policy in Japan, including the Fifth Science and Technology Basic Plan.

Why do we need a 'vision-driven' approach?

The objective of the indicators for policymakers, researchers, and citizens is that they be able to grasp the current situation and promote change in their behaviour, towards building an ideal relationship between science and society. The relationship between STI and society is multi-layered and cannot be wholly encapsulated in a single indicator. There is, moreover, no established framework for what to measure and how to measure it. We need to understand the relationship between STI and society comprehensively, from a bird's eye view, and then determine what type of indices can be used to measure it.

What are 'vision-driven' indicators?

In preparing the draft indicator, we aimed to bring in 'perspective from society and individuals' and aimed not to focus on its relationship with STI too much. Therefore we also covered the visions and targets whose relationship with STI is not so explicit.

Framework

In discussing this project, we first considered the logic models that are widely used for policy evaluation, beginning with a consideration of the desired vision. Due to the complexity of the relationship between society and science, we decided to use a more simplified framework, as shown in Figure 1.

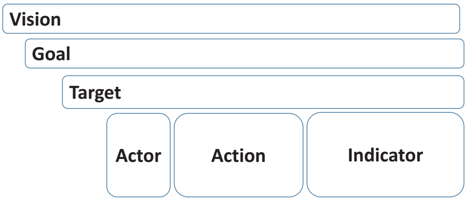


Figure 1. Framework of vision, goal, target, action, and indicator

We also referred to the frames used by the United Nations in the SDGIs, which set 169 sustainable development targets for 17 goals towards 2030 and listed 232 indicators for monitoring these goals (United Nations, 2019). To make the proposed indicators become more specific for each actor, we expanded the frame of SDGIs and introduced actors and actions in addition to goals and targets.

Vision

We first discussed the direction of social changes and social visions along with a discussion on how individual values and perspectives are changing, and categorized the social visions into:
- Smart and resilient
- Flexible towards changes, and accepting/utilizing diversities
- Rewarding challenges

Then, with these visions, we defined 'desirous interactions between science and society' as a state in which the following situations hold:
- STI is promoted in a form responsible to society
- Individuals and society can adapt to STI's transformative changes

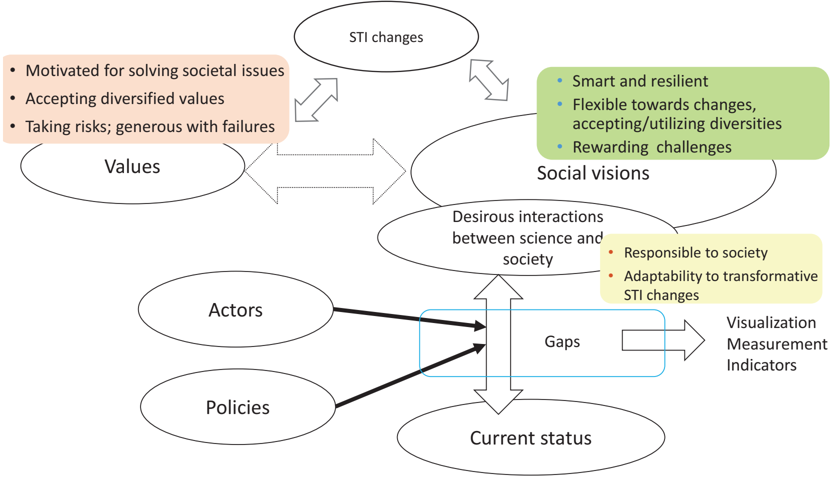

STI changes
Motivated for solving societal issues
Accepting diversified values
Taking risks; generous with failures
Values
Smart and resilient
Flexible towards changes, accepting/utilizing diversities
Rewarding challenges
Social visions
Desirous interactions between science and society
Responsible to society
Adaptability to transformative STI changes
Actors
Gaps
Visualization
Measurement
Indicators
Policies
Current status

Figure 2: Concept of the whole structure

Goals and targets

To achieve the social visions and accompanying 'desirous interactions between science and society' (in SDGIs, it is 'sustainable development'), we established six 'goals' and 21 'targets' to achieve goals (Figure 3). Goals and targets were expressed as states and conditions to be satisfied.

How society faces STI:
A. Citizens understand the uncertainty and ethical aspects of STI, and take an appropriate and integrated approach to addressing science abuse, misuse, and deviation: In this state, trust has been formed among citizens, government, and experts (trusting experts and delegating), while at the same time, the number of people deceived by science abuse / pseudoscience is decreasing.

B. Citizens enjoy science and innovation culture, and are active as knowledgeable players: There are diverse pathways of scientific inquiry in daily life, and citizens' scientific literacy and interest are increasing; there is development of a culture that supports STI.

How to reflect the opinions of citizens and experts in decision-making:
C. Citizens' expectations and concerns, and scientists' advice are properly woven in the policymaking process: Citizens and scientists are concerned about social problems and involved in policy evaluation, formation, and implementation.

Targets for Goals A, B and C, regarding social infrastructure:
I. No regional gap in accessing knowledge and culture: While keeping uniqueness of rationalities, access to knowledge and culture is secured without regional bias.

II. Culture that boosts challenges and follows up on them: This is a social condition under which society is able to fail and also encourage various challenges.

III. Society that accepts diversity, allowing citizens to enjoy life: Society accepts and respects diversity, and there is a foundation for securities and a work-life balance.

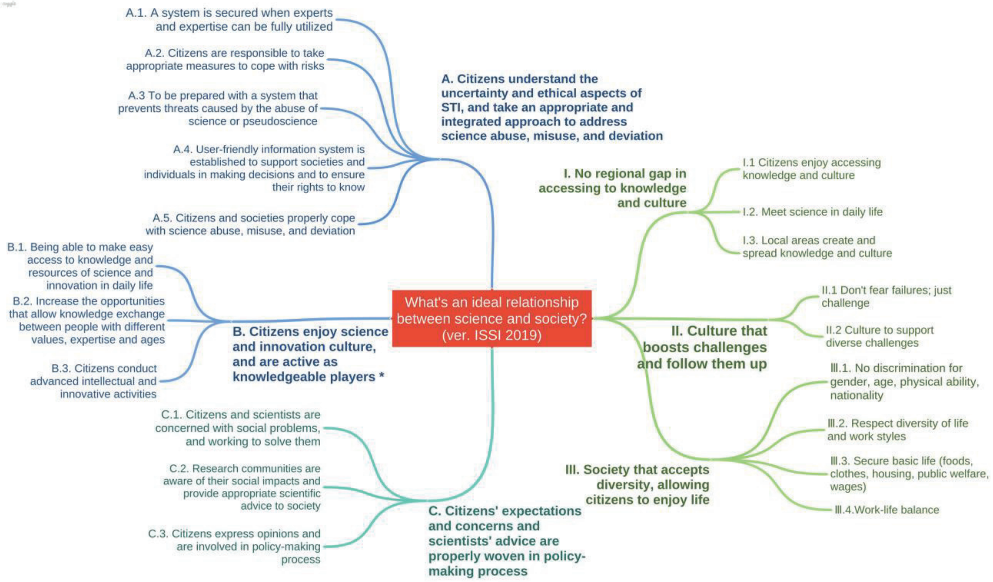

What's an ideal relationship between science and society? (ver. ISSI 2019)
A. Citizens understand the uncertainty and ethical aspects of STI, and take an appropriate and integrated approach to address science abuse, misuse, and deviation
A.1. A system is secured when experts and expertise can be fully utilized
A.2. Citizens are responsible to take appropriate measures to cope with risks
A.3 To be prepared with a system that prevents threats caused by the abuse of science or pseudoscience
A.4. User-friendly information system is established to support societies and individuals in making decisions and to ensure their rights to know
A.5. Citizens and societies properly cope with science abuse, misuse, and deviation
B. Citizens enjoy science and innovation culture, and are active as knowledgeable players *
B.1. Being able to make easy access to knowledge and resources of science and innovation in daily life
B.2. Increase the opportunities that allow knowledge exchange between people with different values, expertise and ages
B.3. Citizens conduct advanced intellectual and innovative activities
C. Citizens' expectations and concerns and scientists' advice are properly woven in policy-making process
C.1. Citizens and scientists are concerned with social problems, and working to solve them
C.2. Research communities are aware of their social impacts and provide appropriate scientific advice to society
C.3. Citizens express opinions and are involved in policy-making process
I. No regional gap in accessing to knowledge and culture
I.1 Citizens enjoy accessing knowledge and culture
I.2. Meet science in daily life
I.3. Local areas create and spread knowledge and culture
II. Culture that boosts challenges and follow them up
II.1 Don't fear failures; just challenge
II.2 Culture to support diverse challenges
III. Society that accepts diversity, allowing citizens to enjoy life
III.1. No discrimination for gender, age, physical ability, nationality
III.2. Respect diversity of life and work styles
III.3. Secure basic life (foods, clothes, housing, public welfare, wages)
III.4.Work-life balance

Figure 3: Goals and targets

Actions and Indicators

Finally, we set 'actions' necessary for each actor. Actors comprise citizens, government, the scientific community (research and education), and media and business corporations. We ended up listing 192 actions with 246 indicators (Figure 4). Both qualitative and quantitative indicators were examined. To enable the development of future indicators, we decided include indicators which were not currently available.

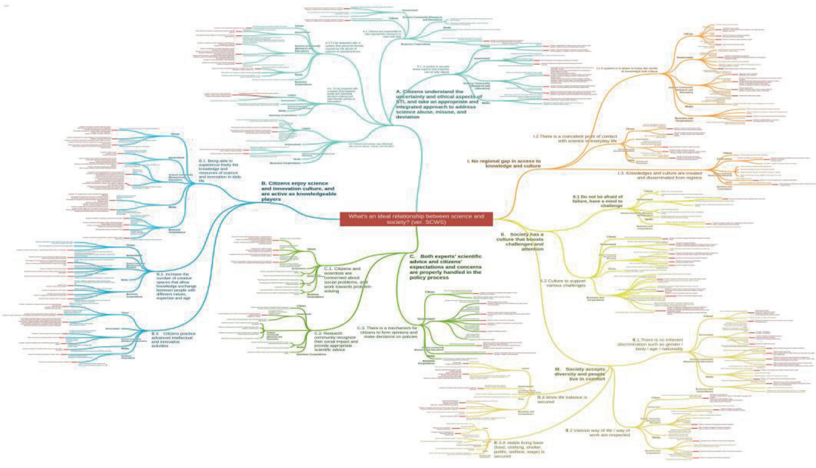

What's an ideal relationship between science and society? (ver. SCWS)
A. Citizens understand the uncertainty and ethical aspects of STI, and take an appropriate and integrated approach to address science abuse, misuse, and deviation
B. Citizens enjoy science and innovation culture, and are active as knowledgeable players
C. Both experts' scientific advice and citizens' expectations and concerns are properly handled in the policy process
C.1. Citizens and scientists are concerned about social problems, and work towards problem-solving
C.2. Research community recognize their social impact and provide appropriate scientific advice
C.3. There is a mechanism for citizens to form opinions and make decisions on policies
I. No regional gap in access to knowledge and culture
I.2.There is a coincident point of contact with science in everyday life
II. Society has a culture that boosts challenges and attention
II.1 Do not be afraid of failure, have a mind to challenge
II.2 Culture to support various challenges
III. Society accepts diversity and people live in comfort

Figure 4: An illustration of goals, targets, actions, and indicators

The whole list showing goals, targets, actions, and indicators is available at:
https://coggle.it/diagram/XPZ9Ds29Yi-bA_PZ/t/what's-an-ideal-relationship-and-society-ver-issi-2019/0aab7f88d4d7b96654300a1c63adb117fca8744a398d8e7f4b9aca21f07bede5

Table 1 shows examples of a comparison between STI and society and SDGIs, to facilitate the understanding of the indicator structures.

Table 1. Comparison between 'vision-driven' indicators for STI and society and SDGIs

	'Vision-driven' indicators for STI and society	*SDGIs*
Superordinate concept / Vision	Desirous interactions between science and society	Sustainable development
Goal	B. Citizens enjoy the science and innovation culture, and are active as knowledgeable players	Goal 1. End poverty in all its forms everywhere
Target	B.2. Increase the opportunities that allow knowledge exchange between people with different values, expertise, and ages	1.1 By 2030, eradicate poverty, currently measured as people living on less than $1.25 a day, for all people everywhere
Actor and action	Citizen Have platforms that allow sharing inventions within the community via internet	Not mentioned
Proposed indicator	E.g., Number of creative share houses, manufacturing-type share houses, and fab labs	1.1.1 Proportion of population below the international poverty line, by sex, age, employment status, and geographical location (urban / rural)

We also created a concordance table with SDGIs and MoRRI indicators to identify similarities and differences in our proposals.

How do we make 'vision-driven' indicators?

There was no universal method applied to the process of deriving vision, goal, target, actor, action, and indicator, so we followed an exploratory step. We used a qualitative method of holding multiple workshops (four invitation-only workshops and four open discussions), through which we formed concepts and ideas with the help of discussions with participants. As there are diverse points of contact between science and society, we thought it important to first elucidate the concepts themselves, clustering participant responses on them, and then use these to form an image of the overall structure. There was repeated divergence and convergence throughout the process, eventually resulting in the creation of a draft indicator plan.

Limitations

Methods based on discussions with participants have limitations, particularly a lack of diversity of participants and an absence of theory in responses. Our study is a trial effort and features a limited number of stakeholders, but we intend to use it as a basis for a more systematic and comprehensive framework. In order to make such a framework more universal, valid, and convincing, it will be necessary to involve more stakeholders with diverse perspectives.

Monitoring the level of action allows us to assess whether the progress of each actor promotes a positive relationship between STI and society. Doing so entails verifying theoretically and empirically whether each action is linked to an ideal relationship and whether the ideal relationship is consistent with the direction of social vision / social change aimed for. The framework should be completed after continuous theoretical and empirical verification of a trial.

Further steps

By introducing a process to converge from divergence, it is necessary to prioritize and create indicators that contribute to policy formation. To this end, it is important to understand the needs of policymakers. At the same time, it is important to scrutinize what the indicator can and cannot measure, such as what kind of a result the action will lead to, the contribution to the achievement of the target, and whether it is really appropriate as a measure of a particular action. For this reason, it is necessary to continue collaboration with researchers, including experts, on indicators. On the other hand, to capture the various points of contact between STI and society, it is also necessary to create an interactive platform that allows various actors in society to discuss widely on this issue, to review the vision, targets, and actions. In particular, it is also important to devise a method to reflect the ideas of people who are not interested in science (and social relations). Visualization with these indicators may help for this purpose.

Conclusions

As the values of people and society, in addition to the latter's structure, are changing, it has been more difficult to identify what the policy should be aimed at and what indicators will monitor those policies. In SDGs and RRI, indicators have been developed to monitor specific visions. In our project, to establish the desirous relationship between science and society, we propose what kind of behavioural changes are necessary for actors, such as citizens, policymakers, researchers, the media, and industry, to make. We take an experimental approach to develop the indicator plan by holding multiple workshops. Our efforts can be called a 'participatory indicator development'. The results obtained therein are still limited, and the methodology is not rigorous. At the same time, there are merits and demerits in setting indicators as policy goals in policy formulation; it will be necessary to reconsider the implications. We hope that our efforts will bring in suggestions for establishing a new framework for the development of indicators on social implications of STI in the future.

Acknowledgments

This work was supported by Science for RE-designing Science, Technology and Innovation Policy (SciREX) Program in Japan. The authors would like to thank all the members of 'Measurement of STI and Society' project at the SciREX Center as well as the participants of the workshops and open-forum discussions for their contributions. The views and opinions expressed here are solely those of the authors, and they do not, in any way, represent those of the institution to which they are affiliated.

References

Okamura, Asako. (2017). Measurement of cultural and social relevance of science: construction of indicators for the relationship between STI and society, STI 2017 - Science, Technology and Innovation indicators, Poster Session.

Peter, Viola et al. (2018). Monitoring the evolution and benefits of responsible research and innovation in Europe: Summarising insights from the MoRRI Project, May 2018. Retrieved from https://www.technopolis-group.com/report/final-report-summarising-insights-from-the-morri-project-d13/

United Nations. (2019). The E-Handbook on Sustainable Development Goals Indicators. Retrieved from https://unstats.un.org/wiki/display/SDGeHandbook/Home

The Citations of Papers with Conflicting Reviews and Confident Reviewers

Jiangen He[1] and Chaomei Chen[2]

[1] jiangen.he@drexel.edu [2] chaomei.chen@drexel.edu
Department of Information Science, Drexel University, Philadelphia, United States

Abstract

Disagreement is essential for knowledge growth in science. However, disagreement in peer review is usually regarded as a sign of unreliability in existing studies. The predictive role of disagreement for potentially impactful discoveries was rarely explored. Reviewer expertise has been proved to have effects on review strictness and outcomes, but reviewers' added-value on paper quality remains unclear. In this paper, we examined the predictive effects of disagreement and confidence of reviewers on citations of reviewed papers. Using a dataset of 489 papers submitted ICRL 2017, we found that predictive effects of disagreement and confidence on citations. Among accepted papers, the papers with higher review ratings tend to receive more citations and the disagreement cannot predict citations. However, among rejected papers with more than ten citations, the ratings they received cannot predict their citations, but the disagreement level of their reviewers does. Similar opposite findings were also seen regarding confidence. Accepted papers reviewed by confident reviewer tend to have more citations but rejected papers might not benefit from confident reviewers.

Introduction

Both peer review and citation are important and commonly used methods for research evaluation implemented by peers. Peer review is a fundamental mechanism in science by which manuscripts or proposals are critically evaluated by one or several reviewers to guarantee and improve the quality of research, but it is not perfect in fairness, reliability, validity, and effectiveness (Lee et al., 2013; Siler, Lee, & Bero, 2015). The quantitative and objective measurement of impact provided by citations can offer complement and supportive tools to peer review (van Raan, 1996). Although citations also have a controversial role as a measure of research quality, citations at least provide a reliable measurement of impact (Bornmann & Daniel, 2008c). In the studies of peer review, citations were often used as a proxy for impact to learn the validity and effectiveness of peer review results, i.e., to examine 'Does peer review system select the scientific work potentially to be impactful?' (Rinia, van Leeuwen, van Vuren, & van Raan, 1998; Siler et al., 2015).

The link between citations and peer review results have been examined by many studies. The comparisons between citation counts and peer-review results (e.g. acceptance/rejection and ratings) have been made in different fields, such as medicine (Siler et al., 2015), chemistry (Bornmann & Daniel, 2008a), and interdisciplinary fields (Jirschitzka, Oeberst, Göllner, & Cress, 2017). Most of comparative studies concluded that editors and reviewers generally fulfilled the task of selecting high-quality articles, but some studies also found many articles later shown to be highly cited have encountered initial rejection by referees and/or editors. The resistance might be explained by the radical novelty contributed by the articles that challenged current views or theories in science (Benda & Engels, 2011; Campanario & Acedo, 2007). In summary, the outcomes of citation and peer review are somehow correlated, but the difference between them cannot be ignored, especially the difference regarding highly cited articles that might describe paradigm-shifting discoveries, which may be regarded as the observed subordination of promoting important innovations to the quality control aspects of peer review (Armstrong, 1991). Besides learning how peer-review outcomes predict citations, it is necessary to examine how variables behind each peer-review outcome, including the rating, content, reviewer characteristics of independent peer-review reports, to reach a deeper understanding the gatekeeping role of peer review.

In this study, we studied how disagreement among reviewers and reviewer confidence predict the citations of accepted and rejected paper. The link between the two variables and citations are rarely studied before, though disagreement and reviewer expertise have been studied well.

Disagreement among reviewers is common in the peer review of both manuscripts (Bornmann & Daniel, 2008c) and grant applications (Pier et al., 2018). Inter-reviewer disagreement is also shown by the varying points of view on the issues described in the peer-review reports (Fiske & Fogg, 1990). Unlike previous studies use the degree of agreement among reviewers to investigate the reliability of peer-review judgements, we assumed that the disagreement among reviewers may serve as a sign of controversial and even highly novel scientific work. This raised our first research question, whether inter-reviewer disagreement can predict the citation counts, especially for papers that were rejected but published elsewhere.

Confidence, a variable of reviewer characteristics we investigated, is a self-reported confidence level of a reviewer. Confidence is mainly determined by the intellectual distance between reviewers' knowledge and the knowledge embodied in scientific work (Boudreau, Guinan, Lakhani, & Riedl, 2016), i.e., their expertise regarding the topic of evaluated work. Reviewers' areas of expertise were proved to have systematically effects on their evaluative strictness or leniency in the peer review of grant applications (Boudreau et al., 2016), which challenges the fairness and effectiveness of peer review (Lee et al., 2013). The varying expertise of reviewers on the topics of manuscripts may help the quality improvement of manuscripts at different levels. In this study, we investigate whether the confidence levels of reviewers have effects on their ratings and the citation outcomes of manuscripts.

Data

We investigated open peer reviews of the conference track in the International Conference on Learning Representations 2017 (ICLR 2017). ICLR is one of leading conferences in machine learning established in 2013. The review data are open-access through OpenReview. OpenReview was designed to test and promote openness, especially peer review process, in the scientific community, thus, it allows flexible review policies and collects data about the consequences of different policies. Although ICLR managed their reviewing process by OpenReview since 2013, the year 2017 was the earliest year for which review ratings for conference track are available on OpenReview.net.

489 submissions were included in this study. 198 (40.4%) submissions were accepted by conference track and 291 (59.6%) submissions were rejected or invited to workshop track. For each submission, we collected the integer scores of reviewer ratings ranging from 1 to 10, referee confidence scores, and chair's decision that is acceptance or rejection on OpenReview. Regarding the citation data, we retrieved citation counts from Microsoft Academic by January 27, 2018. We also examined each rejected submission to see if it eventually published in other venues and where it published. However, publishing is not necessary for being cited.

The peer review process of ICLR is double-blind, which means that authors are not aware of reviewer identities and reviewers are not aware of author identities. Most submissions were reviewed by three reviewers (89.6%) and only a few by two, four, or five reviewers (1.4%, 8.2%, and 0.8%, respectively). Besides, each submission received a meta-review from the conference chair about the final decision.

Method

Our central concerns are to measure the relationship between disagreement and citations, and the one between confidence and citations. We devised means of measuring these key variables and identified several well-studied variables affecting citations. An overview of definitions and summary for the main variables are provided in Table 1.

Table 1 Definitions of Main Variables

Variable	*Description*
(1) *CITATIONS*	Citation counts of each paper retrieved from Microsoft Academic on Jan 30, 2019, if available. Use 0, if not available.
(2) *AVG_RATING*	Mean score for the rating scores received by each paper.
(3) *DISAGREEMENT*	Population standard deviation for the rating scores received by each paper.
(4) *AVG_CONF*	Mean score for the confidence scores of reviewers for each paper.
(5) *SD_CONF*	Population standard deviation for the confidence scores of reviewers for each paper.

We built Negative Binomial (NB) and Zero-Inflated Negative Binomial (ZINB) regression models to validate the effects of inter-reviewer disagreement and reviewer confidence on citation impact. Both models were commonly used to model citation counts. A ZINB model includes a NB model for citation count and a logit model for modelling excess zero citations. We conducted a *Vuong test* to learn which model is superior for each dataset and chose the superior one to conduct our analysis. The results of NB models can be interpreted by using estimates of incidence rate ratios (IRRs).

Results

Statistical Description

Table 2 and Table 3 show the description of studied variables and the correlations between variables respectively.

Table 2 Description of variables.

Variables	*Accepted (N=200)*			*Rejected (N=291)*			*All (N=489)*		
	Mean	*SD*	*Median*	*Mean*	*SD*	*Median*	*Mean*	*SD*	*Median*
AVG_RATING	6.89	0.78	7.00	4.86	0.96	5.00	5.69	0.13	5.67
DISAGREEMENT	0.69	0.48	0.47	0.73	0.48	0.71	0.71	0.46	0.47
AVG_CONF	3.80	0.50	4.00	3.77	0.59	3.75	3.78	0.55	3.80
SD_CONF	0.52	0.40	0.47	0.53	0.45	0.53	0.52	0.43	0.47
CITATIONS	69.93	95.37	23.00	16.37	55.93	2.00	38.27	79.05	7.00

Table 3 Correlation matrix of variables.

	AVG_RATING	*DISAGREEMNET*	*AVG_CONF*	*SD_CONF*	*CITATIONS*
AVG_RATING					
DISAGREEMENT	0.05 (0.30)				
AVG_CONF	-0.08 (0.09)	0.07 (0.12)			
SD_CONF	0.04 (0.42)	0.06 (0.15)	**-0.43 (0.00)**		
CITATIONS	**0.39 (0.00)**	-0.02 (0.64)	0.06 (0.16)	0.00 (1.00)	
ACCEPTANCE	**0.75 (0.00)**	-0.04 (0.39)	0.02 (0.61)	-0.00 (0.80)	**0.33 (0.00)**

All Papers

At first, we examined the predictive effects of variables on citations of all the papers submitted to ICRL 2017. We grouped the papers into three sets according to their citations. Vuong test results indicated that NB model is superior to ZINB model for the set of *All*. ZINB is not applicable for the sets of *Citation>10* and *Citation>20*, because their dependent variables are non-zero. Thus, we only reported the results of NB models in Table 4. The results are consistent over three sets. Only *AVG_RATING* tends to have predictive effects on *CITATIONS*. The outcomes of citation and peer review are consistent. The results are in line with findings in previous studies (Bornmann & Daniel, 2008b; Siler et al., 2015).

Table 4 Predictive effects of rating and confidence on citations for all papers

Variable	*All*		*Citation > 10*		*Citation > 20*	
	Estimate	*p*	*Estimate*	*p*	*Estimate*	*p*
AVG_RATING	**0.752**	**0.000**	**0.270**	**0.000**	**0.180**	**0.004**
DISAGREEMENT	0.102	0.524	0.094	0.519	0.046	0.751
AVG_CONF	0.242	0.113	0.176	0.217	0.095	0.573
SD_CONF	0.012	0.949	0.168	0.387	0.001	0.004
N	489 (100%)		211 (43.1%)		142 (29.0%)	
Dispersion parameter	0.370		1.068		1.688	
2*log-likehood	-3,866		-2,278		-1,619	
AIC	3,855		2,290		1,631	
Vuong test (NB > ZINB)	*z*=11.042, ***p*=0.000**		-		-	

Although *AVG_RATING* have consistent effects over sets with different citations, we can see quite different relationships between *CITATIONS* and *AVG_RATING* for accepted and rejected papers in Figure 1. The data were separated by acceptance and rejection, plotted and the fit via simple linear regression. Figure 1 visualized the effects of acceptance and rejection on the relationship between *AVG_RATING* and *CITATIONS*. The citations of rejected papers are less sensitive to review ratings, which suggest that acceptance can be important in modifying the relationship between peer-review outcomes and citations. This raises the need for examining accepted and rejected papers separately.

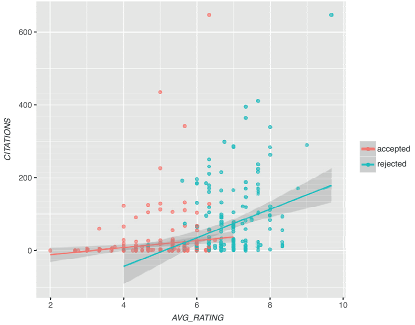


Figure 1 Accepted and rejected papers scatterplots of *AVG_RATING* and *CITATIONS*. The shaded area represents 95% confidence level.

Accepted Papers

For accepted papers, *AVG_RATING* still appears to have the most stable and consistent effects on *CITATIONS*. Accepted papers tend to have more citations if they were reviewed by confident reviewers. However, the benefits are not for all papers. For papers with more than 20 citations, citations were not promoted by *AVG_CONF*. The possible explanation is that the quality improvement brought by reviewer expertise was limited. The core innovation and research content of a scientific manuscript are rarely altered substantially through peer review (Siler et al., 2015). For low-impact papers, the improvement can be observed through citations, but for high-impact papers, the limited improvement may not be reflected by citations.

Table 5 Predictive effects of rating and confidence on citations for accepted papers

Variable	*All accepted*		*Citation > 10*		*Citation >20*	
	Estimate	*p*	*Estimate*	*p*	*Estimate*	*p*
AVG_RATING	**0.432**	**0.000**	**0.290**	**0.005**	**0.242**	**0.014**
DISAGREEMENT	-0.171	0.385	-0.081	0.635	-0.072	0.649
AVG_CONF	**0.535**	**0.006**	**0.373**	**0.039**	0.168	0.367
SD_CONF	0.028	0.906	0.035	0.874	-0.136	0.523
N	200 (100%)		142 (71%)		105 (52.5%)	
Dispersion parameter	0.631		1.150		1.897	
2*log-likehood	-2031		-1569		-1199	
AIC	2043		1581		1211	
Vuong test (NB > ZINB)	*z*=10.45, ***p*=0.000**		-		-	

Rejected Papers

289 of 489 submitted papers were rejected by the conference track of ICRL 2017. The fates of these rejected papers are shown in Table 6. Only 58 (20.0%) of them eventually published in other conferences or journals. Seven conferences accepted at least three rejected papers and they accepted 34 papers in total. All of these seven conferences are leading conferences in artificial intelligence, machine learning, and computational linguistics. About half of unpublished papers were posted on arXiv.

Unpublished papers could be citable. Submitting to ICRL means agree go through the open peer review and all the papers on OpenReview.net can be cited. Besides, papers posted on arXiv can increase the visibility to be cited. For example, the most cited paper among all the submitted papers (including the accepted ones) is an unpublished paper posted on arXiv (Iandola et al., 2016). Thus, we included all the unpublished papers in this study.

Table 6 The fates of rejected papers.

Venue	*Number*	*Percent*
arXiv	124	42.6%
ICML	12	4.1%
ACL	8	2.7%
NeruIPS	3	1%
EMNLP	3	1%
KDD	3	1%
AAAI	3	1%
IJCAI	3	1%
Other	24	8.2%

Unpublished	108	37.1%

There no desk-rejection in the peer review process of ICLR. All the submitted papers have equal chances to receive suggestive and constructive feedback. We examined the review length of rejected and accepted paper to see if reviewers tended to give more comments to accepted papers than rejected papers or if more reviewers were assigned to accepted papers. Reviews of rejected papers (*n* = 463) are on average 1,917 characters, compared for with 1,876 for reviews of accepted papers (n = 684; p=0.91). Both rejected papers and accepted papers were assigned on average 3.15 reviewers. Thus, there is no significant difference between reviews of accepted and rejected papers in terms of quantity.

In Table 7, we cannot see significant effects of *AVG_CONF* on citations of rejected papers, i.e., rejected paper may not have benefited from receiving feedback from reviewers with higher level expertise. The influence of peer review with regard to improving the quality of papers may be more limited to rejected papers. Authors whose papers were rejected might be reluctant to change their articles based on the negative comments from reviewers when they knew their papers were rejected.

AVG_RATING is still effective to predict citations when without restriction on citation counts. The reviewer ratings were still able to distinguish the paper quality to some extent because most of rejected papers were low-quality. When excluding the majority of low-quality papers by citation counts, we had quite different results. For rejected papers with more than 10 citations, their citations were not associated with the ratings they received but tended to increase if they received more conflicting reviews. The results can be reproduced for rejected papers with more than 20 citations.

Table 7 Predictive effects of rating and confidence on citations for rejected papers

Variable	*All rejected*		*Citation > 10*		*Citation > 20*	
	Estimate	*p*	*Estimate*	*p*	*Estimate*	*p*
AVG_RATING	**0.791**	**0.000**	0.171	0.279	0.173	0.330
DISAGREEMENT	0.297	0.239	**0.730**	**0.007**	**0.596**	**0.049**
AVG_CONF	0.100	0.670	-0.195	0.406	0.007	0.983
SD_CONF	-0.085	0.775	0.226	0.535	0.262	0.567
N	289		69		37	
Dispersion parameter	0.259		1.151		1.431	
2*log-likehood	-1,768		-1,569		-412	
AIC	1,780		1,581		424	
Vuong test (NB > ZINB)	*z*=5.271, ***p*=0.000**		-		-	

Conclusions

Our research suggests that conflicting reviews can serve as a sign to be impactful ideas. However, the predictive effects of disagreement can only be observed among rejected papers with more than ten citations in our case. How to detect the disagreement that potentially indicates future impact should be valuable to improve the effectiveness of peer review systems. We also found that reviewer expertise may add value to a reviewed paper if it has been accepted. However, the value may not be taken by authors whose papers have been rejected.

Our study was limited by several issues. The open peer-review data of one conference in a single year were analyzed in this study. The relatively small size of the data may weaken the generalizability and reliability of this study. Citation is a variable affected by many other variables, such as author impact and paper novelty. In the future, we will examine peer-review data from different disciplines and with different levels of openness. We will also include more control variables to reach more reliable results concerning the effects of disagreement and confidence in peer review process.

Acknowledgments

The work is in part supported by the National Science Foundation (Award Number: 1633286). Jiangen He is winner of the 2019 Eugene Garfield Doctoral Dissertation Scholarship Award.

References

Armstrong, J. S. (1991). Does the Need for Agreement Among Reviewers Inhibit the Publication of Controversial Findings ? *Behavioral and Brain Sciences*, *14*(November 1990), 136–137.

Benda, W. G. G., & Engels, T. C. E. (2011). The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. *International Journal of Forecasting*, *27*(1), 166–182. https://doi.org/10.1016/j.ijforecast.2010.03.003

Bornmann, L., & Daniel, H.-D. (2008a). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology*, *59*(11), 1841–1852. https://doi.org/10.1002/asi.20901

Bornmann, L., & Daniel, H. D. (2008b). The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at Angewandte chemie. *Angewandte Chemie - International Edition*, *47*(38), 7173–7178. https://doi.org/10.1002/anie.200800513

Bornmann, L., & Daniel, H. D. (2008c). *What do citation counts measure? A review of studies on citing behavior*. *Journal of Documentation* (Vol. 64). https://doi.org/10.1108/00220410810844150

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. *Management Science*, *62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Campanario, J. M., & Acedo, E. (2007). Rejecting highly cited papers: The views of scientists who encounter resistance to their discoveries from other scientists. *Journal of the American Society for Information Science and Technology*, *58*(5), 734–743. https://doi.org/10.1002/asi

Fiske, D. W., & Fogg, L. F. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. *American Psychologist*, *45*(5), 591.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *ArXiv*, 1–13. https://doi.org/10.1007/978-3-319-24553-9

Jirschitzka, J., Oeberst, A., Göllner, R., & Cress, U. (2017). Inter-rater reliability and validity of peer reviews in an interdisciplinary field. *Scientometrics*, *113*(2), 1059–1092. https://doi.org/10.1007/s11192-017-2516-6

Lee, C. J., Sugimoto, C. R., Zhang, G., Cronin, B., Mihovsky, T., & Naydenova, G. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., … Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, *115*(12), 201714379. https://doi.org/10.1073/pnas.1714379115

Rinia, E. J., van Leeuwen, T. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. *Research Policy*, *27*(1), 95–107. https://doi.org/10.1016/s0048-7333(98)00026-2

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, *112*(2), 360–365. https://doi.org/10.1073/pnas.1418218112

van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3), 397–420. https://doi.org/10.1007/BF02129602

Method for comparison of the number of citations from papers in different databases

Gerson Pech and Catarina Delgado

[1]*pech@uerj.br*
Department of Nuclear Physics and High Energies, Rio de Janeiro State University, Rua São Francisco Xavier, 524, Rio de Janeiro (Brasil)

[2]*cdelgado@fep.up.pt*
LIAAD INESCTEC and Faculty of Economics, University of Porto, Dr. Roberto Frias, 4200-464 Porto (Portugal)

Abstract

Citation analysis has been used to compare researchers, fields, institutions and countries. However, not much has been done to compare citations of papers belonging to different databases and published in different years. This comparison could play a relevant role in many systematic literature reviews concerned with the growth, development, and changes of a particular scientific subject. This study aims to examine whether we can use the percentile approach to compare the number of citations from papers in different databases. We argue that this method can convert citations from different databases when there are same articles belonging to more than one database. We apply the method on Thomson Reuters' Web of Science and Elsevier's Scopus databases because they are the leading databases of scholarly impact. In this study we use two different Scopus subject area: Engineering – Industrial and Manufacturing Engineering; and Arts and Humanities –Archaeology. The analysis comprises articles published for the time period 1987–2017, of journals in the Scopus top 10%, corresponding to approximately 152,000 papers.

Introduction

Citation analysis plays a key role in Scientometry. Many researchers had covered a long journey since the arguments stated by Garfield (1972) highlighted that the results of citation analysis have great potential for management of library journal collections. Garfield (1972) also pointed out that data on citation frequency could be correlated with subscription costs, providing a solid basis for cost-benefit analysis in the management of subscription budgets. Besides the fact that the number of citations is the simplest and most direct indicator of a publication impact (Milojević, Radicchi, & Bar-Ilan, 2017), this metric may provide information on the impact and performance of individual publications, research groups, institutions, countries, and journals (Sangwal, 2013; Waltman, 2016). Therefore, citations can be used for grading the importance of research results, because citation counts seem to correlate with expert assessments (Brito & Rodríguez-Navarro, 2018). Consequently, citations are used for formal and informal evaluations of academics (Thelwall & Wilson, 2014) and, Journal Impact Factor (Garfield, 1972) and h-index (Hirsch, 2005) are widely recognized and used citation impact indicators. However, although many authors have focused their studies on how to apply citation analysis to compare researchers, fields, institutions, and countries (e.g., Fairclough & Thelwall, 2015; Radicchi & Castellano, 2012; Rodríguez-Navarro & Brito, 2018; Waltman, 2016; Zhang, Cheng, & Liu, 2014), little work has been done to investigate how to compare citations of papers belonging different databases and published in different years. This comparison method could play a relevant role in many literature systematic analysis concerned with the growth, development, and changes of a particular scientific subject, and so, is the core of our study. Furthermore, in the last years, the papers that confront citations in different databases are mainly focused on two issues: the coverage that each database provides for the scientific disciplines studied (Li, Burnham, Lemley, & Britton, 2010; Martín-Martín et al., 2018; Winter, Zadpoor, & Dodou, 2014); and a longitudinal comparison involving a very limited period (Moed, Bar-Ilan, & Halevi, 2016; Harzing & Alakangas, 2016). Indeed, as manifested freshly by Martín-

Martín et al. (2018) about citation counts, “there is no recent or systematic evidence about the differences between different databases.”

In response, this article aims to address that issue, by using a percentile-based approach to compare (and convert) the number of citations from papers in different databases, when there is a subset of articles in both databases. However, as we have shown in this paper, the extension of the method to compare papers published in different years and, simultaneously, also in different databases must be analysed through the linear regression coefficients that correlated the percentiles from different databases.

While this article looks at the number of citations from three decades, it also provides some hints as to how this parameter has changed over the years and thus contributes to a better understanding of the complexity of the citation analysis. The intricate meaning of the citations is still an open topic in the scientometric literature, and the question of what citation counts measure must be investigated carefully (Bornmann & Daniel, 2008). Indeed, in recent years, several authors have questioned the conceptual clarity of citation analysis. Specifically, scientific citations can be copied from the lists of references used in other papers, so that the rate of citing a paper is proportional to the number of citations it has already received (Simkin & Roychowdhury, 2007; Waltman, 2016). Additionally, citing certain authors provides support for a paper and persuades the scientific community of the validity of the findings, introducing bias on the analysis (Chan, Guillot, Page, & Torgler, 2015). On the other hand, scientific evaluation based on citation impact indicators may be improved by considering how significant (according to mention frequency) each paper is cited (Pak, Yu, & Wang, 2018). Beyond these points, it is important to keep in mind that the reason why an author cites an article varies from scientist to scientist (Bornmann & Daniel, 2008).

This paper is organized as follows: in the next section, we present the common approaches in citation analysis and, in the following sections, the percentile approach and numerical examples. Finally, in the last section, the conclusions are presented.

Citation analysis

The number of citations of a scientific article is a very common measure of the acceptance of that academic publication (Lu & Liu, 2014), and ultimately of the researcher(s), the research group, the institution and the country. The comparison of these (researchers, research groups, institutions, or countries) publishing in different disciplines and periods is only possible with normalized citation scores (Haunschild & Bornmann, 2016). Some popular indicators follow the same formula: $C_{\mathrm{subset}}/C_{\mathrm{set}}$, where C_{set} is the average number of citations of all publications in a dataset (for example, a scientific area) and C_{subset} is the average number of citations of all publications of a subset. For instance, in the normalized citation impact value indicator, the subset is a country’s set of publications on a specific scientific area. If a country has a normalized citation impact value of 1 in a specific subject area, that indicates that the citation impact of papers published by researchers in this country is no more and no less than the average impact of papers in this subject area (Bornmann & Leydesdorff, 2013). In the source normalized impact per paper indicator (SNIP), however, the subset is a journal’s set of publications on a specific scientific area (Moed, 2016). A similar approach is followed by the scaled citation count indicator. This is a normalized indicator in which the number of citations of a publication is divided by the average number of citations of the papers published in the same year of the paper being analysed. A value of 1 for a specific paper indicates that the citation impact of this paper is no more and no less than the average impact of papers, in this scientific field, published in the same year. That is, the same normalization concept used to evaluate a set of papers, has been also applied to evaluate the number of citations of a specific paper. In this case, the normalization follows the relation: $C\mathrm{i}$ / $C\mathrm{set}$, where $C\mathrm{i}$ is the citations of the paper i.

Other indicators try to consider the “exposure time” of publications:

- the citation rate per year (also called citation count per year since publication, or adjusted citation index) is the total number of citations of a paper divided by 1/12 of the number of months since the initial publication up to the month of data collection, which gives the average number of citations that a paper has received each year since it was published (Wilcox et al., 2013);
- the citation density is a normalized citation-based indicator which captures the citation impact in terms of both citations per paper, and citations per citation year (Ahmed et al., 2017), by dividing the total number of citations of a set of articles published in a certain year by both the number of papers in that subset and the number of years after the publication.

A normalized variant of the average number of citations per publication is obtained by dividing the total number of citations of a given set of publications by the expected total number of citations (the average number of citations of all publications in the same field, same year and same document type). Some authors claim that this ratio provides the "desired universality of citation distributions," but others refute that claim (Waltman, 2016). Other alternatives to ratios, when it comes to the normalization include applying a logarithmic transformation to citation counts and to normalize citation counts by calculating z-scores and the transformation of citation counts by a two-parameter power–law function, which seems to be the best to create normalized citation distributions that are identical across fields (Waltman, 2016). Still, for some authors the ratio between the actual number of citations of a publication and the average number of citations of all publications that are in the same field and that have at least one citation is the best indicator (Waltman, 2016).

Although field- and time-normalization of metrics is currently a standard procedure in bibliometric studies (Leydesdorff *et al.*, 2016), most of the citation indicators are still based on simple non-normalized averages, either weighted (or factional) or not, like the average number of citations, using the sum of total citations received by the publications being analysed divided by the number of papers in the sample (Waltman, 2016).

Still, some issues are yet to be addressed properly in the literature: (i) the fact that citation counts grown up over time with the increase of journals, papers and the amount of references in the papers; (ii) the tendency of the citations to grow/decrease over time; (iii) the fact that citation frequency is highly skewed, with many infrequently cited papers and relatively few highly cited papers, so one should not see citation rates as representing the central tendency of the distribution; (iv) the fact that different databases provide different citation number for the same article, by counting only the number of citations that appear on publications already on that database (at that time). This can be a problem when conducting a longitudinal study, using data from articles of the same journal, but published in different years, therefore with some only obtained in a different database. Alternatively, when trying to capture the publications from highly respected journals or some highly cited publications that are not in the "main" database being used. While collecting a database of highly cited publications, it may be interesting to search in other databases for other important papers in the field. Some studies (e.g., Lu & Liu 2014) use a citation paralleling approach – for instance, after collecting the top 100 most cited articles from the field in the "main database", identify the number of citations of the 100th most cited article in the new database (because the same article has a different number of citations in different databases) and then search in this new database for all articles of that field with the same, or more, citations than the 100th most cited paper. However, what if we want to use the number of citations as a variable to analyse the publication's acceptance? We cannot use, in the same analysis, the number of citations from different databases? At least, we cannot use them in their "raw data" form. In the next section, we introduce a two-stage method to address these problems.

The percentile-based approach

Objectives and research questions

This exploratory study will address the following research questions:

[RQ1] Assuming the same research field and for the same year, how is it possible to compare paper citations that belong to different databases?

[RQ2] In a systematic literature review when a longitudinal study is developed for a given research area, how to know if it is possible to use the percentile approach to convert the number of citations that appear in one database to an equivalent value in another one?

Comparison method

To address RQ1 and RQ2 we developed the percentile-based comparison method with two main stages: (i) a conversion of the number of citations of articles published in different years and (ii) a conversion of the number of citations of articles belonging to different databases.

The steps are as follows:

(i) Method to compare the number of citations (received in a particular year) of papers published in different years:

Step 1 - Consider a sub-area or a set of title sources in a specific database.

Step 2 - For each year, develop a cumulative probability function (CPF) for the number of citations. We can use a characteristic probability distribution function (PDF) like Lognormal, or not.

Step 3 - For each paper, set a citation parameter to be the corresponding percentile that was calculated in Step 2.

Step 4 - Use the citation parameter defined in step 3 to rank the articles, published in any year, in terms of the citations received.

(ii) Method to compare the number of citations of papers in a different database:

Step 1 - Consider a sub-area or a set of title sources in a specific database.

Step 2 - For each year, and each database, separately, develop a cumulative probability function (CPF) for the number of citations. We can use a characteristic probability distribution function (PDF) like Lognormal, or not.

Step 3 - For each year, select the papers that belong in the two databases and obtain a linear regression model to describe the relationship between the number of citations of these papers that are in both databases.

Step 4 – use the model obtained in Step 3 to develop, limited by the uncertainties of the model, a function $F(c_i^y)$ that give the citation relationship between the two databases, and the $F(p_i^y)$ that represents the same for the percentiles.

Among the different normalization procedures that could have been used, the percentile rank approach has the advantage that, intrinsically, implies the normalization of citation counting data (Brito & Rodríguez-Navarro, 2018). In the approach applied in this research each paper is weighted based on the percentile to which it belongs in the citation distribution of its field and of its year of publication. The percentile approach has been extensively applied lately in citation analysis for bibliometric evaluations (Bornmann, 2013; Bornmann, Leydesdorff, & Wang, 2013; Waltman & Schreiber, 2013) and also to predict citation counts (Kosteas, 2018).

The key point of the model is to investigate whether the method used for stage (ii) keeps invariant the method used in stage (i), or, in other words, whether the method used to find equivalence for the conversion of citations over the years is not destroyed by the method of conversion between databases. For example, assume that for a database Ψ the conversion method between years (stage (i)) implies that a particular article published in 2005 with 150 citations is equivalent to an article published in 2015 with 30 citations, and that one article from 2005 with 130 citations is equivalent to one from 2015 with 25 citations. Now, suppose 150 citations from the base Ψ, using the method of stage (ii), are equivalent to 130 citations on the base Ω, for the year 2015. Then, also using the method of stage (ii) for conversion of the number of citations for articles published in different years, but belonging to the base Ω, we must reach the same value of 25 citations that was determined by the method of stage (i). Figure 1 illustrates this problem.

$$\begin{tikzcd}
C^{\Psi}(x) \arrow[rr, Leftrightarrow, "\text{stage (i)}"] \arrow[d, Leftrightarrow, "\text{stage (ii)}"'] & & C^{\Psi}(y) \arrow[d, Leftrightarrow, "\text{stage (ii)}"] \\
C^{\Omega}(x) \arrow[rr, Leftrightarrow, "\text{stage (i)}"'] & & C^{\Omega}(y)
\end{tikzcd}$$

Figure 1. The correspondence between the bases Ψ and Ω

Numerical examples

Data collection

In order to provide some numerical examples, we collected two databases from two very different scientific fields: Engineering and Arts and Humanities. From each field, we selected a narrower subject: Industrial and Manufacturing Engineering (127.208 papers in the Scopus database, from 1987 to 2017) and Archaeology (25.144 papers in the Scopus database). Each database was built with the top 10% articles, in terms of citations, from 1987 to 2017.

Table 1 gives a list of the journals analysed in this study together with some of their characteristics: the CiteScore measures the average citations received per document published in the serial; the CiteScore Percentile indicates the relative standing of a serial title in its subject field (a title will receive a CiteScore Percentile for each subject area in which it is indexed in Scopus); the number of papers published in the range of this study (1987-2017); the publisher and the Scopus Subject area. Scopus Subject areas are defined by the All Science Journal Classification codes in Scopus. It is important to notice that titles can be indexed in multiple subject areas. Data were obtained from the file *CiteScore_Metrics_2011-2017* downloaded on Scopus.com on May 25, 2018, using the following 2 filters: in the column Scopus Sub-Subject Area we selected Industrial and Manufacturing Engineering and Archaeology, and in the column Top 10% (CiteScore Percentile) we selected Top 10%.

Table 1 - Journals included in the analysis

Title	Cite Score	Percentile	SJR	Publisher	Area
Additive Manufacturing	7,73	99	2,611	Elsevier	IND
IEEE Industrial Electronics Magazine	7,15	99	1,978	IEEE	IND
Sustainable Materials and Technologies	7,14	99	1,548	Elsevier	IND
Chemical Eng. J.	7,01	98	1,863	Elsevier	IND
Manufacturing Letters	6,83	98	1,313	Elsevier	IND
J. of Industrial Information Integration	6,5	98	0,866	Elsevier	IND
J. of Operations Management	6,13	97	5,739	Elsevier	IND

Int. J. of Machine Tools and Manufacture	5,92	97	2,700	Elsevier	IND
J. of Cleaner Production	5,79	97	1,467	Elsevier	IND
Energy	5,6	96	1,990	Elsevier	IND
Int. J. of Production Economics	5,42	96	2,401	Elsevier	IND
Composites Part B: Eng.	5,41	96	2,039	Elsevier	IND
Virtual and Physical Prototyping	5,35	96	1,438	Taylor & Francis	IND
Critical Reviews in Food Sci and Nutrition	5,15	95	1,596	Taylor & Francis	IND
Reliability Eng. and System Safety	4,65	95	1,665	Elsevier	IND
Food Eng. Reviews	4,6	95	1,639	Springer Nature	IND
Int. J. of Greenhouse Gas Control	4,34	94	1,458	Elsevier	IND
Int. J. of Precision Eng. and Manuf. Green Tech	4,31	94	1,335	Springer Nature	IND
Int. J. of Robust and Nonlinear Control	4,26	94	2,028	Wiley-Blackwell	IND
J. of Manufacturing Systems	4,15	93	1,548	Elsevier	IND
J. of Materials Processing Tech	4,15	93	1,695	Elsevier	IND
Applied Thermal Eng.	4,14	93	1,505	Elsevier	IND
Robotics and Comp-Integrated Manufacturing	4,11	92	1,041	Elsevier	IND
CIRP Annals - Manufacturing Tech	4,09	92	2,034	Elsevier	IND
IEEE Transactions on Industry Applications	4,05	92	1,020	IEEE	IND
J. of Process Control	3,85	91	1,108	Elsevier	IND
Advanced Materials Technologies	3,85	91	1,241	Wiley-Blackwell	IND
Sustainable Production and Consumption	3,52	91	0,739	Elsevier	IND
Chemical Eng. Science	3,44	90	1,043	Elsevier	IND
Hydrometallurgy	3,43	90	1,208	Elsevier	IND
Industrial Management and Data Systems	3,43	90	0,904	Emerald	IND
Industrial & Eng. Chemistry Research	3,4	90	0,978	AMC	IND
Quaternary Science Reviews	4,51	99	2,668	Elsevier	ARC
J. of Archaeological Research	4,5	99	2,159	Springer Nature	ARC
J. of Archaeological Science	2,96	98	1,885	Elsevier	ARC
J. of World Prehistory	2,96	98	2,022	Springer Nature	ARC
Boreas	2,65	98	1,273	Wiley-Blackwell	ARC
J. of Archaeological Method and Theory	2,53	98	2,014	Springer Nature	ARC
Holocene	2,43	97	1,202	SAGE	ARC
Current Anthropology	2,16	97	1,160	Chicago Press	ARC
J. of Agrarian Change	2,15	96	1,403	Wiley-Blackwell	ARC
J. of Cultural Heritage	2,11	96	0,562	Elsevier	ARC
Vegetation History and Archaeobotany	2,05	95	1,206	Springer Nature	ARC
American Antiquity	1,95	96	1,176	Cambridge	ARC
J. of Anthropological Archaeology	1,84	95	1,240	Elsevier	ARC
J. of Social Archaeology	1,81	95	0,936	SAGE	ARC
Heritage Science	1,77	95	0,491	Springer Nature	ARC
World Archaeology	1,74	94	1,349	Taylor & Francis	ARC
Digital App in Arc and Cultural Heritage	1,72	94	0,412	Elsevier	ARC
Radiocarbon	1,7	93	0,959	Cambridge	ARC
Archaeological and Anthropological Sci	1,63	93	1,052	Springer Nature	ARC
J. of Island and Coastal Archaeology	1,54	93	0,845	Taylor & Francis	ARC
Cambridge Archaeological J.	1,47	92	1,121	Cambridge	ARC
Archaeometry	1,43	92	0,587	Wiley-Blackwell	ARC
PalArch's J. of Vertebrate Palaeontology	1,4	92	0,403	PalArchFoundation	ARC
Archaeological Prospection	1,34	92	0,635	Wiley-Blackwell	ARC
Geoarchaeology - An Int. J.	1,32	91	0,823	Wiley-Blackwell	ARC
African Archaeological Review	1,29	91	0,862	Springer Nature	ARC
Int. J. of Paleopathology	1,22	90	0,618	Elsevier	ARC
Antiquity	1,21	90	0,887	Cambridge	ARC
J. of Archaeological Science: Reports	1,21	90	0,659	Elsevier	ARC
Frontiers of Architectural Research	1,2	90	0,404	Elsevier	ARC
Int. J. of Osteoarchaeology	1,15	90	0,652	Wiley-Blackwell	ARC

In figure 2 we have the evolution of the number of papers, and the number of citations in the top 10 journals of two analysed Scopus subject areas, from 1987 to 2017.

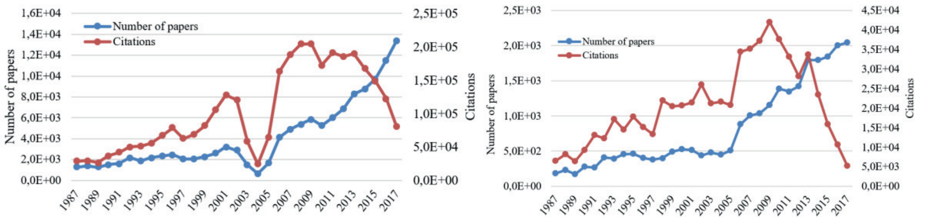


Figure 2. Number of papers and citations in the top 10 journals of the Industrial and Manufacturing Engineering (left) and Archaeology (right) fields.

The search of the articles and their corresponding citation numbers was conducted between August and September 2018, on Scopus and WoS sites. For this, we used the Print-ISSN and e-ISSN codes of each journal listed in Table 1, instead of the title name, to avoid collection errors. We collected the results for the period 1987-2017. On the Scopus website, we used the link View Citation Overview. The citation overview is available as a comma separated file (.csv) with the first 20,000 documents included, that we downloaded separately for each year of the interval. For WoS, after performing the search, we used the Create Citation Report functionality, and we downloaded it using the available export data that only allows 500 records to be downloaded at once. The Scopus database of all 31 years was used for the percentiles analysis. For the analysis of the comparison between Scopus and WoS, the two databases were used for the following years: 1987, 1997, 2005 and 2010. Table 2 shows the number of papers of each database, published in each year, and the number of papers that belong to both databases and therefore were analysed.

Table 2 - Number of papers of each database and number of papers that belong to the two and therefore participated directly in the analysis.

Subject area	Number of papers	1987	1997	2005	2010
IND	Scopus	1270	2054	1691	5255
	WoS	1000	2720	4310	5608
	Scopus & WoS	825	1893	1524	4957
ARC	Scopus	184	378	506	1387
	WoS	212	498	788	1372
	Scopus & WoS	140	329	464	1334

Results

In figure 3, the cumulative probability distribution of citations is shown, in four different periods: (a) 1987 – 1993; (b) 1994 – 2001; (c) 2002 – 2009; and (d) 2010 – 2017. For comparison purposes, the distribution for the first year of each interval – (a); (b); (c) and (d) – is represented by the same symbol and color. The same happens for the distribution of the second, third and subsequent years of each interval.

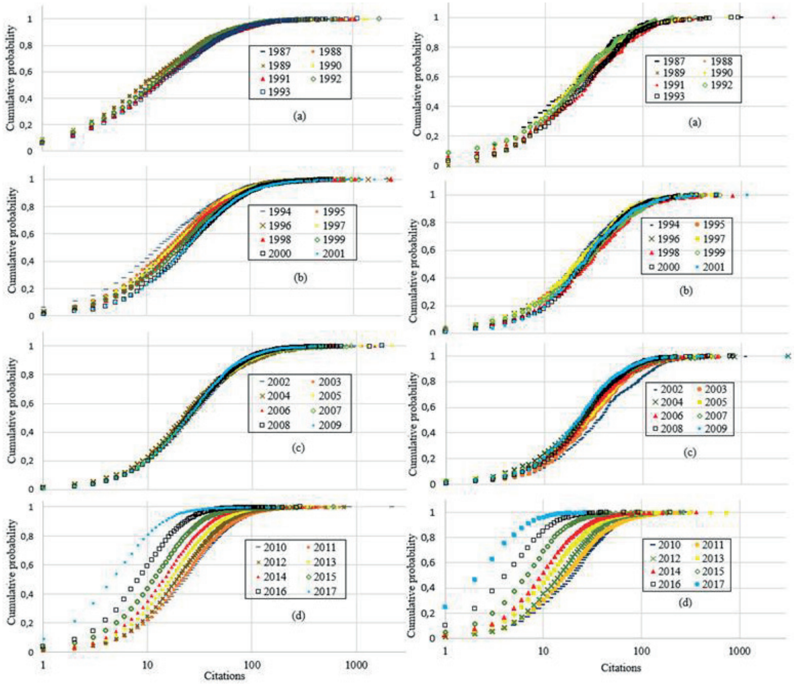


Figure 3. Cumulative probability distribution of citations for 4 intervals of years: (a) 1987 – 1993; (b) 1994 – 2001; (c) 2002 – 2009; and (d) 2010 – 2017 for Scopus subject areas Industrial and Manufacturing Engineering (left) and Archaeology (right).

In figure 4, the number of citations across 31 years is shown for the 10th, the 30th, the 50th, the 60th, the 70th, the 80th, the 90th and the 95th percentiles (respectively, P10, P30, P50, P60, P70, P80, P90, P95) in the Industrial and Manufacturing Engineering and Archaeology fields. The equivalence of the number of citations can be obtained by following each of the curves defined by the points of each percentile. For example, a paper with 15 citations that was published in 1989 is equivalent to a paper published in 2003 today with 30 citations (P60). On the other hand, a paper of 2004 with 150 citations is equivalent to a paper of 2015 today with 45 citations (P95).

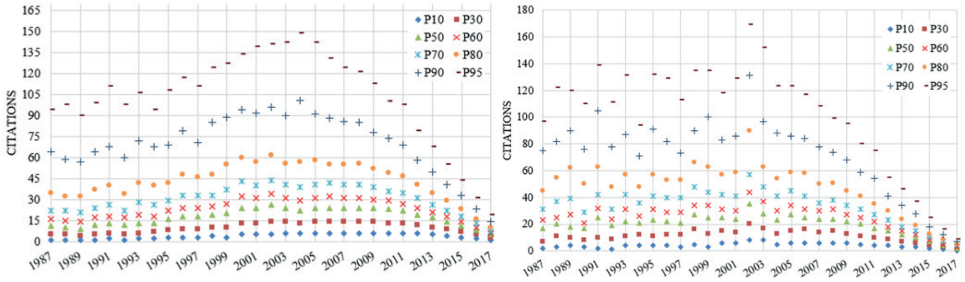


Figure 4. Number of citations across 31 years for the P10, P30, P50, P60, P70, P80, P90 and P95 percentiles in the Industrial and Manufacturing Engineering (left) and Archaeology (right).

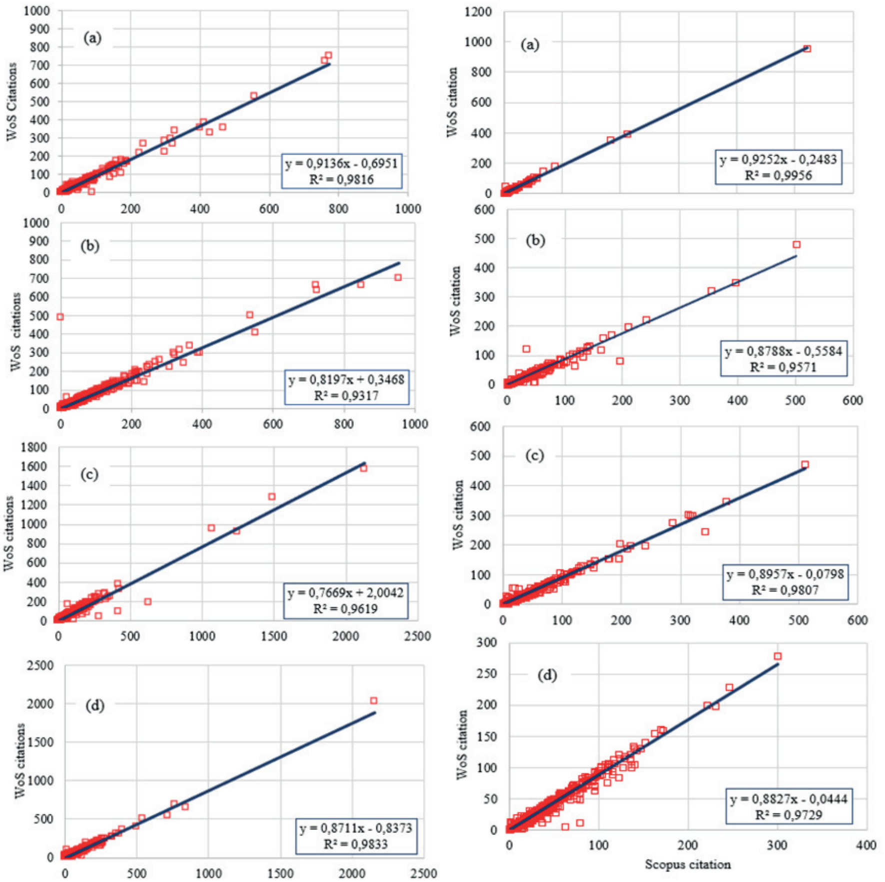


Figure 5. Linear regression for the number of citations of articles belonging to both the Scopus and the WoS databases (Industrial and Manufacturing Engineering (left) and Archaeology (right)), published in the following years: (a) 1987; (b) 1997; (c) 2005; and (d) 2010.

A linear regression model was obtained (figure 5) for the number of citations of articles belonging to both the Scopus and the WoS databases, published in the following years: (a) 1987; (b) 1997; (c) 2005; and (d) 2010. The line describes a model able for converting the number of citations from one database into another for each year separately.

A linear regression model was also obtained (figure 6) for the percentiles of citations of articles belonging to both the Scopus and the WoS, published in the following years: (a) 1987; (b) 1997; (c) 2005; and (d) 2010. The size of the points represents the number of papers that have the same values of *x* and *y* in the graph. The angular coefficient close to 1 and the linear coefficient close to zero show that, for these examples, even though the number of citations in the two databases is different, the percentiles are seemingly invariants between the Scopus and WoS databases. This invariancy is being investigated further by us and will be the subject of an upcoming paper.

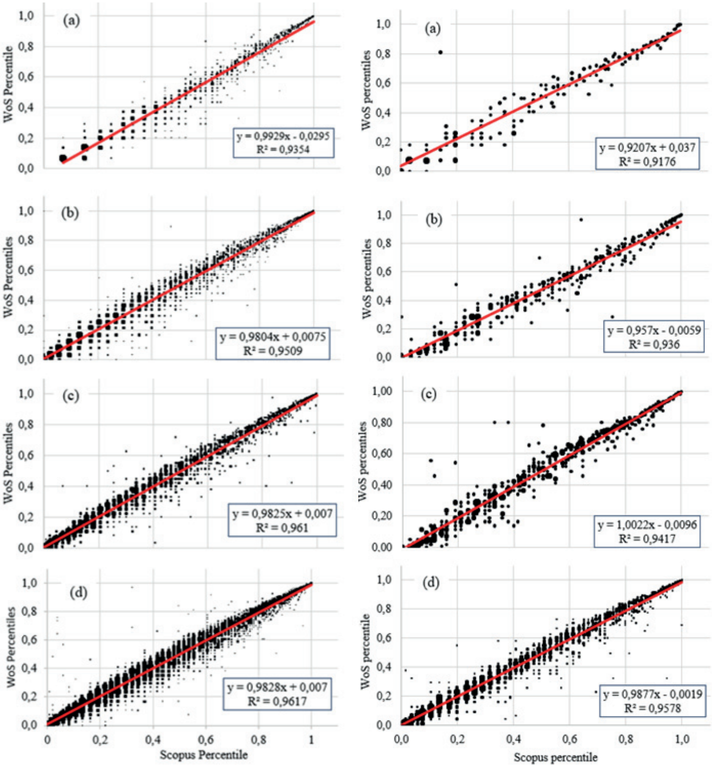


Figure 6. Linear regression for the percentiles of citations of articles belonging to both the Scopus and the WoS (Engineering (left), Archaeology (right)), in (a) 1987; (b) 1997; (c) 2005; and (d) 2010.

Conclusions

In this paper, a percentile-based technique has been introduced to address the problem of having to use, in bibliometric analysis or the data collection stages in systematic literature reviews, citation numbers for publications belonging to different databases (e.g., WoS, Scopus, Google Scholar). When a publication is in different databases, it usually presents a different citation number in each database. We propose a percentile-based method to establish a comparison between the citation numbers of those articles common to both databases, in order to obtain a model that could help us to predict the citation number of the articles that cannot be found in one of the databases. The evidence from the two fields selected (Industrial and Manufacturing Engineering and Archaeology) show that such a model can be derived. However, this is still an exploratory study, and although the results cannot be generalized, they confirm findings from some earlier studies and support the presented technique for comparing and converting citation numbers between different databases. Another contribution of this study is the comparison of models in different fields and different years, suggesting the possibility of a unified conversion model, by including field-related and year-related variables, to capture those influences.

References

Ahmed, K. K. M., Dhawan, S. M., Gupta, B. M. and Bansal, M. (2017). Highly cited publications output by India in clinical pharmacology during 2000-14: A scientometric assessment. *J. Young Pharm.*, *9*(2), 145–157. DOI: 10.5530/jyp.2017.9.30.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Bornmann, L. (2013). How to Analyze Percentile Citation Impact Data Meaningfully in Bibliometrics: The Statistical Analysis of Distributions, Percentile Rank Classes, and Top-Cited Papers. *Journal of the American Society for Information Science and Technology, 64*(3), 587–595.

Bornmann, L. and Leydesdorff, L. (2013). Macro-Indicators of Citation Impacts of Six Prolific Countries: InCites Data and the Statistical Significance of Trends. *PLoS ONE*, *8*(2), e56768. https://doi.org/10.1371/journal.pone.0056768

Bornmann, L., Leydesdorff, L., & Wang, J. (2013). Which percentile-based approach should be preferred for calculating normalized citation impact values? An empirical comparison of five approaches including a newly developed citation-rank approach (P100). *Journal of Informetrics, 7*, 933– 944.

Brito, R., & Rodríguez-Navarro, A. (2018). Research assessment by percentile-based double rank analysis. *Journal of Informetrics, 12*(1), 315–329.

Chan, H. F., Guillot, M., Page, L., & Torgler, B. (2015). The inner quality of an article: Will time tell?, *Scientometrics, 104*(1), 19–41.

Winter, J. C. F. de, Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics, 98*(2), 1547–1565.

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics, 9*(4), 895–906.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.

Harzing, A-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787–804.

Haunschild, R., & Bornmann, L. (2016). Normalization of Mendeley reader counts for impact assessment. *Journal of Informetrics*, *10*(1), 62-73.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Kosteas, V. D. (2018). Predicting long-run citation counts for articles in top economics journals. *Scientometrics,115*, 1395–1412.

Leydesdorff, L., Wouters, P. and Bornmann, L. (2016). Professional and citizen bibliometrics: complementarities and ambivalences in the development and use of indicators—a state-of-the-art report. *Scientometrics*, *109*, 2129–2150. DOI 10.1007/s11192-016-2150-8.

Li, J., Burnham, J. F., Lemley, T., & Britton, R. M. (2010). Citation Analysis: Comparison of Web of Science®, Scopus™, SciFinder®, and Google Scholar. *Journal of Electronic Resources in Medical Libraries, 7*(3), 196–217.

Lu, L. Y. and Liu, J. S. (2014), The Knowledge Diffusion Paths of Corporate Social Responsibility – From 1970 to 2011. *Corporate Social Responsibility and Environmental Management*, 21(2), 113-128.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E.D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics, 12*(4), 1160-1177.

Milojević, S., Radicchi, F., & Bar-Ilan, J. (2017). Citation success index − An intuitive pair-wise journal comparison metric. *Journal of Informetrics, 11*(1), 223–231.

Moed, H. F. (2016). Comprehensive indicator comparisons intelligible to non-experts: the case of two SNIP versions. *Scientometrics*, *106*(1), 51–65. DOI: 10.1007/s11192-015-1781-5.

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus, *Journal of Informetrics, 10*(2), 533–551.

Rodríguez-Navarro, A., & Brito, R. (2018). Double rank analysis for research assessment. *Journal of Informetrics, 12*(1), 31–41.

Pak, C. M., Yu, G., & Wang, W. (2018). A study on the citation situation within the citing paper: citation distribution of references according to mention frequency. *Scientometrics, 114*(3), 905–918.

Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE, 7*(3), e33833. DOI:10.1371/journal.pone.0033833.

Sangwal, K. (2013). Citation and impact factor distributions of scientific journals published in individual countries. *Journal of Informetrics, 7*(2), 487–504.

Simkin, M. V., & Roychowdhury, Vwani P. (2007). A mathematical theory of citing. *Journal of the American Society for Information Science and Technology, 58*(11), 1661–1673.

Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics, 8*(4), 824–839.

Waltman, L, & Schreiber, M. (2013). On the Calculation of Percentile-Based Bibliometric Indicators. *Journal of the American Society for Information Science and Technology, 64*(2), 372–379.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*, 365–391.

Wilcox, M.A., Khan, N.R., McAbee, J.H., Boop, F.A. and Klimo, P. Jr. (2013). Highly cited publications in pediatric neurosurgery. *Childs Nerv Syst*, *29*, 2201–2213.

Zhang, Z., Cheng, Y., & Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories. *Scientometrics, 101*(3), 1679–1693.

Demographic Differences in the Publication Output of U.S. Doctorate Recipients

Wan-Ying Chang, Karen E. White, and Cassidy R. Sugimoto[1]

[1]wchang@nsf.gov, kewhite@nsf.gov, and csugimoto@nsf.gov
National Science Foundation, 2415 Eisenhower Ave, Alexandria, Virginia 22314 (United States)

Wan-Ying Chang and Karen E. White are with the National Center for Science and Engineer Statistics, and Cassidy R. Sugimoto is with the Directorate for Social, Behavioral and Economic Sciences.

Disclaimer: The views expressed in this document are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Abstract

In this paper, we investigate race, ethnicity and gender differences in the publication output of U.S. doctorate recipients using the Survey of Doctorate Recipients (SDR) matched to the publication database Web of Science (WoS). Our research shows the probability of publishing is related to field of doctorate, employment sector and engagement in R&D activity. A doctorate recipient's training is also significant, as those who graduated from doctoral universities with very high research activity are more likely to publish. After controlling for these factors, differences in the probability of publishing are significantly related to demographic variables, including race/ethnicity, gender and U.S. citizenship status at the time of graduation. The ability to examine bibliometric data with a broad range of demographic variables is unique to the SDR-WoS dataset. Of the demographic variables, race/ethnicity has the strongest impact on likelihood to publish. Readers are cautioned that this summary represents a research-in-progress.

Introduction

Diversity in the scientific workforce is essential for a robust scientific system (Sugimoto et al., 2019). Despite this, men and majority populations are disproportionately represented in science and engineering (S&E). In the United States, women and underrepresented minority groups—blacks or African Americans, Hispanics or Latinos, and American Indians or Alaska Natives—are underrepresented among doctoral graduates (National Science Foundation, National Center for Science and Engineering, 2019). Strong disparities are also observed in publication output, a key indicator of involvement in scientific research. Bibliometric studies have examined disparities by gender (Larivière et.al., 2013, Science-Metrix, 2017, and West et.al., 2013), race/ethnicity (Freeman and Huang, 2015) and gender and race/ethnicity combined (Bauer et.al., 2019, Begum et.al., 2017, Marschke et.al., 2018). These studies largely rely upon gender and race disambiguation algorithms (e.g., NamSor, Ginni, Ethnicolr, OriginsInfo), which estimate the probability of a race or gender from given names (or, in the case of Face++, from images). The accuracy of these algorithms varies dramatically by country: gender disambiguation algorithms, for example, perform better for western countries, and particularly poorly for Asian countries (Karimi et al., 2016). The algorithms are also dependent upon full name information. Given that Web of Science did not record full names until 2006, large-scale bibliometric studies of gender and race have only recently become available.

The most accurate source of gender and race data would be self-reported. However, there are few datasets that combine both sociodemographic data and research activity. One creative solution was to use biosketches in grant proposals to the National Institutes of Health (NIH) in the U.S. as the main source of data. These biosketches contain selected publications as well as self-reported race/ethnicity and gender (Ginther et.al. 2018). Unfortunately, the publication list is incomplete and not all respondents choose to disclose demographic information. Tax data has also been used as a source of data for sociodemographic information, which has been matched to patenting data (Bell, et.al. 2019). However, neither of these provide a full analysis of research activity by gender and race. The present analysis overcomes the need for proxies by using self-reported gender, race/ethnicity, and U.S. citizenship status contained within the NCSES Survey of Doctorate Recipients. These data are then matched to publications within the Web of Science to provide comprehensive bibliometric data. These matched data provide a novel approach to a large-scale and cross-disciplinary analysis of the role of race and gender in scientific success.

Source Data

Survey of Doctorate Recipients

The National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation has conducted the longitudinal Survey of Doctorate Recipients (SDR) biennially since 1973, producing cross-sectional data on individuals who have earned a science, engineering or health doctorate degree from a U.S. academic institution and are less than 76 years of age (https://www.nsf.gov/statistics/srvydoctoratework/). The SDR provides data useful in assessing the supply and characteristics of U.S.-trained science, engineering and health (SEH) doctorates employed in educational institutions, private industry, and professional organizations, as well as U.S. federal, state and local government and non-U.S. government (NCSES InfoBrief, 2017, https://nsf.gov/statistics/2017/nsf17319/). The SDR collects demographic information along with educational and occupational histories; questions on scientific collaboration and research outcomes are added periodically to collect additional information on scientific productivity. Unlike most (unobtrusive) bibliometric data, these data are self-reported. Periodically, respondents are asked to provide information on their paper and patent productively for a reference period of five years prior to the survey date.

Key SDR variables include demographics (e.g., age, sex, race, ethnicity, citizenship), employment status, field of degree, principal employer, occupation and academic position, faculty rank and tenure status when applicable. The SDR sample is weighted to represent the U.S.-trained SEH doctorate population. Nonresponse weighting adjustments are applied to reduce potential nonresponse bias by using the NCSES Survey of Earned Doctorates, an annual census of individuals receiving a research doctorate from an accredited U.S. institution. Beginning in 2001, the SDR was expanded to include those graduates from U.S. institutions who move abroad. The matching operation included the cumulative sample of 80,974 SDR respondents from the 1993–2013 surveys and covers the cohort from 1961 to 2011.

Web of Science database

Web of Science (WoS) was used as the source of scientific articles. Publications dated January 1990 to December 2012 were identified for potential matches to SDR respondents through a contract with Thomson Reuters (now Clarivate Analytics). Matching for publications was limited to the years 1990–2012, since 1990 is five years before the 1995 SDR survey wave that first included questions on scientific publications.

Data Matching Methods

The SDR respondents are matched to the authors of publications indexed by the WoS using a machine learning approach. The matching algorithm incorporates name commonality, research field, education and employment affiliations, co-authorship network and self-citations to create matches from the SDR respondents to the WoS. The overall procedure consists of five steps. In step 1, a gold standard data set was constructed for use in training of prediction models and for validation of predicted matches. In step 2, candidate publications were identified using a last name, first initial blocking rule. In step 3, the round one matching is conducted by Random Forest$^{\mathrm{TM}}$ (RF) classification models trained to identify publications which could be matched to

SDR respondents with a high degree of confidence. The high confident matches are called the 'seed publications' and used to increase the amount of data available for the subsequent matching. In step 4, data was extracted from the seed publications and combined with survey data used in round one to enrich the RF models for increased recall and make the final predictions. In Step 5, the final matched data set was refined to ensure that no respondent was matched to more than one authorship on a single publication and that those with an exact match by email were considered matches.

Out of 80,974 sampled respondents that participated in one or more of the 1993–2013 SDR survey waves, about 70% (n=56,928) were matched to publications, yielding close to 1.5 million respondent/publication pairs matched in total. Respondents had an average of 18.2 publications in the 1990–2012 period. The overall matching results were evaluated using a small gold standard set. The gold standard set is constructed by matching SDR respondents to two sources of pre-compiled publication lists: WoS ResearcherIDs and Google Scholar profiles. The Google Scholar and ResearcherID publication lists were merged and restricted to only include publications indexed in the WoS and published between 1990 and 2012. A set of 251 respondents in the gold standard set was used to evaluate the overall matching results. The precision and recall of the matches were 88% and 96.5% respectively (the level of precision indicates a need for further research on the data matching before strong conclusions can be drawn from the data).

Analytic Sample

Comparisons across SDR survey waves is complicated because of the high level of sample overlap between waves in the longitudinal design, and the survey coverage also has changed over time. The present analysis focuses on the 2013 SDR cross-sectional sample, which represents the sample with the most comprehensive coverage. Of the 35,265 individuals who responded to the 2013 SDR, 26,455 were matched to at least one WoS publication, corresponding to a total 596,811 respondent/publication pairs. SDR respondents may be associated with one or more publications and there may be more than one SDR respondent on a given paper. The 2013 SDR respondents are further subset to the cohort of doctorate recipients who graduated between 1995 and 2009. This cohort overlaps with the queried publication time window sufficiently to allow investigation of publication output that occurred from 5 years before doctorate award year to at least 3 years post-graduation. We define a dichotomous indicator to identify respondents matched to at least one WoS publication classified as either an article or a conference preceding paper during 1990–2012. The indicator is used to infer the probability of publishing. Given the very high recall and the relatively low precision rate of the overall matching results, we expect this definition to be more robust against false positive matches. For comparison purposes, the doctoral recipients are grouped into two cohorts: those who receive their doctoral degree in 1995–2000 and those in 2001–2009.

Among the broad SEH doctoral fields, there were higher rates of matching among the physical sciences and biological and agricultural sciences and lower rates among social sciences and psychology (table 1). This is consistent with previous studies analysing publication patterns among doctoral students. Using a sample of rare names on ProQuest dissertations from five disciplines, Waaijer et al. (2016) found that very few Economics or Psychology doctoral students ever published, whereas the majority of students in Chemistry and Astrophysics had published at least once in their career. Similar results were found in a study of Quebec doctoral students: publication was more prevalent among students in the natural and biomedical sciences than in the social sciences (Larivière, 2011). This suggests that the matching is not due to error rates, but rather different publishing proclivity across disciplines.

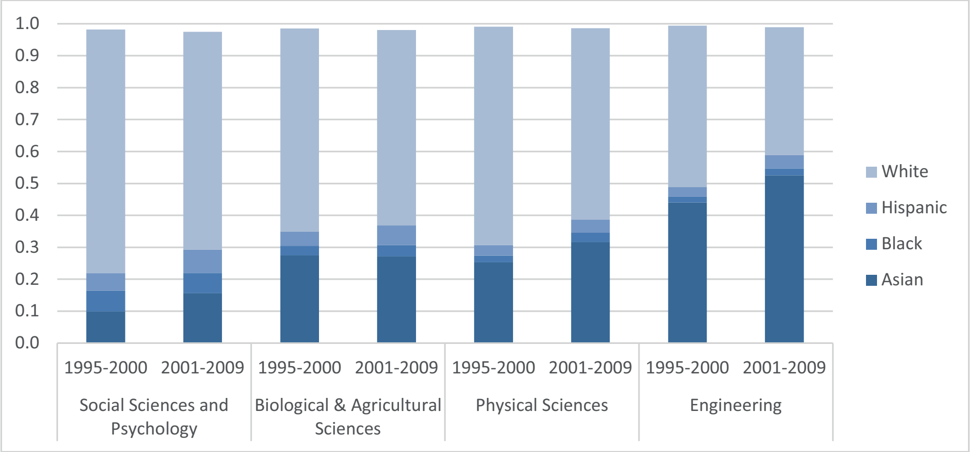
Table 1. Percent of doctorate recipients and percent of matched to WoS by S&E field of doctorate, 1995–2009 Cohort (SDR 2013, weighted estimates)

Fields of Doctorate	*Doctoral Population 1995–2009 Cohort*	*Percentage of Doctoral Recipients Matched to WoS*
Biological and Agricultural Sciences	25.0	89.2
Computer Science	3.9	83.1
Engineering	20.7	81.1
Health	5.6	83.7
Mathematics and Statistics	4.5	77.3
Physical Sciences	15.1	88.6
Psychology	12.3	64.6
Social Sciences	13.0	65.8

Sources: National Science Foundation, National Center for Science and Engineering Statistics, Survey of Doctoral Recipients, 2013 and Web of Science publication data prepared by Thomson-Reuters, 2016.

Focusing on the 4 largest groupings of degree fields—1) biological and agricultural sciences, 2) engineering, 3) physical sciences, and 4) social sciences and psychology—shows changes in the race/ethnicity among PhD recipients when comparing 1995–2000 cohorts to those who obtained their PhD in 2001–2009. Across the four largest fields of science, the proportion of white doctoral recipients has decreased over time. In engineering, a minority group (Asian) has now become the majority population (figure 1). The changes in part are due to the improved SDR coverage of U.S.-trained PhDs residing outside of the U.S. for the 2001–2009 cohort: starting from the 2003 survey wave, the SDR coverage is expanded to include non-U.S. citizens reporting the intent to leave the U.S. after receiving their doctorates. This coverage expansion yielded complete coverage of non-U.S. citizens graduated in 2001 or later.

Figure 1. Share of doctorate recipients by selected doctoral field, race/ethnicity and cohort group (SDR 2013)



Note: Totals do not sum to 1 because "other" racial/ethnic category is excluded
Sources: National Science Foundation, National Center for Science and Engineering Statistics, Survey of Doctoral Recipients, 2013.

Determinants of Publication

The matched SDR-WoS dataset permits a unique exploration into the likelihood of a U.S. doctorate recipient publishing in a peer-reviewed publication or conference proceeding based upon not only training and employment, but also demographics of gender, race/ethnicity and citizenship status at the time of degree and whether they resided in the U.S. in 2013. Training variables include field of study and whether the doctoral institution has a Carnegie classification of a "very high" research university. Variables reflecting employment characteristics in 2013 include whether the primary work activities are in R&D (as determined by the respondent) and institutional type of employment. The level of commonality of the respondents' names, a factor of the matching results, is also considered. The relative importance of these variables in predicting the likelihood of publishing is explored using logistic models. Given the changes of survey coverage and trends in scientific publications, separate models are fit for each of the three cohort groups depending on when they received their doctoral degree: 1995–2000; 2001–2005; and 2006–2009, in addition to an overall model.

Consistently across all four models, the doctorate field and employment sector are the dominating factors. The frequency of the combined last name and first initial, an indicator of the commonality of a respondent's name, also appeared to be a strong predictor, suggesting potential bias introduced by the matching algorithm may be associated with name commonality. The analysis shows, after controlling for training, employment factors and name commonality, significant differences in likelihood of publishing across cohort groupings by gender, race/ethnicity and to a lesser extent U.S. citizenship status. The choice of reference levels in calculating the estimated odds ratios are listed in table 2. Among the demographic variables, females, blacks and Hispanics have statistically significant lower odds of publishing. The publication probability of an Asian PhD recipient is not statistically significantly different from a white PhD recipient in the more recent cohorts. Among the citizenship variables, some evidence of differences exists but the pattern is unclear. It is likely due to the association between race and U.S. citizenship.

Table 2. Estimated odds ratio of doctorate recipients publishing

Categorical factor (specified level versus reference level)	*1995–2000 Cohort*	*2001–2005 Cohort*	*2006–2009 Cohort*	*All Cohorts 1995–2009*
<u>Demographic</u>				
Female	0.878	0.822*	0.84*	0.857*
Male (reference level)				
Age	0.947*	0.943*	0.943*	0.945*
Asian	0.728*	0.812	1.098	0.875
Black	0.719*	0.603*	0.698*	0.682
Hispanic	0.561*	0.516*	0.416*	0.474*
Other	0.986	0.716	0.791	0.799*
White (reference level)				
<u>Citizenship</u>				
U.S. Naturalized citizen	0.887	1.201	0.798	0.923
Permanent Resident	0.602*	0.921	0.777	0.703*

Temporary Resident	0.668*	0.957	0.992	0.939
Unknown	0.892	0.727	0.621*	0.685*
U.S. Native citizen (reference level)				
Field of Degree				
Computer Science	0.714	0.495*	0.927	0.758
Math and Statistics	0.264*	0.563*	0.332*	0.362*
Health	0.773	0.721	1.273	0.888
Physical Science	0.816	0.962	1.005	0.923
Social Science	0.275*	0.209*	0.151*	0.215*
Psychology	0.304*	0.273*	0.303*	0.305*
Engineering	0.534*	0.596*	0.633*	0.597*
Biology and Agricultural Science (reference level)				
Very High Research PhD Institution				
No	0.566*	0.722*	0.641*	0.662*
Yes (reference level)				
Employment Sector				
4-year College/University	3.301*	3.881*	2.379*	3.136*
2-year College	0.909	0.846	1.265	1.041
Self employed	0.818	1.364	0.973	1.102
Business, non-profit	1.857*	1.632*	1.605*	1.721*
Federal Government	1.892*	2.187*	1.606*	1.920*
State/Local Government	0.919	1.364	1.001	1.117
Non-U.S. Government	0.982	1.810*	1.179	1.365
Not Working	0.516*	0.977	0.894	0.851
Business, for-profit (reference level)				
Residing in the U.S.				
No	1.538*	0.885	0.988	1.057
Yes (reference level)				
Primary Work Activities				
Other than R&D	0.421*	0.557*	0.512*	0.495
R&D related (reference level)				
Cohort				
2001–2005				0.933
2006–2009				0.689*
1995–2000 (reference level)				
Name Commonality				
frequency in [0, Q1]	2.077*	1.156	0.926	1.315*
frequency in (Q1, Q2]	7.113*	3.817*	3.149*	4.615*
frequency in (Q2, Q3]	1.771*	1.410*	1.488*	1.566*
frequency > Q3 (reference level)				

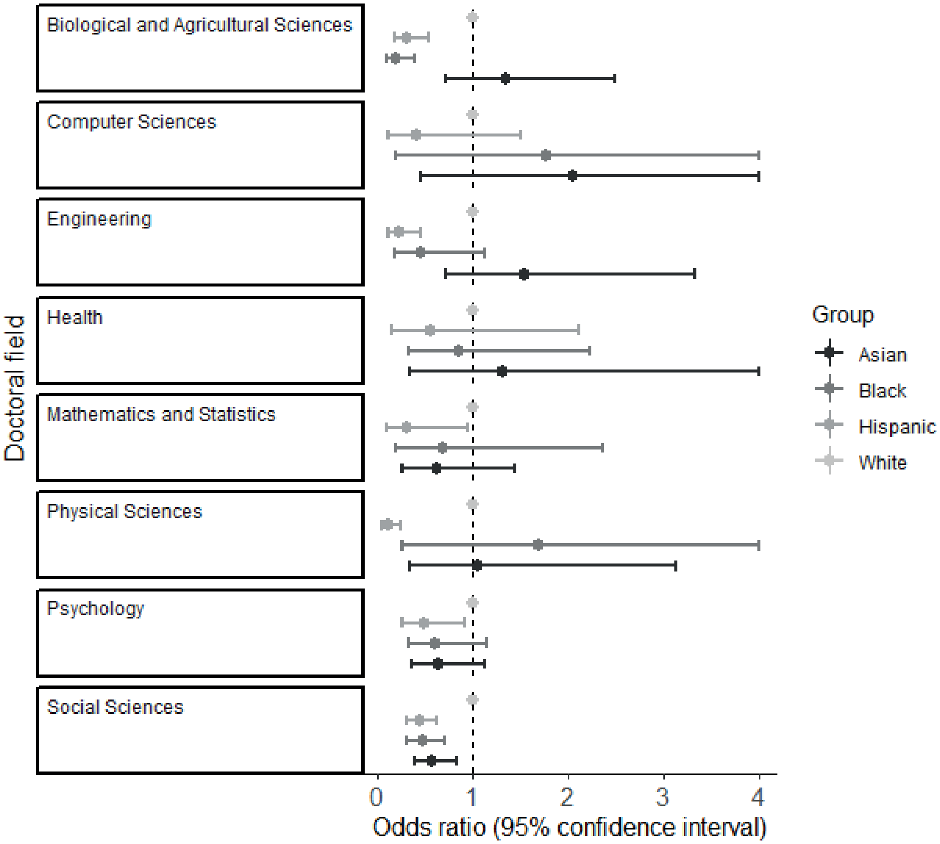
(* = $p < 0.05$)

Notes: An odds ratio greater than 1.00 indicates the SDR respondent group has higher odds of publishing compared to the reference level. An odds ratio of less than 1.00 indicates the SDR respondent group has lower odds of publishing relative to the reference level. The logistic regression models were fitted using SAS 9.4 procedure SURVEYLOGISTIC

Sources: National Science Foundation, National Center for Science and Engineering Statistics, Survey of Doctoral Recipients, 2013 and Web of Science publication data prepared by Thomson-Reuters. 2016.

The analysis is repeated for those employed in four-year college or university, medical school, or university-affiliated research institute so that additional factors such as faculty rank, tenure status, and indicator of receiving U.S. federal support can be considered for this subpopulation. Overall, doctorate field and name commonality remain strong predictors. Those on tenured track or tenured faculty and those supported by U.S. federal contracts or grants have statistically significant higher odds of publishing. After controlling for factors related to training and employment, gender differences are no longer significant; however, race/ethnicity becomes a more prominent predictor. To further examine the likelihood of publishing by race/ethnicity, we fit separate models for each broad doctorate field. The estimated odds ratios of publishing by race/ethnicity group are summarised in figure 2.

Figure 2. Estimated odds ratio of publishing of doctorate recipients employed in academia by race/ethnicity



Notes: An odds ratio greater than 1.00 indicates the SDR respondent group has higher odds of publishing compared to the reference level. An odds ratio of less than 1.00 indicates the SDR respondent group has lower odds of publishing relative to the reference level. The logistic regression models were fitted using SAS 9.4 procedure SURVEYLOGISTIC. The upper limit of estimated odds ratios is truncated at value 4.

When compared to Whites, Blacks and Hispanics tend to have lower odds of publishing, particularly for those in biological and agricultural sciences and social sciences.

Discussion and Limitations

This paper provides a proof-of-concept for linking the SDR and WoS publication databases for further research into demographic, employment and training of U.S. doctoral recipients and their research output as measured by peer-reviewed publications. This represents a novel dataset with high potential for investigating research outcomes in both bibliometrics and science policy. However, these data are not without limitations. The matching quality of this SDR-WoS match set relies heavily on the quality of input data. Among the key attributes used for matching, names from the SDR generally are of very high quality based on doctoral institutions' administrative records and updates of name changes reported in the SDR. Also, though the longitudinal feature of the SDR makes it possible to obtain a more complete profile of the respondent's affiliations, some SDR respondents have not responded to all survey waves. Those with more complete affiliation history provide better data for matching. Similarly, the quality of author information in WoS can vary by time, journal type and field.

The overall precision and recall rates reported by the Thomson Reuters project team is based on a small sample of SDR respondents who maintained ResearcherID with WoS. The training data for the matching models also come from a sample of individuals with WoS ResearcherID. Because the models are trained by such a sample, the matching algorithm is expected to perform well for those sharing similar characteristics of the authors maintaining ResearcherID. It is necessary to reassess the matching quality using a sample that is more representative of the overall SDR population. To address this, a random sample of 350 SDR respondents stratified by race/ethnicity, employment sector, doctorate year, sex, doctorate field, and reported publications in the SDR was selected and used for evaluation. The truth data of this sample were built using more source of publication information, including university website or CV, PubMed, Google Scholar, ResearchGate and LinkedIn, and the matched records were manually validated using the truth data. The evaluation results yield a recall rate of 93% and a precision rate of 78%. Further analysis of the evaluation results is planned to examine whether the matching quality varies by subgroup defined by demographics, training, or employment characteristics.

Despite these concerns, our initial analyses yield important insights into the relationship between sociodemographic characteristics and the likelihood of U.S. SEH doctorate holders to publish. We demonstrate that, even after controlling for training and sector of employment, there are significant differences in the likelihood of publishing by gender, race/ethnicity and U.S. citizenship status. Specifically, women and underrepresented minorities are significantly less likely to publish than white and male peers. Among academic researchers, we find a strong relationship between publishing and receipt of federal contracts and grants. Among the subpopulation of tenured or tenure-track faculty, we find no significant gender differences in publishing outcomes; however, lower odds of publishing for underrepresented minorities remains.

The unique combination of sociodemographic, training and employment data from SDR, matched with the high quality bibliometric data from WoS provides a unique opportunity for analyses of the scientific workforce. The accuracy of the gender and race data is particularly unique among current datasets and allows for high quality analysis on issues of diversity and productivity in the scientific workforce. These initial analyses provide early insight on the gender and race disparities in publishing and the relationship between publishing and career trajectories. There are, however, several more studies that can be done to leverage this rich matched source. The bibliometric data has presently only been explored in relation to a binary distinction on probability of publishing. This can be expanded to include several other bibliometric indicators including frequency and venue of publishing, extent and nature of collaboration, and scientific impact. Furthermore, several other variables are embedded in the SRD data that have not been explored and that may be important angles to include when discussing issues of gender and race disparities in the scientific workforce. This present manuscript provides a proof-of-concept and validity exercise in matching these data.

Acknowledgments

We thank our NSF colleague, Nimmi Kannankutty, who led the SDR-WoS matching project until January 2016 and worked with the contractors to identify attributes that would be included for matching. We are grateful to Professor Donna Ginther at the University of Kansas, for her help in developing summary research metrics to be included in the final data. We also thank the Thomson Reuter (now Clarivate Analytics) team for their continuing efforts of improving the matching methodology since the initial exploration in 2013. Specially we would like to acknowledge the key project team leads: Chris Baker for leading matching methods development, construction of gold standard set, and precision/recall analysis; Di Cross—for providing analytic support and reviewed project reports; Holly Wolcott—for leading project from July 2015 onwards, supporting methods development, troubleshooting, and supporting quality assurance and quality checks. Thanks also go to Christina Freyman at SRI International, who led the manual validation on a sample and conducted bias analysis. Finally, we thank the anonymous ISSI 2019 referees for their helpful suggestions.

References

Baker, C. & Wolcott, H. (2016). Survey of Doctoral Recipients Matching to Publications and Patents: Final Report. Available upon request to NCSES.

Bell Alex, Chetty Raj, Jaravel Xavier, Petkova Neviana, and Van Reenen John (2019). Who becomes an inventor in America? The importance of exposure to innovation, NBER Digest January 2018, https://www.nber.org/papers/w24062.

Begum Mursheda, Roe Philip, Webber Richard, and Lewison Grant (2017). UK ethnic minority cancer researchers: their origins, destinations and sex, Proceedings of the International Society for Scientometrics and Informetrics Conference, http://www.issi-society.org/publications/issi-conference-proceedings/.

Bauer Hans, Gebresenbet Fikirte, Kiki Martial, Simpson Lynne, and Sillero-Zubiri Claudio (2019). Race and gender bias in the research community on African Lions, Frontiers in Ecology and Evolution 11, https://doi.org/10.3389/fevo.2019.00024.

Freeman, Richard B., and Wei Huang. "Collaborating with People Like Me: Ethnic Coauthorship within the United States." *Journal of Labor Economics* 33, no. S1 (2015): S289-318. https://doi.org/10.1086/678973.

Ginther DK, Basner J, Jensen U, Schnell J, Kington R, Schaffer WT (2018) Publications as predictors of racial and ethnic differences in NIH research awards, PLoS ONE 13(11): e0205929. https://doi.org/10.1371/journal.pone.0205929.

Ginther Donna K, Kahn Shulamit, Schaffer Walter (2016) Gender, race/ethnicity, and National Institutes of Health R01 research awards: Is there evidence of a double bind for women of color?, Academic Medicine, Vol 91, no. 8, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965301/.

Kahn, S. & Ginther, D. (2017). Women and STEM, *NBER Working Paper 23525*. Retrieved February 7, 2019 from: https://www.nber.org/papers/w23525.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring gender names on the web: a comparative evaluation of gender detection methods. arXiv. https://arxiv.org/pdf/1603.04322.pdf.

Larivière, V. (2011). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. Scientometrics, 90(@), 463-481.

Larivière, V., Ni C., Gingras Y., Cronin B., Sugimoto C. R. (2013) Bibliometrics: Global gender disparities in science, Nature, DOI: 10.1038/504211a, https://www.nature.com/news/bibliometrics-global-gender-disparities-in-science-1.14321.

Marianne Bertrand & Sendhil Mullainathan, 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," American Economic Review, American Economic Association, vol. 94(4), pages 991-1013, September.

Marschke, Gerald, Allison Nunez, Bruce A. Weinberg, and Huifeng Yu. *2018*. Last Place? The Intersection of Ethnicity, Gender, and Race in Biomedical Authorship. AEA Papers and Proceedings*,* 108: 222-27. https://doi.org/10.1257/pandp.20181111

National Science Board. 2018. *Science and Engineering Indicators 2018*. NSB-2018-1. Alexandria, VA: National Science Foundation. Accessed February 7, 2019 from: https://www.nsf.gov/statistics/indicators/.

National Science Foundation, National Center for Science and Engineering Statistics (NSF/NCSES). 2019. *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019.* Special Report NSF 19-304. Arlington, VA. https://ncses.nsf.gov/pubs/nsf19304/.

National Science Foundation, National Center for Science and Engineering Statistics (NSF/NCSES). 2015. *Characteristics of Doctoral Recipients:2013.* Public Use Data Files. Alexandria, VA. https://ncsesdata.nsf.gov/datadownload/.

National Science Foundation, National Center for Science and Engineering Statistics (NSF/NCSES). 2015. *Survey of Doctorate Recipients, 2015 Technical Notes.* Alexandria, VA. https://ncsesdata.nsf.gov/doctoratework/2015/sdr_2015_tech_notes.pdf.

National Science Foundation, National Center for Science and Engineering Statistics (NSF/NCSES). 2017. *The 2015 Survey of Doctorate Recipients Expands Its Population Coverage and Reporting on Field of Study.* InfoBriefs NSF 17-319. Arlington, VA. https://nsf.gov/statistics/2017/nsf17319/.

Science-Metrix, 2017. *Development of bibliometric indicators to measure women's contribution to scientific publications*. Montreal, Canada: Science-Metrix. Retrieved 7 February 2019 from: http://www.science-metrix.com/en/gender-report.

Sugimoto, C.R., Ahn, Y.-Y., Smith, E., Macaluso, B., & Larivière, V. (2019). Factors affecting sex-related reporting in medical research: a cross-disciplinary bibliometric analysis. *The Lancet*, *393*(10171), p. 550-559.

Waaijer, C.J.F., Macaluso, B., Sugimoto, C.R., & Larivière, V. (2016). Stability and longevity in the publication careers of US doctorate recipients. PLoS ONE 11(4): e0154741 https://doi.org/10.1371/journal.pone.0154741.

West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT (2013) The Role of Gender in Scholarly Authorship. PLoS ONE 8(7): e66212. https://doi.org/10.1371/journal.pone.0066212.

What Does Citations Measure: Evidence from Citers

Misha Teplitskiy[1], Eamon Duede[2], Michael Menietti[3], Karim Lakhani[4]

[1] *mteplitskiy@fas.harvard.edu*
Laboratory for Innovation Science, Harvard University, Cambridge MA, USA

[2] *eduede@uchicago.edu*
Committee on the Conceptual and Historical Studies of Science, Department of Philosophy,
University of Chicago, Chicago IL., USA

[3] *mmenietti@fas.harvard.edu*
Laboratory for Innovation Science, Harvard University, Cambridge MA., USA

[4] klakhani@hbs.edu
Harvard Business School, Harvard University, Cambridge MA., USA

Abstract

Although citations and related metrics like the H-index are widely used in academia to evaluate research and allocate resources, the referencing decisions on which they are based are poorly understood. In particular, it is unclear whether authors reference works that influenced them most -- the "normative" view -- or those they believe the readers will value most -- the "social constructivist" view. We present preliminary results from a pilot survey of authors of scientific articles in which we asked them about specific references they have made. We find that authors (1) know the content of the papers they cite less well when the references are to famous (highly cited) papers and (2) are influenced (per-capita) equally by highly and sparsely cited works. An experiment in which authors were asked about references with and without signals of the references' `status' (e.g., how highly cited the reference is), we find that positive correlations between citations and perceptions of the quality of a paper, like its validity or significance, are explained by status signals. These findings are inconsistent with the normative view and support the social constructivist view, requiring a radical reassessment of the role of citation in scientific practice.

Introduction

The tremendous pace of scientific publishing outpaces individuals' abilities to thoroughly digest and evaluate each published work. Consequently, scientists, administrators, and policy makers often lean on quantitative metrics like citations to *value* scientific works. The more citations, the more quality, the more influence. Citations and metrics derived from them, like the *h*-index, are ubiquitous and routinely used to search the literature, validate claims, promote or hire individuals, allocate grant funding, and so on. Despite the pervasiveness of metrics, why scientists cite particular works and what can be inferred from these decisions remains poorly understood.

In this preliminary work, we distinguish two perspectives of citing decisions, the *normative* and the *social constructivist*. The normative view holds that scientists and scholars cite works that influenced their research choices, and that they consider to be of high quality. In contrast, the social constructivist view holds that individuals cite papers for rhetorical and strategic reasons that are independent of the individuals' *personal* perceptions of the works' quality. For example, under the social constructivist view, scientists and scholars will cite works that they do not know well and that did not influence their research choices, but that support claims they want to make and are familiar to the intended audience. Consequently, whatever the citation counts signal, they do not signal authors' judgments of the quality or the influence of the work.

Data and Methods.

To assess evidence for each of these views and rigorously determine precisely what can be inferred from citation counts, we fielded a web-based, intelligent, pilot survey of scientists across 6 fields of science and humanities, in which we asked about specific references they made in their papers. While others have attempted to survey researchers about citation practices, none have attempted to survey broadly across disciplines and with systematic sampling of *cited* papers from the entire published literature. We rely on the unique blend of computational techniques with rich data from the complete Clarivate Analytics *Web of Science*, which enables our survey instrument to scale arbitrarily.

We sampled researchers using the following sampling frame. First, we selected one field from each of Web of Science's 6 major categories – the fields were Endocrinology, Ecology, Management, Analytical Chemistry, Religion, and Computer Science - Information Systems. Second, for each field we identified all publications published in 2010 and ranked them according how many citations they accrued by 2015. Third, for each field, we randomly selected a paper from each percentile of the field's citation distribution and asked up to ten individuals who cited the paper in 2015 to evaluate its `quality', `validity', `novelty' and other attributes, along with how much the paper influenced their research choices and how well they know their contents. Additionally, we experimentally manipulated the information respondents observed when evaluating papers: the treatment group was shown how much the paper had been cited ("status signal") while the control group was shown no information regarding the paper's citations ("no status signal"). This pilot survey included responses from 731 respondents representing (~17% response rate).

Results.

We present two sets of findings, which combine data responses from all 6 sampled fields. First, authors (1) know the content of the papers they cite less well when the references are highly cited and (2) are influenced (*per capita*) equally by highly and lowly cited works. Over 60% of respondents indicate that the papers they cited had only "minor" or "very minor influence" on their research choices. Second, without an explicit signal of a paper's status in the citation distribution (control condition), respondents perceive the quality, influence, validity, novelty and significance of highly and lowly cited papers to be equal, on average. With an explicit status signal (treatment condition), a positive correlation appears between a paper's citation count and its citers' perceptions of `quality', `influence', `significance' and other attributes of the papers. Positive correlations between citations and perceptions of the quality of a paper, like its validity or significance, are thus explained entirely by status signals. Nevertheless, scientists do rate the works they cite as being above a certain threshold of quality.

Conclusion.

We argue that the evidence is most consistent with a "citation decision function" that combines normative and social constructivist elements. Authors do not cite works they perceive to be below a miminum threshold value of quality, supporting the normative view. However, above this threshold, frequency of use is unrelated to quality. Instead, usage is determined by social constructivist elements: scientists tend to cite works they are not influenced by and that they do not know particularly well. Although normative considerations play a role, the threshold-nature of the role makes it invalid to infer differences in perceived quality between highly and lowly cited items. In sum, our findings elucidate what drives citation decisions, severely undermine the normative view of citation practices, and require a radical reassessment of the role of citations in evaluative contexts.

The impact of air transport availability on research collaboration

Adam Ploszaj[1], Xiaoran Yan[2] and Katy Börner[3]

[1] *a.ploszja@uw.edu.pl*
Centre for European Regional and Local Studies EUROREG, University of Warsaw, Warsaw, Poland

[2] *yan30@iu.edu*
Indiana Network Science Institute, Indiana University, Bloomington, Indiana, United States of America

[3] *katy@indiana.edu*
School of Informatics, Computing, and Engineering, and Indiana Network Science Institute,
Indiana University, Bloomington, Indiana, United States of America

Abstract

This paper analyzes the impact of air transport connectivity and accessibility on scientific collaboration. Numerous studies demonstrated that the likelihood of collaboration declines with increase in distance between potential collaborators. These works commonly use simple measures of physical distance rather than actual flight capacity and frequency. Our study addresses this limitation by focusing on the relationship between flight availability and the number of scientific co-publications. We distinguish two components of flight availability: 1) direct and indirect air connections between airports; 2) distance to the nearest airport from palaces where authors of scientific articles have their professional affiliations. We provide evidence that greater flight availability is associated with more frequent scientific collaboration. More flight connections (connectivity) and proximity of airport (accessibility) increase the number of co-authored scientific papers. Moreover, direct flights and flights with one transfer are more valuable for intensifying scientific cooperation than travels involving more connecting flights.

Introduction

Numerous studies demonstrated that the likelihood of collaboration declines with growing distance between prospective collaborators (e.g., Katz, 1994). This effect is observed both at the micro level of buildings or campuses, as well as at the macro level of collaboration networks among cities, regions, and countries (for a comprehensive review see Olechnicka et al., 2019). The distance between collaborating units in spatial scientometrics studies is usually measured as geographical distance along the surface of the earth ("as the crow flies"), between points which are defined by geographical coordinates: latitude and longitude (Frenken et al., 2009). The actual accessibility is taken into account surprisingly rarely in empirical studies of scientific collaboration. To our best knowledge, only following empirical works considered actual transport accessibility as a covariate of scientific collaboration. Andersson and Ejermo (2005) included road travel time in their case study of Swedish patent co-authorship network. Ejermo and Karlsson (2006) studied road and air travel time impact on co-patenting in Sweden. Ma et al. (2014) hypothesized that high-speed railway accessibility can be one of the factors explaining the intensity of scientific cooperation between Chinese cities. The hypothesis was supported with evidence from instrumental variable regression study designed by Dong et al. (2018). Hoekman et al. (2010) argued that European regions with a major international airport are more likely to develop intensive international scientific collaboration. Against this background, the study of Catalini at al. (2016) stands out as the authors used a quasi-experimental design to examine the impact of introducing a new, low fare, air route on the probability of scientific cooperation. Their analysis shows that the introduction of new routes significantly increases the likelihood of collaboration among US chemistry scholars.

Our study extends prior work by analyzing the relationship between scientific collaboration and worldwide air transport availability. We distinguish two components of flight availability: (1) direct and indirect air connections between airports (connectivity), and (2) distance to the nearest airport (accessibility) from cities and towns where scientific articles are affiliated. We test the hypothesis that better air transport connectivity and accessibility—ceteris paribus—is positively associated with scientific collaboration.

Empirical strategy and descriptive statistics

The analysis is based on a sample of combined ego-networks of four campuses of US public research-intensive universities: Arizona State University at Tempe (ASU), Indiana University Bloomington (IUB), Indiana University-Purdue University Indianapolis (IUPUI) and University of Michigan at Ann Arbor (UMICH). Only the main campuses of the universities are included in the study. Selection of the research sample satisfies following criteria: comparable size and research intensity of universities, various levels of passenger traffic, and the possibility of an unambiguous assignment of a university to a single airport. ASU is served by Phoenix Sky Harbor International Airport (PHX) and UMICH by Detroit Metropolitan Airport (DTW). Both airports are important hubs. According to Federal Aviation Administration data, PHX was the 11th US airport in terms of number of passengers in 2016, while DTW took 18th position. IUB and IUPUI constitute a specific case. The two campuses are served by the same airport, Indianapolis International Airport (IND). IND is an airport with considerably less passenger traffic than PHX and DTW. In 2016, IND was 46th US airport regarding the number of passengers.

The number of co-authored papers is the dependent variable in this study. Co-authorship were identified on the basis of the co-occurrence of author affiliations in articles published in years 2008-2013 and indexed in the Web of Science database. We employed the full counting method, i.e. each co-authored paper is counted as one for a given ego-alter relation, regardless of the number of authors, organizations, or countries involved. The advantage of this approach—as compared to fractional counting—is the intuitive interpretation of results, as well as the possibility of using well-established statistical models for event counts data (Long, 1997).

The dependent variable is measured for each of four institutions—ASU, IUB, IUPUI, and UMICH—as the number of co-authored papers between the given campus and various geographical units across the globe (henceforth called as ‘destinations’). To ensure coherence and international comparability geo-locations are merged into 2,245 town/city/metropolitan/regional entities, such as European NUTS2 regions and US Metropolitan Statistical Areas. For each of four selected universities a separate egocentric co-authorship network was constructed. In consequence, we obtained four ego-networks, in which an ego was ASU, IUB, IUPUI or UMICH, and alters (destinations) were spatial units from around the world (for the details on data sources and data processing, please refer to the supplementing information available on GitHub: https://github.com/everyxs/FlightCoauthor).

To measure air transport availability we employed a number of variables grouped into two categories: commercial air transport connectivity and transport accessibility to the nearest airport. The accessibility variable is measured as the geographical distance from the center (centroid) of a destination to its nearest airport with commercial flights. To account for connectivity, we tested three approaches. The simplest variable is a ‘Minimum number of stops to reach destination’. This factor variable is based on a minimum number of connecting flights needed to travel from ego’s nearest airport to the airport nearest to the centroid of destination geographical unit. It is measured up to 4 connecting flights (or 3 stops) and takes values: 0 (for direct flights), 1, 2, or 3. Second measure ‘LinesXstop’ takes into account number of flights between ego and destination airports. ‘Lines0stop’ accounts for direct flights only. ‘Lines1stop’ measures direct and indirect flights up to one stop (i.e., up to two connecting flights). ‘Lines2stop’ considers direct and indirect flights up to two stops, while ‘Lines3stop’ adds connections requiring 3 stops. To take into account the preference for flights with fewer transfers, weights are applied: 1 for direct flights, 0.5 for one stop connections, 0.33 for two stop, and 0.25 for three stops. The use of concurrent connectivity variables aims to better understand the relationship between air transport and scientific collaboration. Two questions are particularly interesting in this case. First, are direct connections more important than connecting flights? Second, are indirect flights with fewer stops more important than those with more stops?

Two control variables are used in this study. 'Geographical distance' between an ego-institution and a destination is measured along the surface of the earth. We assume that geographical distance alone should explain a lot of scientific collaborations. However, we hypothesize that models accounting simultaneously for geographical distance and flights availability variables will fit the data better. The second control variable is the 'Number of papers at destination'. This variable can be seen as the equivalent of a mass term in the gravity model approach. We assume that probability and intensity of collaboration between ego and destination depend primarily on the scientific capacity of a destination. Collaboration with city, region, or country that have virtually no research activities is improbable. While collaboration with global knowledge hubs can be intensive, despite the geographical distance.

Our full dataset of 8,980 observations consists of four institutional sub-datasets, each comprising 2,245 observations (see Table 1). An observation is defined as a multidimensional link (co-authorships, geographical distance, air links, etc.) between university campus in question—one of the four ego-institutions—and one of 2,245 geographical entities around the world that have at least one paper affiliated as identified by Mazloumian et al. (2013). The number of co-authored papers between ego-institution and defined geographical entities—the dependent variable in this study—ranges from 0 to 3433, with the mean value of the variable equal to 15.4. It means that the four analyzed institutions co-authored on average 15.4 papers per possible relationship between the institution and one of the defined geographical units.

Table 1. Descriptive statistics – full dataset

Variable	Observations	Mean	Std. Dev.	Min	Max
Number of co-authored papers	8980	15.4	89.5	0	3433
Geographical distance (mi)	8980	4232.3	2669.4	20.4	11171
Number of papers at destination	8980	5373.3	13866	1	201693
Distance to airport at destination (mi)	8980	24.8	25.4	0.4	327
lines0stop	8980	0.1	0.7	0	15
lines1stop	8980	3.8	6	0	55
lines2stop	8980	18	16.8	0	127
lines3stop	8980	114.6	91.6	0	822
Min. number of stops to destination	8980	1.5	0.7	0	4

Modelling approach

To model the impact of air transport availability on scientific collaboration we employed zero-inflated model (Zero-inflated Negative Binomial Regression model implemented in STATA). This class of models is designed for event count data where the sample is drawn from a zero-inflated probability distribution—i.e., one that allows for frequent zero-valued observations. Our research dataset fits the requirements for using these models perfectly—about 45% of the outcome variable equals zero. That is, during the observed period, the four ego-institutions had no co-authorships with 45% localizations that are identified as having published at least one scientific paper (according to data from Mazloumian et al., 2013). The zero-inflated model assumes that zero outcome can result from two different processes. First, the absence of collaboration can be due to the lack of research capacities at the destination. In this case, the expected outcome is zero. Second, if the destination has some research capacities, it is then a count process. Zero outcome is still possible (e.g. due to different research profiles), but numerous co-authorships are very likely.

Consequently, the zero-inflated model has two components: “inflate” part that accounts for excess zeros (the equivalent of logit model) and a proper “count” part. To construct inflate part we used a single predictor: ‘Number of papers at destination’. This decision is based on the assumption that the adequate critical mass of scientific capacity determines the emergence of scientific collaboration, regardless of geographical distance and transport accessibility. In the count part, we used both control variables—i.e. ‘Geographical distance’ and ‘Number of papers at destination’—and independent variables for air transport connectivity and accessibility.

To account for expected curvilinearity, additional quadratic terms have been used in the case of four variables: ‘Geographical distance’, ‘Number of papers at destination’, ‘LinesXstop’, and ‘SeatsXstop’. We assume that the impact of enumerated variables on scientific collaboration is not uniformed across their possible values. In particular, the impact can be more pronounced at low values and gradually less distinct at high values (diminishing returns).

Because air transport makes little sense for short distances, observations in which geodistance variable was less than 100 miles were excluded from the further empirical analysis. In total, 55 observations were omitted. As a result, a restricted dataset used as a basis for estimations consisted of 8,925 observations, multidimensional links (co-authorships, geographical distance, air links, etc.) links between four universities and theirs possible research collaborators.

Results

Table 2 presents estimation results of models with air transport connectivity and accessibility (model 2-5) compared to the reference model that does not include any transport variables (1).

Table 2. Research collaboration and air transport connectivity and accessibility

Dep. var.: Number of co-authored papers	(1)	(6)	(7)	(8)	(9)
Count part					
Geographical distance (thous mi)	-0.342***	-0.271***	-0.196***	-0.225***	-0.248***
Geographical distance squared (thous mi)	0.016***	0.010***	0.006*	0.008**	0.010***
Number of papers at destination	0.129***	0.117***	0.110***	0.109***	0.108***
Number of papers at destination squared	-0.001***	-0.001***	-0.000***	-0.000***	-0.000***
lines0stop		0.342***			
lines0stop squared		-0.026***			
lines1stop			0.079***		
lines1stop squared			-0.001***		
lines2stop				0.030***	
lines2stop squared				-0.000***	
lines3stop					0.005***
lines3stop squared					-0.000***
Distance to airport at destination (mi)		-0.012***	-0.013***	-0.013***	-0.013***
Constant	2.052***	2.154***	1.756***	1.606***	1.568***
Inflate part					
Number of papers at destination	-3.787***	-3.487***	-3.438***	-3.436***	-3.438***
Constant	-0.104	-0.193**	-0.224**	-0.242**	-0.249***
Constant lnalpha	0.827***	0.796***	0.773***	0.773***	0.771***
Statistics					
Observations	8925	8925	8925	8925	8925
AIC	40998.1	40785.7	40626.8	40608.1	40590.9
BIC	41054.9	40863.8	40704.9	40686.2	40668.9

Significance levels: * p<0.05; ** p<0.01; *** p<0.001.

As expected, the basic model (1) with no air transport availability variables does significantly worse than all other models with transport variables included. This is evidenced by the fact that model (1) has the highest values of Akaike Information Criterion (AIC) and Bayesian information criterion (BIC). The difference in AIC and BIC between the model (1) and the second worst specification, model (2), highly exceeds 10 and can, therefore, be considered significant (Raftery, 1995). The addition of air connectivity and availability variables (models 2-5) noticeably improves the fit of the model (significant decrease in both AIC and BIC). These results plainly indicate that not only the physical distance influences the intensity of scientific collaboration, but also, the actual transport accessibility plays a significant role.

The relationship between air connectivity and the number of co-authored papers is not linear. All the squared air connectivity variables are significant in specifications (2)-(5). Negative coefficients of the quadratic terms suggest that at some point, the connectivity is so high that its further increase (e.g. adding one more flight between given airports) has far less impact on collaboration than the similar increase at low levels of the overall connectivity.

Further analysis of the compared models reveals, firstly, that direct connections have a stronger impact on the probability of scientific cooperation than flights requiring transfers—see specification (6) with dummy variables for direct and connecting flights presented in Table 3. In the case of destinations that have no direct flight connection and requires minimum one stop, the number of expected co-publication decreases by a factor of 0.49 as compared to destinations that can be reached with a single flight. Secondly, the greater the number of transfers required, the weaker the effect on the dependent variable. This is evidenced by the fact that the model with only direct flights—specification (2)—have the highest coefficient of air transport variable (Lines0stop). In turn, models with up to one, two or three stops show decreasing values of air transport coefficient (Lines1stop, Lines2stop, and Lines3stop, respectively). This result is in line with expectations. Direct flights and those requiring fewer transfers are more convenient for passengers than connections requiring many stops. At the same time, not only air transport connectivity matters but also the distance between the location of the co-authors and their nearest airport. The results of the estimation confirm the common sense of expectations that the proximity of the airport is advantageous, at least in the case of long-distance cooperation.

Table 3. Research collaboration and air transport—direct and connecting flights

Dependent variable: Number of co-authored papers	(6)
Count part	
Geographical distance (thous mi)	-0.122***
Geographical distance squared (thous mi)	-0.000
Number of papers at destination	0.113***
Number of papers at destination squared	-0.000***
Minimum number of stops to reach destination (compared to direct flight):	
1 stop	-0.705***
2 stops	-1.274***
3 stops	-1.617***
Distance to airport at destination (mi)	-0.012***
Constant	2.743***
Inflate part	
Number of papers at destination	-3.528***
Constant	-0.225**
Constant lnalpha	0.766***
Statistics	
Observations	8907
AIC	40522.8
BIC	40607.9

Significance levels: * p<0.05; ** p<0.01; *** p<0.001.

Conclusions

The paper makes two contributions. First, we show that air transport availability is an important factor for scientific collaboration, even when controlling for geographical distance and research capacities of collaborators. Second, both air transport connectivity (direct and indirect air connections between airports) and accessibility (distance to the nearest airport) are important correlates of scientific collaboration. Presented estimation results provide evidence that more flight connections increase the number of co-publications. Also, proximity of airport at collaborating destination is positively related to the expected number of co-authored papers. Moreover, direct flights and flights with one transfer are more valuable for intensifying scientific collaboration than travels involving more connecting flights. One additional direct flight rise the expected number of co-publications by a factor of 1.41, while additional connection requiring up to two stops rises the number by a factor of 1.03. The results of our study are in line with conclusions from broader research corpus highlighting the importance of air transport for the economic development of cities and regions (Conventz & Thierstein, 2015). In particular, the availability of direct flights is seen as a significant predictor of a city's fortunes (Campante & Yanagizawa-Drott, 2017).

Acknowledgments

This work uses Web of Science data by Clarivate Analytics provided by the Network Science Institute and the Cyberinfrastructure for Network Science Center at Indiana University. Katy Börner is partially supported by the National Institutes of Health under awards P01AG039347 and U01CA198934 and National Science Foundation awards AISL1713567 and OAC1445604. Adam Ploszaj is partially supported by grant 2011/03/B/HS4/05737 from the Polish National Science Center. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Polish National Science Center.

References

Andersson, M., & Ejermo, O. (2005). How does accessibility to knowledge sources affect the innovativeness of corporations?—evidence from Sweden. *The annals of regional science*, *39*(4), 741-765.

Campante, F., & Yanagizawa-Drott, D. (2016). Long-range growth: economic development in the global network of air links. *The Quarterly Journal of Economics*.

Catalini, C., Fons-Rosen, C., & Gaulé, P. (2016). Did Cheaper Flights Change the Geography of Scientific Collaboration? *MIT Sloan Research Paper*. No. 5172-16.

Conventz, S., & Thierstein, A. (2015). *Airports, cities and regions* (Routledge advances in regional economics science and policy). London, New York, NY: Routledge.

Dong, X., Zheng, S., & Kahn, M. E. (2018). *The Role of Transportation Speed in Facilitating High Skilled Teamwork* (No. w24539). National Bureau of Economic Research.

Ejermo, O., & Karlsson, C. (2006). Interregional inventor networks as studied by patent coinventorships. *Research Policy*, *35*(3), 412-430.

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, *3*(3), 222-232.

Hoekman, J., Frenken, K., & Tijssen, R. J. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*(5), 662-673.

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, 31(1), 31-43.

Long, S. J. (1997). *Regression models for categorical and limited dependent variables*. Advanced quantitative techniques in the social sciences, vol. 7. Beverly Hills, CA: Sage.

Ma, H., Fang, C., Pang, B., & Li, G. (2014). The effect of geographical proximity on scientific cooperation among Chinese cities from 1990 to 2010. *PloS one*, *9*(11), e111705.

Mazloumian, A., Helbing, D., Lozano, S., Light, R. P., & Börner, K. (2013). Global multi-level analysis of the 'Scientific Food Web'. *Scientific reports*, *3*, 1167.

Olechnicka, A., Ploszaj, A., & Celińska-Janowicz, D. (2019). *The Geography of Scientific Collaboration*. Routledge.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111-163.

International Postdoctoral Mobility and Career Effect in Italian Academia – 1986-2015

Massimiliano Coda Zabetta[1] and Aldo Geuna[2]

[1] *mcodazabetta@u-bordeaux.fr*
University of Bordeaux, GREThA – UMR CNRS 5113, Avenue Léon Duguit, 33608, Pessac (France)

[2] *aldo.geuna@unito.it*
University of Turin, Department of Economics and Statistics "Cognetti de Martiis", Lungo Dora Siena 100/A, 10122, Turin (Italy) & Collegio Carlo Alberto, Piazza Albarello 8, 10122, Turin (Italy)

Abstract

This paper investigates the effect of international postdoctoral mobility on academic career. International postdoctoral appointments may help to expand researchers' scientific and technical human capital while at the same time ensuring career stability. We use duration models on individual data to predict promotion from assistant professor to associate or full professor positions. Using a panel dataset of 18 thousands Italian academics in all disciplines over 30 years, we find that international postdoctoral appointments have a positive effect on career outcomes and reduce the waiting time for promotion. This provides evidence that early stage international mobility is beneficial for academics' career in the long-term. We use bibliometric indicators to measure different dimensions of social capital which affect researchers' career, namely: localism, home country linkages and expanding the scientific network. The article contributes to a better understanding of the role of meritocratic and non-meritocratic factors in achieving scarce and highly competitive job positions in academia.

Introduction

International mobility of highly skilled workers is a growing phenomenon, with important implications for human resource management, innovation and policy. International mobility is increasingly part of a broader phenomenon of globalization of the careers of the highly skilled, involving also the expansion of mass higher education, growth in the number of international students and increasing international collaborations (Freeman 2010). A large majority of movements are not permanent and involve more than one destination (Newland 2009). If migrants do not remain in the host country, in some cases, they return to their country of origin, and in others, they move to a third country (Van Bouwel 2010).

In this paper we focus on the effect of international research appointments and social capital on career outcomes. International mobility, in fact, may help or harm in speeding career progression and ensuring career stability. We focus on international mobility due to its growing importance and pervasiveness in structuring public policies (Stephan 2012). In fact, the evidence regarding the ability of international mobile academics to provide benefits to their own countries in terms of spill-overs (Ackers, 2005; Saxenian, 2005) fosters policy initiatives aimed at encouraging national academics to go abroad and migrant academics to return (Hunter et al. 2009).

Given the inadequacy and lack of appropriate data to assess the phenomenon of researcher mobility (CEC 2004; Ackers 2005; Fontes 2007), the major limitation of previous studies is that they usually rely on cross-section survey data, covering a limited span of time and scientific areas. We built a database of doctorate holders in all disciplines from Italian universities who obtained their degree in the period from the first cycle (1986) until 2006. The doctorates who pursued an academic career in Italy have been identified by matching with academics in the official archives of the Italian Ministry of Education and followed in their career until 2015. From this matched databased we identified those academics that undertook a postdoctoral appointment before entering the Italian academic system (about 44%). To classify mobility in the postdoctoral period we used affiliation information reported on scientific publication data from Scopus[1]. Postdocs have been classified in either internationally mobile (about 8%) or not (about 36%). This database allows to shed light on the employment and career outcomes of researchers active in Italian universities in a time period of 30 years.

Conceptual framework

The relationship between mobility and careers is complex and worth to be studied: mobility, in fact, may have different effects on careers and on knowledge production depending on the type of mobility and the career stage at which it occurs (Fernandez-Zubieta et al., 2015).

Especially at the early stages of a career, international mobility can provide training in leading research groups which can either result in the establishment of a career in the new institution and country (Becher and Trowler 2001) or in the acquirement of specialist tacit knowledge that can then be applied at the sending institution or in the home country (Stephan 2012). Indeed, Musselin (2004) finds that academics participating in postdoctoral fellowships perceive their international mobility to be a personal strategy aimed at improving their career prospects back home.

Here we focus on the concept of scientific and technical human capital (STHC), which can help to further explain the link between mobility and promotion. Bozeman et al. (2001) describe the notion of STHC as: "the sum of scientific, technical and social knowledge, skills and the resources embodied in a particular individual". Job mobility contributes to scientist's STHC to the extent that it increases the number of collaborations and strengthens existing relationships. For this reason it is possible to expect a positive relationship between mobility and career success. However Sabatier et al. (2006), using a sample of 583 French scientists in the field of life science in one of the national research centre, and Heining et al. (2007), using a sample of 243 German professors in the field of economics, do not find any clear evidence. In particular, Heining et al. (2007) explain this result by suggesting that "moving destroys (or at least weakens) the ties in social networks which could turn out important for the tenure decision".

In the empirical analysis, in addition to international mobility, we take into account different dimensions of the notion of social capital and their impact on scientific careers, that hereafter we briefly summarize.

First we investigate the dimension of social capital related to inbreeding and localism, the former to be intended as the tendency of a university to recruit new staff among the ranks of local graduates, the second as the more general tendency to fill professorial position through internal careers, as opposed to attracting scientists from other institutions.

Hargens and Farr (1973) look at the number of years it takes for an assistant professor to be promoted to an associate position, and find that inbred scientists wait for longer than others, even after controlling for differences in terms of productivity. Perotti (2002) documents a number of instances in which Italian selection committees preferred local candidates to much better qualified external ones. More generally, localism is denounced as a factor of backwardness in the academic systems (Abbot, 2006, Godechot and Louvet, 2008).

Secondly, we explore the importance of professional knowledge networks at the international level, which corresponds both to a professional need and to a factor shaping the mobility of researchers. Mahroum (2000) defines "scientific mobility as a process of networking and extending one's social space […] stimulated by a need for professional socialization". When academic factors are dominant in the decision to move, migration can be temporary and return can naturally follow through purposefully created linkages. Following this line of reasoning,

[1] We are aware that, due to different publication practices, using this approach underestimate postdoctoral mobility in social and human sciences, in the econometric estimation we will use a reduced sample of only Science, Technology, Engineering, Mathematics and Medicine (STEMM) + Economics fields researchers. Results are robust to the inclusion of Human and Social Sciences and are available upon request.

home country linkages might increase the probability of becoming aware of opportunities and make it easier to find an opportunity and the necessary information and support when a researcher returns (Ackers, 2005). Furthermore, it has been empirically shown that these linkages may be necessary for reintegration in the national work market (Gill, 2005; Morano-Foadi, 2005).

Furthermore, when universities decide to fill a vacancy or offer a promotion may give positive consideration to the size and reach of candidates' personal network, since the latter may add to the university's visibility and access to resources (Gonzalez-Brambila et al., 2006). As individual performances are often hard to evaluate only on the basis of past scientific production and citations (especially when junior scientists are considered, whose publication list is necessarily short), prospective recruiters or promoters may look for other signals of quality, and the ability of expanding the scientific network is one of these. Expanding this notion, new social ties an individual may have established in universities and research labs, by moving across different institutions, can be considered as a relevant form of social capital.

Data

We have collected information from three primary sources: the National Library of Florence (BNCF), the Italian Ministry of education (MIUR) and the bibliographic Scopus database. From BNCF We retrieved all doctoral dissertations discussed in Italian universities from I cycle (1986) to 2006. BNCF online public access catalogue provides information on: author, title of the thesis, supervisor, PhD university, scientific field and year of degree. From MIUR we obtained administrative data on academic positions, disciplinary areas, university affiliation and personal information, such as birth year and gender, for all academics working in Italian universities from 1990 to 2015. Using these two sources of information, we identified PhD holders who pursued a scientific career in Italian academia.

The identification of academics who hold an Italian doctoral degree was pursued through the record linkage between academics from the MIUR data and doctorate holders from Italian universities from BNCF data. We performed the matching relying on four fields: name, gender, scientific area and year of PhD. We were thus able to identify the population of researchers with doctoral degree from Italian universities who have worked at least for one year in Italian academia. For further details on the retrieval process from BNCF, the record linkage procedure and its results see Coda Zabetta (2018). To reduce the potential selection bias in our empirical analysis, we use data on 18,039 doctorate holders, who entered Italian academia as assistant professors within 10 years after the PhD, and are active in 2015.

For these researchers, we retrieved from Scopus all scientific articles published in international journals since their first publication to 2015. Within this group of researchers, 15,385 (85%) published at least one paper on Scopus journals. In total we gathered 285 thousands publications.

We use the following procedure in order to identify authors. Using Scopus API, we downloaded all available personal information for the academics in our data. This information includes: affiliation, scientific research area and Scopus Author-ID, the latter is a unique identifier for each author inside Scopus database.

A recent study (Kawashima et al., 2015), evaluated the accuracy of the Author ID in the Scopus bibliographic database. They matched bibliographic records between Scopus and an open database which manages all the information of the largest public fund for academic researchers, then they calculated recall and precision of the Scopus Author ID for researchers. They found that recall and precision were around 98% and 99% respectively.

Then, we assigned all academics in our data and authors' record downloaded from Scopus to a broad disciplinary category. In order to attribute comparable disciplinary categories for authors and individuals, we aggregate disciplines defined by MIUR and Scopus disciplinary areas into the following categories: Agriculture; Chemistry; Biology; Physics; Mathematics and Computer Science; Architecture and Engineering; Medicine and Veterinary; Economics and Management; Humanities and Law, Sociology and Political Science. Finally, in each broad disciplinary category we matched authors with academics in our data using the information on their surnames, names and affiliation.

After filtering, duplicates and incomplete records were deleted obtaining a consistent database of 285,283 scientific publications with at least one Italian author. We then employed a matching procedure to assign the corresponding author identifying codes to each research product (it might be possible that one paper is co-authored by two or more different individuals belonging to Italian academia).

We proxy early career mobility using the affiliation reported in the publication collected from Scopus. In this way we are able to identify those researchers who, after the PhD and before the first appointment in Italian academia, spent a research period at least in a postdoctoral position (we do not take in to account short research stays, which usually do not resolve in a publication). A crucial point that has to be made here is that bibliometric research allows us to track mobility only to the extent that researchers publish and that their affiliation is stated on their publication in a way that can be traced back to them.

A number of studies lend some qualified support to the use of these data for tracking mobility. Laudel (2003) and Conchi and Michels (2014) compared scientist mobility records derived from bibliometric data with those derived from alternative data sources, including CV and self-reported data from scientist surveys. Moed et al. (2014) evaluate the potential and limitations of the bibliometric approach in terms of author profile accuracy and interpretation, looking at the coherence between related statistics and scientist mobility as implied in Scopus publication records for authors in 17 countries. The authors conclude that the bibliometric approach is promising since error rates for units of assessment with indicator values based on sufficiently large numbers are estimated to be fairly below 10%.

Using affiliation data from Scopus we then identified affiliation with a single address per author (in order not to take into consideration virtual mobility) and categorized the country of the reported institution. In this way, we are able to identify researchers' mobility if: i) the researcher publishes; ii) the affiliation is reported in the publication; iii) authors are single-affiliated (we do not take into account multiple affiliations per a single author). To identify and disambiguate affiliation reported on publication data we used GRID database. Table 1 and Table 2 show some exploratory information for the international mobile academics.

Methods

We estimate a duration model of career promotion as a function of international research appointments. We assume that each academic is subject to the probability of being promoted conditional on her status as an assistant professor. In the duration analysis an academic is at risk of being promoted from the first appointment as assistant professor.

We make use of the Cox-proportional hazard model where the dependent variable is the time that elapses from first appointment until promotion to associate or full professor position. This model is written for any individual *i*:

$$h(t) = h_0(t) \times \exp(\alpha_1 PD_{Abroad_i} + \alpha_2 PD_{Italy_i} + \beta SocCap_i + \gamma \boldsymbol{X}_{i,t})$$

where h_0 is the baseline hazard, PD_Abroad_i is a dummy which takes value one if the researcher spent a postdoctoral period abroad, PD_Italy_i is a dummy variable which takes value one if the researchers did a postdoc in Italy, $SocCap_i$ is a set of variables that aim to capture the social capital/network effect and $\boldsymbol{X}_{i,t}$ is a vector of individual characteristics, some of them time-variant. Age and its squared term are included to control for a possible age effect on promotion. Gender, PhD, and university type indicators are used as controls. Performance measures are included to assess the importance of merit for promotion. All regressions also include year, university and scientific area dummies.

Table 1. Number of PhD, international mobiles and share by gender and cohort

	All	*M*	*F*	*Cohort 86-96*	*Cohort 97-07*
# PhD	18039	10358	7681	4908	13131
# postdoc abroad	1375	906	469	442	933
% postdoc abroad	8%	9%	6%	9%	7%

Table 2. Share of international mobiles for the top 10 destinations by gender and cohort

Country	*All*	*M*	*F*	*Cohort 86-96*	*Cohort 97-07*
United States	34%	33%	35%	37%	32%
United Kingdom	15%	14%	15%	14%	15%
France	11%	11%	11%	11%	11%
Germany	10%	10%	8%	8%	10%
Switzerland	6%	6%	5%	7%	5%
Spain	4%	4%	3%	2%	4%
Netherlands	3%	3%	4%	5%	3%
Belgium	2%	2%	2%	2%	2%
Canada	2%	2%	2%	3%	2%
Sweden	2%	33%	35%	2%	2%

Standard models that control for confounding factors may fail if the treatment, postdoctoral mobility in our case, is time-variant (Robins, 1999). For example, controlling for past values of productivity, which affect later research appointments and promotion, can lead to biased estimates. To address this problem of reverse causality between research visits and promotion, we use coarsened exact matching (CEM; Iacus et al., 2012) to match each academic to a peer who has not participated in a research visit based on pre-visit observable characteristics (gender, birthyear, PhD year, scientific area, rank of PhD university, publications and citations during the PhD).

This strategy considers research visits as a treatment with a lasting effect on academics' careers. Research appointments are usually undertaken by junior academics and can serve as a treatment affecting future career paths. We thus divide the sample into a treated group and an untreated control group (i.e. academics who participate in research visits and similar academics who do not). We then apply Cox proportional hazard model to this matched sample. Hereafter we briefly discuss the explanatory variables related to individual characteristics, STHC and social capital.

According to the Italian legislation, scientific productivity ought to be the key determinant for career advancement. It is most common to measure productivity by counting publications and citations in international scientific journals. Therefore, we extracted from Scopus all the scientific articles published between the first appointment and 2015, authored by at least one individual in our sample, together with their citations in 2015. We also include the dummy *Precocity* for those who published during the PhD. We constructed the two variables cumulative number of publications by year (*CumPub*) and cumulative number of average citations by year (*CumAvgCit*). Other variables of interest at the individual level are *Age* (which we treat as a proxy for the scientist's career length) and gender (*Female*).

Among the variables related to STHC, we started by including a variable that identifies if the focal academic got her promotion in her Alma Mater (*Inbred*) and interacted it with postdoctoral mobility to assess whether promotion was speeded up by a tight social network.

For what concerns the home country linkages, we try to capture them by a variable based upon information on researchers' affiliation, as derived from the Scopus records. Each Scopus record lists, in separate fields, the authors' names, and their affiliations, with a one-to-one correspondence between names and affiliations. Thus, we are able to derive from publication records the exact affiliation of each scientist. With this information we built the variable *PD_Abroad_Coauth_ITA*, if more than 75% of the co-authors with whom the researcher has published while abroad are affiliated to an Italian university, which signals a strong connection with Italian academia for researchers who are abroad.

Furthermore, since we downloaded publications for the academics on our data since the PhD years, we are able to built the variable *PD_Abroad_Coauth_NEW*, if more than 75% of coauthors with whom the focal researcher has published while abroad were not among her previous coauthors, in order to verify whether the researcher is actually expanding his/her scientific network, thus acquiring more STHC.

The literature surveyed above pays particular attention to the prestige of the PhD-granting institution. We identified the top universities in Italy (according to ARWU ranking[2]) and created the dummies *Top_Uni_PhD* for the PhD granting institution, and *Top_Uni*, for the university at which the focal researcher was employed before being promoted.

Finally we control for disciplinary differences in the availability of new jobs and promotion opportunities, by inserting in all regressions a dummy variable for each university, scientific area and calendar year.

Given the strategy used to identify international postdoctoral mobility (based on publications), we focus our analysis on the Science, Technology, Engineering, Mathematics and Medicine (STEMM) + Economics scientific areas (11,404 academics), as in those fields international mobility is more common and using Scopus to trace mobility is more reliable as journal publications is the normal way of communicating research results.[3]

Table 3 show that on average international mobile academics are more frequently promoted but that they are not significantly promoted sooner than they peers. Table 4 describes the variables used in the empirical analysis and Table 5 presents the summary statistics.

Table 3. Promotion and number of years until promotion (by groups of academics)

Variable	*Promoted*		*Years to promotion*	
	No	*Yes*	*No*	*Yes*
PD_Abroad	0.51	0.55***	8.11	7.78**
PD_Italy	0.58	0.44***	8.08	8.05
Inbred	0.54	0.49***	7.65	8.34***
STEM	0.37	0.72***	7.93	8.20***
Cohort8696	0.57	0.41***	7.81	8.60***
Female	0.52	0.50**	7.76	8.24***
Precocity	0.45	0.53***	7.69	8.14***
Top_Uni_PhD	0.51	0.55***	8.11	7.78**

Significance test of mean differences with "No" group.*p<0.10, **p<0.05, ***p<0.01

[2] We have also created three other alternative rankings of Italian universities based on: a) national competitive funding success, b) national excellence program success and c) national Research Assessment. We obtained similar results, which are available upon request.

[3] For example, both in the UK REF and in the Italian RAE, STEMM and Economics fields were considered bibliometric fields (Geuna and Piolatto, 2016).

Table 4. Description of the main variables

Variable	*Description*
PD_Abroad	1: Researcher spent a postdoctoral period abroad
PD_Abroad_USA	1: Researcher spent a postdoc period in the US
PD_Abroad_EUR	1: Researcher spent a postdoc period in a EU-country
PD_Abroad_OTH	1: Researcher spent a postdoc period in a non-US/EU country
PD_Abroad_Coauth_ITA	1: Researcher has >75% IT co-authors during postdoc abroad
PD_Abroad_Coauth_FOR	1: Researcher has ≤75% IT co-authors during postdoc abroad
PD_Abroad_Coauth_NEW	1: Researcher has >75% new co-authors during postdoc abroad
PD_Abroad_Coauth_OLD	1: Researcher has ≤75% new co-authors during postdoc abroad
PD_Italy	1: Researcher spent a postdoctoral period in Italy
Inbred	1: Researcher is employed at the PhD university
Precocity	1: Researcher has published a scientific article during the PhD
CumPub	Cumulative number of publications by year
AvgCumCit	Cumulative number of average citations by year
Top_Uni	1: Current employing university is listed in ARWU
Top_Uni_PhD	1: PhD university is listed in ARWU
Female	1: Researcher is female
Age	Researcher’s age

Table 5. Descriptive statistics of main variables.

Variable	*Mean*	*SD*	*Min*	*Max*
Years since first appointment	7.71	3.65	1	24
Promoted to AP or FP	0.51	0.50	0	1
PD_Abroad	0.11	0.32	0	1
PD_Italy	0.50	0.50	0	1
PD_Abroad_USA	0.05	0.21	0	1
PD_Abroad_EUR	0.06	0.24	0	1
PD_Abroad_OTH	0.01	0.09	0	1
PD_Abroad_Coauth_ITA	0.02	0.15	0	1
PD_Abroad_Coauth_FOR	0.09	0.29	0	1
PD_Abroad_Coauth_NEW	0.08	0.27	0	1
PD_Abroad_Coauth_FOR	0.04	0.19	0	1
Inbred	0.59	0.49	0	1
CumPub	17.82	35.75	0	943
AvgCumCit	21.18	26.08	0	878
Precocity	0.64	0.48	0	1
Female	0.40	0.49	0	1
Age	42.96	5.39	29	63
Top_Uni_PhD	0.81	0.40	0	1
Top_Uni	0.69	0.46	0	1

Results and discussion

To provide a first impression of the survival process, Figure 1 depicts the hazard curve (left) and the Kaplan-Meier survival estimate (right) for the observations split by our main variable of interest, namely the dummy *PD_Abroad*, which takes value 1 for researchers who spent a postdoctoral period abroad.

We can observe that academics who spent a research period abroad exhibit a higher hazard of being promoted (left) and a steeper survival curve (right) which means that their probability of surviving (not getting a promotion) is decreasing faster with respect to those without international experience.

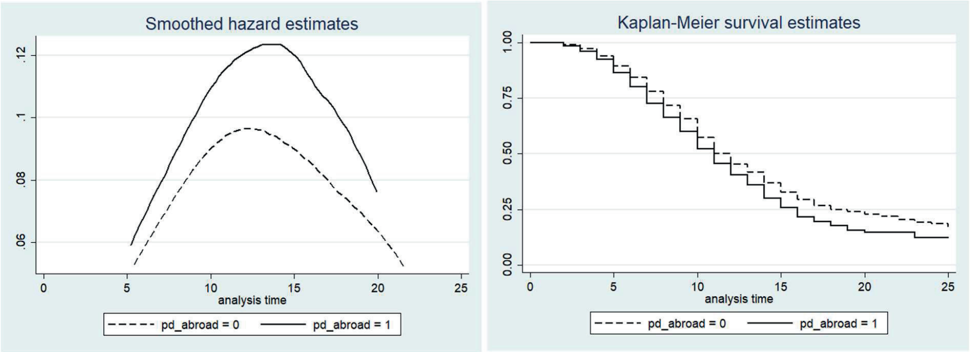


Figure 1. Kaplan-Meier survival estimate (left) and hazard curve (right).

Table 6 shows the results of the Cox model estimations for promotion from assistant professor to associate or full professor positions, investigating the effect of having spent a postdoctoral period abroad. The baseline Cox results for promotion (column 1) show that postdoctoral appointments have a strong positive effect, indicating that academics benefit from their additional experience in terms of career advancement and are promoted faster.

In column 2 we split international postdoc positions into appointments to the USA, which are assumed to be the most valuable due to the global status of their institutions, Europe and other countries. Indeed we find that visits to the USA increase the likelihood of being promoted more than visits elsewhere.

In column 3 we relate inbreeding (i.e. working at the PhD awarding institution) with postdoctoral appointments. We expect a greater effect from postdoctoral appointments for inbred academics who can take advantage of institutional links since their PhD training. Results in columns 3 show that the main effect for inbreeding is negative, thus, time to promotion is longer for inbred academics than for non-inbred academics. This results is in line with the US-based evidence discussed above, but contradicts previous studies focusing on promotion in Italian academia (such as Perotti, 2002). The interaction term with postdoctoral mobility abroad is positive, following our expectation that this type of research appointments is particularly important for inbred academics, but not significant.

Column 4 shows that researchers whose co-authors in the postdoctoral period are for the large majority Italian, are at a higher risk of promotion with respect to researcher who are more involved in collaborations with foreign researchers. Hence, maintaining contacts with the home country pays off in terms of having a faster career.

Column 5 shows that, for researchers who have the vast majority of new co-authors acquired during the postdoctoral period, the effect on promotion is positive, significant and larger in magnitude with respect to the baseline estimation. The control variables follow our theoretical expectations. In particular, *Precocity* suggest that having published a scientific article during the PhD increases that hazard of obtaining a professorial position by 18%. Finally, the prestige of the PhD granting institution has a positive and significant impact on career advancement, confirming the US-based empirical evidence discussed above.

Table 6. Survival analysis: risk of being promoted in t

	(1)	(2)	(3)	(4)	(5)
	Baseline	By Dest.	Inbreeding	ITA Coauth.	New Coauth.
PD_Abroad	1.629***		1.573***		
	(0.080)		(0.106)		
PD_Italy	1.211***	1.211***	1.216***	1.211***	1.211***
	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
PD_Abroad_USA		1.751***			
		(0.116)			
PD_Abroad_EUR		1.562***			
		(0.095)			
PD_Abroad_OTH		1.462***			
		(0.215)			
Inbred			0.804***		
			(0.025)		
PD_Abroad×Inbred			1.053		
			(0.085)		
PD_Abroad_Coauth_ITA				1.700***	
				(0.155)	
PD_Abroad_Coauth_FOR				1.611***	
				(0.086)	
PD_Abroad_Coauth_NEW					1.777***
					(0.100)
PD_Abroad_Coauth_OLD					1.414***
					(0.099)
CumPub	1.035***	1.035***	1.035***	1.035***	1.035***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
CumAvgCit	1.006***	1.006***	1.006***	1.006***	1.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Precocity	1.176***	1.175***	1.184***	1.176***	1.182***
	(0.040)	(0.040)	(0.041)	(0.041)	(0.041)
Gender	0.697***	0.698***	0.700***	0.698***	0.697***
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Age	1.147***	1.147***	1.148***	1.147***	1.146***
	(0.048)	(0.048)	(0.048)	(0.048)	(0.048)
Age^2	0.998***	0.998***	0.998***	0.998***	0.998***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Top_Uni_PhD	1.123***	1.124***	1.069*	1.123***	1.123***
	(0.045)	(0.045)	(0.042)	(0.045)	(0.045)
Top_Uni	0.915***	0.914***	0.999	0.915***	0.916***
	(0.030)	(0.030)	(0.034)	(0.030)	(0.030)
University dummies	Yes	Yes	Yes	Yes	Yes
Scientific area dummies	Yes	Yes	Yes	Yes	Yes
Calendar year dummies	Yes	Yes	Yes	Yes	Yes
Individuals	11404	11404	11404	11404	11404
Log likelihood	-47034.8	-47033.4	-47009.5	-47034.7	-47030.5
Chi-squared	6268.6	6271.4	6319.2	6268.9	6277.3

Exponentiated coefficients. * p<0.10, ** p<0.05, *** p<0.01

Table 7. Survival analysis: risk of being promoted in t (CEM sample)

	(1)	(2)	(3)	(4)	(5)
	Baseline	By Dest.	Inbreeding	ITA Coauth.	New Coauth.
PD_Abroad	1.412***		1.229*		
	(0.132)		(0.144)		
PD_Italy	1.016	1.025	1.005	1.015	1.015
	(0.102)	(0.103)	(0.101)	(0.102)	(0.102)
PD_Abroad_USA		1.590***			
		(0.176)			
PD_Abroad_EUR		1.307***			
		(0.132)			
PD_Abroad_OTH		1.567**			
		(0.295)			
Inbred			0.674***		
			(0.067)		
PD_Abroad×Inbred			1.244*		
			(0.158)		
PD_Abroad_Coauth_ITA				1.435***	
				(0.182)	
PD_Abroad_Coauth_FOR				1.404***	
				(0.137)	
PD_Abroad_Coauth_NEW					1.477***
					(0.146)
PD_Abroad_Coauth_OLD					1.295**
					(0.147)
University FE	Yes	Yes	Yes	Yes	Yes
Scientific area FE	Yes	Yes	Yes	Yes	Yes
Calendar year FE	Yes	Yes	Yes	Yes	Yes
Individuals	2028	2028	2028	2028	2028
Observations	19098	19098	19098	19098	19098
Log likelihood	-6427.0	-6424.6	-6418.6	-6427.0	-6426.0
Chi-squared	1113.9	1118.6	1130.6	1113.9	1115.8

Exponentiated coefficients. * p<0.10, ** p<0.05, *** p<0.01

Table 7 shows the previous results replicated for the restricted sample of matched academics. Since the matching is done considering postdoctoral mobility abroad (which is our main phenomenon of interest) as a treatment, the variable regarding postdoctoral mobility within Italy loses its explanatory power and is no longer significant, so coefficients should be interpreted carefully. We do not report coefficients for control variables, which are consistent with those reported in Table 6, in order to make the table more readable (the results are available upon request). It is possible to notice that all previous results hold also for the restricted matched sample. We thus conclude that are our findings are robust.

Conclusions

In this paper we have examined the effect of international research appointments and social capital on career outcomes in Italy in terms of the length of time until promotion. We have assembled data on affiliations, productivity and careers of researchers active in Italian academia in between 1986 and 2015. We focused on international postdoctoral research appointments, which may help to expand existing scientific and technical human capital while at the same time ensuring career stability.

In addition to international mobility, we have considered both individual and social determinants of promotion to professorial positions for assistant professor.

As for individual determinants (such as productivity, gender, and precocity), our results are in line with the US-based evidence, although some differences were found with previous literature which investigated Italian academia.

Coming to social determinants, we focused on social capital that contributes to enhance an individual scientific potential (scientific and technical human capital). We distinguished three different dimensions of the notion of social capital and we have produced individual and bibliometric indicators that try to capture their specificities.

We found that expanding the own scientific network during the postdoctoral period abroad accelerates academic careers in Italy. In particular, the ability of expanding the scientific network in universities and research labs, by moving across different institutions, is a relevant form of social capital an valuable in the long term. We also found that maintaining connections to Italian academia while being abroad beneficial in terms of time to promotion. In particular, Italian potential candidates to professorial positions benefit from intensity of collaboration with professors in their home country during the postdoctoral period abroad. However, we found no effect of localism: international returnees who work at their PhD granting institution are not promoted sooner than their peers. We used coarsened exact matching to match each academic to a peer who has not participated in international mobility based on pre-move observable characteristics, obtaining results which confirm the robustness of our findings.

These results present some interesting insights into the role of research visits for career advancement. Our findings suggest that early career international research appointments avoid some of the barriers to job mobility: career insecurity, instability, and difficulty of re-entry, and are therefore more likely to lead to promotion. This makes a case for governments to provide better incentives for employing organisations to also reward other types of mobility. A better understanding of individual scientists' career incentives and constraints, of the type we tried to provide with our study, may help to evaluate recent reforms in Italy, which modified many aspects of academic careers, including recruitment and promotion.

References

Abbot, A. (2006). Saving Italian Science. *Nature* 440, 264–265.

Ackers, L. (2005). Moving people and knowledge: Scientific mobility in the European Union. *International Migration*.

Becher, T., & Trowler, P. R. (2001). *Academic Tribes and Territories. Cultures*.

Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: an alternative model for research evaluation. *International Journal of Technology Management*.

CEC (2004) Commission staff working paper. Second implementation report on "A mobility strategy for the European Research Area'' (SEC(2004)412). Brussels: Commission of the European Communities.

Coda Zabetta, M. (2018). Essays on Career Progression in Italian Academia. Ph. D. thesis, University of Turin.

Conchi, S., & Michels, C. (2014). Scientific mobility - An analysis of Germany, Austria, France and Great Britain. *Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis*, (41).

Fernandez-Zubieta, A., Geuna, A., & Lawson, C. (2015). *What Do We Know of the Mobility of Research Scientists and of its Impact on Scientific Production. SSRN.*

Fontes, M. (2007). Scientific mobility policies: How Portuguese scientists envisage the return home. *Science and Public Policy*.

Geuna, A., & Piolatto, M. (2016). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while). Research Policy, 45(1), 260–271.

Gill, B. (2005). Homeward bound? The experience of return mobility for Italian scientists. *Innovation*.

Godechot, O., Louvet, A. (2008). Le localisme dans le monde académique: un essai d'évaluation. La Vie Des Idées.

Gonzalez-Brambila, C.N., Veloso, F., Krackhardt, D. (2006). Social Capital and the Creation of Knowledge. Mimeo.

Hargens, L. L., & Farr, G. M. (1973). An Examination of Recent Hypotheses About Institutional Inbreeding. *American Journal of Sociology*.

Heining, J., Jerger, J., Lingens, J. (2007). Success in the Academic Labour Market for Economists – The German Experience. University of Regensburg, Regensburger Diskussionsbeiträge zur Wirtschaftswissenschaft, no. 422.

Hunter, R. S., Oswald, A. J., & Charlton, B. G. (2009). The elite brain drain. *Economic Journal*.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*.

Kawashima, H., Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of scopus author ID based on the largest funding database in japan. *Scientometrics*.

Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help? In *Scientometrics*.

Mahroum, S. (2000). Scientific mobility: An agent of scientific expansion and institutional empowerment. *Science Communication*.

Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, *101*(3), 1987–2001.

Morano-Foadi, S. (2005). Scientific mobility, career progression, and excellence in the European Research Area. *International Migration*.

Musselin, C. (2004). Towards a European academic labour market? Some lessons drawn from empirical studies on academic mobility. In *Higher Education*.

Newland, K. (2009). "Circular Migration and Human Development," MPRA Paper 19225, University Library of Munich, Germany.

Perotti, R. (2002). The Italian University System: Rules and Incentives. ISAE, Rome.

Robins, J. (1999). Association, Causation, and Marginal Structural Models. *Synthese*, 121(1/2), 151-179. Retrieved from http://www.jstor.org/stable/20118224

Sabatier, M., Carrere, M., & Mangematin, V. (2006). Profiles of academic activities and careers: Does gender matter? An analysis based on french life scientist CVs. *Journal of Technology Transfer*.

Stephan, P., 2012. How Economics Shapes Science. Harvard University Press, Cambridge.

Saxenian, A. L. (2005). From brain drain to brain circulation: Transnational communities and regional upgrading in India and China. *Studies in Comparative International Development*.

Van Bouwel, L (2010). Return rates of European graduate students in the US: How many and who return, and when?. *Belgeo*, 4 , 395-405.

Citing Alike, Writing Alike: Comparing Discourse- and Bibliographic Coupling-Based Science Maps

Bradford Demarest[1], Cassidy R. Sugimoto[2], and Vincent Larivière[3]

[1] *bdemares@indiana.edu*
Indiana University, School of Informatics, Computing, and Engineering, Department of Information and Library Science, 611 N. Woodlawn, 101B, Bloomington, IN 47408

[2] *sugimoto@indiana.edu*
Indiana University, School of Informatics, Computing, and Engineering, Department of Information and Library Science, 919 E. 10th St., 263, Bloomington, IN 47408

[3] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7, Canada

Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC. H3C 3P8, Canada

Abstract

This study proposes a way to map the sciences based on social and epistemic cultural features in writing that can expose heretofore unexposed connections between disciplines. A network based on social and epistemic term frequencies in 1,269,146 journal articles from 14 disciplines is created and compared to a network of the same articles based on bibliographic coupling at the discipline level. The two networks are found to correlate moderately (0.577) with a p-value of 0.0002, and hierarchical clustering conducted on the networks show connections between Health and Clinical Medicine based on bibliographic coupling, and between Health, Psychology, and Social Sciences among others, based on social and epistemic terms in writing.

Introduction

A conflict exists in contemporary science regarding interdisciplinarity. On one hand, interdisciplinary research is widely heralded as critical to solving complex global issues such as climate change (Rylance, 2015). interdisciplinary research has the capacity for great success; Larivière, Haustein, and Börner (2015) find that in a study of 9.2 million papers from 2000 to 2012, a majority of co-cited interdisciplinary research papers result in higher relative citation counts for citing papers, with the highest relative citation counts reserved for interdisciplinary papers that draw from distant disciplines.

On the other hand, interdisciplinary research has been found to have consistently lower success in acquiring funding than disciplinary research (Bromham, Dinnage, & Hua, 2016); this would seem to reflect the perspective of some scholars that interdisciplinary research suffers when evaluated from traditionally disciplinary perspectives (Rylance, 2015). One reason that these evaluations may be hard on interdisciplinary research is due to differing social and epistemological norms in different disciplines, leading evaluators to see interdisciplinary work as an unsatisfactory version of scholarship from the evaluator's discipline rather than a culturally related but distinct product.

Mapping the socio-epistemic cultures of the disciplines is, then, an important first step toward accounting for such disciplinary clannishness and thus eventually opening the door for more productive interdisciplinary innovation. The current study begins this work, using a computational linguistics method (i.e., discourse epistemetrics (DE), per Demarest & Sugimoto, 2015) to extract social and epistemological disciplinary cultural information from scholarly article abstracts. This information is summarized as a pairwise distance metric, which we then use to derive a network of disciplines. As a basis for comparison, we also create disciplinary networks from references for the same papers. We compare the two networks using Quadratic Assignment Procedure (QAP), graphically based on heatmaps, and with hierarchical clustering.

Literature Review

A wide variety of scholarship in sociology of science bears out the multitudinous ways in which new knowledge is created and verified (Becher & Trowler, 2001; Whitley, 1984). Furthermore, the writing of different scientific communities reflects these different disciplinary identities through social and epistemic language (Argamon, Dodick, & Chase, 2008; Cronin, 2005; Hyland, 2000). While science has been mapped using other measures of similarity including bibliographic coupling (Kessler, 1963; Boyack & Klavans, 2010), co-citation (Small, 1973; Boyack & Klavans, 2010, White & McCain, 1998), and co-authorship analysis (Glänzel & Schubert, 2005), no studies so far have attempted to map science based on social and epistemic written discourse terms. Leveraging the discourse epistemetrics method we previously established as both accurate and interpretable (Demarest & Sugimoto, 2015), the current study undertakes this mapping effort.

Methods

To study disciplinary social and epistemic features in academic writing, this method uses a sample of journal article abstracts from the Web of Science, taken from a single publication year. These abstracts are transformed into frequency vectors of lexical features that previous scholars have found to be indicative of different types of stance. After this transformation, support vector models (SVMs) are then generated for each disciplinary pair, with the accuracy of the model used as a measure of socio-epistemic distance between the disciplines. These accuracy measures are then used to describe the collection of disciplines as a network, which can be compared with a network of disciplines created based on patterns of references. Each of these aspects, including the specifics of sample, features, and model parameterization, are discussed in further detail below.

Sample

The current study utilizes abstracts and references for 1,269,146 English-language scholarly articles from the Web of Science from 2011. Articles from a single year with available abstracts were chosen to avoid any temporal effects on disciplinary socio-epistemic cultures and writing. For article counts by discipline, see Table 1.

Discourse Epistemetrics Features

Each abstract is first converted to a vector of relative frequencies of 568 social and epistemic terms collected from previous scholarship of social and epistemic stance in writing (Biber, 2006; Biber & Finegan, 1989; Hyland, 2005). These terms were found by the scholars to serve one of several functions. Hedging terms mitigate the certainty of an assertion; examples include “perhaps”, “approximately”, or “seem”. Conversely, boosting terms amplify assertions, e.g., “obviously”. Terms that frame an assertion emotionally or judgmentally are affective markers, including terms such as “unfortunately” and “surprisingly”. Aside from these, two other sets of socio epistemic terms exist – those that refer to the author herself (self-references such as “I”, “we”, or “the author”), and those that refer to the reader directly or implicitly (such as “the reader”, and “you”, as well as imperative verbs). For a full list of features, please contact the first author.

Discourse Epistemetrics Model Parameterization

After preparing the data, pairs of disciplines or specializations were then used to train and test SVMs. The LinearSVC from Python’s scikit learn toolkit (Pedregosa et al., 2011) was employed, and thus a linear kernel, such that feature weights could be analyzed. Per Varma and Simon (2006), we used a grid search approach to hyperparameter optimization of C (the total error value). In order to avoid bias deriving from uneven sample sizes, balanced error values by category size were used. Finally, 10-fold cross validation was employed, with accuracy values averaged across the 10 cycles, to minimize variation due to assignment of samples to the training or test data sets. The resulting average accuracy measures for each disciplinary (or specialization) pair was then used as a distance metric – the higher the accuracy of the optimized model, the more distinct the two disciplines are from one another in terms of the social and epistemic discourse they use.

Table 1. Counts of Web of Science articles by discipline.

Discipline	**Articles**
Arts	1731
Biology	93765
Biomedical Research	153166
Chemistry	129685
Clinical Medicine	340574
Earth and Space	70018
Engineering and Technology	172949
Health	28343
Humanities	13673
Mathematics	42685
Physics	121702
Professional Fields	34590
Psychology	25802
Social Sciences	40463
Total	**1269146**

Disciplinary categories are taken from the U. S. National Science Foundation (NSF) field classification (Hamilton, 2003).

Bibliographic Coupling

To form a reference-based network at the disciplinary level, a matrix of reference counts per discipline was collected, with each row reflecting counts for a given referring discipline, and each column reflecting number of papers for a given discipline referenced by the row-discipline. Cosine distance was then used to calculate distance between each pairwise combination of disciplines. The process was repeated at the specialization level.

Findings

The findings presented here constitute summaries and visualizations of the study's data; for item-level information (such as cosine distance or accuracy for a given disciplinary pair), please contact the first author. Table 2 presents summary statistics for discipline-level networks based on discourse epistemetrics (for which the numbers are accuracy rates) and on bibliographic coupling (for which values reflect cosine distance).

Table 2. Summary statistics for discipline-level networks.

	Maximum Value	Minimum Value	Median
DE (accuracy)	0.988	0.612	0.887
BC (cosine distance)	0.983	0.132	0.898

Notably, pairwise models based on interactive metadiscourse term frequencies achieve accuracy rates of as high as 98.8%, and even the lowest accuracy models improve upon the baseline of 50% accuracy by 11%. Cosine distance based on bibliographic coupling by discipline reflects a wider range. However, even taking this difference between distributions into account, we ran a 5000-iteration Quadratic Assignment Procedure analysis via UCINET (Borgatti, Everett, & Freeman, 2002) that compared discourse and reference-based distance matrices that yielded a Pearson's Correlation of 0.577 (p= 0.0002), suggesting that moderate correlation does exist between disciplines that cite alike and those that write alike.

Figure 1 presents a heatmap of disciplines based on discourse epistemetrics measures. Of the disciplines shown in Figure 1, the closest disciplines (i.e., those with the lowest DE accuracy scores) are Clinical Medicine and Biology; Social Sciences and Professional Fields; Biology and Biomedical Research; Biomedical Research and Clinical Medicine; and Physics and Engineering. Disciplines with the highest DE accuracy scores (and thus furthest apart) are all paired with Arts: Biomedical Research, Physics, Engineering and Technology, Biology, and Clinical Medicine.

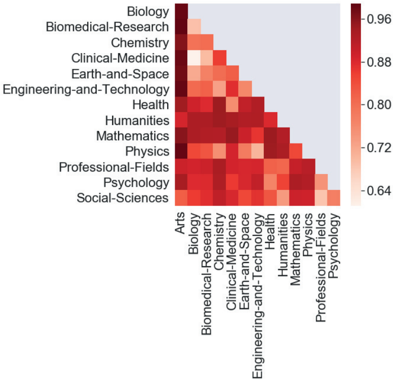


Figure 1. Heatmap of distances between disciplines (DE, accuracy).

Figure 2 presents a disciplinary heatmap showing cosine distance based on discipline-level bibliographic coupling.

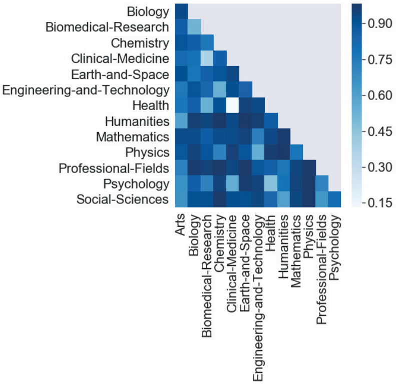


Figure 2. Heatmap of distances between disciplines (Bibliographic Coupling, cosine).

In Figure 2, the closest disciplines (i.e., with the lowest cosine distance) are Health and Clinical Medicine, followed by Biomedical Research and Clinical Medicine; Health and Psychology; Biology and Biomedical Research; and Biomedical Research and Health. Discipline pairs that are furthest apart by bibliographic coupling measure are Chemistry and Humanities, Physics and Humanities, Earth and Space and Humanities, Earth and Space and Professional Fields, and Chemistry and Professional Fields.

Hierarchical Clustering

Using the accuracy values from the DE modeling in one case and the cosine distances from the bibliographic coupling in the other, we next used the scipy implementation of hierarchical clustering (Jones, Oliphant, & Peterson, 2014) using Ward distance for each of the networks. Figures 3 and 4 below show the resulting dendrograms.

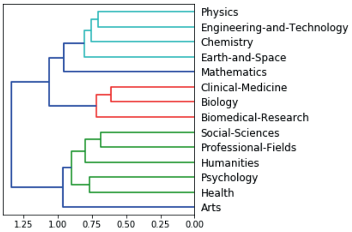


Figure 3. Disciplines clustered using hierarchical clustering (discourse epistemetrics, accuracy)

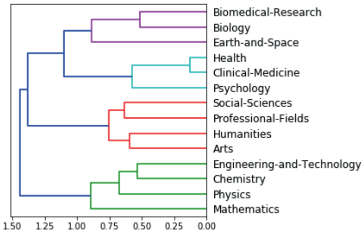


Figure 4. Disciplines clustered using hierarchical clustering (bibliographic coupling, cosine distance)

Figure 3 shows three clusters at the threshold of 1.00 – one for physical sciences, one for the biological sciences, and the last containing human-oriented and applied fields. The last of these clusters notably contains Psychology as well as Health, while the second cluster contains Biology, Clinical Medicine, and Biomedical Research. In contrast, Figure 4 contains four clusters at the same threshold. As before, a physical science cluster and a humanities-social science-professional cluster exist, but Biology, Biomedical Research, and Earth and Space disciplines occupy a separate cluster from Health, Clinical Medicine, and Psychology.

Discussion and Conclusions

This study has established that the discourse epistemetrics method can serve as a useful tool for mapping disciplines in recognizable constellations, and that differences in discourse networks and bibliographic coupling networks expose meaningful differences. Foremost among these is the distinction between Health as a discipline that writes most similarly to fields such as Social Sciences and Psychology, while citing similarly to the biomedical fields; this lays bare the interstitial nature of the Health field in particular, and the pipeline of the biological sciences from research fields (Biology and Biomedical Research) to Clinical Medicine, and then on to Health (with its emphasis on public policy). In consideration of the paradox of interdisciplinary research, it is hoped that this line of research will help to clarify differences when disciplines cite alike but write (and work) differently.

References

Argamon, S., Dodick, J., & Chase, P. (2008). Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. Scientometrics, 75(2), 203–238.

Becher, T., & Trowler, P. R. (2001). Academic Tribes and Territories: intellectual enquiry and the cultures of disciplines (2nd edition). Retrieved September 6, 2012, from http://eprints.lancs.ac.uk/3714/

Biber, D. (2006). University language: a corpus-based study of spoken and written registers. Amsterdam ; Philadelphia: J. Benjamins.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. Text, 9(1), 93–124.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for Windows: Software for social network analysis.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61(12), 2389–2404. https://doi.org/10.1002/asi.21419

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. Nature, 534(7609), 684–687. https://doi.org/10.1038/nature18315

Cronin, B. (2005). The Hand of Science: Academic Writing and Its Rewards. Scarecrow Press.

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. Journal of the Association for Information Science and Technology, 66(7), 1374–1387. https://doi.org/10.1002/asi.23271

Glänzel, W., & Schubert, A. (2005). Analysing Scientific Networks Through Co-Authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems (pp. 257–276). https://doi.org/10.1007/1-4020-2755-9_12

Hamilton, K. (2003). Subfield and Level Classification of Journals (CHI Report No. 2012-R) (No. CHI Report No. 2012-R). Cherry Hill, NJ: CHI Research.

Hyland, K. (2000). Disciplinary discourses: Social interaction in academic writing. Retrieved from http://www.lavoisier.fr/livre/notice.asp?id=OASW3KAR6OKOWH

Hyland, K. (2005). Metadiscourse: Exploring interaction in writing. Retrieved from http://books.google.com/books?hl=en&lr=&id=jgfgHpEqPN8C&oi=fnd&pg=PR8&dq=epistemic+metadiscourse+discipline&ots=60Ofr4zJRL&sig=ZYs8Z8BASLtXs3GGExodyDm07h0

Jones, E., Oliphant, T., & Peterson, P. (2014). SciPy: Open source scientific tools for Python.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1), 10–25.

Larivière, V., Haustein, S., & Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. Plos One, 10(3), e0122565.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, *12*(Oct), 2825–2830.

Rylance, R. (2015). Grant giving: Global funders to focus on interdisciplinarity. Nature News, 525(7569), 313. https://doi.org/10.1038/525313a

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24(4), 265–269. https://doi.org/10.1002/asi.4630240406

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science, 49(4), 327–355.

Whitley, R. (1984). The intellectual and social organization of the sciences. New York, NY: Clarendon Press.

Does the PageRank method improve the citations count?

Abdelghani Maddi[1] *and Damien Besancenot*[2]

[1] abdelghani.maddi@hceres.fr
The High Council for Evaluation of Research and Higher Education (Hcéres); CEPN, Université Paris 13.

[2] damien.besancenot@parisdescartes.fr
LIRAES and Université Paris 5, Sorbonne Paris Cité.

Introduction

In the assessment of research, counting of citations is ubiquitous. Articles' influence are valued according to the number of citations they received (Van Noorden, et al. 2014). Literature classifies the impact of researchers' work through their number of citations in academic journals (Medoff, 1996). Citations influence academic careers and researchers' rewards (Diamond, 1986), contribute to journals reputations (Bollen et al, 2006, Ritzberger, 2008) and allow the assessment of research departments, universities and countries (Dusansky and Vernon 1998).

However a citation doesn't exhibit the same worth depending on the citing article. In economics, being referenced by an article published in The American Economic Review gives a paper a higher visibility than being cited in a second or a third tier journal and, logically, bibliometric indicators take into account the heterogeneity of the pool of citations.

In the literature, the citation worth is assessed according to two major approaches. It is first measured as a function of the number of articles citing the citing article (see for instance Bollen et al 2006 and, in economics, Laband and Piette 1994). It may also reflect the quality of the publishing medium in which citing articles are published (the benchmark method here is close to the Eigenfactor approach; see Bergstrom et al, 2008). Focusing on the pro and con of the two approaches, a vast strand of research considers the most efficient way to measure the value of a given citation. In these settings, this work has two main objectives: first, to suggest alternative ways of introducing the value of a citation in the assessment of articles influence and second, to measure the sensibility of the resulting indicators to the way in which the weight of citations is taken into account.

Data and Method

Our work considers all articles published in five of the best economic journals: The American Economic Review, Econometrica, Journal of Political Economy, Reviews of Economic Studies and Quarterly Journal of Economics over the period January 2000 - December 2010. Following Hu et al (2011), hereafter, these articles will be referred to as the generation zero. We then identified every article citing these articles and listed in the Web of Science (the first generation of citing articles) and each article citing these citing articles (the second generation). The Dataset includes 3142 benchmark articles, 57.244 citing articles in the first generation and 191.000 in the second generation. In order to normalize the time period in which citations were recorded, we restricted these citations to the four year window following the publication of the cited articles.

With this database, we defined three families of indicators: the first assesses the influence of an article by the raw number of citations it received from the first generation (T_C). The second considers that a citation presents a higher value if the citing article in the first generation is itself frequently cited (Ph, Pg). The third approach pays attention to the journals where the first generation of citing articles were published (J_h^{CLh}, J_h^{CLm}, J_g^{CLh}, J_g^{CLm}). We also computed these indicators considering only citations published in the 600 journals ranked by Combes and Linnemer (2010) (T_C^R, Ph^R, Pg^R). Following a methodology introduced by Schubert (2009), indicators are built by analogy to the h and g-indexes proposed by Hirsch (2005) and Egghe (2006) to measure researchers' influence.

In order to assess the proximity between the ten rankings arising from our indicators, for each indicator, we built a specific ranking of the benchmark papers and computed the Spearman's coefficient for each couple of rankings.

Otherwise, in order to focus more precisely on the relationship between our various indicators, we then developed a Principal Components Analysis (PCA) using graphical representations in two dimensions for the two periods 2000-2005 and 2006-2010.

Results

Table 1 gives the mean values of the spearman's rho in our database, the lowest coefficient is equal to 0.7 and the mean value is 0.84. According to this table, there is a strong correlation between the different rankings. Even if the 10 indicators measure the worth of a citation according to rather different criteria, the ranking they induce appears to be weakly sensitive to them.

Table 1: Spearman's Rank correlation matrix

	P_h	P_g	J_h^{CLh}	J_g^{CLh}	J_h^{CLm}	J_g^{CLm}	P^R_h	P^R_g	T^R_C
T_C	0.967	0.862	0.912	0.952	0.984	0.775	0.874	0.777	0.891
	P_h	0.899	0.923	0.982	0.974	0.768	0.858	0.772	0.874
		P_g	0.923	0.917	0.899	0.697	0.775	0.701	0.788
			J_h^{CLh}	0.932	0.952	0.731	0.818	0.735	0.835
				J_g^{CLh}	0.969	0.762	0.850	0.766	0.866
					J_h^{CLm}	0.770	0.869	0.774	0.889
						J_g^{CLm}	0.940	0.998	0.926
							P^R_h	0.940	0.996
								P^R_g	0.927

Figure 1 presents the various indicators on the correlation circle for the two sub periods. Each indicator is associated with a dot whose coordinates are given by the correlation between the factor and the indicator.

Note that, in both circles figures, dots locations are very close to the edge of the circle and are therefore very representative of the indicators on the plane. The proximity between the vectors – as measured by the angles they form pair wise - allows us to determine whether evaluation methods are similar.

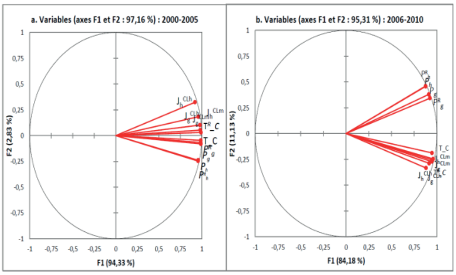


Fig 1: PCA Axis

Figure 1 therefore illustrates the high correlation between the different variables. For articles published during period 2000-2005. We note a slight difference between the J- indicators (J_h^{CLh}, J_h^{CLm}, J_g^{CLh}, J_g^{CLm}), above axis 1, that take into account the quality of the citing journal, and the P-indicators (Ph, Pg, Ph^R, Pg^R) below the same axis and focusing on the influence of the citing article. Both types of indicators seem also highly correlated with the two raw indicators: T_C and T_C^R (respectively the total number of citations received by each article and the number of citations from a journal surveyed in Combes and Linnemer, 2010). Note that the two last indicators present the highest correlation with axis 1 – respectively 0.945 and 0.946 – meaning that this axis is closely related to the raw number of citations. Equivalent results may be observed while considering period 2006-2010.

The PCA allows emphasizing the strong link between the various indicators. Statistically, the assessment of articles' influence leads to very similar results if one counts only the raw number of citations or if one considers more subtle measures taking into account the influence of the citing articles or the quality of the citing journals.

Conclusion

The main result of our study is that the information conveyed by the different indicators is not significantly different whatever the way citation worth is taken into account. When we rank our benchmark articles according to the various indicators, rankings do not appear statistically different (Spearman's rank correlation coefficients are high whatever the rankings considered). More formally, PCA allows us to show that there are only minor differences between the three families of indicators. According to Occam's razor principle, the use of the raw number of citations as a measure of articles influence seems efficient – at least for articles published in the five considered journals.

References

Bergstrom, C.T., J.D. West and M.A. Wiseman (2008), The Eigenfactor™ metrics, Journal of Neuroscience, 28(45), p. 11433–11434.

Bollen, J., Rodriguez, A., and H. Van de Sompel, (2006). Journal Status, Scientometrics, 69(3).

Combes, P-P. and L. Linnemer (2010), "Inferring missing citations. A quantitative multi-criteria ranking of all journals in economics". GREQAM Working Article.

Diamond, A., (1986), "What is a Citation Worth?", The Journal of Human Resources , 21(2).

Dusansky, R. and C. Vernon, (1998), "Rankings of U.S. Economics Departments", Journal of Economic Perspectives, 12(1), p. 157-170.

Egghe, L. (2006), "Theory and practise of the g-index", Scientometrics, vol.69, n° 1, p. 133.

Hirsch, J., (2005), "An index to quantify an individual's scientific research output", PNAS Proceedings of the National Academy of Sciences of the United States of America, 102(46), p. 165-169.

Hu, X., Rousseau, R. and J. Chen, (2011), "On the definition of forward and backward citation generations," Journal of Informetrics, vol. 5(1).

Laband D., and M. Piette, (1994). – « The Relative Impacts of Economics Journals: 1970- 1990 », Journal of Economic Literature, 32 (2), p. 640-66.

Medoff M., (1996), "A Citation-Based Analysis of Economists and Economics Programs", The American Economist, 40(1), p. 46-59 .

Ritzberger, K. (2008), "A Ranking of Journals in Economics and Related," German Economic Review 9(4), p. 402–430.

Schubert, A., (2009), "Using the h-index for assessing single publications", Scientometrics, 78(3), 559-565.

Van Noorden, R., Maher, B., and R. Nuzzo, (2014), "The top 100 articles", Nature, 514(7524).

A glance on the status of Library and Information Science discipline in the world ranking systems of universities

Amir Reza Asnafi[1] and Maryam Pakdaman Naeini[2]

[1] *a_asnafi@sbu.ac.ir*
Shahid Beheshti University, Psychology and Education School, Information Science and Knowledge Department, Velenjak, Tehran (I.R.Iran)

[2] *m.pakdaman@gmail.com*
International Institute of Earthquake Engineering and SeismologyTehran, South Arghavan Avnenue (I.R.Iran)

Introduction

Nowadays, the status, position and function of higher education institutions are being studied, compared with University Ranking Systems. The rankings of higher education institutions, which are based on different educational and research indicators, provide a log of universities compared to each other. In other words, the rankings of universities are a set of systematic rules for judging the quality of higher education institutions. Higher education institutes use these rankings as a means of promoting their performance in order to demonstrate their educational and research excellence internationally. Universities that rank in the world ranking systems will enjoy popularity. The results of academic ranking systems play an important role in scientific policies and the allocation of national and regional resources. It also raises rating systems; discussions about the performance of universities will be drawn to the general public. The result of this is that the creators of the ranking system or specialists are weighed and how data is collected for each indicator.

The first nationwide ranking of universities began in 1983 by U.S. News & World Report in the United States, with the annual publication of the best universities in the country. Currently, many countries in the world have national systems to assess the performance and ranking of educational institutions. In the early years of the third millennium, the internationalization of higher education and the increased scientific mobility of students and scholars changed the ranking systems of universities from a national approach to an international approach. The first international academic ranking system in the world was launched in 2003 by the Shanghai Jiao Tong University in China, under the title Academic Ranking of Universities of the World. So far, we have seen the creation of various ranking systems internationally; each using different indicators and methods has introduced the top institutions of higher education.

In recent years, in Iran, the attention of universities and educational and research policy makers has become increasingly high on global rankings. Each university strives to position itself in ranking systems to showcase its scientific and research power in the international arena and benefit from it. Given the country's rapid pace of development, it is expected that more universities from Iran will have better rankings in global ranking tables. The purpose of this study was to identify the status of the library discipline in the world's prestigious ranking systems, which would determine the status of this field from the national and global levels in these systems.

Literature Review

Performance evaluation of ranking systems has been the subject of extensive research at home and abroad. The literature on this subject can be divided into five categories:

- Researches on the introduction, evaluation of indicators and methodology of rating systems have been carried out. Lukman, R., Krajnc, D., & Glavič, P. (2010). Noormohammadi, H.A. & Safari, F. (2014).
- Comparative research that studies the similarities, differences, and statistical relation between the results of different ranking systems. Aguillo, I. F., Bar-Ilan, J., Levene, M., & Ortega, J. L. (2010). Pavel, A. P. (2015).
- Researches that critically examine the indicators, the sources used to collect data, the methods of scoring and normalization, as well as statistical analyses of data in ranking systems. Billaut, J. C., Bouyssou, D., & Vincke, P. (2010). Cheng, S. K. (2011).
- Researches that have investigated the research indices used in international ranking systems using the science-based approach. Lin, C.-S., Huang, M.-H., & Chen, D.-Z. (2013). Lazaridis, T. (2010).
- Research that evaluates the performance of educational institutions of one or more countries based on international ranking systems. Khosrowjerdi, M., & Kashani, Z. S. (2013). [14]

Kumar, M.J. (2015).

Research method

In current study, eight ranking systems of universities and research institutes of the world were studied to observe the status of the field of Library and Information Science in Iran and the world. The study period was September and November. Thematic sections of the ranking systems were reviewed. These systems were: Shanghai, Times, Taiwan, ISC, QS.

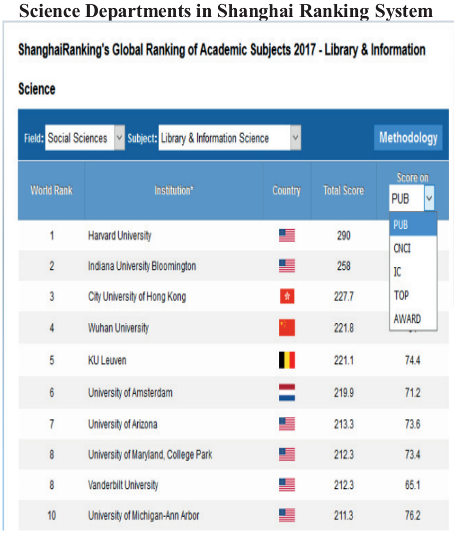
Findings

Study on Shanghai ranking system indicated that Iran's universities are not well-positioned in Library and Information Science in this system. But Top Asian universities in Shanghai rankings system include:

1. Hong Kong International Union
2. Wuhan China
3. Taiwan National University
4. Singapore Nangah
5. Younes South Korea
6. New South Wales Australia
Malaysia Malaysia
8. Fudan China
9. South Korea's Advanced Science and Technology Institute
10. Hong Kong Polytechnic

Hong Kong and Wuhan Chinese CT Universities have ranked third and fourth in terms of library and information science in the Shanghai ranking system and ranked first and second among Asian universities. Figure 1 indicates that Iranian universities had not any position in this ranking system.

Figure 1- Presence of Library and Information Science Departments in Shanghai Ranking System



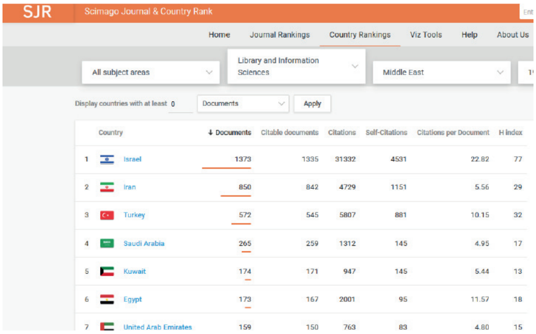
ShanghaiRanking's Global Ranking of Academic Subjects 2017 - Library & Information Science

Field: Social Sciences Subject: Library & Information Science Methodology

World Rank	Institution*	Country	Total Score	Score on PUB
1	Harvard University	🇺🇸	290	
2	Indiana University Bloomington	🇺🇸	258	
3	City University of Hong Kong	🇭🇰	227.7	
4	Wuhan University	🇨🇳	221.8	
5	KU Leuven	🇧🇪	221.1	74.4
6	University of Amsterdam	🇳🇱	219.9	71.2
7	University of Arizona	🇺🇸	213.3	73.6
8	University of Maryland, College Park	🇺🇸	212.3	73.4
8	Vanderbilt University	🇺🇸	212.3	65.1
10	University of Michigan-Ann Arbor	🇺🇸	211.3	76.2

The study on the library and information field in the Scimago ranking system indicated that Iran has the second rank in the Middle East. Among the countries of the world and from the total of 174 countries, Iran ranked 29th. However, this system has been emphasized on citation indicators such as Hirsh index and documents based on Iranian articles in the field of library and information science.

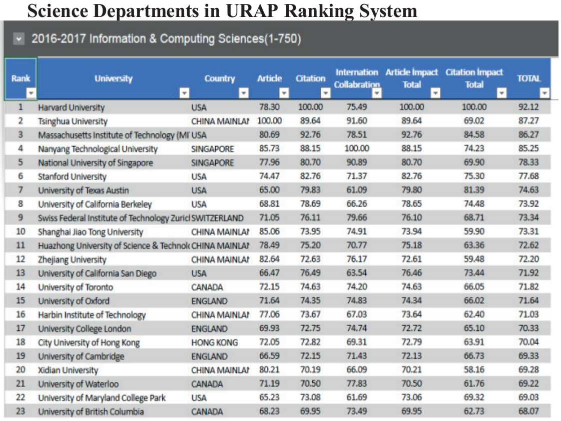
Figure 2- Position of Iran Universities in the Library and Information Science in Scimago Ranking System



	Country	↓ Documents	Citable documents	Citations	Self-Citations	Citations per Document	H index
1	Israel	1373	1335	31332	4531	22.82	77
2	Iran	850	842	4729	1151	5.56	29
3	Turkey	572	545	5807	881	10.15	32
4	Saudi Arabia	265	259	1312	145	4.95	17
5	Kuwait	174	171	947	145	5.44	13
6	Egypt	173	167	2001	95	11.57	18
7	United Arab Emirates	159	150	763	83	4.80	15

The study on the URAP system in Turkey revealed that there is only the field of Information & Computer Science that a look at the top Iranian universities in this field suggests that it might not have been the subject of librarianship. Because most industrial universities are ranked in this ranking.

Figure 3- Presence of Library and Information Science Departments in URAP Ranking System



2016-2017 Information & Computing Sciences(1-750)

Rank	University	Country	Article	Citation	Internation Collabration	Article Impact Total	Citation Impact Total	TOTAL
1	Harvard University	USA	78.30	100.00	75.49	100.00	100.00	92.12
2	Tsinghua University	CHINA MAINLAI	100.00	89.64	91.60	89.64	69.02	87.27
3	Massachusetts Institute of Technology (MI	USA	80.69	92.76	78.51	92.76	84.58	86.27
4	Nanyang Technological University	SINGAPORE	85.73	88.15	100.00	88.15	74.23	85.25
5	National University of Singapore	SINGAPORE	77.96	80.70	90.89	80.70	69.90	78.33
6	Stanford University	USA	74.47	82.76	71.37	82.76	75.30	77.68
7	University of Texas Austin	USA	65.00	79.83	61.09	79.80	81.39	74.63
8	University of California Berkeley	USA	68.81	78.69	66.26	78.65	74.48	73.92
9	Swiss Federal Institute of Technology Zuricl	SWITZERLAND	71.05	76.11	79.66	76.10	68.71	73.34
10	Shanghai Jiao Tong University	CHINA MAINLAI	85.06	73.95	74.91	73.94	59.90	73.31
11	Huazhong University of Science & Technolo	CHINA MAINLAI	78.49	75.20	70.77	75.18	63.36	72.62
12	Zhejiang University	CHINA MAINLAI	82.64	72.63	76.17	72.61	59.48	72.20
13	University of California San Diego	USA	66.47	76.49	63.54	76.46	73.44	71.92
14	University of Toronto	CANADA	72.15	74.63	74.20	74.63	66.05	71.82
15	University of Oxford	ENGLAND	71.64	74.35	74.83	74.34	66.02	71.64
16	Harbin Institute of Technology	CHINA MAINLAI	77.06	73.67	67.03	73.64	62.40	71.03
17	University College London	ENGLAND	69.93	72.75	74.74	72.72	65.10	70.33
18	City University of Hong Kong	HONG KONG	72.05	72.82	69.31	72.79	63.91	70.04
19	University of Cambridge	ENGLAND	66.59	72.15	71.43	72.13	66.73	69.33
20	Xidian University	CHINA MAINLAI	80.21	70.19	66.09	70.21	58.16	69.28
21	University of Waterloo	CANADA	71.19	70.50	77.83	70.50	61.76	69.22
22	University of Maryland College Park	USA	65.23	73.08	61.69	73.06	69.32	69.03
23	University of British Columbia	CANADA	68.23	69.95	73.49	69.95	62.73	68.07

Current study showed that in the Times ranking system, there is no area in the field of library and information science. In the citation database of the Islamic world, there was also no thematic option for librarianship and information science. Leiden, Taiwan, QS and Round also lacked such an option. In the U.S. ranking system, only the American

Library and Information Science Schools of ranked and the rest of the countries are not ranked.

Conclusions

Each ranking organization measures institutions in different ways, using different criteria, and different weightings of similar criteria. Rankings can take into account research quality and revenue, surveys of academics and employers, staff-student ratios, and statistics on demographics such as the number of international students. Universities use ranking systems to improve their performance to show their academic and research excellence internationally .Universities that rank in the world ranking systems will enjoy a high degree of popularity. Such an operation will attract widespread media attention and scientific communities around the world. Many students use rating results as a guide to the selection of higher education institutions. Therefore, it seems that Iranian Library and Information Departments have to find suitable strategies for their desirable status and ranking in global rankings, and have fundamental reviews in their missions and programs. Publishing articles in internationally recognized journals, indexing internal journals in valid citation databases, interacting with scholars of this field in different countries of the world, attracting foreign students, and wider communication with the community and industry, are some factors that Iranian Library and Information Science departments should pay close attention to them.

References

Aguillo, I. F., Bar-Ilan, J., Levene, M., & Ortega, J. L. (2010). Comparing university rankings. *Scientometrics*, *85*(1), 243–256.

Billaut, J. C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shanghai ranking? *Scientometrics*, *84*(1), 237–263.

Cheng, S. K. (2011). World university rankings: take with a large pinch of salt. *European Journal of Higher Education*, *1*(4), 369–381.

Khosrowjerdi, M., & Kashani, Z. S. (2013). Asian top universities in six world university ranking systems. *Webology*, *10*(2), 1–9.

Kumar, M.J. (2015). Global university rankings: What should India do?. *IETE Technical Review*, 32(2), 81-83.

Lin, C.-S., Huang, M.-H., & Chen, D.-Z. (2013). The influences of counting methods on university rankings based on paper count and citation count. *Journal of Informetrics*, *7*(3), 611–621.

Lazaridis, T. (2010). Ranking university departments using the mean h-index. *Scientometrics*, *82*(2), 211–216.

Lukman, R., Krajnc, D., & Glavič, P. (2010). University ranking using research, educational and environmental indicators. *Journal of Cleaner Production*, *18*(7), 619–628.

Noormohammadi, H.A. & Safari, F. (2014). Introducing world university rankings and their indicators. Journal of Science & Technology Policy, 2(3), 71-86 [Persian].

Pavel, A. P. (2015). Global University Rankings-A Comparative Analysis. *Procedia Economics & Finance*, *26*(15), 54–63.

Implementation of Altmetrics in Central Library of Islamic Azad University, Science and Research Branch of Tehran

Amir Reza Asnafi[1] and Firoozeh Dookhani[2]

[1] a_asnafi@sbu.ac.ir
Shahid Beheshti University, Psychology and Education School, Information Science and Knowledge Department, Velenjak, Tehran (I.R.Iran)

[2] fdokhani5@gmail.com
Information Science and Knowledge, Islamic Azad University, Poonak, Tehran, (I.R.Iran)

Introduction

In the use of social networks in libraries, the topic of resource observation and knowledge sharing in the organization are addressed by Altmetrics tools. Librarians play an important role in this field. Librarians should familiarize themselves with the concept of Altmetrics and its tools, and help researchers to make informed choices of scientific resources. The main purpose of this study is to study the use of virtual media resources by users of the Central Library of Islamic Azad University, Tehran Science and Research Branch, based on the implementation of Altmetrics. Zahedi (2015) in an attempt to assess the presence and use of Iranian international papers in Mendeley, 43 Iranian magazines indexed in the database have reviewed citation reports. The findings show that about half of the statistical population is stored in Mendeley, and in terms of the citation effect, the publications that are stored in Mendeley have a higher citation rating than those that were not used; in terms of the relationship By citation, there is a positive, but weak correlation between the citation and the preservation of papers in Mendeley between the publications being investigated. Harrison et al. (2017), in the context of the significant growth of libraries from social media, in aphenomenological and institutional theory study of the discovery of social media in six public and private university libraries in both the central and western states of the United States. The results indicated that there is a similar gap between these three issues, libraries spend more time creating local communication rather than posting content and information about the library environment. Konkiel & Scherer (2013) referred to the need to use Altmetrics tools in libraries' tanks. According to his plan, the viewing of pages, downloads, and references to resources is monitored through Altmetrics tools in the library.

Mohammadi, Thelwallm, Haustein & Larivière (2015) in their research found that little detailed information is known about who reads research articles and the contexts in which research articles are read. Using data about people who register in Mendeley as readers of articles, this article explores different types of users of Clinical Medicine, Engineering and Technology, Social Science, Physics, and Chemistry articles inside and outside academia. The majority of readers for all disciplines were PhD students, postgraduates, and postdocs but other types of academics were also represented. In addition, many Clinical Medicine articles were read by medical professionals. The highest correlations between citations and Mendeley readership counts were found for types of users who often authored academic articles, except for associate professors in some sub-disciplines. This suggests that Mendeley readership can reflect usage similar to traditional citation impact if the data are restricted to readers who are also authors without the delay of impact measured by citation counts. At the same time, Mendeley statistics can also reveal the hidden impact of some research articles, such as educational value for no author users inside academia or the impact of research articles on practice for readers outside academia.

Research Methods

In order to investigate the status of the use of scientific resources of virtual media by users of the target library, the method of scientometrics has been used with the Altmetrics approach. The needed data to create a target system for the production of Altmetrics donuts was collected from the Scopus database. With the help of software, a digital library was created. Then the desired articles (first, 100 randomly obtained cluster articles) were loaded into the csv file at this site. The codes for the Altimetrix button (Donut) are loaded for each article. The required metadata (author, title, DOI, etc.) was loaded into the system metatags. Using the Eddis software report analysis, the target audience's academic exchanges were investigated using virtual media. The use of virtual librarians to provide services after the installation of virtual media types and Altmetrics was investigated.

Findings

The present study intends to answer the following basic questions. How is the use of virtual media resources by users after installing Altmeterics donuts in the library? To answer this question, we will examine the use and sharing of scientific articles embedded in the system by all the beneficiaries of this system.

Table 1-The frequency of using social media to share articles

Social Media	Frequency	Percent
Mendeley	22	40
ResearchGate	9	16.36
Pinterest	6	10.9
Diigo	3	5.45
Google+	3	5.45
LinkedIn	3	5.45
Telegram	2	3.63
Facebook	2	3.63
CiteULike	2	3.63
Reddit	1	1.81
Twitter	1	1.81
Copy Link	1	1.81
Total	100	100

The largest social network used by users in the Mendeley system with 22 uses (40.0) and then the ResearchGate with 9 uses (16.36). The smallest tool used to subscribe to twitter articles and rankings is with 1 use (81/1). 57/85 percent of users use the desktop and 14.42 percent of the mobile to log into the system and use the articles. The articles that have been most commonly shared, Article No. (1) has been shared by the Department of Social Sciences 15 times. The most commonly used tool for this article is Mendeley (8 items). The article number (2) is from the Social Sciences Department, which has been distributed eight times and is the most widely used instrument of Mendeley (2 items). Subsequently, the article number (3) is related to the medical group that has been shared 8 times. And the Mendeley ResearchGate, Facebook and the citeulike (each case 2 times) have been used to the same extent. In this way, lower scores are related to the engineering sciences (4 items) and basic sciences (3 items). What follows from the review of this table is that Mendeley tool is further used by users to access articles. The largest shareholder is the Social Sciences Department.

Conclusion

The use of Altmetrics indicators in the organizational reserves of libraries, on the one hand, helps the library to collect specific information on the online user interaction of content generated by researchers within the organization, on the other hand, leads the faculty members to be informed And students are academically acquainted with the latest scientific achievements in their field of study and the position of organization and their position in the organization. Altmetrics informs the library of current topics and popular patterns of behaviour for its scientific activities. In the organizational reservoir, the Altmetrics allows libraries to better manage their collections and see if the library staffs have the necessary expertise to participate in scientific areas and exchange information - added value, monitoring It is based on the user experience and the library library's capabilities for future library digitization initiatives.

References

Harrison, A., Burress, R., Velasquez, S., & Schreiner, L. (2017). Social media use in academic libraries: A phenomenological study. *The Journal of Academic Librarianship*.

Konkiel, S., & Scherer, D. (2013). New opportunities for repositories in the age of altmetrics. *Bulletin of the Association for Information Science and Technology*, *39*(4), 22-26.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, 65(8), 1627-1638.

Palmer, Stuart. (2014). 'Characterizing University Library Use of Social Media: A Case Study of Twitter and Facebook from Australia', *The Journal of Academic Librarianship*, 40(6): 611-619.

Lazaridis, T. (2010). Ranking university departments using the mean h-index. *Scientometrics*, *82*(2), 211–216.

Zahedi, Zohreh (2015). *Study of the use of English-language publications published in Iranian International Magazines in Mendeley*. Presented at the first national conference on science, assessment and pathology of scientific output. Iran University of Isfahan. 7-8 March 2015.

INNOVATION INDICATORS BY MINING WEB DATA

Vilius Stanciauskas[1] and Lukas Pukelis[2]

[1] *vilius@ppmi.lt*
PPMI Public Policy and Management Institute UAB, Gedimino pr. 50, LT-01110 Vilnius (Lithuania)

[2] *lukas.pukelis@ppmi.lt*
PPMI Public Policy and Management Institute UAB, Gedimino pr. 50, LT-01110 Vilnius (Lithuania)

Introduction

Gathering comparable large-scale data on innovation activities and outcomes is of paramount importance to policymakers and research-funding bodies. For this reason, pan-European "Community Innovation Survey" (CIS) is carried out every-two years to collect data on enterprise innovation activities. Furthermore, the European Commission (EC) updates two of its other initiatives, namely the European Innovation Scoreboard and Regional Innovation Scoreboard on a yearly basis. However, due to the scope and complexity of the endeavour, significant time delays can occur when processing CIS data. For this reason, the EC has commissioned an exploratory study to find out to which extent can the CIS data be replicated using solely data from the web. This poster presents our research progress achieved in this area up to date.

Methodology

Our aim was to develop a methodology which could not only produce valid and precise indicators but would also lend itself to a high degree of automation thus enabling large sample sizes and rapid delivery of results. The approach employed a combination of web crawling and supervised machine learning to calculate innovation indicators from the data on company websites. Similar approach was outlined in ZEW Discussion Paper by Kinne and Axenbeck (2018).

We started by closely following the approach of CIS, explicitly targeting small and medium European enterprises. Overall, our sample consisted of around 30 000 companies. However, since some company data was outdated and some websites inaccessible, we collected data from around 50 200 companies. A special algorithm was used to 'crawl' company websites and collect text data, and a number of measures were adopted in the process to respect and correctly handle sensitive or personal data. We also limited the overall database size by installing an upper limit, which capped the scraping activities at 20 000 pages per website. With this approach we collected over six million distinct webpages for the companies in our sample.

During the next stage, we selected the website contents of 1 000 randomly selected companies and each page in the company website by hand. By exploring multiple coding options and frameworks, we settled on a final list of indicators. Prior to employing machine learning algorithms, we pre-processed and vectorised the texts from company websites (following doc2idx approach), see Table 1.

Table 1. Illustration of doc2idx approach

Normal text	"A fool thinks himself to be wise, but a wise man knows himself to be a fool."
Tokens	'fool', 'thinks', 'be', 'wise', 'wise', 'man', 'knows', 'be', 'fool'
Dictionary	'fool': 1, 'thinks': 2, 'be': 3, 'wise': 4, 'man': 5, 'knows': 6
Vector	1,2,3,4,5,6,3,1

In developing a dictionary, we filtered the extreme values out. Then we selected 50 000 of the most common items and trained a deep-learning model to analyse each text from the company website to enable it to predict whether the text belongs to one of the grid categories. We employed a convolutional neural network for text classification similar to the one developed by Kim (2014). Results for each indicator were aggregated to a company level. If at least one innovation or commercialisation text was identified on a company website, it was considered that the company has produced an innovation or undertaken commercialisation activities.

Conceptual framework

Data available on company websites

Although our initial aim was to replicate the indicator scheme of the Community Innovation Survey, the manual analysis of the company websites proved that this will not be possible. This was primarily due to a weak tendency of companies to report data on various elements that are relevant for CIS indicator framework. Nevertheless, we identified other areas where relevant information was reported by companies.

We noticed a strong tendency of companies to report good news which relate to the awards and recognitions that they have won or socially/environmentally responsible initiatives in which they have participated. We also observed that companies tend to report IP-related information. In addition to the data on patent or trademark applications filed or granted, companies also present announcements on patent or trademark licensing agreements in where they either license their IP to other companies or agree to manufacture items on behalf of other companies. Companies also tend to report the information on the formation of business partnerships as well as company mergers and acquisitions. Finally, we noticed that companies tend to make announcements if they have attracted significant investors, raised significant amounts of capital or attracted public funding.

Table 2. Simplified indicator framework for internet company data (indicators under development in italics)

Broad category	**Medium category**
Innovations	Product innovation
	Product innovation type
	Product innovation name/description
	Process innovation count/type
Innovation activities	New IP developed by company
	Prototypes
	Utilising IP from outside
	Partnerships
	Funding attracted

Preliminary aggregated results

Based on the research conducted so far, we produced preliminary calculations for various indicators, including the share of innovative enterprises, see Table 3. Since our indicators were developed on the basis of the information that was available on company websites, they might not completely align with the existing ones. Nevertheless, we attempted to make this comparison.

Table 3. Benchmarking our measurement to CIS indicators

CIS Indicator[1]	Web indicator
37.4%	20.13%

Source: Our calculations, CIS data (2016)

Our measure for the share of companies that have introduced product or other types of innovations is lower than the most comparable CIS indicator[2]. This is due to the fact that our analysis relied on the first round of data collection which detected only explicitly announced innovations. To detect new additions to the product lines which are not explicitly labelled as innovations, we will process the data from the second round of data collection, compare the newer data to the existing records, identify new text on company websites and process it in a similar manner.

Below we present the preliminary results broken down by NACE sector of economic activity. The results are consistent with CIS data.

Table 4. Score breakdown by NACE sector, based on data extracted from company websites

NACE (rev.2) code	**Share of innovative enterprises**
J (Information and communication)	38,3%
M (Professional, scientific and technical activities)	26,9%
G (Wholesale and retail trade; repair of motor vehicles and motorcycles)	21,4%
C (Manufacturing)	20,6%
L (Real estate activities)	17,5%
N (Employment activities)	17,5%
H (Transporting and storage)	17,2%
B (Mining and quarrying)	15,9%
D (Electricity, gas, steam and air conditioning supply)	14,2%
F (Construction)	13,3%
E (Sewerage)	12,7%
I (Accommodation and food service activities)	7,6%

Next steps

We will expand the current data collection and analysis effort over the next two years by:

•Increasing the linguistic coverage to include all EU languages;

•Increase the sample size to ~ 200 000 enterprises;

•Expand the indicator framework to include more rich and detailed indicators mined from the hypertext-data (links between the enterprises, network embeddedness, proper product names, classifications, etc.).

References

Eurostat (2017) Community Innovation Survey < https://ec.europa.eu/eurostat/web/productseurostat-news/-/DDN-20170124-2>

Kim, Y., 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kinne, J. and Axenbeck, J., 2018. Web mining of firm websites: A framework for web scraping and a pilot study for Germany. *ZEW-Centre for European Economic Research Discussion Paper, (18-033).*

[1] (“Product and/or process innovative enterprises, regardless of organisational or marketing innovation (including enterprises with abandoned/suspended innovation activities)”

[2] Ibid

Does Environmental Economics lead to patentable research?

Xiaojun Hu[1], Ronald Rousseau[2,3] and Sandra Rousseau[4]

[1]*xjhu@zju.edu.cn*
Medical Information Center, and Department of Neurology of Affiliated Hospital 2, Zhejiang University School of Medicine, Hangzhou 310058 (China)

[2]*ronald.rousseau@kuleuven.be*
KU Leuven, MSI, Facultair Onderzoekscentrum ECOOM, Naamsestraat 61, 3000 Leuven, (Belgium)

[3]*ronald.rousseau@uantwerpen.be*
University of Antwerp, Faculty of Social Sciences, Middelheimlaan 1, 2020 Antwerpen, (Belgium)

[4] sandra.rousseau@kuleuven.be
KU Leuven, CEDON, B-1000 Brussel (Belgium)

Introduction

Patents are generally related to fields such as material sciences, mechanics, computer technology, biotechnology, pharmacy and other 'hard' fields of science. This leads to the question: what is the contribution of social sciences and humanities to the intellectual property system as covered by patents. In this contribution we focus on a small subfield, namely environmental economics (Turner et al., 1994) and in particular to the role played by journals as sources of non-patent references. Our work can be seen as a partial validation study of academic research, in particular of environmental economics.

This work is in part inspired by a similar publication by Halevi & Moed (2012) which focused on the use of journals from the field of Library Science in patents. These authors used TotalPatentTM as their data source. Their most important conclusion was that it was Library Science research that informed and inspired the development of information retrieval solutions, sometimes years before the technology was available to translate these ideas into technical devices or computer algorithms.

Methods: journals and used database

In this study we operationalize the field of environmental economics by a set of leading journals. As there is no Subject Category called Environmental Economics in the WoS/JCR, we take the journals used by Sandra Rousseau in (Rousseau, 2008). These journals are shown in the first column of Table 1. All journals are multidisciplinary in the sense that they belong to more than one WoS category.

In this investigation we restrict patents to those included in the USPTO database, the database of the United States Patent and Trademark Office. Data were collected in December 2018. We recall that, contrary to articles, patents often have the same title and largely the same content. They may differ only in the claim.

Results

Results on article level

The oldest cited article dates from 1929. It was cited in 2009 and published in the *American Journal of Agricultural Economics* (Hardie & Strand, 1929). The second oldest one dates from 1936 and was published in the *Journal of Land & Public Utility Economics* (Burton, 1936). The notion of "highly-cited" environmental economics articles in patents turned out to be very relative, especially as the few cases of (relative) high numbers of citations in patents result from re-citations. Ahlheim & Schneider (2002) was cited 26 times, but by the same group of inventors.

Results on journal level

This set of journals dealing with environmental economics received a total of 195 citations to 85 different articles. Even more than for journal article citations, non-patent citations often have re-citations, i.e. the same author citing the same article (White, 2000). An extreme case is the *Australian Journal of Agricultural and Resource Economics* which has only two cited articles: one which is cited nine times by an inventor and another which is cited three times by (another) inventor. Table 1 shows the number of citations for each journal and the number of different cited articles. Two journals were never cited in the USPTO, namely *Environment & Development Economics* and *Natural Resources Journal*. They are omitted from Table 1.

Conclusion

In this feasibility study, the impact of academic research from social sciences and humanities on

Table 1. Number of citations for each journal and number of different cited articles.

Journals	Number of citations	Different cited articles
American Journal of Agricultural Economics	33	20
Australian Journal of Agricultural and Resource Economics	12	2
Ecological Economics	20	11
Energy Journal	39	19
Environmental & Resource Economics	30	5
Journal of Agricultural and Resource Economics	19	8
Journal of Environmental Economics and Management	12	8
Land Economics	27	12
Resource and Energy Economics	3	3
TOTAL	195	85

technological innovation is explored through a study of citations patterns of journal articles in patents. Specifically we focused on citations of journals from the field of environmental economics in patents included in an American patent database (USPTO). Three decades of patents have led to a small set of journal articles (85) that are being cited from the field of environmental economics. While this route of measuring how academic research is validated through its role in stimulating technological progress may be rather limited (based on this first exploration), it may still point to a valuable and interesting topic for further research. A more detailed version of this investigation can be found in (Hu et al., 2019).

Questions for further research

-) What is the relation (on journal level) between the journal's synchronous impact factor (different periods could be considered) and the (relative) number of patent citations. We expect though that a correlation would be low as the indicators measure something different.

-) Would making a distinction between citations by the applicant and by the examiner lead to different results?

-) For a complete investigation a better proxy for the field of environmental economics is necessary. If it would be possible to describe the field on an article basis, instead on a journal basis, that would certainly lead to a better result.

-) What are the major factors affecting the pattern of citation links between an article and a patent? Does geographical distance influence the citation behavior of inventors and examiners?

-) Are patents citing contributions in environmental economics related to the economy in general, or do they contribute to a "green economy"?

-) Can journal citations in patents help to measure the impact of research performed at academic institutions?

-) Clearly, for a thorough investigation a larger database, not just the USPTO, is necessary.

-) Finally, whatever the topic of a patent investigation, it could be discussed in view of the struggle for scientific and technological leadership between the USA, Europe and China. Indeed, in 2017 China moved already into the second position (country level) as a source of international patent applications filed via WIPO (World Intellectual Property Organization).

References

Ahlheim, M., & Schneider, F. (2002). Allowing for household preferences in emission trading - A Contribution to the Climate Policy Debate. Environmental and Resource Economics, 21(4), 317-342.

Burton, J.E. (1936). Guaranteed certificated mortgages in New York. The Journal of Land & Public Utility Economics, 12(2), 191-193.

Halevi, G., & Moed, H.F. (2012). Patenting library science research assets. Research Trends, 27, 11-14.

Hardie, I. & Strand, I. (1929). Measurement of Economic Benefits for Potential Public Goods. American Journal of Agricultural Economics, 61(2), 313-317.

Hu, XJ., Rousseau, R. & Rousseau, S. (2019). Does Environmental Economics lead to patentable research? https://arxiv.org/abs/1905.02875.

Rousseau, S. (2008). Journal evaluation by environmental and resource economists: A survey. Scientometrics, 77(2), 223-233.

Turner, R.K., Pearce, D. & Bateman, I. (1994). Environmental Economics. An elementary introduction. New York: Harvester Wheatsheaf.

White, H.D. (2000). Toward ego-centered citation analysis. In: B. Cronin & H. B. Atkins, (Eds.), The Web of Knowledge. A Festschrift in Honor of Eugene Garfield (pp. 475-496). Medford (NJ): Information Today.

Williams, G.S., Raper, K.C., DeVuyst, E.A., Peel, D.S., & McKinney, D. (2012). Determinants of Price Differentials in Oklahoma Value-Added Feeder Cattle Auctions. Journal of Agricultural and Resource Economics, 37(1), 114-127.

A New Perspective of Evaluating Journals Impact: Altmetrics and Citation Indicators

Rongying Zhao[1], Xu Wang[2], Zhaoyang Zhang[3], Yongkang Qi[4], Ruru Chang[5]

[1]*zhaorongying@126.com*
[2] *xuwang@whu.edu.cn*
[3]*windboy727@vip.qq.com*
[4]*qiyongkang@whu.edu.cn*
[5]*Changrr@whu.edu.cn*

School of Information Management, Wuhan University, Wuhan, Hubei Province (P.R.China)

Introduction

Journals evaluation has a long history, and since the Journal Impact Factor was put forward (Garfield 1970, 1955), the researches on literature and scientific development trends from the perspective of citation have provided quantitative references for journals evaluation. However, under the new media environment, the traditional journals evaluation methods based on citation indicators have been unable to fully evaluate the impacts of academic journals. Just at the time, altmetrics (alternative metrics) came into being (Priem et al. 2010), and has become an important metrics to evaluate articles or journals impact.

In order to evaluate the impact of journals from a multi-indicators fusion perspective, this study attempts to further explore the following questions:

- Is it feasible to evaluate journals impact by using altmetrics indicators and combining traditional citation indicators?
- Are there any relations among the evaluation indicators of academic impact, societal impact and the evaluation results of the two dimensions?
- How to use the two-dimensional rectangular coordinate system to evaluate journals impact?

Data and processing

As shown in figure 1, we use the way of multi-indicator fusion to construct the impact evaluation framework of journals. Taking 64 international LIS journals as an example, we divide the journals impact type into academic impact and societal impact. In terms of the evaluation of academic impact of journals, first, we select 8 indicators to evaluate the academic impact of journals from JCR (2017). Second, with Altmetric.com as the data source (Type of output: Articles, Publication Date: 2015-01-01 to 2016-12-31, Retrieval time: 2018-11-11), ten indicators are selected to evaluate the societal impact of journals. Then, the indicators data is normalized by using the formula: $y_{ij}=x_{ij}/max(x_j)$, and the correlation/reliability analysis, validity analysis, factor analysis are used to evaluate the academic impact and societal impact of journals. After that, we analyse the correlation of the evaluation scores of two dimensions of journal impact. Then, we map the evaluation scores of journals' academic impact and societal impact to the two-dimensional rectangular coordinate system to evaluate and classify the sample journals.

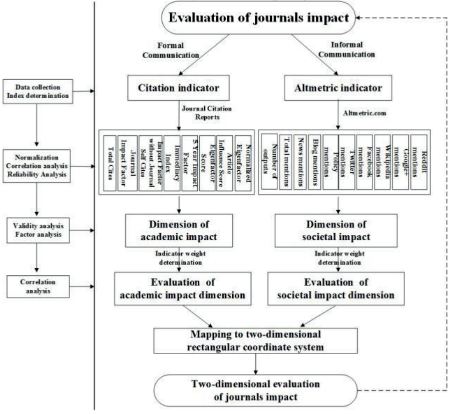


Figure 1. Framework

Results and discussion

Evaluation of academic impact

Table 1. Correlation of academic impact indicators provided by JCR

	TC	JIF	IFJSC	5YIF	IX	ES	AIS	NE
TC	1							
JIF	0.798**	1						
IFJSC	0.784**	0.978**	1					
5YIF	0.821**	0.967**	0.956**	1				
IX	0.624**	0.778**	0.747**	0.724**	1			
ES	0.877**	0.862**	0.869**	0.862**	0.683**	1		
AIS	0.808**	0.909**	0.917**	0.943**	0.717**	0.879**	1	
NE	0.939**	0.875**	0.878**	0.888**	0.687**	0.963**	0.903**	1

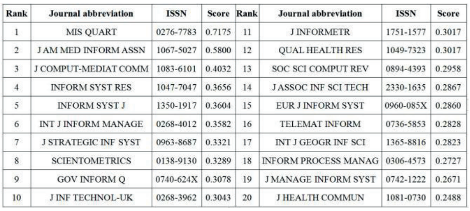
Throngh correlation analysis, reliability analysis, validity analysis, factor analysis, we obtain evaluation model formula of the academic impac:

F$_1$= 0.161*Total Cites+0.356*Journal Impact Factor+0.320*Impact Factor without Journal Self Cites+0.222*5-Year Impact Factor+0.385*Immediacy Index-0.176*Eigenfacetor Score+0.101*Article Influence Score-0.196*Normalized Eigenfactor；

F$_2$=0.371*Total Cites-0.150*Journal Impact Factor-0.110*Impact Factor without Journal Self Cites-0.006*5-Year Impact Factor-0.227*Immediacy Index+0.386*Eigenfacetor Score+0.110*Article Influence Score+0.405*Normalized Eigenfactor.

$$F_A = \frac{44.958\%}{88.572\%} * F_1 + \frac{43.614\%}{88.572\%} * F_2$$

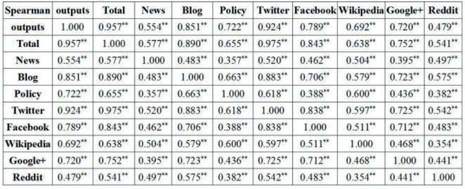
Table 2. Evaluation results of academic impact for international LIS journals (Top20)



Rank	Journal abbreviation	ISSN	Score	Rank	Journal abbreviation	ISSN	Score
1	MIS QUART	0276-7783	0.7175	11	J INFORMETR	1751-1577	0.3017
2	J AM MED INFORM ASSN	1067-5027	0.5800	12	QUAL HEALTH RES	1049-7323	0.3017
3	J COMPUT-MEDIAT COMM	1083-6101	0.4032	13	SOC SCI COMPUT REV	0894-4393	0.2958
4	INFORM SYST RES	1047-7047	0.3656	14	J ASSOC INF SCI TECH	2330-1635	0.2867
5	INFORM SYST J	1350-1917	0.3604	15	EUR J INFORM SYST	0960-085X	0.2860
6	INT J INFORM MANAGE	0268-4012	0.3582	16	TELEMAT INFORM	0736-5853	0.2828
7	J STRATEGIC INF SYST	0963-8687	0.3321	17	INT J GEOGR INF SCI	1365-8816	0.2823
8	SCIENTOMETRICS	0138-9130	0.3289	18	INFORM PROCESS MANAG	0306-4573	0.2727
9	GOV INFORM Q	0740-624X	0.3078	19	J MANAGE INFORM SYST	0742-1222	0.2671
10	J INF TECHNOL-UK	0268-3962	0.3043	20	J HEALTH COMMUN	1081-0730	0.2488

Evaluation of societal impact

Table 3. Correlation of societal impact indicators provided by Altmetrics.com



Spearman	outputs	Total	News	Blog	Policy	Twitter	Facebook	Wikipedia	Google+	Reddit
outputs	1.000	0.957**	0.554**	0.851**	0.722**	0.924**	0.789**	0.692**	0.720**	0.479**
Total	0.957**	1.000	0.577**	0.890**	0.655**	0.975**	0.843**	0.638**	0.752**	0.541**
News	0.554**	0.577**	1.000	0.483**	0.357**	0.520**	0.462**	0.504**	0.395**	0.497**
Blog	0.851**	0.890**	0.483**	1.000	0.663**	0.883**	0.706**	0.579**	0.723**	0.575**
Policy	0.722**	0.655**	0.357**	0.663**	1.000	0.618**	0.388**	0.600**	0.436**	0.382**
Twitter	0.924**	0.975**	0.520**	0.883**	0.618**	1.000	0.838**	0.597**	0.725**	0.542**
Facebook	0.789**	0.843**	0.462**	0.706**	0.388**	0.838**	1.000	0.511**	0.712**	0.483**
Wikipedia	0.692**	0.638**	0.504**	0.579**	0.600**	0.597**	0.511**	1.000	0.468**	0.354**
Google+	0.720**	0.752**	0.395**	0.723**	0.436**	0.725**	0.712**	0.468**	1.000	0.441**
Reddit	0.479**	0.541**	0.497**	0.575**	0.382**	0.542**	0.483**	0.354**	0.441**	1.000

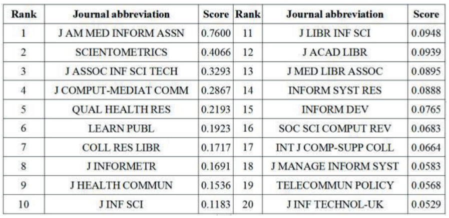
Through correlation analysis, reliability analysis, validity analysis, factor analysis, we obtain evaluation model formula of the societal impact:

F_3=-0.053*Number of mentioned outputs+0.111*Total mentions+0.299*News mentions+0.043*Blog mentions-0.242*Policy mentions+0.099*Twittermentions+0.229*Facebook mentions-0.117*Wikipedia mentions+0.211 *Google+mentions+0.254*Reddit mentions;

F_4=0.288*Number of mentioned outputs+0.094*Total mentions-0.208*News mentions+0.156*Blog mentions+0.438*Policy mentions+0.108*Twitter mentions-0.085*Facebook mentions +0.32*Wikipedia mentions-0.074 *Google+mentions-0.16*Reddit mentions

$$F_S = \frac{48.417\%}{82.055\%} * F_3 + \frac{33.638\%}{82.055\%} * F_4$$

Table 4. Evaluation results of societal impact for international LIS journals (Top20)



Rank	Journal abbreviation	Score	Rank	Journal abbreviation	Score
1	J AM MED INFORM ASSN	0.7600	11	J LIBR INF SCI	0.0948
2	SCIENTOMETRICS	0.4066	12	J ACAD LIBR	0.0939
3	J ASSOC INF SCI TECH	0.3293	13	J MED LIBR ASSOC	0.0895
4	J COMPUT-MEDIAT COMM	0.2867	14	INFORM SYST RES	0.0888
5	QUAL HEALTH RES	0.2193	15	INFORM DEV	0.0765
6	LEARN PUBL	0.1923	16	SOC SCI COMPUT REV	0.0683
7	COLL RES LIBR	0.1717	17	INT J COMP-SUPP COLL	0.0664
8	J INFORMETR	0.1691	18	J MANAGE INFORM SYST	0.0583
9	J HEALTH COMMUN	0.1536	19	TELECOMMUN POLICY	0.0568
10	J INF SCI	0.1183	20	J INF TECHNOL-UK	0.0529

Result analysis of impact evaluation

We conduct Spearman correlation analysis on the score of F_A, F_S, and map them into the two-dimensional rectangular coordinate system for comprehensive evaluation.

Table 5 Correlations of the scores

Spearman	F_A	F_S
F_A	1.000	0.566^{**}
F_S	0.566^{**}	1.000

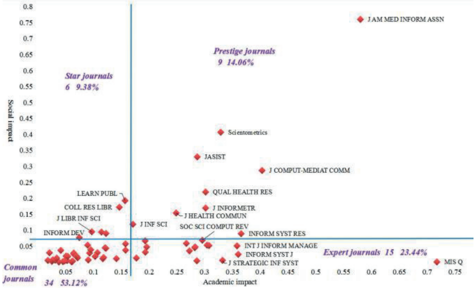


Figure 2. 2-dimensional evaluation of LIS journals impact.

Conclusion

We get the following conclusion: 1) We get the 2-dimensional evaluation results and according to LIS journals impacts, we divide the roles into four categories: "Prestige journals", "Star journals", "Common journals" and "Expert journals". 2) It is found that 8 traditional citation indicators based on JCR, present strong positive correlations (range from 0.624 to 0.978), and the indicators are highly consistent overall and internally. 3) There are moderate or high positive correlations among 10 Altmetrics (range from 0.354 to 0.957) based on Altmetrics Explorer, and there are also significant consistency among the indicators as a whole and internally. 4) The correlation coefficient between F_A and F_S is 0.566 with a moderate positive correlation. It indicates that the evaluation of LIS journals' societal impact based on Altmetrics indicators has a good supplement to the evaluation of LIS journals' academic impact based on citation.

Acknowledgments

This research is funded by the National Social Science Found Major Project of China (18ZDA325), the National Social Science Found Project of China (16BTQ055) and Humanities & Social Science Independent Research Projects of Wuhan University.

References

Garfield, E. (1955). Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, 122(3159), 108-111.

Garfield, E. (1970). Citation indexing for studying science. *Nature*, 227(5260), 870.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). altmetrics: a manifesto.

Topic Evolution and Emerging Topic Analysis Based on Open Source Softwares

Xiang Shen[1] and Li Wang[2]

[1] *shenx@mail.las.ac.cn*

[2] *wangli@mail.las.ac.cn*

National Science Library, Chinese Academy of Sciences, Beijing 100190 (China)
Department of Library, Information and Archives Management, University of Chinese Academy of Sciences, Beijing 100190 (China)

Introduction

On the purpose of scientific discovery and technology foresight, mapping the evolution of topics and detecting emerging topics in science and technology, has been of interest to governments, companies, and individual scientists for many years. VOSviewer is an open source computer program for creating, visualizing, and exploring bibliometric maps of science (Van Eck & Waltman, 2010). It has been widely used for analysing and mapping kinds of bibliometric network data, such as documents co-citation analysis, researchers collaboration analysis and keywords co-occurrence analysis. However, few researches are mentioned on utilizing the text mining (Van Eck & Waltman, 2011) and thesaurus functionality of VOSviewer for analysis. Furthermore, VOSviewer could only visualize some general and vague terms about topic evolution and could not reveal emerging topic vividly.

Herein, through natural language processing techniques, information extraction techniques and thesaurus functionality provided by VOSviewer and some general functionalities provided by Microsoft Office softwares, we develop a flexible approach for topic evolution and emerging topic analysis on all kinds of data-tagged text context. It can be applied for researchers in the bibliometric and scientometric community, while the computer programming skills are not necessarily required.

Method and application

The general approach we have explored to carry out topic evolution and emerging topic analysis to a specific research domain is showed in the following subsections, and the perovskite solar cells field is taken for an example.

Data collecting

Data collecting was conducted within December 23, 2018 using Scopus database. The central theme in this study was research articles containing 'perovskite solar cells (PSCs)' in the title, abstract and keywords. The query string used for the search was: TITLE-ABS-KEY ("perovskite solar cells") AND DOCTYPE (ar OR re) AND PUBYEAR < 2019. This query string resulted in 7125 documents.

Data slicing

In order to construct time series text data for trend analysis, the text data should be sliced to several subperiods by an appropriate time period. The time period for slicing can be selected flexibly according to the amount of literature and the purpose of research. In the example of PSCs field, the articles published before 2013 are sliced as a subperiod since the amount are small, and the articles published between 2014 and 2018 are sliced each year as a subperiod.

Thesaurus construction

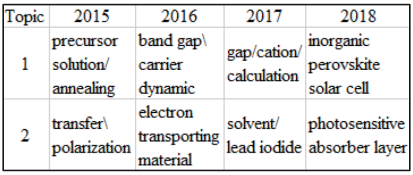
A pre-text mining procedure should be performed by VOSviewer to establish a thesaurus file for merging terms. Considering the distribution of keywords in the whole text data is quite different from each time-sliced text data, the pre-text mining is performed on each subperiod text data, and the thesaurus file is used and updated each time till the final version.

Text mining and map files generation

With the final version of thesaurus file, the series of subperiod text data are processed by natural language processing, information extraction, co-occurrence term analysis, clustering and visualization techniques based on VOSviewer, and terms are extracted from the title and abstract. A series of map files are generated and saved for further analysis.

Map files treatment and analysis

Each cluster in map file of each subperiod text data could be considered as a topic. By analysing the total links strength of term in each cluster, labelling the cluster by the most dominant and meaningful terms as the topic. The top 2 cluster topics from 2015 to 2018 are shown in Figure 1.



Topic	2015	2016	2017	2018
1	precursor solution/ annealing	band gap\ carrier dynamic	gap/cation/ calculation	inorganic perovskite solar cell
2	transfer\ polarization	electron transporting material	solvent/ lead iodide	photosensitive absorber layer

Figure 1. The topic evolution of PSCs field

To identify the emerging terms of each subperiod, the 'LOOKUP' function of Excel (Microsoft Office, 2019) is applied. By searching and comparing the column of 'label' in the map files of the previous subperiod, and returning the information in the column of 'weight<Occurrences>', the emerging terms are detected. The '#N/A' means the label or the term doesn't exist in the subperiod text data compared. As examples are shown in Table 1, 'halide perovskite' and 'module' are one of the emerging terms in 2016 and 2015 respectively.

Table 1. Detecting the emerging terms by 'LOOKUP' function of Excel

label	Weight<Occurrences>					
	2018	2017	2016	2015	2014	before 2013
halide perovskite	173	135	107	#N/A	#N/A	#N/A
module	54	55	21	11	#N/A	#N/A

Visualization

The analysis results can be visualized by the Microsoft Office PowerPoint and Excel or another open-source or free available softwares. Figure 2 shows the Top 10 emerging terms in PSCs field in descending order of occurrence from 2015 to 2018, which are picked by researcher and visualized by PowerPoint.

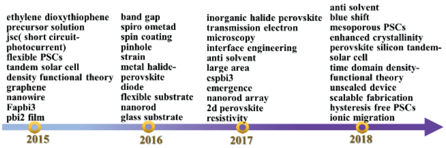


Figure 2. Top 10 emerging terms in PSCs field

In addition, the variations of term occurrences can also be accounted and visualized by Excel, and the future trend of terms can be forecasted by analysing the increasing or decreasing trend of term occurrences. In PSCs research domain, these topics involved 'module', 'photodetector' and 'flexible perovskite solar cells' may be paid more attentions by researchers in the near future, while the topic involved 'carbon nanotube' may be paid less attentions, as shown in Figure 3.

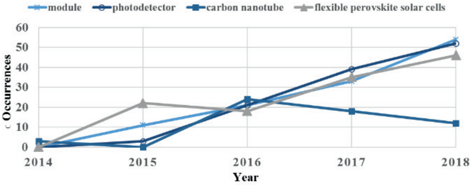


Figure 3. Variations of term occurrences in PSCs field

Verification and discussion

The emerging terms detecting by our method is verified in the way of comparing to the searching result in Scopus database by combining the 'perovskite solar cells' and the emerging term. Generally, few or only a small percentage of articles have been published before the year we detect the term through this method. Taking the emerging term 'flexible perovskite solar cells' in 2015 for example, there was only one paper published in 2014, and the number of papers published containing 'flexible perovskite solar cells' has rose up to 21 in 2015.

For a better analysis performance for emerging term and topic evolution through this approach, these following pathways can be considered: (1) set an appropriate threshold of minimum occurrences of a term to be analysed, and the smaller the better. (2) construct a detailed thesaurus file, and the more detailed the better. It would take some time to merge the synonyms and abbreviations of the same term, and it is very difficult to clear up the relationship of the upper and lower levels of a term, even for the specific scientist in the research domain. Certainly, with the development of natural language processing and text mining techniques, the analysis performance for topic evolution and emerging term would be improved.

Conclusion

In this paper, a flexible approach for topic evolution and emerging topic analysis and visualization based on VOSviewer and Microsoft Office softwares has been presented. The example of perovskite solar cells research domain has been given of application. Through natural language processing, information extraction, thesaurus, co-occurrence term analysis and clustering techniques provided by the open source bibliometric software VOSviewer, this approach can be used to analyse a large amounts of text data, not only scientific texts, but also non-scientific texts (e.g., project files, policy files, newspaper articles). It is useful and effective for those interdisciplinary researchers, especially without computer programming skills.

Acknowledgments

We would like to thank all researchers who contribute to open source tools, especially Nees Jan van Eck and Ludo Waltman considering this work.

References

Van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84, 523–538.

Van Eck, N.J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. ISSI Newsletter, 7, 50-54.

Microsoft Office. LOOKUP function. Retrieved January 9, 2019 from https://support.office.com/en-us/article/LOOKUP-function-446D94AF-663B-451D-8251-369D5E3864CB.

Library and Information Science papers as Topics on Twitter: A network approach to measuring public attention

Robin Haunschild [1], Loet Leydesdorff [2], and Lutz Bornmann[3]

[1] *r.haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)

[2]*loet@leydesdorff.net*
Amsterdam School of Communication Science (ASCoR), University of Amsterdam, PB 15793, 1001 NG Amsterdam (The Netherlands)

[3]*bornmann@gv.mpg.de*
Administrative Headquarters of the Max Planck Society, Division for Science and Innovation Studies, Hofgartenstr. 8, 80539 Munich (Germany)

Introduction

Can the impact of research on society (beyond science) or the attention research receives from society be measured using Twitter data? In the UK Research Excellence Framework (REF) the case-study approach was used for societal impact measurements and altmetrics have been proposed to measure impact or attention quantitatively (Bornmann, Haunschild, & Adams, 2019). The use of altmetrics data for networks seems to be a promising opportunity to measure public discussions. Hellsten and Leydesdorff (in press) analyzed Twitter data and mapped the co-occurrences of hashtags (as representation of topics) and usernames (as addressed audiences).

The resulting networks enable us to show the relationships between three different types of nodes, i.e. authors, addressees, and topics. The maps demonstrate how audiences and topics are co-addressed in science-related communications. Our method operationalizes Wouters, Zahedi, and Costas (2018) proposal to use social media data in research evaluation.

Recently, Haunschild, Leydesdorff, Bornmann, Hellsten, and Marx (2019) added a network-oriented approach to using Twitter data in research evaluation. The approach can be used to measure and map public discussions about fields or topics. In this study, we apply our method to all papers published in the Web of Science (WoS, Clarivate Analytics) Subject Category “Information Science & Library Science” (LIS). Which are the publicly and scholarly discussed topics?

Methods and datasets

We use WoS data from the in-house database of the Max Planck Society (MPG) licensed by Clarivate Analytics. In this database, 86,657 papers which were assigned to the WoS subject category “Information Science & Library Science” and published during the period 2010-2017. A DOI could be obtained for 33,335 (38.5%) of these papers, from our WoS database or from CrossRef (see also Bornmann, Haunschild, & Marx, 2016).

We match the papers via the DOIs with altmetrics data from a locally maintained database at the Max Planck Institute for Solid State Research. The following information was thereafter appended to the papers: (1) links to the tweets in which these papers were mentioned, (2) the numbers of tweets in which the respective papers were mentioned, and (3) the numbers of mentions in news outlets of this same paper. Among the papers with DOI, 20.0% (n=6,676) were mentioned in 48,966 tweets; 14.2% (n=4,726) of the papers were mentioned in at least two tweets. However, only 0.7% (n=232) were also mentioned in news outlets.

We downloaded the 48,966 tweets mentioning a LIS paper and further processed these as described in Haunschild, et al. (2019). Altmetrics began to monitor Twitter in 2011. The period 2011-2017 is analyzed in this study.

How and to which extent do author keywords in publications and hashtags in tweets match? We use a cosine-normalized term-term co-occurrence matrix for the analysis of the most frequently occurring author keywords and hashtags using a dedicated routine available at https://www.leydesdorff.net/software/twitter.

Four sets of author keywords are distinguished: (1) author keywords of all LIS papers, (2) author keywords of not-tweeted papers, (3) author keywords of papers tweeted at least twice, and (4) author keywords of papers tweeted at least twice and additionally mentioned in news outlets at least once. The latter set enables us to identify public discussions which are also triggered by the news sector. In total 721 different author keywords occurred in LIS papers tweeted at least twice and mentioned in news outlets at least once; 77 of these author keywords occurred at least twice in the news. We use these top-77 author keywords in order to compare networks of the same size.

The resulting files (containing cosine-normalized distributions of terms) were laid-out using the algorithm of Kamada and Kawai (Kamada & Kawai, 1989) in Pajek and then exported to VOSviewer for visualizations using the community-searching algorithm in VOSviewer. The sizes of nodes and thickness of lines indicate the frequency of co-occurrence of specific terms.

Results

A map of the top-77 author keywords of LIS papers published between 2011 and 2017 can be retrieved at https://tinyurl.com/ybb29ox2. The corresponding semantic network of not-tweeted LIS papers is available at https://tinyurl.com/ya9h4fkg. Figure 1 shows the semantic map of the top-77 author keywords of tweeted LIS papers during the time period 2011-2017.

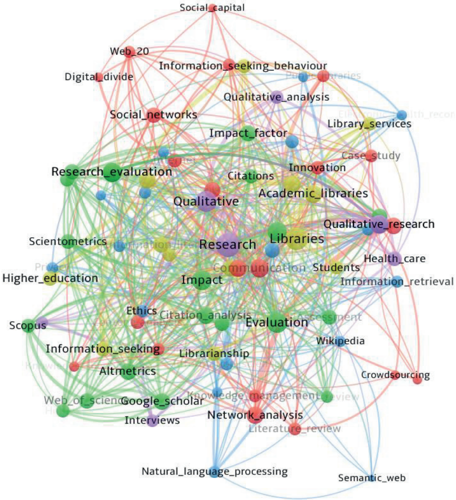


Figure 1: Top-77 author keywords of LIS papers tweeted and published between 2011 and 2017.
(An interactive version of this network can be found at: https://tinyurl.com/yc4b7uz9)

The corresponding figure showing the semantic network of the top-77 hashtags is available at: https://tinyurl.com/yb8warcn. The semantic map of the top-77 author keywords of LIS papers tweeted and mentioned in news outlets can be found at: https://tinyurl.com/ybugpg7z.

Discussion and Conclusions

Our results show that LIS papers are well-represented on Twitter. Most relevant keywords for bibliometrics, altmetrics, social networks, and libraries can be found in the maps of both tweeted and not-tweeted papers. Significant differences were found between networks of tweeted papers and papers which were both tweeted and mentioned in the news. We suggest that the latter can be considered as representations of public discourse. This public discourse is oriented towards digital and electronic health care more than the papers which are tweeted. This focus of public discourse is also visible in the network of hashtags.

Acknowledgments

The bibliometric data used in this paper are from in-house database developed and maintained in cooperation with the Max Planck Digital Library and prepared by Clarivate Analytics. The Twitter and news data were retrieved from a locally maintained database at the Max Planck Institute for Solid State Research with data shared with us by the company Altmetric (see www.altmetric.com) on June 10, 2018.

References

Bornmann, L., Haunschild, R., & Adams, J. (2019). Do altmetrics assess societal impact in the same way as case studies? An empirical analysis testing the convergent validity of altmetrics based on data from the UK Research Excellence Framework (REF). *Journal of Informetrics, 13*(1), 325-340. doi: 10.1016/j.joi.2019.01.008.

Bornmann, L., Haunschild, R., & Marx, W. (2016). Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? *Scientometrics, 109*(3), 1477-1495. doi: 10.1007/s11192-016-2115-y.

Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author keywords and hashtags. *Journal of Informetrics, 13*(2), 695-707. doi: 10.1016/j.joi.2019.03.008.

Hellsten, I., & Leydesdorff, L. (in press). Automated Analysis of Topic-Actor Networks on Twitter: New approach to the analysis of socio-semantic networks. *Journals of the Association for Information Science and Technology*. doi: 10.1002/asi.24207.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*(1), 7-15. doi: 10.1016/0020-0190(89)90102-6.

Wouters, P., Zahedi, Z., & Costas, R. (2018). Social media metrics for new research evaluation. Retrieved from https://arxiv.org/abs/1806.10541

Unsupervised Keyphrase Extraction in Academic Publications Using Human Attention

Yingyi Zhang and Chengzhi Zhang[*]

yingyizhang@njust.edu.cn, zhangcz@njust.edu.cn
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094，China

Introduction

In bibliometric research, publication keywords have been commonly utilized to reveal the knowledge structure of research domains. Since some academic publications lack of author assigned keyphrases, the technology of keyphrase extraction is in demand. The premise for annotator to annotate keyphrases is to read the corresponding content. Intuitively, features estimated from human reading behaviour can be leveraged to assist keyphrase extraction. Previous studies on keyphrase extraction have ignored these features. Thus, this paper aims to integrate the reading behaviour into keyphrase extraction frameworks.

When human read, they do not pay the same attention to all words. The reading time of per-word is the indicative of textual processing, which reflects human attention on various content. To obtain human attention during reading, this paper estimates eye fixation duration from eye-tracking corpus. The modern-day eye tracking equipment result in a very rich and detailed dataset. Thus, we utilize open source eye-tracking corpora and do not require eye-tracking information of the target datasets.

In this paper, we explore the idea of using human attention, as estimated from eye-tracking corpus as external information on TextRank (Mihalcea & Tarau, 2004). Human attention is leveraged to normalize word and edge weights of the TextRank. Experimental results demonstrate that our model yield a better performance than TextRank. We are, to the best of our knowledge, the first to integrate human attention into keyphrase extraction.

Dataset

GECO corpus

This paper estimates human attention from GECO corpus (Cop et al., 2017). In GECO, participants read a part of the novel "The Mysterious Affair at Styles by Agatha Christie". Six males and seven females whose native language is English participated in and read 5,031 sentences. There are various features in GECO, including First Fixation Duration (FFD) and Total Reading Time (TRT). In this paper, we use the TRT feature, which represents total human attention on words during reading.

Human attention is correlating with word frequency. Thus, ATRT is normalized by the word frequency of the British National Corpus[1] (BNC). Before normalizing, BNC is log-transformed per million and inversed (INV-BNC), such that rare words get a high value. ATRT and INV-BNC are min-max-normalized to a value in the range 0-1. ATRT is multiplied with INV-BNC to get normalized ATRT (N-ATRT).

Keyphrase extraction datasets

This paper uses two academic datasets, i.e., *Inspec* and *KP20k*. *Inspec* was built by Hulth (2003). It contains 2,000 abstracts of research articles and 19,254 manually annotated keyphrases. *KP20k* dataset was built by Meng et al. (2017). Its testing dataset contains 20,000 scientific articles in computer science. Although this paper aims to extract keyphrase from academic publications, we use other types of datasets to evaluate whether the genres of eye-tracking corpus will affect the performance. Hence, we use *DUC2001* dataset, which was built by Wan and Xiao (2008). This dataset contains 308 news articles.

Human Attention based TextRank Keyphrase Extraction Algorithm (HATR)

First, the original text is divided into sentences, and only the noun and adjective words are retained. Assuming that words composed of a sentence can be expressed as w_1, w_2, ..., w_n, there is an edge $\mathrm{e}(w_j, w_i)$ between two words w_j and w_i if these two words co-occur in same word windows. Based on the graph composited by word vertices and edges, the importance of each word vertices can be calculated by Eq. (1), in where λ is a damping factor range from 0 to 1, and $|V|$ is the number of vertices. The damping factor indicates the probability of each vertex performing random jump to any other vertexes within the graph. In this paper, we let the word window size be 4 and λ be 0.85.

$$\mathrm{R}(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e(w_j,w_i)}{O(w_j)} R(w_j) + (1-\lambda)\frac{1}{|V|} \quad (1)$$

[*] Corresponding Author
[1] http://www.kilgarriff.co.uk/bnc-readme.html

In the TextRank algorithm, $R(w_j)$ and $e(w_j, w_i)$ are initialized unprivileged. In our models, we utilize human attention to normalize the initialized value of $R(w_j)$ and $e(w_j, w_i)$. The initialized value of $R(w_j)$ depends on the N-ATRT value of itself. The initialized value of $e(w_j, w_i)$ depends on the N-ATRT value of $R(w_j)$ and $R(w_i)$.

As for $R(w_j)$, if the word exists in the GECO corpus, the initialized value of $R(w_j)$ is assigned as N-ATRT. Otherwise, the initialized value of $R(w_j)$ is assigned as the mean value of N-ATRT. As for $e(w_j, w_i)$, the initialized value depends on the average of initialized values of $\mathrm{R}(w_j)$ and $R(w_i)$.

After obtaining word scores using HATR, we extract phrases with the pattern $(adjective) \times (noun) +$, which represents zero or more adjectives followed by one or more nouns. The ranking score of a candidate keyphrase is computed by summing up the scores of all words within the phrase: $R_{(phrase)} = \sum_{w_i \in phrase} R_{(w_i)}$. Then candidate keyphrases are ranked in descending order of ranking scores. The top M candidates are selected as keyphrases. Note that if a phrase is a part of other phrases, this phrase will be ignored.

Result

In this paper, we compare the F1 score of TextRank and human attention based TextRank (HATR). We use three real life datasets, i.e., DUC2001, Inspec, and KP20k. The quantity of keyphrases we select range from 5 to 15. We use the F1 score to evaluate the performances of keyphrase extraction models. Table 1 report the F1 scores on three datasets. We have following observations.

Table 1. F1 values of DUC2001, Inspec and KP20k datasets.

Dataset	Num	Textrank	HATR
DUC2001	5	18.26	21.29
	10	22.52	24.14
	15	23.29	**24.48**
Inspec	5	23.65	24.30
	10	29.28	30.05
	15	30.01	**30.90**
KP20k	5	11.78	**11.88**
	10	11.12	11.24
	15	10.10	10.14

Human attention estimated from eye-tracking corpus is helpful in improving the performances of unsupervised keyphrase extraction algorithm. As shown in Table 1, all F1 scores of HATR are higher than those of TextRank. It indicates that the attempt of integrating human reading behaviour into TextRank is feasible.

The genres of eye-tracking corpus have an impact on the performances of keyphrase extraction. In Table 1, we can find that the degree of improvement varies from the DUC2001 news dataset to the Inspec and KP20k academic datasets. The impact of human attention is more significant on news datasets than academic datasets. The genres of available eye-tracking corpus may be the cause. In this paper, we use the GECO eye-tracking corpus consisting of the text from novels. The genres of novels is informal and narrative, which is more similar to it of news.

The word overlap between test datasets and GECO corpus affect the performances of keyphrase extraction. The word overlap represents the ratio of words in the test dataset that exist in the GECO corpus to all words in the test dataset. It is 9.28%, 11.06% and 2.07% of the DUC2001, Inspec and KP20k dataset, respectively. It can be observed that the overlap and F1 scores are all low in the KP20k. Thus, low word overlap may result in low F1 scores.

Conclusion and Future Work

In this paper, we consolidate the TextRank with Total reading time (TRT) estimated from GECO open-source eye-tracking corpus. The proposed model have proved to yield a better performance on three datasets. In the future, first, we try to leverage more eye-tracking features and integrate them with more keyphrase extraction algorithms. Then, we attempt to explore more specific human attention features when reading academic publications.

Acknowledgement

This work is supported by Major Projects of National Social Science Fund (No. 16ZDA224), Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, ISTIC (No. ZD2018-07/01), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX18_0351).

References

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2016). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. Behavior Research Methods, 49(2), 1-14.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 216-223)

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., & Chi, Y. (2017). Deep keyphrase generation. In proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 582-592).

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. In proceedings of the conference on empirical methods in natural language processing (pp.404-411).

Wan, X., & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In proceedings of the National Conference on Artificial Intelligence (pp. 855-860).

Applying the Author Affiliation Index to Rank Chinese Library and Information Science Journals

Qing Ke， Ming Li and TingTing Zhu

keqing@nju.edu.cn，njulm@nju.edu.cn ，841140449@qq.com
Information Management School, Nanjing University, Nanjing(China)

Introduction

As an important feature, author affiliation is traditionally widely accepted and applied for assessing the prestige of individual researchers, universities, and academic departments (Fry & Donohue, 2014). The ranking of a journal could also base on the reputation of the affiliations of the authors that publish in that journal. This is the basic premise for using Author affiliation index (AAI) as a journal ranking metric. For the first time, Cronin and Meho (2008) apply the AAI to LIS journals to determine whether faculty at the top-10 North American LIS programs have a disproportionate presence in the premier journals of the field. The study finds that LIS may be both too small and too interdisciplinary a domain for the AAI to provide reliable results. The scant coverage points to a need for an extension of research efforts on the application of AAI in some certain subjects. Here, we attempt to apply AAI to the ranking of Chinese library and information science journals to explore its potential utility to this field, and to find feasible ideas for more rational use of AAI metrics.

Method

Data Gathering

Our journal sample comprises 18 of the CSSCI-LIS journals (2015-2016) and relevant data of each journal was collected from CNKI and Wangfang database. To unify time window, we count all the published article number from each journal in the year 2016 when choosing the sample papers, excluding the news bulletins, book reviews, conference proceedings et al., and only retains the research papers. Previous studies have shown that AAI results tend to be stable when the number of equivalent articles exceeds 50(Gorman & Kanet, 2011). The number of equivalent articles published in 18 journals in 2016 is ranged from 50 to 472.

Selection of the top-notch institution set

This paper chooses four prestigious journals in LIS, namely *Journal of Library Science in China, Journal of The China Society for Scientific and Technical Information, Journal of Academic Libraries and Library and Information Service*. In order to test the impact of the size of top-notch institutions on AAI calculation results, we established a large (top 90)and a small(top 20) top-notch institution set.

Calculating the AAI

Following the formal definition of AAI by Gorman and Kanet (2005), we calculate the AAI based on Top-20 (AAI1) and Top-90 (AAI2) institution set. Following a modified AAI model by Fry and Donohue (2014), we set different weights to institutions according to the number of articles published in prestigious journals（i.e., the four journals we have set above）.

Results and Discussion

AAI Ranking Results

The AAI scores of 18 Chinese LIS journals are showed in Table 1 based on three AAI methods.

Table 1 AAI of 18 LIS Journals

Journal Title	AAI1（Top-20）	AAI2（Top-90）	AAI-weighted
Journal of Library Science in China	0.57	0.86	1.32
Document, Informaiton & Knowledge	0.52	0.68	1.05
Journal of The China Society for Scientific and Technical Information	0.52	0.81	1.22
New Technology of Library and Information Service	0.48	0.69	1.08
Information and Documentation Services	0.48	0.73	1.05
Library and Information Service	0.46	0.73	1.08
Information studies: Theory & Application	0.40	0.61	0.93

Library and Information	0.38	0.53	0.87
Information Science	0.36	0.55	0.85
Journal of Academic Libraries	0.34	0.81	1.08
Library Tribune	0.32	0.59	0.81
Researches in Library Science	0.30	0.48	0.70
Journal of Intelligence	0.23	0.53	0.72
Library Journal	0.21	0.55	0.75
Journal of the National Library of China	0.21	0.59	0.75
Library	0.19	0.37	0.54
Library Development	0.18	0.45	0.60
Library Work and Study	0.12	0.23	0.32

Table 1 shows the size of Top20 institution set would lead to a low AAI score and it is easy to omit some institutions that have important contributions to journals. The results of AAI2 show that with the expansion of top-notch institution set, the AAI scores of journals have increased. It suggests that the large top-notch institution set is a better choice for LIS journals rankings.

Table 1 also shows that most journals (i.e., 14 journals) fluctuate within three positions comparing AAI1 with AAI2.

The weighted AAI ranking

The weighted AAI score of each journal increased again compared with AAI1 and AAI2. So, when the size of top-notch institution set is expanded, its ranking declines. But if we give higher weights to some prestigious institutions, the journal AAI ranking rises again.

Comparing AAI with Impact Factors

To evaluate the AAI's reliability and validity, three AAI ranking results were compared with journal impact factor ranking (2016 edition). The greatest difference between AAI ranking and impact factor ranking is 9,9 and 10 for three AAI methods (AAI1,AAI2,AAI-weighted). And the numbers of journals changing within three ranking positions are 11,15 and 14 individually. We find that there are 6 journals including *Journal of Library Science in China*、*Library Development*、*Information and Documentation Services*、*Library Tribune*、*Library and Library Work and Study* which have the same ranking positions for the four ranking methods. The correlation coefficient between AAI-weighted and impact factor is highest among the three(r=0.738), then followed by AAI2(r=0.713) and AAI1sequently(r=0.567).

Table 2 Spearman Correlation Coefficients between AAI and IF

	AAI1	*AAI2*	*AAI- weighted*
Impact Factors	0.567	0.713	0.738
Sig.	0.014	0.001	0.000

This finding suggests that as a supplement, AAI provides an additional choice of journal rankings by overcome the limitations of journal citation analysis.

Conclusion

Based on the application of AAI in ranking LIS journals, this current research proves a more reasonable and effective way to use AAI. Our suggestions are as follows:

- including non-academic authors' articles when selecting paper samples;
- keep a consistent time period rather than a consistent number of equivalent articles;
- if the author affiliations are diverse, it would rather use self-determined top-notch institution set than existed open-published top-notch institution set;
- the size of top-notch institution set should not be too small;
- give different weights to institutions can enhance the accuracy of AAI.

It appears that the advantages of AAI metric are partly recognized in the LIS literature. The present study serves as a foundation to encourage more such research in the field. AAI appears to be in a relatively nascent form with good prospects, but there are remaining problems to be solved. The present study has contributed to the practice of AAI in journal rankings in LIS discipline. The results suggest that the application of AAI in journal evaluation is far from reaching any kind of peak like other bibliometric metrics. The present study has provided some guidelines and support needed in these areas. However, more efforts are needed for researchers as active promoters of AAI so that journal evaluation and ranking can progress in research and in practice.

References

Cronin, B.&Meho, L.I.(2008). Applying the author affiliation index to library and information science journals. *Journal of the American Society for Information Science & Technology*,59,1861–1865.

Fry, T. D.&Donohue, J.M.(2014). Exploring the Author Affiliation Index. *Scientometrics*,98,1647-1667.

Gorman, M. F.&Kanet, J.J.(2005). Evaluating operations management–related journals via the author affiliation index. *Manufacturing & Service Operations Management*,7,3-19.

Using Citation Contexts to Evaluate Impact of Books

Qingqing Zhou[1] and Chengzhi Zhang[2]

[1] *breeze7zhou@163.com*
Nanjing Normal University, Nanjing (China)

[2] *zhangcz@njust.edu.cn* (corresponding author)
Nanjing University of Science and Technology, Nanjing (China)

Introduction

Traditional researches on assessing book impacts focuses on citation frequencies, and prove that more citations denote higher impacts (Barilan, 2010; Krampen, Becker, Wahner, & Montada, 2007). However, these methods ignore citation contexts, which cannot mine semantic information or identify citation intentions for fine-grained analysis. Hence, in this paper, we conducted fine-grained citation context analysis automatically for book impact assessment, and determined the most influential metrics. In order to prove the validity of our method, correlation analysis between our book impact results and other evaluating metrics for book impact were undertaken.

Methodology

Data collection

We first compared the Chinese book category provided by Amazon with Chinese discipline category to identify book disciplines. Then, we screened out books sold and commented in Amazon of identified disciplines, and got 6006 candidate books. Thirdly, we matched 6006 books' titles, authors and publication years in Baidu Scholar to obtain metadata of their citing literatures. To ensure the accuracy of citation context extraction, we annotated citation contexts manually, and 500 of 6006 books were selected as final book set. We downloaded full texts of all citing literatures about the 500 books via full-text databases, and extracted citation contexts about the books in these literatures. As some citing literatures have no citation mark in the texts, finally, we got 2288 citation contexts of 370 books.

Method

The primary purpose of this paper is to specify how and whether it is feasible to assess book impact via citation contexts. The overall framework is shown in Figure 1. We first collected full texts of citing literatures to get citation contexts about each book. Secondly, we identified citation intensities in citing literatures (Ding, Liu, Guo, & Cronin, 2013). Meanwhile, we conducted supervised machine learning to classify citation functions, such as background citation, use citation and comparison citation (Hernández-Alvarez, Soriano, & Martínez-Barco, 2017). Finally, to prove the validity of our method, correlation analysis between book impact metrics via our method and other assessment metrics was undertaken.

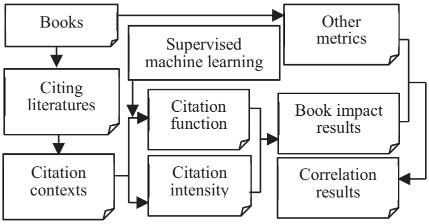


Fig. 1. Book impact assessment by citation context analysis

Citation function

To identify citation functions, we classified citation function automatically, as shown in Figure 2. First, we conducted text representation, so as to convert citation contexts into vectors for machine learning. Then, we annotated part of citation contexts to train the classification model. Finally, we got citation function labels of all citation contexts, and we used macro-average precision, macro-average recall and F_1 value (Salton & Mcgill, 1983) to evaluate classification performance.

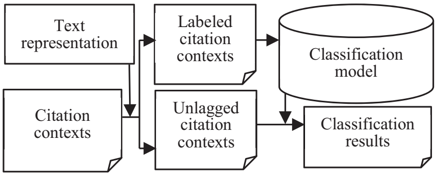


Fig. 2. Citation function classification

According to processing above, we got categories of all citation contexts, and then we computed scores of citation function metric using equation (1) and (2) (Hernández-Alvarez et al., 2017).

$$fun_i = \frac{\sum_{j=1}^{n} fun_{ij}}{n} \quad (1)$$

$$fun_{ij} = \begin{cases} 1,\ \text{Background citation} \\ 2,\ Comparison\ citation \\ 3,\ Use\ citation \end{cases} \quad (2)$$

Where, fun_i denotes citation function score of book i, fun_{ij} means citation function score of the

jth citation context about book i. n is the total citation times in the texts of cited literatures.

Citation intensity

Citation intensity denotes citation counts in a citing literature about a given book. Higher citation intensity indicates a higher impact. In this paper, we calculated citation intensity via equation (3).

$$int_i = \frac{\sum_{j=1}^{n} int_{ij}}{n} \qquad (3)$$

Where, int_i denotes citation intensity score of book i, int_{ij} means citation intensity score of book i in citing literature j, i.e. citation count of book i in the citing literature j. n is citations of book i.

Results

Performance evaluation on citation function classification

We used SVM to conduct citation function classification. Performance evaluation results are shown in Table 1. We can see all three evaluation indicators of our method are about 0.9, which means that the classification results are reliable. Thereby, this paper used the trained model to classify citation functions of all citation contexts. The final classification results are represented in Figure 3. From Figure 3 we can see, most citation functions are background or use citations, few books are cited for comparison.

Table 1. Performance evaluation of citation function classification

Metrics	Macro-P	Macro-R	Macro-F1
Performance	0.8994	0.8983	0.8989

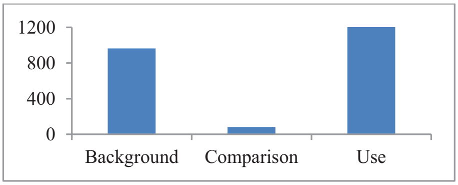


Fig. 3. Classification results of Citation function

Correlation analysis between citation context metrics and other metrics

There are significant Spearman correlations between citation context metrics and other metrics (see Table 2). Specifically, citation function has significant positive correlations with sales, which means that books with higher scores of citation functions may be sold more. Meanwhile, citation intensity has significant positive correlations with citations and sales. It suggests that books with higher scores of citation intensities tend to get more citations and higher sales. In addition, both citation function and citation intensity have no significant correlations with holding metrics. Hence, we can conclude that metrics delivered from citation context can affect books' citations and sales, especially citation intensity. However, they cannot provide decision supports about book ordering for libraries.

Table 2. Spearman correlations between citation context metrics and other metrics

Metrics	Citations	Sales	Holding numbers	Holding regions
Citation function	-0.006	0.114*	-0.090	-0.049
Citation intensity	0.161**	0.122*	-0.052	-0.046

Note: **. Significant at p=0.01 *. Significant at p=0.05

Conclusion

This study introduced a framework for measuring book impact according to citation contexts. In order to verify the reliability of our method, correlation analysis was conducted. The weak but often significant relationships suggest that citation contexts could be a particularly helpful reference for assessing book impact. Meanwhile, citation intensities have higher correlation values than citation functions. It reveals that citation intensities of books are particularly more important in improving the impact of academic books than citation functions.

Acknowledgement

This work is supported by Major Projects of National Social Science Fund (No. 17ZDA291), Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Institute of Scientific and Technical Information of China (No. ZD2018-07/01) and Qing Lan Project.

References

Barilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics, 82*(3), 495-506.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Hernández-Alvarez, M., Soriano, J. M. G., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering*, 1-28.

Krampen, G., Becker, R., Wahner, U., & Montada, L. (2007). On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. *Scientometrics, 71*(2), 191-202.

Salton, G., & Mcgill, M. J. (1983). *Introduction to modern information retrieval*: McGraw-Hill.

Insight Into Research Hot Topics and Research Groups of Sustainable Urbanization

Danni Liang[1] , Lili Wang[2] and Bowen Song[3]

[1] *liangdanni1992@mail.dlut.edu.cn*
Dalian University of Technology, Faculty of Humanities and Social Sciences, Science of Science and S&T Management, Linggong Road 2, 116024 Dalian (China)

[2] *wangll@dlut.edu.cn*
Dalian University of Technology, Faculty of Humanities and Social Sciences, Linggong Road 2, 116024 Dalian (China)

[3] *bowensong333@163.com*
Dalian University of Technology, Faculty of Humanities and Social Sciences, Science of Science and S&T Management, Linggong Road 2, 116024 Dalian (China)

Introduction

In line with the transformation of global urbanization, sustainable urbanization plays a critical role in boosting prosperity of cities and societies in the 21st century (Rasoolimanesh, Badarulzaman & Jaafar, 2012) and has been considered as the core of scholarly discussion. With the proposition of the concept of sustainable development in 1987, sustainability became the most important index in evaluating the quality of sustainable urbanization. And four dimensions which were social sustainability, economic sustainability, environmental sustainability and institutional sustainability were pointed out in The Localize Agenda 21 in 1996 (Spangenberg, Pfahl & Deller, 2002). Therefore, sustainable urbanization can be seen as a balanced and dynamic process which presented the process of sustainable development goals, the synchronous process of constructing sustainable cities and rural areas, and the high quality and complex combination of factors involving science, technology, and policy.

This paper aims at presenting a review of the available literature in the field of sustainable urbanization in the light of summarizing the most valuable findings about hot topics, core literature and research groups in this field.

Data collection and research methodology

Data collection

The data used in this research was retrieved from Web of Science Core Collection, the most frequently used data source, which ensures the authoritative and representative of the retrieved academic journals relevant to the field of sustainable urbanization. The topic was (“sustainable urbanization” OR “urban sustainability” OR “sustainable urban development”) and the time was (“from 1990 to 2019”), which were for the search strategy. After data cleaning, 11851 articles sourced from Web of Science Core Collection from 1990 to 2019 were collected in total.

Research methodology

Co-word analysis and citation analysis were both used in this paper. In science and technology studies, co-word analysis is associated with content analysis, which can be mainly used to analyse the research status and research trend of certain subjects, and to explore the relationship among keywords extracted by co-occurrence analysis of technical terms (Guo, Chen & Long et al., 2017). Citation analysis and time series analysis can help to explore the relationship between different literature in the field of sustainable development, obtain the context of knowledge changing with time, and summarize the main research groups.

Results

Topics analysis

The results indicate that a total of 32628 keywords are obtained, and among them there are 82 keywords with high frequency greater than 120. The total word frequency (23581) accounts for 26.2% of the frequency of all keywords (90067). Specifically, management, policy, model, and economic growth, environment and ecology are paid more attention. Besides, governance plays a more and more important role in sustainable urbanization.

The results also reveal that applied research and methodological research have increased in recent years and become a remarkable feature in scientific research. More scholars are devoted to figure out how to make the urbanization sustainable and how the economy, society, environment and governance interact, while systematic theoretical research is less.

Hot topics analysis

In this paper, 479 high frequency keywords with frequency greater than 30 times are selected for clustering. From Figure 1, there are mainly four types of research as a whole: (1) The realization of systems, policies and governance and the study of social sustainability; (2) The study of urban sprawl and land use and their impacts in China and its cities; (3) The relationships between energy consumption, air pollution and economic growth; (4) Modes, evaluation indexes, management systems and strategy of sustainable urbanization. Among them, each part contains various contents which form a connection network around the central nodes and aggregate into their own research clusters. In addition, there is no obvious dividing line between the clusters and there are 37649 lines between the nodes, indicating that the connections between nodes have closely linked and the four regions are related to each other.

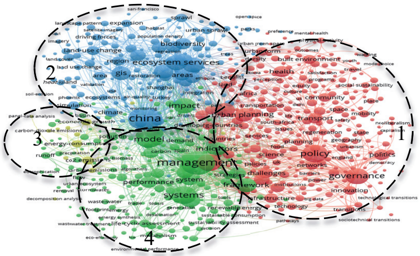


Figure 1. The cluster map of keywords with high frequency of 479.

Research cliques

From Figure 2, three main research groups are identified according to their research characteristics. Besides, the core literature of the citation network were also identified and summarized.

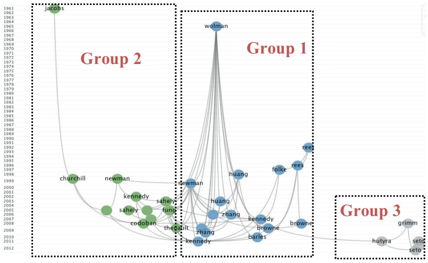


Figure 2. Citation network and clustering map of literature.

(1) Group of Theory and Construction (Group 1). An article, called *The metabolism of the city*, written by Wolman Albel in 1965, provides a theoretical basis. Representatives are Newman, PWG, William Rees and Kennedy and they mainly focus on the construction of conceptual models and urban metabolism theory. (2) Group of Practice and Deconstruction (Group 2). It was inspired by a book called *The death and life of great American cities*, written by Jacobs.J in 1961. Representatives are Jacobs.J, Newman, P, Halla R Sahely, and Natalia Codoban and they mainly focus on estimating the urban metabolism and developing sustainability criteria for urban infrastructure systems. (3) Group of Risk and Challenge (Group 3). Scholars in this group mainly emphasize on the ecology of cities, global urban land expansion, the impacts on biodiversity, as well as the contradictions of sustainable development.

Conclusion and Discussion

First, sustainable urbanization is a dynamic process in which society, economy and environment are sustainable development by means of scientific and innovative governance. Management, policy and systems are the main approaches to promote sustainable urbanization and governance has also played an important role in the balance of society, economy and environment recently. Second, four main topics are identified by cluster analysis, focusing on how to improve the development of modes, systems, policies and governance, urban sprawl and land use and their impacts in China, the relationship between energy consumption, air pollution and economic growth, and the evaluation of sustainable urbanization. Notice that they are not completely fragmented. Besides, three research groups are concluded by citation analysis, which are supplementary to each other and draw lessons from each other.

Acknowledgments

This research was supported by the National Social Science Fund project (14BGL124) and the Fundamental Research Funds for the Central Universities (DUT17RW228).

References

Rasoolimanesh, S. M., Badarulzaman, N., & Jaafar, M.. (2012). City development strategies (cds) and sustainable urbanization in developing world. *Procedia-Social and Behavioral Sciences, 36*(none), 623-631.

Spangenberg, J. H., Pfahl, S., & Deller, K.. (2002). Towards indicators for institutional sustainability: lessons from an analysis of agenda 21. *Ecological Indicators, 2*(1), 61-77.

Daoyan, G., Hong, C., Ruyin, L., Hui, L., & Qianyi, L.. (2017). A co-word analysis of organizational constraints for maintaining sustainability. *Sustainability, 9*(10), 1928-.

Albel, W. (1965). Metabolism of cities. *Scientific American, 213*(3), 179-190.

Jacobs, J. 1961. The death and life of great American cities, New York: Random House.

Historical bibliometrics using Google Scholar: the case of Roman law, 1500-2016

Janne Pölönen[1] and Björn Hammarfelt[2]

[1] *janne.polonen@tsv.fi*
Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki (Finland)

[2] *bjorn.hammarfelt@hb.se*
Swedish School of Library and Information Science, University of Borås, Allégatan 1, Borås, 503 32 (Sweden)

Introduction

Bibliometrics can be a useful resource for social sciences and humanities (SSH) research beyond its role in research evaluation and funding-schemes (Scharnhorst & Garfield, 2010). Predominately research in the field of bibliometrics focus on contemporary developments using datasets that rarely provide historical perspectives. Still, historical approaches are not unheard of, with de Solla Price's seminal studies of the growth of science being one key example (de Solla Price, 1986). Yet, the field of bibliometrics has not fully explored the potential of what Hérubel (1999) calls "historical bibliometrics". There have been several attempts of going beyond established databases to study for example Catalan literature (Ardanuy, Urbano, & Quintana, 2009), Swedish literature (Hammarfelt, 2012) and Venetian histography (Colavizza, 2018). Yet, the approaches and methods used are often time-consuming and not easily transferred to other contexts and materials.

Considering the lack of coverage in established citation databases (Web of Science and Scopus), and the limitations of local and specific approaches, in this paper we investigate the potential that Google Scholar (GS) data has for studying the development of research fields from a historical perspective using Roman law as an example. Roman law (RL) has constituted an international research field within academia since the 12th century (Stein, 1999). After Latin ceased to be the *lingua franca*, there remain five international RL publishing languages: English, French, German, Italian and Spanish. Thus, Roman law literature provides a good case for probing the historical and linguistic coverage of GS.

Methods and materials

All publication records including in the title words denoting "Roman law" in English, French, German, Italian and Spanish, published between years 1500 and 2016, were retrieved from Google Scholar in August 2017, in blocs not exceeding 1000, using the Publish or Perish interface. The publication records were copied to Excel in RIS format, and processed with BibExcel tool-box (Persson, Danell, & Schneider, 2009).

The dataset of Roman law publications is analyzed to establish the number of publications and authors, differentiating between the five language groups, from 1500 to 2016. The growth of the field is estimated on basis of the development of the absolute number and average yearly number of publications, as well as the number of authors involved in producing them, in different periods. Also bibliometric measurements are performed on the data to investigate its properties and consistency. These include the average number of publications per author (publication productivity), as well as the concentration of publications and citations.

Findings

The data retrieved from GS contains a total of 21300 publications published between years 1500 and 2016 and including the title words "Roman law" in the five languages (Table 1). The oldest publication year in French is 1727, in German 1730, in English 1772, in Spanish 1796 and in Italian 1833. Largest group of records consists of 9983 French publications that account for 47 % of all records. English language publications make up 18 %, Italian publication 13 %, Spanish publication 13 % and German publications 9 % of the records.

Table 1. GS records for publications 1725-2016

Period	En	Fr	De	It	Es	All
1725-1749	0	3	7	0	0	10
1750-1774	4	1	4	0	0	9
1775-1799	3	8	14	0	1	26
1800-1824	3	23	41	0	1	68
1825-1849	8	354	85	5	26	478
1850-1874	32	2540	116	22	32	2742
1875-1899	106	5592	213	148	56	6115
1900-1924	259	162	218	214	48	901
1925-1949	406	276	168	415	125	1390
1950-1974	616	414	359	629	301	2319
1975-1999	928	294	359	613	721	2915
2000-2016	1347	300	363	708	1352	4070
No date	71	16	53	49	68	257
Total	3783	9983	2000	2803	2731	21300

The number of RL publications has increased from around 10 publications in the earliest periods to 4000 publications in 2000-2016 (the latest timeframe is only 17 years). The early 19th century is a period when the number of publications begins to increase in all language groups (Table 1).

The largest number of RL publications is attested in the late 19th century, when there is a very large number of French publications (mostly thesis and dissertations). This can be related to requirements of French legal education. Following the introduction of Code Napoleon in 1804 and the reform of law schools, between 1808 and 1895 doctoral thesis in law consisted of two dissertations, one of which had to be based on Roman law (Imbert, 1984).

According to the GS data, the 21300 Roman law publications have a total of 11420 different authors. The largest number of authors is attested in 1875-1899, vast majority being related to the French publications. The average number of publications per author in the GS dataset has somewhat increased (Figure 1).

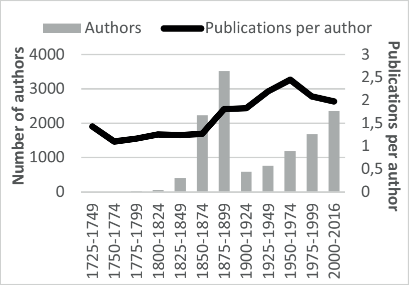


Figure 1. Publications per author 1725-2016

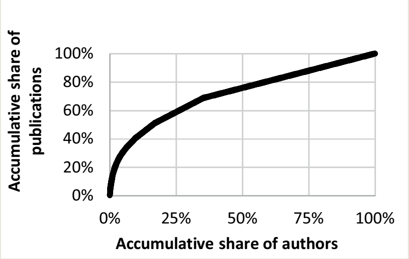


Figure 2. Concentration of publication to authors

Publication are unevenly distributed among the authors: one-half of all publications is produced by 16 % of the most prolific authors (Figure 2). Citations are even more unevenly distributed: 73 % of the publications have received no citations recorded in Google Scholar, and only 1 % of the most highly cited publications account for one-half of all the citations.

Discussion and conclusions

We find Google Scholar to be a promising data source for historical bibliometrics: it is accessible, has broad coverage and has quite a historical depth. At the same time there are distinct disadvantages: the quality of data is low, and the database is continuously updating which renders it difficult to reproduce earlier searches and data collections. Still, the possibilities for historical bibliometrics will most likely increase as the digitisation of older materials progress. Hence, while the approach taken here is a probing one, with many difficulties to solve, we find that employing Google scholar data for historical studies of fields and disciplines is a promising path for the future, and it is likely that such a path might attract travellers among bibliometricians as well as historians and other digital humanists.

References

Ardanuy, J., Urbano, C., & Quintana, L. (2009). A citation analysis of Catalan literary studies (1974–2003): Towards a bibliometrics of humanities studies in minority languages.

Colavizza, G. (2018). Understanding the History of the Humanities from a Bibliometric Perspective: Expansion, Conjunctures, and Traditions in the Last Decades of Venetian Historiography (1950–2013). *History of Humanities*, 3(2), 377–406.

De Solla Price, D. J. (1986). *Little science, big science... and beyond.* Columbia University Press New York.

Hammarfelt, B. (2012). Harvesting footnotes in a rural field: citation patterns in Swedish literary studies. *Journal of Documentation*, 68(4), 536–558.

Hérubel, J.-P. V. M. (1999). Historical Bibliometrics: Its Purpose and Significance to the History of Disciplines. *Libraries & Culture*, 34(4), 380–388.

Imbert, J. (1984). Passé, présent et avenir du doctorat en droit en France, *Revue d'histoire des Facultés de droit*, 1, 11–35.

Persson, O., Danell, R., & Schneider, J. W. (2009). How to use Bibexcel for various types of bibliometric analysis. In F. Åström, R. Danell, B. Larsen & J. Schneider (eds.), *Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at His 60th Birthday*. International Society for Scientometrics and Informetrics, 9–24.

Scharnhorst, A., & Garfield, E. (2010). Tracing scientific influence. arXiv Preprint *arXiv*:1010.3525.

Stein, P. (1999). *Roman Law in European History*. Cambridge: Cambridge University Press.

Identification of Milestone Papers in Physics via Reference Publication Year Spectroscopy

Yu Liao[1]，Zhesi Shen[2]，Liying Yang[3]

[1] liaoyu@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences, Beijing 100190, P. R. China (China)
Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, P. R. China(China)

[2] *shenzhs@mail.las.ac.cn,*[3]*yangly@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing 100190, P. R. China (China)

Introduction

Research progresses are often built on previous scientific works, especially those milestones that break the boundaries of knowledge. With the citation data becoming more available and citation analysis approaches and tools being developed, we can efficiently trace the idea flows to these breakthroughs that significantly boost the advancement of physics via the citation links. Detecting these historical roots plays important roles in providing insights for understanding the context of knowledge creation and guidance for future directions of development. Citation analysis offers an interesting perspective to describe and understand the evolution of significant achievements. Redner analysed the citation statistics of publications from 110 years of *Physical Review* series and the citation patterns of highly-cited papers (Redner, 2005). Khelfaoui and Gingras turn to a journal-level perspective and quantitatively analyse the changing position of *Physical Review* from the periphery to the center of the physics journal citation network among its more than 120 years history (Khelfaui, 2019). Among the citation analysis approaches proposed from various perspectives, the recently proposed Reference Publication Year Spectroscopy (RPYS) method (Marx, 2014), aiming at exploring the historical roots, attracts scholars' attention. RPYS investigates the yearly distribution characteristics of the cited references of a set of publications in a pre-selected field (Comins, 2015). RPYS has been hitherto applied by some researchers to identify milestone works in several research fields and topics. RPYS has shown promising potential in identifying seminal works, but in most of the tests the authors consider themselves also as experts to interpret the results, resulting to the validation of RPYS less objective. Comins (2017) compared RPYS's performance with independent expert-opinion in the research topic of basal cell carcinoma. In this work, we attempt to apply RPYS analysis in the discipline of physics to identify historical roots and compare them with milestones selected by experts to see its performance beyond topic level.

Data and Methods

APS data set. In this work we use the papers published in American Physical Society (APS) journals to test the effectiveness of RPYS in identifying milestone papers.

PACS codes. Physics and Astronomy Classification Scheme (PACS) codes is a classification system of fields in physics. In this work, we used PACS code category the papers into 29 classes focusing on similar research topics.

Milestone papers as Golden standard. In 2008 to celebrate the 50th anniversary of Phys. Rev. Lett., a collection of Milestone Letters that "have made long-lived contributions to physics, either by announcing significant discoveries, or by initiating new areas of research" are selected by the APS editors. These milestone letters are further used as golden standard to validate the performance of RPYS. For details of the milestone letters, please see (https://journals.aps.org/prl/50years/milestones).

RPYS. RPYS analysis is based on the reference publication year analysis and the procedure can be briefly described into four steps. 1) Gather related publications together with their references for an interested research topic or area. Here we use the PACS codes to construct our analyzing publication set. 2) Aggregate the cited references according to their published years to form a *total citation v.s. reference publication year* plot. 3) Calculate the deviation of the number of cited references of each reference publication year compared with the median in a 5-year window. Large positive deviation implies important publications. 4) Select the years with large deviation as peak years based on some criteria; and then select the highly cited publications within these peak years as milestones. Here for simplicity, we use absolute peak criterion to select the peak years. Figure1 shows the RPYS analysis on the publications belonging to 05(Statistical Physics).

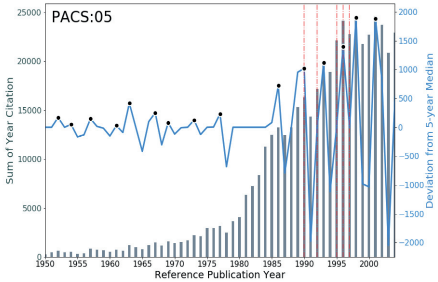


Figure1. Example of RPYS analysis on the field of PACS-05. Gray bars represent the citations of each reference publication year, blue solid line represents the deviation within a five-year time window, black dots represent the selected peak years and red dashed lines highlight the publication year of milestones in gold standard.

Results

The performance of RPYS on the 30 milestone papers is shown in Table 1. Among the 30 milestones, in total, 21 milestone papers are successfully identified by RPYS, that is the identification rate is 70%. Diving into each PACS class, we find that the class of PACS-02, 04, 11, 13, 68, 72, 78, 96 and 97 achieve 100% identification rate. However, for some classes, no milestones are successfully identified, e.g., PACS-06, 12, 71 and 87. The average identification rate over all classes is 64.25% and the median is 100%.

Table 1. The performance of RPYS in identifying milestone papers. The PACS column indicates category number classified by PACS, *N* column for the number of milestones in this class, *Sn* for the number of successfully identified milestones by RPYS and *R* for recall of milestones.

PACS	*N*	*Sn*	*R*	PACS	*N*	*Sn*	*R*
02	1	1	100%	68	1	1	100%
03	5	1	20.0%	71	2	0	0.0%
04	2	2	100%	72	1	1	100%
05	6	3	50.0%	73	2	0	0.0%
06	2	0	0.0%	74	1	0	0.0%
11	2	2	100%	75	2	1	50.0%
12	1	0	0.0%	78	1	1	100%
13	2	2	100%	82	1	0	0.0%
14	7	2	28.6%	84	2	0	0.0%
26	2	1	50.0%	87	1	0	0.0%
31	1	0	0.0%	89	3	1	33.3%
32	5	3	60.0%	95	4	1	25.0%
41	3	1	33.3%	96	1	1	100%
42	11	3	27.3%	97	2	2	100%
64	1	0	0.0%				

Conclusion and discussion

In this work, we investigate the RPYS analysis and its application in identifying milestones in the field of physics. Comparing with the golden standards selected by experts, 21 out of 30 milestones are successfully identified (70% identification rate). However the success rates across the fields of physics show large difference.

The RPYS analysis shows promise in detecting the historical roots and identifying fundamental papers. There are still some open questions need to be investigated. In the PRYS analysis, cited references are gathered and compared based on their publication years, while there is a huge difference in time to accumulate citations between articles published in January and articles in December of the same year. Mariani (2016) pointed that comparing the target paper *i* within a relative publication window [*i*-Δ/2, *i*+Δ/2] centered on paper *i* is better than comparing within an absolute year. It will be an open question of how to modify RPYS in a relative reference time instead of year. In addition, it is important to compare the performance of RPYS with other citation-based approaches.

Acknowledgments

We thank Jinshan Wu for useful discussions. ZS is supported by the China Postdoctoral Science Foundation under Grant 2017 M620944.

References

Comins, J. A., & Hussey, T. W. (2015). Detecting seminal research contributions to the development and use of the global positioning system by reference publication year spectroscopy. *Scientometrics*, 104, 575−580.

Comins, J. A., & Leydesdorff, L. (2017). Citation algorithms for identifying research milestones driving biomedical innovation. *Scientometrics*, 110, 1495 − 1504.

Khelfaoui, M., & Gingras, Y. (2019). Physical review: From the periphery to the center of physics. *Physics in Perspective*, 21, 23–42.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the Historical Roots of Research Fields by Reference Publication Year Spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, 65, 751-764.

Mariani, M. S., Medo, M., & Zhang, Y.-C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, 10, 1207−1223.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, 58, 49–54.

Topic Map Analysis of Deep Learning Patents

Chi-Hsuan Chen[1], Lung-Hao Lee[2], and Yuen-Hsien Tseng[3]

[1] *landy840517@gmail.com*, [3]*samtseng@ntnu.edu.tw*
National Taiwan Normal University, Graduate Institute of Library & Information Studies, Taipei (Taiwan, ROC)

[2] *lhlee@ee.ncu.edu.tw*
National Central University, Dept of Electrical Engineering, Taoyuan (Taiwan, ROC)

Introduction

Deep learning (DL) (LeCun, Bengio, & Hinton, 2015) is a class of machine learning (ML) algorithms that use multi-layers of connected nonlinear processing units inspired by neural networks in animal brains. The characteristics that distinguish DL from traditional ML is that DL techniques can learn the knowledge representation features needed in a task. In the last few years, there are more and more news reported that DL-based systems have outperformed human experts in certain tasks. Due to these successes (or threats), the interest/concern in Artificial Intelligence (AI) has resurged. Major countries have proposed to invest more resources in related research (European Commission, 2018), and many high-tech companies have deployed DL-embedded products or applied for relevant patents to secure their future development.

This paper aims to explore DL patents: 1) to reveal the overall trend; 2) to show the topics or technical fields that have emerged so far; 3) to visualize how these fields are related. An understanding of current status of DL patents would inform the stakeholders to better evolve their investment and development policy in response to the fast-changing world.

Data and Method

The patent data to be analyzed come from the United States Patent and Trademark Office (USPTO), because major companies would file the patents in the USA due to her huge market and because about 98% patent analysis papers used the patents from USPTO (Sharma & Tripathi, 2017). After 40 search trials using various terms and conditions, we finally use "deep learning" or "deep neural networks?" from 1976 to Dec. 31, 2018 to delineate the patent set for analysis, which consists of 1636 patents.

The analysis procedures and guidelines mentioned in Trippe (2015) has been consulted and the analysis tool based on the techniques in Tseng, et al (2007) was used to download and parse the USPTO patent documents. The **tool and the patent set are public available at**: https://github.com/SamTseng/CATAR for free use and result verification.

Result and Discussion

The overview and co-word clustering analyses based on CATAR were applied to this patent set. For overview analysis, the S-curve concept of technology life cycle and Technology Life Cycle Diagram (TLCD) were presented, among all other methods (Gao et al., 2013).

Some initial analyses showed that IBM, Google, Amazon, Microsoft, Samsung, Intel, Siemens Healthcare, Adobe Systems, SAS institute, and State Farm Mutual Automobile Insurance are the top 10 assignees in this patent set. Although Apple and Facebook did not show up in the top 10 list, they are ranked at the 11st and 18th. Together with the top 4 assignees, these six world's largest companies are the founding members of the Partnership on AI (https://www. partnershiponai.org/faq/), which more or less verifies the validity of the patent set delineated by our query, because deep learning resurges the interest in AI and they are now considered as interchangeable terms in some fields.

The number of deep learning patents per year indicates that DL is at the rising stage of the S-curve of technology life cycle. Before 2000, there is no such patent. The number of deep learning patents start to increase since 2013. For detailed patent numbers, Figure 1 shows a coordinate plane where the X axis denotes the number of different assignees while the Y axis denotes the number of patents per year. The plane is a kind of TLCD in which the year line will up-wards curve back to the original point if the technology life cycle is finished. The S curve and TLCD diagram show that the DL technology seems to shift from the emerging stage to the growth stage at around the year of 2015.

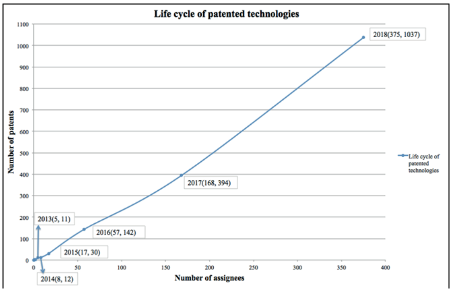


Figure 1. Numbers of different assignees (X axis) vs number of patents (Y axis) per year.

For co-word analysis, patents were automatically grouped into topics by the Multi-Stage Clustering (MSC) procedure provided in CATAR. For succinct and interpretable results, only the patent titles are used for co-word clustering in this study. The clustered topics were then further visualized in a 2-dimensional map based on Multi-Dimensional Scaling (MDS) to show their relatedness. Finally, 16 salient topics emerge from the clustering.

Figure 2 shows the visualized topics, where the size of the circle denotes the relative number of patents in the corresponding topic and the distance between the circle centers denotes the relatedness between the topics. The topic descriptors (terms) shown in Figure 2 are derived based on the occurrence of the terms appear inside and outside of the topic.

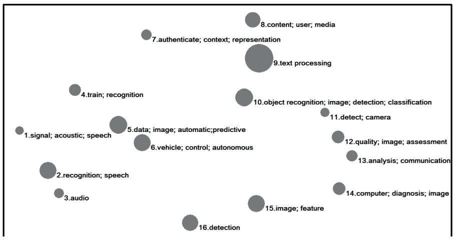


Figure 2. Topic map of the DL patents.

Topic 1, 2, and 3 are about speech recognition and audio signal processing. This can be verified by their common cited sources: ICASSP (International Conference on Acoustics, Speech, and Signal Processing) and IEEE SAP (IEEE Trans. on Speech and Audio Processing). The major assignees are non-surprisingly Google, Microsoft, Amazon, and IBM.

Topic 4 is about sequence training and recognition. Topic 5 is about data, image, and predictive techniques, where SAS Institute has most patents in this topic. Topic 6 focuses on autonomous vehicle control and unexpectedly the Israel company Mobileye and the State Farm Mutual Automobile Insurance (SFMAI) company own the highest numbers of patents (instead of Tesla or other automobile companies).

From the left hand-side of the topic map, if we imagine autonomous vehicle as an integrated application of various techniques, it would involve audio, image, data, and sequence processing, recognition, and prediction.

Topic 7 is about generating representation or context for various contents. Topic 8 is about content, user, and media. Topic 9 covers many sub-topics where text and linguistic information are major ones.

Topic 10 focuses on object recognition. Topic 11 covers two sub-topics, one for detecting physiological responses or objects, the other about camera systems, in which the major assignee Facense is an Israel company who develops smart glasses with tiny thermal and CMOS sensors.

Topic 12 is about assessment of images or medical images, in which the German company Siemens Healthcare owns 10 out of 41 DL patents. Topic 13 includes two sub-topics: communication and scene analysis. Topic 14 is about computer diagnosis and image processing, in which the South Korean company Samsung has 8 patents. Topic 15 is about image feature processing. Topic 16 covers detection of various objects, including malware, traffic, pedestrian, events, keywords, accidents, etc.

The 1636 patents actually cover many applications and topics. At first, we had tried to analyze the Cooperative Patent Classification (CPC) and International Patent Classification (IPC), but failed to recognize concrete topics due to the vague descriptions of the CPC/IPC categories. In comparison, our topic map analyses not only complemented the vagueness of the CPC or IPC analyses, but also revealed the relationship among these topics.

Conclusions

A set of deep learning patents from USPTO has been analyzed with a freeware CATAR developed by the authors. Our results show that DL patents are increasing in an extremely fast pace and major players in this technology not only include those AI partnership companies, but also those with niche technology or market. Our analyses imply that it will impact more industrial sectors in the future.

Acknowledgments

This work is supported in part by the "Institute for Research Excellence in Learning Sciences" of NTNU from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) and the MOST projects in Taiwan.

References

European Commission. (2018). Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines. http://europa.eu/rapid/press-release_IP-18-3362_en.htm

Gao, L., Porter, A. L., Wang, J., Fang, S., Zhang, X., Ma, T., . . . Huang, L. (2013). Technology life cycle analysis method based on patent documents. Technological Forecasting and Social Change, 80(3), 398-407.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436.

Sharma, P., & Tripathi, R. C. (2017). Patent citation: A technique for measuring the knowledge flow of information and innovation. World Patent Information, 51, 31-42.

Trippe, A. (2015). Guidelines for Preparing Patent Landscape Reports. https://www.wipo.int/edocs/pubdocs/en/wipo_pub_946.pdf

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text Mining Techniques for Patent Analysis. Information Processing and Management: An International Journal, 43(5), 1216-1247.

Are corresponding authors reflecting collaboration’ degree in interdisciplinary program such as Cancer Bioinformatics?

Pauline Couffignal[1] and Philippe Gorry[2]

[1] *couffignalpauline@gmail.com*
Faculty of Sociology, University of Bordeaux, 3ter Pl. Victoire, 33076, Bordeaux, (France)

[2]*philippe.gorry@u-bordeaux.fr*
GREThA, UMR CNRS 5113, University of Bordeaux, Av. Leon Duguit, 33608, Pessac, (France)

Introduction

The diffusion of knowledge is a complex process facing numerous barriers, and interdisciplinarity is believed to be a key issue in translational research (Marcovich & Shinn, 2012). This fact is particularly valid in cancer research. In the era of big data, the sharing of tumour databases and new developments in genomics research, the use of bioinformatics in cancer research is becoming unavoidable.

Purpose

The aim of this study is to get a better understanding of collaboration degree in “Cancer Bioinformatics” research using a bibliometric approach (Rafols & Meyer, 2010) and focusing the analysis on corresponding authors (Mattson, 2011).

Methods

Literature search on “Cancer Bioinformatics” was run in Scopus database up to 2018. Search query was designed according to Glänzel et al (2009) and Liu & al. (2014). We extracted from the resulting corpus (n=19,425 documents) one paper by different corresponding authors (n=11,232). We drew a random sample of 824 documents, giving us a 99% confidence level and a margin of error between 4 and 5%. From this sample, we identified and categorized each authors’ profession and each source in two groups “Computer Science” and “Health Science” using various methods (affiliation, Scopus author profil, author website, CV, ...). Bibliometric and descriptive statistics analyses were performed. The journal impact factor (IF) was based on the indicator (SJR) taken from SCImago®. The co-occurrence of words was calculated for each group of authors ("Computer scientists" or "Health scientists"), or journals ("Computer science" and "Health sciences") using VOSviewer® network analysis software. An analysis of cited references for each groups’ publication set were run using CRexplorer®.

Results

Corresponding author's profession analysis shows that there are less “computer scientists” who signed “Cancer Bioinformatics” publications than "Health scientists" (Table 1). This category represents 75.5% of the corresponding authors in our sample.

Table 1. Descriptive statistics of “Cancer Bioinformatics” corresponding authors' publications random sample.

	Corresponding authors’ groups		
	Health Sc.	**Comput. Sc.**	**Total**
Nb articles	619	201	820
%	75,5	24,5	100
Mean nb of authors/pub.	4	4	4
Mean nb of affiliations /pub.	8	6	8
Mean nb of affiliations/author /pub.	2.03	1.69	1.97
Mean SJR 2017	2.551	2.972	2.654
Max SJR 2017	34,.896	25.137	34.896
Min SJR 2017	0,.02	0.109	0.102
STD. SJR 2017	3.503	3.871	3.599

There is no correlation between the publication year and the distribution of the two corresponding authors’ groups. “Computer scientists” collaborate with as many people as "Health scientists". The latter publish in journals with an average SJR slightly lower than the "Computer scientists".
The United States are leader in “Cancer Bioinformatics” research field in terms of publications number and its corresponding authors distribution follows the overall results. India is the only country where “Computer scientists” are almost equally represented as corresponding authors in regard of the “Health scientists” (data not shown). “Cancer Bioinformatics” publications have been published in 444 different journals. There is a majority (81.9%) of “Health Science” journals (n=383) compared to “Computer Science” journals (n=61) (data not shown). Corresponding authors publish mainly in journals specialized in their respective research fields. it is important to emphasise that top 10 “Health Science” journals of

have an average SJR in 2017 of 2.962 whereas the average SJR of "Computer Science" journals are 3.207 (data not shown).

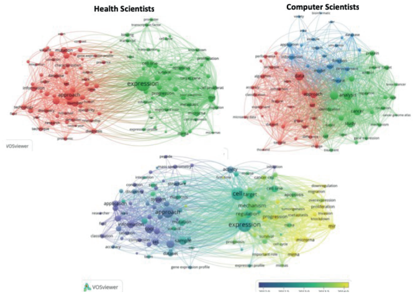
"Health scientists" corresponding authors principally publish in "Health Science" journals rather than in "Computer science" journals (Table 2). "Computer scientists" corresponding authors, publish almost equally in the both journals. For both corresponding authors' groups, the average impact factor is higher when corresponding authors publish in their non research field journals.

Table 2. Distribution of articles according to corresponding authors' and journal categories

	Corresponding authors groups	Health scientists	Computer scientists
Health Science Journals	Nb	565	100
	%	91,3	49,8
	Mean SJR 2017	2,519	3,503
Computer Science Journals	Nb	54	101
	%	8,7	50,2
	Mean SJR 2017	2,883	2,446
Total	Nb	619	201
	%	100	100
	Mean SJR 2017	2,551	2,972

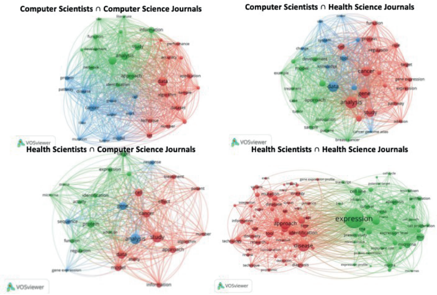
Co-occurrence of words analysis of publications in the two categories shows that there is a homogeneous structure of words used by "Computer scientists" whereas "Health scientists" used two types of knowledge. It can be explained by co-occurrence words over the time analysis which demonstrates two specific waves of knowledge before and after 2013-2014 (Figure 1).

Figure 1. Topic network maps by corresponding authors' groups, and by publication year.



Analysis of co-occurrence of words in both categories of journals by both corresponding authors' categories shows that "Computer Scientists" use a homogeneous structure of words whatever the journal category. However, "Health scientist" borrow the knowledge structure of "Computer scientists" when they publish in "Computer science" journals category (Figure 2).

Figure 2. Topic network maps by corresponding authors & journal categories.


Computer Scientists ∩ Computer Science Journals
Computer Scientists ∩ Health Science Journals
Health Scientists ∩ Computer Science Journals
Health Scientists ∩ Health Science Journals
expression
VOSviewer
VOSviewer
VOSviewer
VOSviewer

Discussion

Our preliminary results support that there are less computer scientists invested in "Cancer Bioinformatics" corresponding authorship than physicians or biologists. These last ones publish more in their field journals whereas "Computer scientists" equally publish in both journal categories. The structure of knowledge used by the two corresponding authors' groups is clearly different. The choice of journals to value their research seems to be the result of publication strategies that are independent of the author's areas of professionalization. These results reinforce our understanding of interdisciplinary program and corresponding author analysis is a valuable indicator of collaboration degree in interdisciplinary research.

References

Glänzel, W., Janssens, F. & Thijs, B. (2009). A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics, *Scientometrics,* 79, 109-129.

Liu, A.Y., Li, S.Y., & Guo, Y.Q. (2014). Characteristics of research on bioinformatics in China assessed with Science Citation Index Expanded, *Scientometrics*, 99, 371-391.

Marcovich, A. & Shinn, T. (2012). Regimes of science production and diffusion: towards a transverse organization of knowledge. *Scientiae Studia*, 10, 33-64.

Mattsson, P., Sundberg, C.J., Laget P. (2011). Is correspondence reflected in the author position? A bibliometric study of the relation between corresponding author and byline position. *Scientometrics,* 87, 99–105.

Rafols, I. & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82, 263-87.

Exploring the Lotka's Phenomenon in Sense Complexity of English Word

Si Shen[1], Hao Sun[2] , Zihe Zhu[3] and Dongbo Wang[4]

[1]*shensi@njust.edu.cn*
Nanjing University of Science and Technology(China)

[2]*2117107010889@njust.edu.cn*
Nanjing Agricultural University(China)

[3]*zihe.zhu@qq.com*
Nanjing University of Science and Technology(China)

[4]*db.wang@njau.edu.cn*
Nanjing Agricultural University(China); KU Leuven(Belgium)

Introduction

Lotka (1926) proposed the Lotka's law, which states that in a certain period of time, the percentage of the authors publishing x articles to all the authors is inversely proportional to the square of x. We observe an intriguing relationship between English words' sense complexity and the amount of words, which can be seen as an analogy to the Lotka's law. Here, this phenomenon regarding the distribution of words in sense is known as Lotka's phenomenon. This discovery suggests that in addition to the Zipf's law describing the rank-frequency distribution of words (1949), there exist another rule governing the distribution of sense of English word. On this basis, we will help the linguistic researchers understand more thoroughly the sense distribution of the English language(McCarthy, 2016). Besides, the findings can be used to promote the building of English corpus for natural language processing as well as semantic knowledge mining. For example, according to English words' sense complexity, we can know how many words are ambiguous on the semantic level. They provide the basis for semantic analysis of sentences after combining the words into sentences.

Data source

The five dictionaries used as corpus are New Oxford Dictionary of English, Cambridge Dictionary, Collins COBUILD Dictionary of English, Oxford Advanced Learner's Dictionary (7th Edition), and Longman Dictionary of Contemporary English (4th Edition). These dictionaries contain English words and their English senses. Detailed information of the five dictionaries is presented in Table 1.

Table 1 Basic data of five dictionaries

Dictionary	Language	Word amount	Number of senses	Mean number of senses per word
New Oxford Dictionary of English	English	91379	184479	2.02
Cambridge Dictionary	English	36521	67936	1.86
Collins COBUILD Dictionary of English	English	37831	68105	1.80
Oxford Advanced Learner's Dictionary (7th Edition)	English	123811	183517	1.48
Longman Dictionary of Contemporary English (4th Edition)	English	108824	158258	1.45

Results and discussion

This Fig.1 shows the numbers, proportions in all the papers, temporal trend lines of single-country papers, two-country collaboration papers, three-country collaboration papers, collaborative papers with four and more countries, domestic collaboration papers and international collaboration papers in the field of artificial intelligence. Here, we define English words' sense complexity as follows: In a given dataset U, if the total number of senses that a word A has is n, then the sense complexity of A in U is n. Sense complexity is then calculated as the following example. In Cambridge Bilingual Dictionary, the word "absurd" has two senses, so the sense complexity of "absurd" is 2. The word "adjust" has three senses, namely, "to change something so that it works better", "to make familiar with something" and "to change in order to work or do better in a new situation". Therefore, the sense complexity of “adjust” is 3. According to the above definition, the word “adjust” is more complex than the word “absurd”.Based on the whole results, the words are placed in rows and senses in columns, and the words- number of senses table is generated.

$$t_i = \begin{cases} 1 & \text{(The number of senses is i)} \\ 0 & \text{(The number of senses is not i)} \end{cases}$$

$$a_i = \sum t_i$$

According to the definition of words’ sense complexity, s sense complexity ai is calculated using formulas. The words having i senses are calculated and the sense complexity of the i-th word is obtained. Words having the same sense complexity are grouped into one category. The table of words’ sense complexity showing the relationship between sense complexity and the number of words is generated. The complexity is arranged in a decreasing order of the word amount. When performing data analysis of words’ sense complexity and unary linear regression analysis, the data on the category comprising less than 20 words is removed.

Although the Lotka’s law does not directly apply to the relationship between words’ sense complexity and word amount, an analogy can be made to Lotka’s law. This phenomenon is made even more pronounced after taking the logarithm of the word amount. So it is called Lotka’s phenomenon regarding the words’ sense complexity. An apparent linear tendency can be observed by taking the logarithm of the word amount that leads to a linear scatter plot, as shown in Fig. 1.

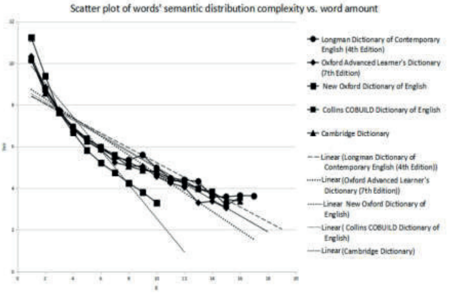


Fig. 1 Scatter plot of words’ sense complexity vs. word amount

Fig. 1 shows the scatter plots of sense complexity vs. logarithm of the word amount (ln) for five dictionaries. X is sense complexity, and lnn is the logarithm of the word amount. A linear tendency can be observed from the Fig.4, which is basically consistent with the graphic representation of Lotka’s law. Below is a unary linear regression analysis between sense complexity and logarithm of the word amount. Data of the five dictionaries are processed using SPSS and the correlation coefficients of unary linear regression are calculated in Table 2.

Table 2 Unary linear regression between sense complexity and word amount for the five dictionaries

	Longman Dictionary of Contemporary English (4^{th} Edition)	Oxford Advanced Learner's Dictionary (7th Edition)	New Oxford Dictionary of English	Collins COBUILD Dictionary of English	Cambridge Dictionary
Word amount	37373	42480	91294	37831	36521
Number of senses	69655	69428	115380	68105	67936
constant	8.771	9.193	10.671	8.925	8.805
coefficient	-0.939	-0.954	-0.959	-0.947	-0.943
R Square	0.881	0.910	0.920	0.897	0.889
F sig	0	0	0	0	0
T sig	0	0	0	0	0

As seen from Table 2, R square is 0.881, 0.910, 0.875, 0.897 and 0.889 for the five dictionaries, respectively. Based on Table 5, the linear regression equations for the five dictionaries are lnn=-0.939x+8.771, lnn=-0.954x+9.193, lnn=--0.959x+10.671, lnn=-0.947x+8.925 and lnn=-0.943x+8.805, respectively. The relationships are of high significance based on F sign and T sign. Combining Fig. 1 and Table 2, Lotka’s phenomenon in words’ sense complexity is manifested in each dictionary, though to varying extent.

Conclusion

We calculate the words’ sense complexity based on five English dictionaries and present the definition of words’ sense complexity. After generating the data table of words’ sense complexity, we perform a thorough analysis on the words’ s sense complexity. The average words’ sense complexity of the five dictionaries is calculated. By making an analogy to Lotka’s law in bibliometrics, the relationship between words’ sense complexity and whole vocabulary is called Lotka’s phenomenon and fitted by unary linear regression. This relationship is of statistical significance based on the results.

Acknowledgments

Researches are supported by the National Natural Science Foundation of China (Grant Number: 71673143).

References

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington academy of sciences*, 16(12), 317-323.

McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word sense clustering and clusterability. *Computational Linguistics*, 42(2), 245-275.

Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Cambridge MA edn. Reading: Addison-Wesley.

International collaboration in the field of artificial intelligence: global trends and networks at the country level

Haotian Hu[1], Dongbo Wang[2] and Shuiqing Huang[3]

[1] *hhtdlam@126.com*
Nanjing Agricultural University(China)

[2] *db.wang@njau.edu.cn*
Nanjing Agricultural University(China); KU Leuven(Belgium)

[3] *sqhuang@njau.edu.cn*
Nanjing Agricultural University(China)

Introduction

In order to gain a deeper understanding of international collaboration in the field of artificial intelligence (AI), by combining the bibliometric analysis with social network analysis, this paper investigated the international collaboration trends, networks, and core groups of artificial intelligence at the country level (Larivière et al. 2015). This study selected 26,808 papers from the Web of Science (WoS) core collection database from 1985 to 2017. The results show that 84% of the papers in the field of artificial intelligence belong to joint publications. Two-country papers are the main patterns of international collaboration at the country level (Leydesdorff et al. 2013). Through social network analysis, this paper found that the country collaboration network in the global AI field has reached a certain maturity in past 33 years.

Data source

Unlike other research fields, a considerable portion of papers in the field of artificial intelligence are published as conference papers, so the artificial intelligence field dataset used in this study is a collection of journal papers and conference papers from the Web of Science (WoS) Core Collection Database with the search term of 'Artificial Intelligence'. Not only does this database have high impacts, but also journal papers and conference papers included in this database, such as 'Expert Systems with Applications', 'Ai Magazine', 'Artificial Intelligence', and 'Proceedings of the 28th Chinese Control and Decision Conference (2016 CCDC)', have high prestige in the field of artificial intelligence. The data used in this study was downloaded from Thomson Reuters' web of knowledge (WoK). The entire dataset can be downloaded in batches in text format, and up to 500 papers in each batch. Through manual download, this research has downloaded a total of 30,156 records. After analyzing the downloaded data, this paper found that some of the publications do not have the author's address information field (C1). The dataset collection work was completed on November 4, 2017. After deleting the problem records, the resulting dataset contains 26,808 papers. In order to obtain collaboration information, the citation data tags selected in this paper include publication year (PY), source publication (SO), author's full name (AF), author's address information (C1), group author (CA). In order to fully understand the global collaboration in the field of artificial intelligence, this paper used Python3.6.2, Microsoft Excel 2016 for Windows, Matlab 2011b and Pajek 5.01 for data extraction and analysis, the establishment of collaborative networks, and the calculation of various indicators.

Results and discussion

This Fig.1 shows the numbers, proportions in all the papers, temporal trend lines of single-country papers, two-country collaboration papers, three-country collaboration papers, collaborative papers with four and more countries, domestic collaboration papers and international collaboration papers in the field of artificial intelligence.

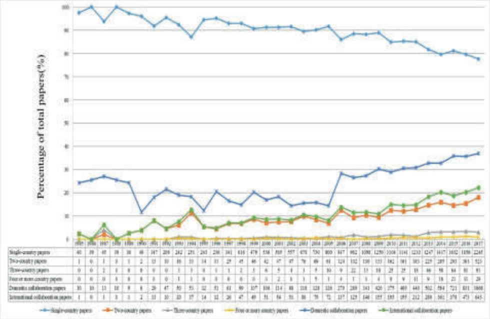


Fig. 1 The temporal trend lines of collaboration at the country/region level in artificial intelligence field

The statistical results show that in the global AI field, the proportion of domestic collaboration papers is 28.1%, while the proportion of international collaboration papers is 14.6%, which is only half of the proportion of domestic collaboration papers. As shown in Fig. 1, the collaboration papers in the global artificial intelligence field have shown an upward trend between 1985 and 2017. In 1985, the proportion of collaboration papers accounted for 26.8% of the total number of papers, while the proportion of collaborative papers rose to 59.2% by 2017. Accordingly, the proportion of non-collaborative papers dropped from 73.2% to 40.8% over the past 33 years.

The social network analysis method is used to explore the international collaboration network in the global AI field from the macro and micro perspectives. According to the above method, excluding the time factor of collaboration, this paper found that there are 133 countries and regions in the world participating in international collaboration between 1985 and 2017. After further analysis, it is found that the weights of a large number of edges in the entire network are equal to 1, indicating that the collaboration intensity between many countries is relatively low. In order to obtain a stable and centralized network, edges with weights less than 3 and isolated nodes were deleted. The resulting country collaboration network contains 74 nodes and 397 edges. For easy reading, this paper defines country collaboration network as CCN. The CCN network diagram was drawn using the Pajek software and the Kamada-Kawai algorithm. As shown in Fig. 2 below, the nodes with the same color represent they have the same degree.

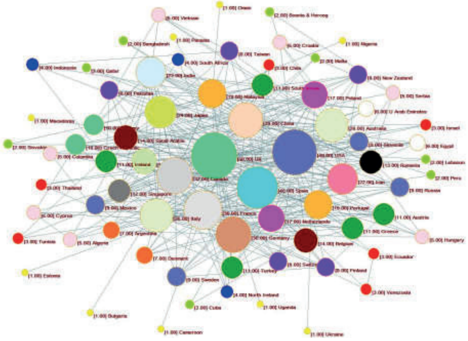


Fig. 2 The country collaboration network(CCN) in the global artificial intelligence field between 1985 and 2017

Through the analysis with Pajek software, this paper found that the average degree of CCN nodes is 10.7, indicating that each country or region has an average of 10-11 partners. As can be seen from Fig. 2, the largest component of the entire network covers all countries and regions involved in the collaboration. This phenomenon indicates that the CCN has reached a certain maturity over the past 33 years. As can be seen from the size and color of the nodes, USA and UK are more notable than other countries and they are the hubs of the CCN network. Other countries with a broader partnership include Spain, France, Canada, Germany, Italy, P. R. China, Australia. These countries dominate the entire CCN network. While countries such as Lebanon, Oman, Qatar, Slovenia, Panama are located in the periphery of the network. In these countries, for example, Algeria has 25 collaboration publications in the CCN. However, because it only collaborates with France (19 collaboration publications) and the Cyprus (6 collaboration publications), it is also placed in the periphery of the network.

Conclusion

For the given field of AI, this paper not only reveals the upward trend of international collaboration, but also reveals the characteristics of the discipline. First, as an interdisciplinary field, the international collaboration degree of artificial intelligence is between natural science and humanities/social science (Haddow et al. 2017). This finding confirms the interdisciplinary nature of artificial intelligence from the perspective of scientific collaboration. Second, research results at the country and institutional levels suggest that the CCN has reached a certain maturity degree in the global AI field over the past 33 years. In the future, we will capture the dynamic changes in international collaboration over a long time span. In addition, we will track and investigate subtle patterns and author orders in international collaboration at the institutional level, as well as partnerships between country and institutional scientific collaboration.

Acknowledgments

Researches are supported by the National Social Science Foundation of China and National Natural Science Foundation of China (Grant Number: 15ZDB127 and 71673143).

References

Haddow, G., Xia, J., & Willson, M. (2017). Collaboration in the humanities, arts and social sciences in Australia. *Australian Universities' Review*, 59(1), 24.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: collaboration and scientific impact since 1900. *Journal of the Association for Information Science & Technology, 66*(7), 1323–1332.

Leydesdorff, L., Wagner, C., Han, W. P., & Adams, J. (2013). International collaboration in science: the global map and the network. *Profesional De La Informacion*, 22(1), 87-94.

Visualizing gender representation by field of research at institutions in the United Kingdom

Hélène Draux[1], Simon Porter[2], Ricarda Beck[3], Suze Kundu[4] and Stacy Konkiel[5]

[1] *h.draux@digital-science.com*
Digital Science Consultancy, 90 York Way, London, N1 9AG (United Kingdom)

[2] *s.porter@digital-science.com*
Digital Science, 90 York Way, London, N1 9AG (United Kingdom)

[3] *r.beck@digital-science.com*
Digital Science, 90 York Way, London, N1 9AG (United Kingdom)

[4] *s.kundu@digital-science.com*
Digital Science, 90 York Way, London, N1 9AG (United Kingdom)

[5] *s.konkiel@digital-science.com*
Digital Science, 2nd Floor, 625 Massachusetts Ave, Cambridge, MA, 02139 (United States)

Introduction

The gender imbalance in academic publishing practices has been widely documented. In many disciplines, women publish less often than their male counterparts (Kyvik, 1990; Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013; Stack, 2012; West, Jacquet, King, Correll, & Bergstrom, 2013). Male authors have also been found to hold more prestigious first and last author positions (Larivière et al., 2013; West et al., 2013). Fewer than 6% of countries have achieved gender parity in terms of number of papers published (Larivière et al., 2013).

How wide is the STEM gender gap within UK research institutions specifically, and how does this compare to the arts and humanities? Have initiatives implemented to address this gender imbalance been successful? To answer these questions, we created an interactive tool to visualise a vast array of UK authorship, discipline, and institutional data in terms of gender. We describe the capabilities of the tool using a single UK research institution (University College London) and the field of Education research as case studies.

Methodology

Data collection

We queried the interlinked research information database Dimensions[1] for publications written between 2012 to 2017 that had at least one author affiliated to UK-based institution (N = 302,000). The resulting authors institutional affiliation information was disambiguated and normalized using GRID identifiers[2].

Data analysis by gender

Given a first name as input, the gender-guesser Python package[3] categorizes the name as male, female, or 'unknown'.

Using the gender-guesser tool on our sample, we found that 47.5% (n = 143,502) researchers had first names that were mostly or very likely male names, 32.6% (n = 98,666) researchers had first names that were mostly or very likely female names, and the remaining 19.9% (n = 60,245) could not be identified either way (i.e. "androgynous" or "unknown").

Data analysis by subject area

Dimensions classifies published research according to Field of Research (FOR) codes. Research outputs can be assigned multiple FOR codes at once. For this study, we have used the first and broadest level of research area classification to compare the proportion of each gender in these fields. FOR coverage varies between fields; we compensate for this variance by aggregating FOR codes at the researcher level.

[1] https://www.dimensions.ai
[2] https://grid.ac/
[3] https://pypi.org/project/gender-guesser/

Visualising the data

Using the study data, we built an Interactive Data Visualizer tool[4]. The tool allows the user to search by institution or by field of research (FOR) and displays gender information based on authors' first names as categorised by gender-guesser. We use data associated with University College London (UCL) and the Education field of research to illustrate the capabilities of the Visualizer tool.

Results

Psychology and Cognitive Science at UCL has the highest percentage of women researchers (49%). 34% of UCL researchers in that discipline have a male name, and 9% are of unknown gender. Following Psychology and Cognitive Science, UCL's most gender balanced fields of research are Language and Communication (49% female researchers), Education (47%), Environmental Sciences (44%), Law and Legal Studies (42%), and Studies in Human Society (42%).

Education research at UCL

A scatter graph within the Visualiser illustrates the breakdowns of researchers' gender within various subject areas across UCL. Of Education researchers at UCL, 47% are women and UCL is in the 100th percentile of Education research institutions in the UK, meaning that no institutions carrying out research in the field of Education have a greater number of women publishing than UCL.

Education research across institutions

An alternative means of analysing the Education field of research in the Visualiser is to search for the subject area across all UK institutions. In total, there are 65 UK institutions carrying out research in Education.

Overall, the field of Education has a good representation of women researchers, with women outnumbering men at more than twenty UK research institutions. The distribution of institutions does however indicate that not all Education research institutions have more women than men, as a histogram visualization tails off to the left with 20 institutions having below 40% female researcher representation in the subject.

Gender balance in UK research, across all disciplines

The gender balance trend seen in UK Education research is not reflected in Science, Technology, Engineering and Maths (STEM) subjects, which have a lower representation of women researchers. Across the UK, representation of researchers with female names is only greater than male names in Psychology and Cognitive Sciences, Language, Communication and Culture, and Education. Arts and Humanities subjects tend to have a higher representation of women researchers than do STEM subjects. However, the Arts and Humanities do not achieve gender parity.

STEM subjects such as Physical Sciences, Technology and Information, and Computing Sciences display distributions across all institutions that do not extend beyond more than 40% women, with both Technology and Physical Sciences showing peak representation of women in these subjects between 5% and 15%.

Subjects such as Medical and Health Science and Psychology and Cognitive Sciences have a wide distribution range of representation of women within their fields of research, with the former subject spanning zero representation of women all the way up to 70% representation.

There is also an interesting trend in the percentage of researchers that could not be identified by the gender-guesser. This phenomenon seems to increase as the percentage of women decreases in a field, and as we move from the Arts and Humanities towards the STEM subjects.

Future work

Future work will include the creation of interactive data analysis and visualization tools for other countries, updates to the existing tool's data over time, and improvements in the capability of "guessing" gender for non-Western names.

References

Kyvik, S. (1990). Motherhood and Scientific Productivity. *Social Studies of Science*, *20*(1), 149–160. https://doi.org/10.1177/030631290020001005

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213. https://doi.org/10.1038/504211a

Stack, S. (2012). Gender, Children and Research Productivity. *Research in Higher Education*, *45*(8), 891–920.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PloS One*, *8*(7), e66212. https://doi.org/10.1371/journal.pone.0066212

[4] https://www.digital-science.com/gender-representation-in-research-tool/

Changing Dynamics in an Emerging Field: Tracking Authorship Developments in the Journal ‘Political Psychology’ 1985-2015

Sabrina J. Mayer[1] and Justus M. K. Rathmann[2]

[1] *sabrina.mayer@uni-due.de*
Institute of Political Science, University of Duisburg-Essen, Lotharstrasse 65, 47057 Duisburg, Germany

[2] *rathmann@soziologie.uzh.ch*
Institute of Sociology, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland

Introduction

The field of political psychology focuses on the explanation of political phenomena by using psychological theories and instruments. Researchers often rely on established psychological concepts from social, cognitive and personality psychology and apply them to the explanation of political issues (Houghton, 2014). Political psychology combines primarily political science and psychology, but also uses elements from sociology, social anthropology, and history (Houghton, 2014).

Although political psychology as such has only been institutionalized in the 1970s, when the International Society for Political Psychology (ISPP) was founded, the sub-discipline emerged predominantly in the United States and Europe in the late 1940s (e.g. Polo et al., 2015). In the last two decades, the field of political psychology has become increasingly popular, 642 journal articles indexed in Web of Science (WoS) contain ‘political psychology’ in the abstract, title or author keywords. About half of the articles each are classified as belonging to the WoS categories for psychology and political science (with International Relations). Most articles were published in the journal *Political Psychology*, which has been founded by the ISPP in 1979. Today it is among the top-20% journals in political science as well as social psychology.

Even though the field of political psychology shows a very positive trend for publication number development (Krampen, von Eye & Schui, 2011), it is still deemed an emerging field and bibliometric analyses of the discipline are scarce (e.g. Houghton, 2014). The only other study, that already took the journal *Political Psychology* in its focus, analyzed plainly download, submission, and citation numbers without focus on the temporal dimension, team dynamics and diversification. Our contribution focuses on trends in the field of political psychology and observes authorship developments from the beginning in 1979 up to 2015. We ask if these changes mirror developments in most disciplines (Waltman, 2012).

Data & Methods

An in-house data base of the WoS is used. We queried all items published in the journal *Political Psychology* between 1985 and 2015. However, the publication years 1979-1984 are missing in the data base and will be added manually at a later point. This results in a data set of 1,830 documents of which 1,011 are original articles in the time-span 1985 to 2015. Additionally, to determine the gender of authors, we applied a gender identification algorithm based on the names of authors. However, for the years before 2006, we mostly have initials and not full names. The algorithm identified 935 observations from 2006 to 2015, of which 801 unique authors of 435 unique articles were identified. A full set of first and last names for earlier publications will be added to the data set in the course of the research project.

We use US Social Security Administration data available in the R package ‘gender’. Additionally, we use the package ‘gender.c’ to improve the identification algorithm for names only common in Europe. First names are only classified automatically if they were given to a single gender in 95 percent of cases in 1970.

Preliminary Results

Teams play an increasingly important role in the production of many scientific disciplines (Lariviere et al., 2014). However, top journals in political science and psychology vary when it comes to the development of authorship numbers: whereas in political science the average number of authors has increased from 1.5 to 1.9 the last 25 years (and the share of single authors dropped from 63% to 41%), numbers are higher in psychology (2.5 to 4.2 mean authors, 21% to 5% single authors) (Mayer, 2016).

The share of articles by single authors and teams follows a clear trend. While in 1985 almost 100% of articles were written by single authors, in 2015, this share has decreased to less than 50%. The proportion of articles written by a team may be volatile, but the trend is consistent. The number of authors per article ranges from 1 to 10 and has a mean of 1.55. In this sense, *Political Psychology* shows more similar developments to the field of political science.

The average size of author teams in *Political Psychology* increased by 50% from 2 in 1985 to 3 in 2015. A simple linear regression of the publication year onto the team size supports the assumption that teams increase in size (β_{Year} = 0.022; t = 4.134).

When articles are written in collaboration, the rewards of the effort have to be shared among the contributors. In science, this is done mostly by the order of the author names in the article head. Basically, there are two ways to order the names; Author names can be ordered alphabetically, or in a non-alphabetical way, where usually author names are ordered by the amount of contribution to the article (Rauhut, Winter & Johann, 2018). Recently, the share of contribution-based authorship order has increased in most disciplines, but varies: In political science, still approximately 60% of the publications of teams are ordered alphabetically, whereas in psychology, this share is now below 49% (Waltman, 2012). Over the entire observational period only about 36% of the publications of teams are ordered alphabetically. Although this value is relatively volatile in some years, the trend is very steady. Thus, *Political Psychology* clearly differs from the norm in political science and authorship trends more tend towards the field of psychology.

With a Journal Impact Factor (JIF) of 2.089 in 2015, *Political Psychology* is among the top 20 journals in political science. Publications in high-impact journals are particularly important for career progression and the acquisition of third-party funding. However, ceteris paribus, female scholars tend to publish less than their male counterparts, especially in high-impact journals (Mayer & Rathmann, 2018). The development of women's involvement in publications in *Political Psychology* from 2006 to 2015 as the proportion of publications without the participation of at least one female scholar has some outliers, but fluctuates around 30%. The proportion of female authors also remains constant over the years. A linear regression shows that there is no statistically significant effect over time (β_{Year} = 0.008; t = 1.292).

Preliminary Conclusion

These preliminary results support the trend that science nowadays increasingly takes place in teams and that these teams are becoming larger. The share of single authors in the sample is continuously decreasing and the average team size is constantly increasing. Working in a team offers many advantages for scientists but can also create problems; scientists become dependent and are exposed to social team dynamics.

The number of alphabetically ordered articles in *Political Psychology* is clearly below average compared to political science. However, a changing trend is not discernible, even though the proportions fluctuate. When interpreting the data from the 1980s, it should be noted that there were still very few articles by teams, so these few articles are therefore particularly influential. Especially with small teams the 'illusion of equality' comes into play (Rauhut, Winter & Johann, 2018), i.e. articles in alphabetical order do not have to be intentionally ordered alphabetically. The probability that non-intended alphabetically ordered articles are alphabetically ordered decreases exponentially with the number of authors. In political science, small teams are the norm, therefore a correction factor should be included in further research.

The proportion of female scientists publishing in *Political Psychology* could so far only be investigated for the period 2006 to 2015. Although the proportion of women among the authors in the journal has increased slightly, this growth is not statistically significant. Still, almost 70% of the authors in *Political Psychology* are male

Acknowledgments

Justus Rathmann is funded by the Swiss National Science Foundation (SNSF) under the SNSF Starting Grant "CONCISE" (S-64408-01-01). The authors are ordered alphabetically, both authors contributed equally to this research.

References

Houghton, D. P. (2014). *Political Psychology*. Routledge.

Krampen, G., von Eye, A. & Schui, G. (2011). Forecasting trends of development of psychology from a bibliometric perspective. *Scientometrics*, 87(3), 687-694.

Larivière, V., Gingras, Y., Sugimoto, C.R. & Tsou, A. (2014). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, 66(7), 1323-1332.

Mayer, S. J. (2016). Trends and developments in multi-authorship in five social science disciplines from 1991 to 2014. In I. Ràfols, et al. (Eds.), *Peripheries, frontiers and beyond. Proceedings of the 21st International Conference on Science and Technology Indicators*, 924-933.

Mayer, S. J. & Rathmann, J.M.K. (2018). How does research productivity relate to gender? Analyzing gender differences for multiple publication dimensions. *Scientometrics*, 117(3), 1663-1693.

Polo, L., Godoy, J.C., Imhoff, D. & Brussino, S. (2015). Following the tracks of an emerging area: Bibliometric analysis of Latin American political psychology in the 2000-2010 period. *Universitas Psychologica*, 13(5), 2047-2057.

Rauhut, H., Winter, F. & Johann, D. (2018). Does the winner take it all? Increasing inequality in scientific authorship. *Emerging Trends in the Social and Behavioral Sciences*, 1-14.

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, 6(4), 700-711.

Measuring the scientific publications of top universities from Mainland China

Fangfang Wei[1*], Guijie Zhang[2] and Jianben Wu[3]

[1] *weifftju@163.com*
Business School, University of Jinan, Jinan 250002, China
*Corresponding author

[2] *zgjzxmtx@163.com*
School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, China

[3] *hiteriter@gmail.com*
School of Data Science, City University of Hong Kong, Kowloon, Hong Kong 999077 (China)

Introduction

As the world's most populous country and the second largest economy, China's progress in science and technology has received the attention of many scholars (Ma and Li, 2018). China's booming economy has been attributed in part to a conscientious effort to improve the competitive power of China's universities through a massive of programs, the most famous one is the 985 Project (Zhang, et al., 2013). 985 Project was put forward by the ministry of education of China in May 1998, with the purpose of promoting the development of education in China and strengthening global positions of leading Chinese universities (Guskov, et al., 2018). Chinese government has been spending increasingly more funding on research and development (R&D) activities of universities and China's total expenditure on R&D has increased by 23% a year on average over the past decade (Qiu, 2014). Since 985 Project universities are the best among Chinese universities, it is necessary to conduct a comparative analysis to measure their research performance with international standards for the purpose of assessing the quality.

Data Collection

This study used the data derived mainly from the Web of Science Core Collection database. The current time range of data statistics is from 2006 to 2018, and the deadline of data collection is November 5, 2018.

Results

On the whole, the number of SCI papers in these 39 universities shows an increasing trend and these universities rank slightly differently each year. In 2006, the total number of SCI papers issued by the 985 Project universities is 44,686 and that is 151,560 in 2017, which is 3.39 times of the former. The numbers of SCI papers published by 985 Project universities in 2018 are 1.95 to 24.57 times of that of 2006, which correspond to the Naikai University and Minzu University of China, respectively. The number of SCI papers in many universities has increased considerably since 2011, such as Tsinghua University, Peking University and Tianjin University. Some universities have maintained a high scientific research output, such as Tsinghua University, Zhejiang University, Shanghai Jiao Tong University, Peking University and Harbin Institute of Technology. On the contrary, the amount of SCI papers of some of the others are always very small, such as Minzu University of China, Renmin University of China, Ocean University of China, National University of Defense Technology and East China Normal University. The peak of Tsinghua University's scientific research output appeared in 2017, which is 9,124, while Minzu University of China is at its bottom value in 2006, which is only 7. These universities have a great potential for scientific research and obviously, there is a long way to go to make such progress.

On the whole, the number of SSCI papers in these universities is much smaller than the number of SCI papers. In 2006 and 2017, the total number of SSCI papers is 518 and 9,743, respectively, which means the number of SSCI papers has increased more than 18 times. The numbers of SSCI papers of the 985 Project universities in 2018 are 6.37 to 131 times of that of 2006, which correspond to the Peking University and Southeast University, respectively. Some universities have maintained numerous SSCI papers, such as Peking University, Beijing Normal University, Tsinghua University, Zhejiang University and Shanghai Jiao Tong University. On the contrary, the amount of SSCI papers of some of the others are always very few, such as Minzu University of China, National University of Defense Technology, Northwest A&F University, Ocean University of China, Northwestern Polytechnical University and Lanzhou University. We can also find out that the research output of many universities have been very high in recent years, such as Tsinghua University, Peking University, Zhejiang University, Shanghai Jiao Tong University and Sun Yat-sen University. On the contrary, neither the number of SCI nor SSCI papers in some of the others are relatively few, such as Minzu University of China, National University of Defense Technology, Northwest A&F University, Ocean University of China and Lanzhou University.

According to Essential Science Indicators (ESI), highly cited papers refer to those papers that rank in the top 1% by citations for their category and year of publication (Miyairi and Chang, 2012). Highly cited papers play an important role in maintaining the reputation of scholars and research institutions to characterize their world-class scientific contributions (Bauer, et al., 2016). Accordingly, highly cited papers are considered as the key indicators in research performance assessment.

According to the number of highly cited papers, Tsinghua University, Peking University, University of Science and Technology of China, Zhejiang University and Shanghai Jiao Tong University are the top five universities, which occupy 1399, 1114, 868, 864 and 729 highly cited papers, respectively. On the contrary, Northwest A&F University, Ocean University of China, Renmin University of China, University of Defense Technology and Minzu University of China have the least number of highly cited papers, which are 108, 86, 64, 62 and 24, respectively. Since highly cited papers in ESI are defined employing a citation threshold at the 1% level, the expected percentage of a university's highly cited papers in terms of its total scientific output is 1%, and thus a value beyond 1% means a performance better than the expectation (Miyairi and Chang, 2012). There are 29 universities whose proportion of highly cited papers are beyond 1%, such as Hunan University (2.05%), University of Science and Technology of China (1.87%) and Tsinghua University (1.81%).

Except for journals, international conferences are also important as a venue to disseminate research results (Kim 2019). So we also indexed the Conference Proceedings Citation Index (CPCI). The range of the proportion of CPCI-S papers of these 39 universities is from 68.99% (Renmin University of China) to 99.39% (National University of Defense Technology). According to the sum of conference papers, the top five universities are Tsinghua University (22,202), Harbin Institute of Technology (19,191), Beihang University (14,851), Zhejiang University (14,412) and Shanghai Jiao Tong University (13,578). Since the number of conference papers indicates how many international conferences these universities have participated in, the increasing number of conference papers means that the number of international academic exchanges scholars participating in is increasing.

Conclusion

This paper conducts a comparative analysis of the scientific publications of Chinese 985 Project universities. The results of this study clearly indicate that publications by 985 Project universities remarkably increased during the study period. However, publications were unevenly distributed among the 985 Project universities. Scholars from Tsinghua University had the most SCI papers, the highly cited papers and the international conference papers. On the contrary, scholars from Minzu University of China had the least scientific research output. This also indicates that some of them have a great potential of scientific research and there is a long way to go to make progress.

Acknowledgments

This study was supported by a grant from the University Scientific Research Plan Projects of Shandong Province (Nos. J18KA342 and J18RA049). This study was also funded by Social Science Foundation of the University of Jinan (Nos. 18YY02 and 17YB06).

References

Bauer, J., Leydesdorff, L., and Bornmann, L., (2016). Highly cited papers in Library and Information Science (LIS): Authors, institutions, and network structures. *Journal of the Association for Information Science & Technology,* 67, 3095-3100.

Guskov, A. E., Kosyakov D. V., & Selivanova, I V. (2018). Boosting research productivity in top Russian universities: the circumstances of breakthrough. *Scientometrics*, 117, 1053-1080.

Kim, M. (2010). Visibility of Korean science journals. *Scientometrics,* 84 (2): 505-522.

Ma, Q., and Li, W., (2018). Growing scientific collaboration between Hong Kong and Mainland China since the handover: a 20-year bibliometric analysis. *Scientometrics* 117, 1479-1491.

Miyairi, N., and Chang, H., (2012). Bibliometric characteristics of highly cited papers from Taiwan, 2000–2009. *Scientometrics* 92, 197-205.

Qiu, J., (2014). China goes back to basics on research funding. *Nature* 507, 148-149.

Zhang, H., Patton, D., & Kenney, M. (2013). Building global-class universities: Assessing the impact of the 985 Project. *Research Policy*, 42, 765-775.

A holistic and bibliometric view on autonomous driving for the time period 2000 to 2017

Sandra Boric[1], Michaela Hildebrandt[1], Christina Hofer[1], Doris M. Macht[1],
Edgar Schiebel[2] and Christian Schlögl[1]

sandra.boric@edu.uni-graz.at; michaela.hildebrand@edu.uni-graz.at; christina.hofer@edu.uni-graz.at; doris.macht@gmx.net; christian.schloegl@uni-graz.at
[1]University of Graz, Universitätsstraße 15, A-8010 Graz, Austria

edgar.schiebel@ait.ac.at
[2] AIT Austrian Institute of Technology GmbH, Donau City Str. 1, A-1220 Vienna (Austria)

Introduction

Autonomous driving is considered as a megatrend for the mobility in the future and is currently a hot topic in media, politics and economy (Cavazza et al., 2019). The development goes together with electromobility and alternative mobility concepts such as car sharing (Krimmel & Ersoy, 2017, p. 914). As the technical development is progressing very fast, it will change the automotive industry fundamentally in the next few years (Cacilo, Sarah, Philipp, & Al., 2015).

Progress in autonomous driving is of interest particularly to manufacturers and suppliers of vehicles. To be informed about new technologies, research priorities and trends in a timely manner, early technology detection plays an important role for these companies (Krystek, 2007, p. 50).

The aim of this study is to identify the current research status of autonomous driving, the identification of bottlenecks as well as upcoming relevant issues in this technological field, and to investigate whether the scientific community searches for answers to open questions. Therefore, research literature regarding autonomous driving has been investigated by applying the bibliometric method of science mapping which is described below in more detail.

Methodology and Data

We used a science mapping approach to structure disciplines, scientific domains or significant research fronts of autonomous driving (Schiebel, Bianchi, & Vernes, 2016). Subfields were delineated with bibliographically coupled publications (Kessler, 1963) using the first order similarity (Thijs, Schiebel, & Glänzel, 2013).

In this study, the science mapping exercise is performed with a bibliometric software called *BibTechMon* (AIT Austrian Institute of Technology GmbH, 2018). A major feature of *BibTechMon* is that it enables the delineation and visualisation of subfields – this is the focus of this paper.

The data for this research was retrieved from Web of Science (WoS) at two different dates. When searching for the exact wording autonomous driving in the TOPIC field, which considers title, abstract and keywords, 1,202 matching results were found for the time span between 2000 and 2017. The retrieval date was November 23rd, 2017. This data was used for the delineation of the subfields. For the representation of the time series, we conducted the latest possible search to cover a nearly complete year 2017 by December 31st, 2017. 1,445 hits were found and explored with the help of the analyse feature of Web of Science.

All 1,202 resulting documents on autonomous driving were downloaded and imported into *BibTechMon*. Similar documents were clustered to subfields on autonomous driving using bibliographic coupling. The agglomeration of similar documents is represented in a 3D density surface map in which subfields are visible as peaks of the high local distribution of similar documents. The x and y axis are local coordinates and the z axis is the local number of similar documents, each weighted by the Jaccard index of the number of common references.

Results

In recent years, *autonomous driving* has experienced a high rise in the number of published research papers. Figure 1 visualizes the tremendous growth in the research output which underlines the high awareness and interest in the issue of autonomous driving. From a regional point of view, Germany showed the highest number of publications with 24.9% of the 1,445 papers (360 papers). Secondly ranked are the USA with 22.4% (324 papers), followed by the People's Republic of China with 12.2% (176 papers). Japan is in fourth place with 8.8% (127 papers) and South Korea is in fifth position with 8.7% (125 papers).

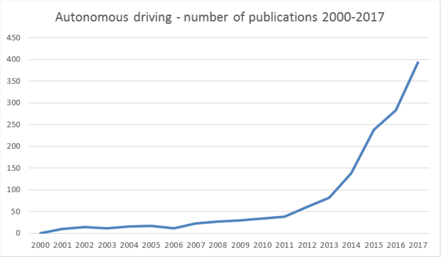


Figure 1: Number of publications on autonomous driving by publication year.

Since national funding of technology reflects the efforts and national strategies to promote technologies, it also makes sense to consider the ranking by funding agencies. This ranking is led by the National Natural Science Foundation of China with 42 papers, followed by the National Science Foundation of the USA with 13 papers.

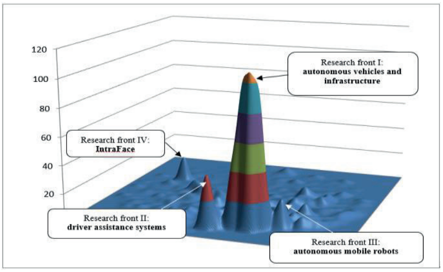


Figure 2: Subfields of autonomous driving

As can be seen in Figure 2, we identified four major, partly closely related research fronts which are *Autonomous Vehicles and Infrastructure*, *Driver Assistance Systems*, *Autonomous Mobile Robots*, and *IntraFace.* The first research front is more general and consequently, the largest one (approximately 100 publications). It deals mainly with the development of autonomous vehicles and their software as well as the necessary infrastructure. The fact that nearly all of the top-publishing organizations are university departments suggests that the related research is still at an early development stage. The second largest research front concerns *Driver Assistance Systems*. This is expressed by keywords like backward driving and self-learning classifier. Most of the publications in this research front focus on technologies for environment detection in automated vehicles, steering, and the prediction of traffic related events. Research front III, which deals with *Autonomous Mobile Robots*, is also closely connected with research front I. Robots and artificial intelligence play a key role in the publications of this research front. This is confirmed by the fact that the three most frequently cited articles were published in the Journal of Field Robotics. Many of the most recent publications deal with planning and control systems of autonomous mobile robots, localization and mapping, and predicting the intentions of other road users. Research front IV (*IntraFace*) is by far the smallest and most specific one. It covers publications which describe the development of algorithms for automatic face analysis. This software enables the automatic identification, tracking, and interpretation of facial expressions and the associated emotions. The complete study will be soon published in a journal.

References

AIT Austrian Institute of Technology GmbH (2018), BibTechMon – A software tool for bibliometric technology monitoring, AIT, Vienna

Cacilo, A., Sarah, S., Philipp, W. et al (2015). (German) Hochautomatisiertes Fahren auf Autobahnen – Industriepolitische Schlußforderungen. *Frauenhofer IAO*.

Cavazza, B. H., Gandia, R. M., Antonialli, F., Zambalde, A. L., Nicolaï, I., Sugano, J. Y., & Neto, A. D. M. (2019). Management and business of autonomous vehicles: a systematic integrative bibliographic review. *International Journal of Automotive Technology and Management*, *19*(1/2), 31. https://doi.org/10.1504/IJATM.2019.098509

Kessler, M. M. (1963). Bibliographic coupling between scientific articles. *Journal of the American Society for Information Science and Technology*, *14*(1), 10–25.

Krimmel, H., & Ersoy, M. (2017). Fahrwerkelektronik. In M. Ersoy & S. Gies (Eds.), *Fahrwerkhandbuch: Grundlagen -- Fahrdynamik -- Fahrverhalten -- Komponenten -- Elektronische Systeme -- Fahrerassistenz -- Autonomes Fahren -- Perspektiven* (pp. 747–792). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15468-4_15

Krystek, U. (2007). Strategische Früherkennung. *Controlling & Management*, *51*(2), 50–59. https://doi.org/10.1365/s12176-012-0165-4

Schiebel, E., Bianchi, D., & Vernes. (2016). Bibliometric field delineation with heat maps of bibliographically coupled publications using core documents and a cluster approach—the case of multiscale simulation and modelling. In *COST Action Knowescape TD1210, Budapest Workshop*.

Thijs, B., Schiebel, E., & Glänzel, W. (2013). Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, *96*(3), 667–677. https://doi.org/10.1007/s11192-012-0896-1

Tuning national performance-based science policy: introducing fractional count

Andrey Guskov[1] and Denis Kosyakov[2]

[1] guskov@spsl.nsc.ru, [2] kosyakov@spsl.nsc.ru
The State Public Scientific Technological Library, Siberian Branch of the Russian Academy of Sciences, Voskhod Str. 15, Novosibirsk (Russian Federation)

Introduction

Since 2012, the scientometric-centric (or performance-based) science policy has been finally approved in Russia. In May 2012, President of Russia V.V. Putin proclaimed that the fraction of Russian research publications indexed by Web of Science in 2015 has to be greater than 2.44 %. At that time, scientometric KPIs had already appeared in various official documents, but it was this decision that became the "point of no return" from the trajectory of a new science policy based on quantitative indicators. In 2018, a commitment to the performance-based science policy was confirmed at the highest state level. The new Decree of the President of the Russian Federation sets the task to take the 5th place in the world in the number of scientific publications by 2024 (in fact, this means a doubling of this number).

In the previous research, we introduced the principles of performance evaluation of Russian scientific organizations (*Kosyakov & Guskov*, 2019). However, an attempt to evaluate the universities with quantitative methods is failing. Among the reasons are: poor data quality (reports from institutions with aggregated indicators are used), lack of well-established mechanisms for data analysis and verification (there is no connection with primary data in citation indices), as well as conscious and unconscious manipulations (associated with the inability to determine the real involvement degree of the institution specified in the author affiliation in the preparation of a scientific publication).

Another significant problem is the phenomenon of synchronous mobility – simultaneously holding scientific positions in different institutions *(Markova, Shmatko & Katchanov, 2016).* Due to the relatively low wages of scientists and teachers since the 1990s and other historical reasons, the practice of combining positions has been developed, when one person is simultaneously an employee of the university (educational activity) and of the scientific institute (scientific activity). This creates the prerequisites for designation both affiliations in their publications, regardless of the real contribution.

Thus, in the existing situation, conditions are created for systemic distortion and formation of false assumptions in the scientific community. The science policy requires adjustment to encourage the conscientious contribution of each researcher and institution to the target indicator – the number of publications indexed in the WoS or Scopus.

This work is devoted to the study of the method of fractional count of publications in the process of national-level research assessment, which allows to reduce these negative effects. A systematic review of the methods of fractional count of publications was performed by (Egghe, Rousseau & Hooydonk, 2000), the advantages of the fractional count were shown in (Huang, Lin, & Chen, 2011).

Method

To conduct the study, data on publications with at least one Russian affiliation from 2000 to 2018 was downloaded from the Scopus (900,000 records). The indicators of the publication activity dynamics of the leading Russian scientific institutions and universities was carried out in three ways:

- PS_i is the number of publications with at least one affiliation of the institution i,
- FS_i is the sum of fractional scores $fs_p(i) \in [0..1]$ of all publications p of the institution i, $\mathbf{fs}_p(i) = \frac{1}{N_p}\sum_{a_j}\frac{Z(i,a_j)}{AF(a_j)}$, where a_j are authors of p; $1 \le j \le N_p$ – the number of authors; $AF(a_j)$ is the number of affiliations for author a_j, $Z(i, a_j)$ is 1 if a_j has an affiliation i, and 0 it has not. E.g., if in publication p author a_1 has affiliation i_1, $a_2 – i_2$, $a_3 – i_2$ and i_3, then $fs_p(i_1)$=1/3, $fs_p(i_2)$=1/2, $fs_p(i_3)$=1/6.
- LFS_i is the sum of the local fractional scores $lfs_p(i)$ of all publications p of the institution i, where $lfs_p(i)$ is $fs_p(i)$, from the count of which foreign affiliations are excluded. E.g., if i_2 is foreign institution, then $fs_p(i_1)$=$fs_p(i_3)$=1/2.

The use of fractional count makes it possible to evaluate more fairly the performance of scientific research than a whole-number count. In fact, 1-2 researchers usually spend disproportionately more effort on preparing an article than a group of 10 people or a large collaboration that publishes them in dozens and hundreds. The disadvantage of using fractional count is its laboriousness, since it requires detailed processing of the entire array with authors and affiliations, aggravated by the problems of their identification ambiguity. It can demotivate research work in real collaborations; therefore, in the Russian context, it is most expedient to apply the ***LFS**ᵢ* indicator to support international collaborations.

Results

This assumption is factually confirmed. Figure 1 shows that the number of publications with single-affiliated authors from 2011 to 2017 has dramatically decreased from 84.4% to 69.8%. At the same time, the number of publications with at least one multi-affiliated author has doubled. Almost 1% of Russian publications having an author that indicates more than three affiliations in one article! This creates a rather strange situation, when the indicators of individual institutions are growing much faster than overall result.

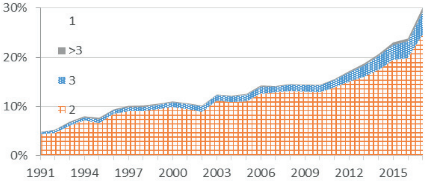


Fig. 1. Share of publications which authors have maximum one, two, three or more affiliations.

We demonstrate the results of fractional count with an example of 10 universities among the leading in the country (Fig. 2). In the transition from whole-number to fractional count, ranking varies significantly. The collaborations around these universities are very different and provide different contributions to publication activity, which discriminates institutions with weak external links. This is especially noticeable in cities where the university is the only serious scientific institution (Southern Federal University and Samara National Research University). The transition to fractional count eliminates this difference and allows to more accurately determine the "own performance".

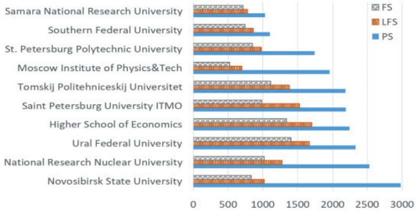


Fig. 2. Compare of PS, LFS and FS for some top Russian universities in 2017.

Let's introduce local collaboration coefficient of the institution $\boldsymbol{LCC_i} \in$ **[0;1]**, which is calculated as $\boldsymbol{LCC_i = 1 - (LFS_i / PS_i)}$. If $\boldsymbol{LCC_i} = 0$, the institution ***i*** does not have any common publications with other Russian institutions. If $\boldsymbol{LCC_i} = 0.5$, this means that among the publications of the institution ***i***, their affiliation contribution is equal to that of all other Russian institutions.

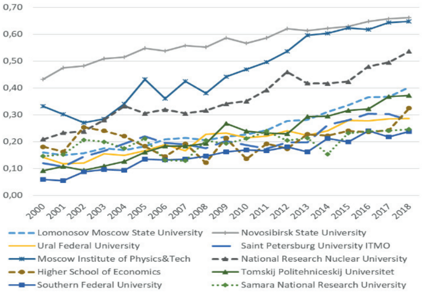


Fig. 3. Dynamics of *LLC* for some top Russian universities.

Fig. 3 shows that since 2000, the dynamics of $\boldsymbol{LCC_i}$ at leading Russian universities is increasing (level of national collaboration is growing). Since 2013, this growth becomes faster for most institutions. The analysis of the 100 most successful Russian organizations in 2017 showed that for them $\boldsymbol{LCC_i \in (0.2\text{-}0.5)}$. The exception is nine universities and research centres in the field of physics ($\boldsymbol{LCC_i \in (0.6\text{-}0.8)}$), which have many publications in large collaborations.

Conclusion

Research performance-based policy has led to a number of distortions. Using local fractional score ***LFS*** allows for a more fair account of the contribution of authors and organizations. Introducing such performance indicator more clearly shows the goals of national science policy for institutions and researchers. The local collaboration coefficient is a stable-in-time indicator that adequately demonstrate the share of the institution contribution to the published results.

References

Egghe, L., Rousseau, R. & Van Hooydonk, G.(2000) Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American society for information science*. N. 2 (51). (pp. 145–157).

Huang, M.-H., Lin, C.-S. & Chen, D.-Z. (2011) Counting Methods, Country Rank Changes, and Counting Inflation in the Assessment of National Research Productivity and Impact // Journal of the American society for information science and technology No. 12 (62). (pp. 2427–2436).

Kosyakov, D. & Guskov, A. (2019) Research assessment and evaluation in Russian fundamental science. *Procedia Computer Science*. Vol. 146. (pp. 11–19). DOI: 10.1016/j.procs.2019.01.072

Markova, Y. V., Shmatko, N. A., & Katchanov, Y. L. (2016). Synchronous international scientific mobility in the space of affiliations: evidence from Russia. SpringerPlus, 5(1). DOI: 10.1186/s40064-016-2127-3

A preliminary scientometric analysis of the Cross-Strait scientific collaboration

Kai Li[1] and Pei-Ying Chen[2]

[1] *kl696@drexel.edu*
Drexel University, College of Computing and Informatics, Philadelphia, PA (USA)

[2] *peiychen@iu.edu*
Indiana University, School of Informatics, Computing, and Engineering, Bloomington, IN (USA)

Introduction

The rise of China in the global scientific system has been an increasingly popular research topic in the scientometric scholarship. In addition to earlier efforts on the number of publications (Chen, Chen, Hwang, & Chou, 2006; Leydesdorff & Zhou, 2005; Youtie, Shapira, & Porter, 2008), another important indicator is China's participation in international scientific collaboration (Yuan et al., 2018; Zhou & Lv, 2015; Zhou, Zhong, & Yu, 2013). Rarely examined is how scientists within Chinese-speaking region collaborate with each other with one notable exception (Ma & Li, 2018). This question, however, could shed fresh light on how scientific collaboration is shaped by socio-economic context and political dynamics.

As the first step of the research agenda, the present work provides a preliminary scientometric profile of publications co-authored by researchers in the Mainland China and Taiwan, the two most productive Chinese-speaking areas. In this study, collaboration is operationalized as *a publication co-authored by at least two researchers, each from an institution in the Mainland China and Taiwan, respectively*. We retrieved bibliographic data indexed by the Web of Science (WoS) database for the initial analysis. Specifically, we focused on the temporal and institutional aspects of the scientific collaborations across the Strait, as a demonstration of how scientific collaboration is subject to influences of political and educational policies.

Materials and Methods

We collected all publication records, on December 24, 2018, from the WoS database through the following query terms: CU=Peoples R China AND CU=Taiwan. This query term included some but not all institutions in Hong Kong and Macau, even though we decided not to do any special treatment for their inclusions. In total, 27,521 publications were selected into our final sample based on the following criteria: 1) works published in 2019 were EXCLUDED; 2) only works written in English were selected; 3) only research articles (Document Type: Articles) and conference papers (Document Type: Proceedings paper) were selected; and 4) all publications that have only one author were EXCLUDED.

Results

Temporal distribution of collaborations

Figure 1 shows the total amount of collaborated works as well as the share of this sub-collection among all publications in the two regions.

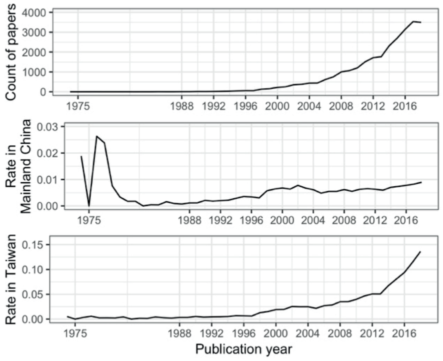


Figure 1. Total amount of collaborated publications (Top), the share of collaborated subset among all publications in Mainland China (Middle) and Taiwan (Bottom)

While the number of collaborative works increased significantly during the past three decades, there are much subtler stories at the local level. Generally speaking, collaborations between both regions have become more important in both regions since the 1980s, even though the speed of increase was rather different. The share of collaborative publications in Taiwan almost tripled from 2013 to 2018. In contrast, despite the growing number of cross-Strait collaborative papers, its share only increased by less than 10% in China's total publications, which echoed an earlier finding that the growth of scientific publications in China is primarily endogenous to international collaboration (Zhou & Glänzel, 2010).

Top institutions in collaborations

In term of the collaborating institutions, a total of 3,877 institutions from the Mainland China and 1,903 institutions from Taiwan were extracted from the collaboration subset. The top 10 institutions from both regions are summarized in Table 1. The most important observation is that majority of institutions on the list are renowned research-intensive universities sponsored by the prestigious nation-wide programs such as "Project 985" in the Mainland China (Zhang, Patton, & Kenney, 2013) and "Plan to Develop First-class Universities and Top-level Research Centers" in Taiwan (Chang, Wu, Ching, & Tang, 2009) to leverage scientific research capabilities in higher education. In fact, all the Taiwanese universities except Feng Chia University are recipients of the competitive grants. Also notable is that *Chinese Academy of Science* (Mainland China) and *Academia Sinica* (Taiwan), both the government-owned, highest research institution in the respective regions, are among the top collaborating institutions. This raised an important question about the relationship between local science/research policies and participation in the Cross-Strait collaboration.

Table 1: Top 10 institutions from the Mainland China and Taiwan

Rank	Mainland China	Taiwan
1	Chinese Academy of Science	National Taiwan University
2	Peking University	Academia Sinica
3	Chinese University of Hong Kong	National Tsing Hua University
4	University of Hong Kong	National Cheng Kung University
5	University of Science & Technology	National Central University
6	Tsinghua University	National Sun Yat-Sen University
7	Zhejiang University	National Chiao Tung University
8	Xiamen University	Chang Gung University
9	Harbin Institute of Technology	China Medical University
10	Shanghai Jiao Tong University	Feng Chia University

Discussion and Conclusion

The paper presents a comprehensive yet preliminary scientometric scanning of cross-Strait scientific collaboration between the Mainland China and Taiwan, including the growth in absolute and relative terms over time and top participating institutions. Overall, our findings point to the importance of socio-political contexts and policy-making in shaping scientific collaboration, a research agenda to be fulfilled in the future.

In our next step, we will focus on how the cross-Strait scientific collaborations are shaped by important socio-political and policy developments. Moreover, we will also use a broader selection of data, including not only publications in English but also in Chinese, the latter of which represents a different and arguably more extensive aspect of scientific collaborations within the region.

Acknowledgments

This project was made possible in part by the Institute of Museum and Library Services for the project titled "Building an entity-based research framework to enhance digital services on knowledge discovery and delivery" (Grant Award Number: RE-07-15-0060-15) and "IDEASc: Integrated doctoral education with application to scholarly communication" (Grant Award Number: RE-02-14-0023-14).

References

Chang, D., Wu, C., Ching, G. S., & Tang, C. (2009). An evaluation of the dynamics of the plan to develop first-class universities and top-level research centers in Taiwan. *Asia Pacific Education Review*, *10*(1), 47–57. https://doi.org/10.1007/s12564-009-9010-7

Chen, T.-J., Chen, Y.-C., Hwang, S.-J., & Chou, L.-F. (2006). The rise of China in gastroenterology? A bibliometric analysis of ISI and Medline databases. *Scientometrics*, *69*(3), 539–549.

Leydesdorff, L., & Zhou, P. (2005). Are the contributions of China and Korea upsetting the world system of science? *Scientometrics*, *63*(3), 617–630.

Ma, Q., & Li, W. (2018). Growing scientific collaboration between Hong Kong and Mainland China since the handover: a 20-year bibliometric analysis. *Scientometrics*, *117*(3), 1479–1491. https://doi.org/10.1007/s11192-018-2916-2

Youtie, J., Shapira, P., & Porter, A. L. (2008). Nanotechnology publications and citations by leading countries and blocs. *Journal of Nanoparticle Research*, *10*(6), 981–986.

Zhang, H., Patton, D., & Kenney, M. (2013). Building global-class universities: Assessing the impact of the 985 Project. *Research Policy*, *42*(3), 765–775.

Zhou, P., & Glänzel, W. (2010). In-depth analysis on China's international cooperation in science. *Scientometrics*, *82*(3), 597–612. https://doi.org/10.1007/s11192-010-0174-z

Analysis of the relationships between academic research fields based on co-occurrence of journal categories

Chizuko Takei[1], Fuyuki Yoshikane[2] and Hiroshi Itsumura[3]

[1] *naoe.chizuko@adm.nagoya-u.ac.jp*
Graduate School of Library, Information and Media Studies, University of Tsukuba, Kasuga 1-2, Tsukuba, Ibaraki, 305-8550 (Japan)

[2] *fuyuki@slis.tsukuba.ac.jp,* [3] *hits@slis.tsukuba.ac.jp*
Faculty of Library, Information and Media Studies, University of Tsukuba, Kasuga 1-2, Tsukuba, Ibaraki, 305-8550 (Japan)

Introduction

Many studies on interdisciplinary research have been conducted, but most of them were based on article citation analysis focusing on the degree of integration (strength of the connection) between fields (eg. Leydesdorff et al., 2013). More recently, diversity (variety, bias, and similarity) of related fields has attracted attention as a new interdisciplinary perspective (eg. Zhang et al., 2016). However, there are not sufficient studies investigating more detailed characteristics of interdisciplinary research such as the specific composition of related fields, which are necessary for effective policies and funding to encourage interdisciplinary research. Moreover, the data-driven article-level classification is emerging, but many funders or institutions still evaluate the research or researchers by fields using databases like Web of Science (WoS) in many countries. Therefore, to clarify the interdisciplinary relationships between academic fields, in this research, we investigated the network structure of all categories in the WoS for research evaluation.

Method

WoS bibliographic data was used to identify the relationships between the categories. We extracted the journal title lists from Arts & Humanities Citation Index, Science Citation Index Expanded, and Social Sciences Citation Index in the 2018 Core Collection. The WoS assigns one or more categories to each journal based on the subject matter. Assuming the categories assigned to the same journal (i.e., co-occurring categories) to be related fields, we constructed a network based on category co-occurrence. That is, each node in the network referred to a category (field), and each edge indicated the connection between categories in a journal. Through observing the connections between categories in this network, we assessed the nature of each category such as the extent to which it is interdisciplinary.

The following indicators were adopted to characterize the individual categories from the viewpoint of network structure:

a. Degree centrality
b. Betweenness centrality
c. Eigenvector centrality

(a) is an index for evaluating the importance of a node by the number of its adjacent edges (i.e., degree), and in this research, it shows the amount of the relationship with other categories. (b) is used to evaluate a node by the extent to which it is located on the shortest paths between nodes. Here, it measures the importance as an intermediary between other categories. (c) also evaluates degree as in (a), however, while (a) simply counts the number of edges, (c) weights edges based on the centrality of their adjacent nodes. That is, (c) reflects the structure of the entire network, including indirect connections between nodes.

In this research, we evaluated interdisciplinarity based on the variety of related fields (degree). Therefore, categories with high variety, that is, high values of degree centrality or eigenvector centrality, were deemed to be highly interdisciplinary. In addition, even though betweenness centrality is not directly related to interdisciplinarity, it will provide clues to the possible development of interdisciplinarity, that is, the possibility of collaboration with fields that have not yet been linked.

Results

Table 1 gives the basic statistics for the network characteristics, and Tables 2 to 4 show the results for the characteristics of the individual categories in the network.

Table 1. Basic quantities regarding the network of journal categories.

No. of nodes	*No. of edges*	*Density*	*Cluster coefficient*
242	2417	0.083	0.417

In Tables 2 to 4, the top ten categories are shown for each indicator. Pharmacology & pharmacy had the highest degree centrality, betweenness centrality, and eigenvector centrality, followed by Biochemistry & molecular biology and Materials science (multidisciplinary). In addition, Public, environmental & occupational health, Computer science (interdisciplinary applications), Environmental sciences, and Mathematics (interdisciplinary applications)—all had high degree centrality and betweenness centrality but were not in the top ten for eigenvector centrality, whereas engineering-related and physics-related categories were in the upper part for eigenvector centrality.

Table 2. Categories with high degree centrality.

	Categories	*Degree*
1	Pharmacology & pharmacy	0.432
2	Public, environmental & occupational health	0.253
3	Computer science (interdisciplinary applications)	0.237
4	Neurosciences	0.228
5	Biochemistry & molecular biology	0.224
6	Environmental sciences	0.207
7	Materials science (multidisciplinary)	0.199
8	Mathematics (interdisciplinary applications)	0.199
9	Economics	0.191
10	Genetics & heredity	0.187

Table 3. Categories with high betweenness centrality.

	Categories	*Betweenness*
1	Pharmacology & pharmacy	2396.058
2	Computer science (interdisciplinary applications)	1524.696
3	Physiology	1135.285
4	History & philosophy of science	1104.630
5	History	1102.921
6	Mathematics (interdisciplinary applications)	1090.447
7	Public, environmental & occupational health	1046.490
8	Environmental sciences	930.006
9	Biochemistry &molecular biology	862.426
10	Materials Science(multidisciplinary)	801.126

Table 4. Categories with high eigenvector centrality.

	Categories	*Eigenvector*
1	Pharmacology &pharmacy	1.000
2	Materials science (multidisciplinary)	0.968
3	Nanoscience &nanotechnology	0.702
4	Engineering, electrical &electronic	0.546
5	Chemistry, physical	0.445
6	Physics, condensed matter	0.417
7	Chemistry (multidisciplinary)	0.380
8	Optics	0.320
9	Biochemistry & molecular biology	0.237
10	Engineering, chemical	0.230

Discussion and Conclusion

Pharmacology & pharmacy, Biochemistry & molecular biology, and Materials science (multidisciplinary), all of which had high degree centrality, betweenness centrality, and eigenvector centrality, were shown to be highly interdisciplinary categories connecting with many categories. Moreover, because of their high betweenness centrality, they also appeared to be the major intermediaries between categories. Relation between categories, including intermediation, can be regarded as useful information for researchers and those who support research, which shows the possibilities for promoting interdisciplinary or collaborative research and searching appropriate journals for submitting papers. A difference was found between degree centrality and eigenvector centrality, with the former being dominated by environmental science-related and computer science-related categories and the latter by engineering-related and physics-related categories. However, in both indicators, the higher ranked categories have multiple connections with others; therefore, these categories are highly interdisciplinary. In particular, as the categories that have high eigenvector centrality tend to be connected to adjacent categories that have multiple connections with others, there is a possibility that they are more variously affected through direct and indirect connections. In addition, the results of the categories whose name has “multidisciplinary“ or “interdisciplinary“ are various (eg. Computer science (interdisciplinary applications), Chemistry (multidisciplinary)). These differences might be due to the differences of interdisciplinarity of fields by the influence of research theme, collaboration, and citation tendencies. To clarify these differences, it is required further investigation. This study has a limitation that only the categories in the WoS are investigated. Although the results of this research could be useful in providing suggestions for possible support for interdisciplinary research and research funding, it is necessary to take into account this limitation and use a clue for evaluate interdisciplinary research.

References

Leydesdorff, L., Rafols, I. & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal citations. Journal of the American Society for Information Science and Technology, 64(12), 2573-2586.

Zhang, L., Rousseau, R. & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: taking similarity between subject fields into account. Journal of the Association for Information Science and Technology. 67(5), 1257-1265.

A Study on the Multidimensional Scientometric Indicators to Detect the Emerging Topics

Haiyun Xu[1], Zenghui Yue[2], Rui Luo[3], Ziqiang Liu[4], Zhao Zhang[5],Yan Qi[6] and Zhengyin Hu[7]

[1] *xuhy@clas.ac.cn;* [7]*huzy@clas.ac.cn;*
Institute of Scientific and Technical Information of China (ISTIC), 100038, Beijing (China)
Chengdu Library and Information Center, Chinese Academy of Sciences, 610041, Chengdu (China)

[2] *yzh66123@126.com*
School of Medical Information Engineering, Jining Medical University, 276826,Rizhao (China)

[3] *luorui@mail.las.ac.cn;* [4] *liuziqiang@mail.las.ac.cn;* [5]*cheungaomail@foxmail.com;*
Chengdu Library and Information Center, Chinese Academy of Sciences, 610041, Chengdu (China)
University of Chinese Academy of Science, 100190, Beijing (China)

[6] *qi.yan@imicams.ac.cn.*
Institute of Medical Information/Medical Library, CAMS & PUMC, 100020, Beijing (China)

Introduction

Nowadays, policymakers and academic researchers have paid increasing attention to the detection of emerging (EM) topics, which are useful for promoting the advancement of potentially promising research. This is because it is important to accurately recognize EM topics and their developing trend in order to support the strategic planning and optimal allocation of innovation resources. Meanwhile, evidence of the increasing attention being paid to methods for discovering EM topics can be found in the growing number of publications related to EM technologies and topics. Scientometric methods are vital to discovering EM topics, and this study continues this approach to EM topic discovery.

Related work on EM topic indicators

Defining EM technologies

To discover EM topics at fine topic granularity in a specific domain, we must first understand the evolution, laws, and trends of EM topics in specific fields at this granularity. Rotolo et al. (2015) defined EM technologies or topics using five key attributes: (i) radical novelty, (ii) relatively fast growth, (iii) coherence, (iv) high impact, and (v) uncertainty and ambiguity. Specifically, they conceived of an EM topic as a radically novel and relatively fast-growing technology characterized by a certain degree of coherence persisting over time. Moreover, it has the potential to exert a considerable impact on socioeconomic domain(s), which can be observed from the composition of the relevant actors, institutions, and their interaction patterns, along with the process of associated knowledge production。

EM topic criteria

Our study attempts to identify EM topics at fine topic granularity in a specific domain, and focuses on how to operationalize the criteria for an EM topic. Porter et al. (2018) conducted a similar study. However, whereas our aim is to discover potential EM topics, Porter et al.'s research is more inclined to find more potential term sets of EM, which we argue is not informed enough. This is because further attempts to find EM topics based on these terms is usually a more complex and error-prone process. In addition, the criteria and multi-indicator methods for identifying EM topics we propose are applied to a specific research domain—stem cells. Our analysis focuses on a specific domain and can help more specific and comprehensive information about the targeted domain to be obtained. Another innovation of this study is that we analyze the attributes of uncertainty and ambiguousness.

Framework for the operationalization of EM topics

There are two steps to discovering high-impact and transformative EM topics: finding meaningful topics and then discovering which of these topics are EM topics. For the second step, a research topic can be considered as an EM topic if it displays five attributes: (i) novelty, (ii) growth, (iii) persistence and coherence, (iv) potential high impact (Social and economic influence), and (v) uncertainty and ambiguity reduction. We hence first briefly review meaningful topic discovery. Then, we focus on determining each attribute of an EM topic as the second step.

Experimental Study

Data set and statistic description

The study of stem cells, which are a type of cell that are capable of self-renewal and multidirectional differentiation, is an important topic in biomedical research (Wei et al., 2017). In this section, the literatures on stem cells were selected as a case study to demonstrate the approach proposed in this paper. The data for this study were collected on October 20, 2018, and 422,101 research articles and 50,556 patents were retrieved.

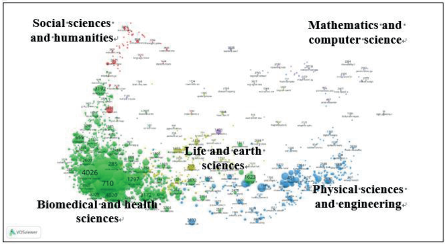


Figure 4. The micro-level classification in stem cell from 2001 to 2018

EM topics in the stem cell field

Finally, 26 topics were identified as EM topics in stem cell research. The features of 10 of them are listed in Table 2 because of space limitations. Then, we conducted an additional thematic analysis of the text sets under each of the CWTS topics to relabel each topic. The topic labels in Table 3 are the subject tags for the 10 EM topics in the stem cell field with the original topic labels from the CWTS classification.

Table 3. EM topics in stem cells domain (partial)

No.	*Topic id*	*Topic labels of stem cells*
1	353	stem cell; cell; surface; mesenchymal stem; extracellular matrix;substrate; differentiation; tissue; hydrogel; biomaterial
2	2276	stem cell; intestinal stem; expression; organoid; cancer; intestinal; colorectal cancer;crypt; model; Lgr5
3	1460	stem cell; scaffold; tissue engineering; cell; extracellular matrix; tissue; mesenchymal stem; VITRO; regeneration; decellularized
4	142	DNA methylation; stem cell; epigenetic; gene; gene expression; EMBRYONIC STEM; expression; human; cell;
5	60	stem cell; mesenchymal stem; ARTICULAR CARTILAGE; scaffold; tissue engineering; chondrocyte; chondrogenic differentiation; bone marrow; growth factor; cartilage repair
6	*921*	*stem cell; dental pulp; expression; periodontal ligament; mesenchymal stem; pulp stem; osteogenic differentiation; human dental; VITRO; growth factor*
7	*727*	*beta catenin; stem cell; expression; Wnt beta; signaling pathway; cell; catenin signaling; protein; gene*
8	*1046*	*stem cell; EZH2; expression; cell; gene; protein; gene expression; differentiation; embryonic; chromatin*
9	*161*	*acute myeloid; myeloid leukemia; stem cell; myelodysplastic syndrome; cell transplantation; patient; AML; hematopoietic stem; leukemia AML; treatment*
10	*221*	*stem cell; retina; transplantation; cell; differentiation; retinal pigment; photoreceptor; pluripotent stem; macular degeneration; progenitor cell*

Discussion and Conclusions

This study proposed a multidimensional and practical bibliometric methodology for the detection of EM topics. In addition, an in-depth analysis in the field of stem cell research was conducted to examine this approach further. In particular, we emphasis more the study of potential high impact and the uncertainty or ambiguity of the EM topic. Therefore, our approach could offer more comprehensive information about EM topics.

Acknowledgments

We appreciate the very constructive instruction and suggestion from Ludo Waltman. This article is an outcome of the project “Study on the Recognition Method of Innovative Evolving Trajectory based on Topic Correlation Analysis of Science and Technology” (No. 71704170) supported by National Natural Science Foundation of China, the project “Study on Regularity and Dynamics of Knowledge Diffusion among Scientific Disciplines” (No. 71704063) supported by National Natural Science Foundation of China, the project “E-Science Application for Knowledge Discovery in Stem Cells” (No. KFJ-EW-STS-032) supported by the Informationization Special Project of Chinese Academy of Sciences and the Youth Innovation Promotion Association, CAS (Grant No. 2016159).

References

Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2018). Emergence scoring to identify frontier R&D topics and key players. Technological Forecasting and Social Change.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? Research Policy, 44(10), 1827-1843.

Wei, L., Hu, Z., pang, H., tan, X., Guo, H., & Fang, S. (2017). Study on knowledge discovery in biomedical literature based on spo predications: A case study of induced pluripotent stem cells. Digital Library Forum(9), 28–34.

Characterization of URLs in scientific documents: the profile of the journal Information Science

Ronnie Fagundes de Brito[1], Milton Shintaku[2], Ingrid Schiessl[3], Diego José Macêdo[4] and Janinne Barcelos[5]

[1]*ronniebrito@ibict.br*
Brazilian Institute of Information in Science and Technology, SAUS 5, Brasília (Brazil)

[2]*shintaku@ibict.br*
Brazilian Institute of Information in Science and Technology, SAUS 5, Brasília (Brazil)

[3]*ingridschiessl@ibict.br*
Brazilian Institute of Information in Science and Technology, SAUS 5, Brasília (Brazil)

[4]*diegomacedo@ibict.br*
Brazilian Institute of Information in Science and Technology, SAUS 5, Brasília (Brazil)

[5]*janninesilva@ibict.br*
Brazilian Institute of Information in Science and Technology, SAUS 5, Brasília (Brazil)

Introduction

One of the elementary resources from the web network is the link or the hyperlink, which allows navigating between addresses and documents available on this network. Furthermore, it is by means of URLs (Uniform Resource Locator) that the links are activated, permitting the access to new content. In the scenario of scientific communication, the link is used in documents that disclose research results, adding information to the documents and referring to research objects as texts of reference, data sets, figures, simulations and other complementary materials.

From a historical perspective, (Isfandyari -Moghaddam, 2010) indicate that in the period from 1995 to 2008 there was a significant increase in the rate of articles containing URLs, proportionally increasing the accessibility of the article on the web. To characterize the behavior of URLs, one can apply the concept of half-life, which describes the estimated number of years for which 50 percent of citations published on the Internet stop working (Dimitrova and Budeja, 2007). Consequently, (Sampath Kumar & Manoj Kumar, 2012) estimated a half-life of 11.5 years for links in a group of open access journals.

In this context, it is aimed to measure the URLs behavior in scientific documents in order to verify if trends and patterns previously identified, specifically the half-life of URLs. For the purpose to carry out this measurement, journals were selected in the area of information science which had their documents collected and analyzed, being extracted the URLs and verified their status.

This research aims to verify the URLs behavior in the scientific literature, especially the papers in the Journal Information Science, in relation to its temporal endurance and the use of technologies to mitigate the instability of web links.

Methods

This is an exploratory study on the URLs behavior of the Journal of Information Science[1]. It was collect the published papers with OCR and access to the PDF archive.

The metadata collection, the download of resources and other procedures were parameterized and implemented in a Jupyter Notebook[2] script (Brito, 2018). The analysis of links was restricted to the protocol http, due to answer standardization of servers.

[1] http://revista.ibict.br/ciinf/

[2] https://jupyter.org/

Results

In 2005, the year that the journal became digital, it is possible to measure the increase of URLs in published papers, with a current average of 15 URLs per published paper.

It is noticed that the URLs started to appear from 1996, when the first paper published with URLs counted with 23 addresses, in which only 5 are still accessible nowadays. Most recently, in 2017, it is possible to measure the availability rate of 77% of the URLs. The figure 1 summarizes the accessibility behavior of URLs over the years.

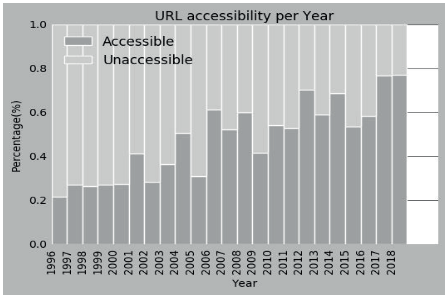


Figura 1. Accessibility of URLs present in published documents by the journal.

It is possible to observe a bigger quantity of URLs (Figure 2) of the type 300, which correspond to redirecting (301 Moved Permanently or 302 Found), indicating the concern of content maintainer to preserve his or her access even through legacy URLs. It is important to notice a big quantity of URLs not found (404 - Not Found).

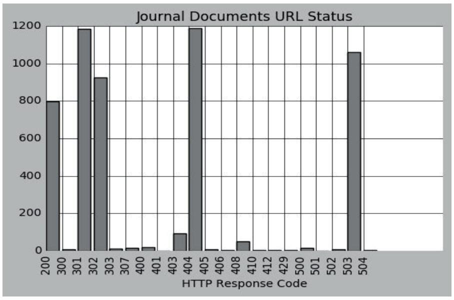


Figura 2. Code of HTTP answer and the URLs analyzed

The quality of the URLs was verified in relation to the use of DOI or Handle, with a significant increase in the year 2016. The half-life was calculated as in Dimitrova & Budeja(2007).

The results were analyzed allowing to state that the patterns of URLs half-life (do not) maintain themselves and that the adoption of identifiers strategies (has not) been adopted and (has not) affected the quality of connections between documents and research objects.

The answer of availability is subjected to the kind of access, and the correct implementation of the communication protocol http. The disciplinary differences were not analyzed as regards the availability of the URLS, even though they exist . It is also suggested to evaluate the half-life variation over time, with measurement at regular intervals and to verify if this alters due to the use of resources as DOI.

The approach focused on the lexical/syntactical level of documents, without differentiation of URL in citations or scattered in the text. There is also a need to implement a more robust URL capture form, using regular expressions and additional resources to ensure the integrity of the URL. Finally, it should be noted that using Internet archiving to recover lost content is a successful approach that affects the process of sending manuscripts to the journal.

References

Dimitrova, D. V., & Bugeja, M. (2007). The half-life of internet references cited in communication journals. *New Media & Society*, *9*(5), 811–826. Retrieved from https://doi.org/10.1177/1461444807081226

Brito, R. F. (2018). A Jupyter Notebook to Describe URLs from Journal documents (Version 1)[Computer software]. Retrieved from http://doi.org/10.5281/zenodo.2362826

Isfandyari-Moghaddam, Al.; Saberi, M. K. & Mohammad Esmaeel, S. (2010). Availability and Half-life of Web References Cited in Information Research Journal: A Citation Study. International Journal of Information Science and Management, 8(2), 57-75.

Sampath Kumar, B.T. & Manoj Kumar, K.S. (2012). Persistence and half-life of URL citations cited in LIS open access journals.Aslib Proceedings, 64(4), 405-422. doi.org/10.1108/000125312112447

Role of structural determinants in development of universities

Angelika Tsivinskaya[1] and Mikhail Sokolov[2]

[1] *atsivinskaya@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, 6/1A
Gagarinskaya Street, 191187, St. Petersburg (Russia)

[2] *msokolov@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, 6/1A
Gagarinskaya Street, 191187, St. Petersburg (Russia)

Introduction

Since 2012, Russian Ministry of Science and Education relies on a complex system of statistical indicators (93 in 2013 and 2014, 162 in 2015-2017) provided by universities themselves to distribute funding and make a decision about the dissolution of institutions that do not meet performance standards.

A recent study by (Mateos-González and Vikki Boliver, 2018) is shown how pre-existing inequalities between universities in Italy influence their performance and propagated by the current evaluation system. The paper by (Sánchez-Barrioluengo and Mabel, 2014) discuss shortcomings of the assumption that universities are "homogeneous institutions with equal capacity to perform" in the case of Spanish universities.

We argue that not only structural determinants influence the performance of universities but as result, current evaluation system becomes a major factor in the fate of universities.

To achieve this purpose, we present our findings for (1) the influence of structural determinants on the performance of universities; (2) the consequences of the implementation of the current system on the survival of universities.

Data

The main data used in this paper is obtained from the Efficiency monitoring initiated by the We used data collected from 2014 to 2017 for 822 universities and 979 branch campuses. The total number of variables used for preliminary analysis was 70, so we had sample size=822. Then we did initial variable selection and exclude universities that were officially in the survey, but the data were not provided. Sample size N=777 universities, 43 variables. The choice of this subset of variables was based on considerations of 1) having enough variability for conducted analysis and 2) avoiding collinearity (mostly between measures dependent on size).

Methods

Our main hypothesis is that the whole trajectory of the university's evolution is heavily determined by its "ascriptive" or "inborn" characteristics.

We considered several classes of independent variables determining niches of universities, such as (a) if it was public or private; (b) if it was localized in a bigger city or a wealthy region, (c) its nominal profile or "family" enjoying certain prestige and guaranteeing influx of highly motivated students (e.g. an agricultural college is not likely to be as attractive, as an institution specializing on law); (d) its ecological situation at a local market for higher education.

Firstly, we performed hierarchical cluster analysis on a subset of variables showing the change (Δ) in performance measures between 2013 and 2017, which can be used to characterize different trajectories of development.

Secondly, we showed the consequences of the implementation of the current evaluation system using survival analysis techniques to estimate "survival" for different groups of universities.

Results

Hierarchical cluster analysis

To characterize the trajectories, we performed a hierarchical cluster analysis by the Ward method, using the Euclidean distance as a measure of distance, normalizing the indicators for individual variables. The optimal number of clusters according to the "elbow" method is 4.

From Figure 1 we can see that the first cluster consist mostly from universities with high numbers in publication activity, international students and grants. The second cluster is characterized by universities with a large area of laboratories, low incomes of the staff and R&D, a small number of students receiving additional education. The third cluster consists of universities with high salaries of academic staff and incomes, a high number of academic staff members with a degree, a low scores of students in the unified state examination, and a small number of international students. The fourth cluster represents universities with low numbers for R&D, number of staff members with the PhD and doctoral degrees, and low numbers for indicators of material resources.

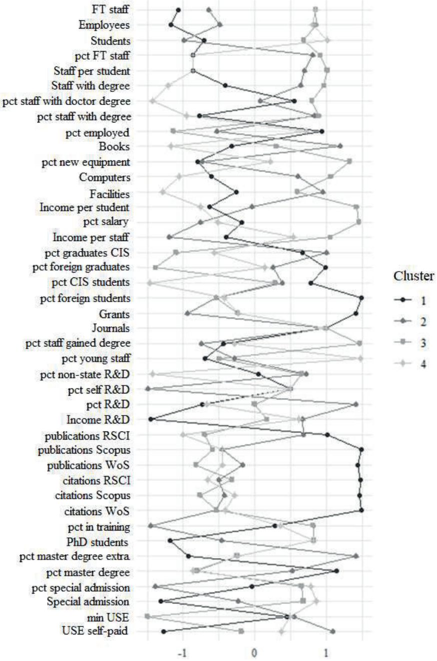


Figure 1. Graph of variables in parallel coordinates for clusters.

Survival analysis

What were the consequences of implementing efficiency metrics not taking into account structural variables which influence a university's performance? To answer this question, we performed survival analysis using the Cox proportional risk model (Table 1).

According to our model, the influence of several factors turned out to be statistically significant, for example, the risk of being closed is 2.5 times higher for branches. Socio-economic universities and private universities have around three times higher risk of being closed compared to the reference category (classical universities), even though they are all recognized as effective institutions. On the contrary, medical universities have great chances to survive, even if they are recognized as ineffective. Ineffective universities have six times the risk of being closed compared to effective ones. According to the socio-economic development of the region, the chances of universities in the regions which economies based on providing raw materials and background regions are less likely to survive (reference category is "world" cities, which are Moscow and Saint Petersburg). In general, we see that the family of the university, its effectiveness and the type (branch or main campus) have the strongest influence on survival.

Table 1. Cox proportional hazard risk model for closure of university

Variables	*Risk ratio*
Branch campus	2.554e+00***
Regional capital	7.895e-01
Competitiveness	1.004e+00*
Effective	1.725e-01***
Academic centre	8.668e-01
Family	
Agricultural universities	6.787e-01
Culture and Arts	1.415e+00
Transport	1.510e+00
Privilege	6.447e-01
Medical	3.634e-07***
Pedagogical	1.582e+00
Technical	1.707e+00**
Socio-economic	2.982e+00***
Other ME	3.353e+00***
Other state	1.252e+00
Municipal and regional	1.126e+00
Private, including religious	3.263e+00***
Regional typology	
Federal centre	1.462e+00
Region providing raw materials	2.350e+00**
Old industrial region	1.561e+00
Everage region	1.739e+00*
Region in crisis	1.735e+00
Special region	7.811e-01
R2	

Conclusion

Overall, it seems that choosing a set of performance indicators without enough regard to the structural sources of variation results in the survival of the best positioned, rather than the fittest or the most effective. The old sociological wisdom says that fair competition among unequal participants just increases the gap and serves only to legitimize the winner's advantage. Without considering the role of these factors, allocation of resources on the bases of universal formal criteria which is practiced in Russia now would inevitably lead to polarization of the higher education system and to degradation of the schools which were initially lacking "inborn" characters essential for success.

References

Mateos-González, José Luis; Boliver, Vikki (2018): Performance-based university funding and the drive towards 'institutional meritocracy' in Italy. In British Journal of Sociology of Education 33 (3), pp. 1–14.

Rebora, Gianfranco; Turri, Matteo (2013): The UK and Italian research assessment exercises face to face. In Research Policy 42 (9), pp. 1657–1666.

Sánchez-Barrioluengo, Mabel (2014): Articulating the 'three-missions' in Spanish universities. In Research Policy 43 (10), pp. 1760–1773.

New Measures of Journal Impact Based on Citation Network

Wataru Souma[1], Irena Vodenska[2], and Lou Chitkushev[3]

[1] *souma.wataru@nihon-u.ac.jp*
College of Science and Technology, Nihon University, Funabashi, 274-8501 (Japan)

[2] *vodenska@bu.edu*
Metropolitan College, Boston University, Boston, MA 02215 (USA)
Center for Polymer Studies, Boston University, Boston, MA 02215 (USA)

[2] *ltc@bu.edu*
Metropolitan College, Boston University, Boston, MA 02215 (USA)

Introduction

The number of citations is the most frequently used measure to quantify the significance of papers. However, the following question is of interest: which paper is the most significant if there are some papers with the same number of citations? In order to answer this question, some measures have been introduced, for example, the PageRank proposed by Brin & Page (1999).

Bollen et al. (2006) showed that the Institute for Scientific Information's (ISI's) impact factor (IF) is a metric of popularity. On the other hand, the PageRank is a metric of prestige. Chen et al. (2007) calculated the number of citations and the PageRank for all publications in the Physical Review family of journals from the year 1893 to 2003. The authors found that there exists a linear relationship between the number of citations and the PageRank, and also observed some outliers from this linear relationship, mainly papers of which the ranking of the PageRank is remarkably high while the citations are only slightly high, which are universally familiar to physicists. Thus, they refer to such papers as scientific "gems". Furthermore, Ma et al. (2008) confirmed that the same structure applies to the case of biochemistry and molecular biology.

These previous studies investigate the citation network of some selected scientific field. However, no study has considered the application of the concept of the PageRank to all papers in all field of science. Thus, one purpose of this paper is to present a method of finding papers based on prestige in all field of sciences (Souma & Jibu, 2018). Furthermore, this paper proposes new measures of popularity and impact of journals by using the number of citations and the PageRank of each paper published in the journal.

The remainder of the paper is organized as follows. In the next section, we explain the dataset used in this paper and also explain the concept of the PageRank. Subsequently, we show the relation between the number of citations and the PageRank. Furthermore, we define new measures of popularity and impact of journals and compare them with the SCImago Journal Rank and ISI's IF rank.

Data and definition of the PageRank

In this paper, we utilize the Science Citation Index Expanded (SCIE) provided by Clarivate Analytics Co., Ltd. This dataset contains bibliographic information of scientific papers published from the year 1900 to present. However, due to limited research budget, we utilize the dataset from the year 1981 to 2015. This dataset contains 34,666,719 papers and 591,321,826 citations.

If we consider papers as nodes and regard citations from a citing paper to a cited paper as directed links, one can create a directed network of citations from the dataset. We refer to such a network as the citation network. The citation network consists of many interconnected components. The giant weakly connected component (GWCC) comprises 34,428,322 nodes which are 99.3% of the total number of papers contained in the dataset, and of 591,177,607 directed links which are 99.98% of the total number of citations contained in the dataset. In what follows, we focus mainly on the GWCC.

Brin & Page (1999) proposed the so-called PageRank to obtain the appropriate ranking of a web page in the World Wide Web. The PageRank of paper i corresponds to the ranking of the Google number, g_i, defined by the recursion formula (Chen et al., 2007):

$$g_i = (1-d)\sum_{i\ \mathrm{nn}\ l} \frac{g_l}{k_l} + \frac{d}{N} \ , \qquad (1)$$

where $N = 34{,}428{,}322$ is the total number of papers contained in the GWCC. The sum is taken over the neighboring nodes l in which a link points to node i. In Equation (1), d is a free parameter that controls the convergence and effectiveness of the recursion formula. Brin & Page (1999) used $d = 0.15$ for the case of the World Wide Web. On the other hand, Chen et al., (2007) used $d = 0.5$ for the case of the citation network. Also, in this paper, we utilize $d = 0.5$.

New measures of journal impact

Figure 1 shows the double-logarithmic scale scatter plot of k_i and g_i. In this figure each black dot corresponds to a paper. The cyan-colored solid line represents the average value of g_i, $\langle g \rangle$, which is calculated for bins with logarithmically equal width and plotted against k_i. Furthermore, the figure shows that the plot of $\langle g \rangle$ versus k is smooth and increases linearly especially in the range $k \geq 500$.

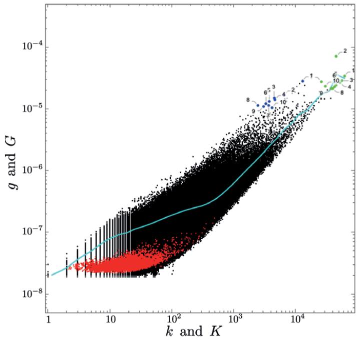


Figure 1. Black dots correspond to the scatter plot of k_i and g_i. Red dots correspond to the scatter plot of K_j and G_j.

For the journal j, we define the mean values of the number of citations, $K_j(y)$, and that of the Google number, $G_j(y)$, for the journal j as follows:

$$K_j(y) = \frac{1}{n_j}\sum_{i \in j}^{n_j \geq y} k_i \ , \quad G_j(y) = \frac{1}{n_j}\sum_{i \in j}^{n_j \geq y} g_i \ , \qquad (5)$$

where, n_j is the number of articles contained in the journal j. We consider journals that publish yearly and contain at least one article as appropriate for this study. Since we are studying 35 years (from the year 1981 to 2015) of data of the SCIE, so we set $y = 35$. Thus, the total number of journals is 15,533. Figure 1 shows the scatter plot of K_j against G_j as red dots.

For the journal j, we denote the ranking of the number of citations as $R_{K,j}$ and that of the PageRank as $R_{G,j}$. Further, we define the average ranking:

$$R_j = \frac{1}{2}\left(R_{K,j} + R_{G,j}\right) \ , \qquad (6)$$

and consider this as a new measure of journal popularity. Figure 2 shows the correlation between R and the ranking of the SCImago Journal Rank, R_{SJR}. In this case, the Spearman's rank correlation coefficient is found to be $\rho_{\text{SJR}} = 0.6963$. If we investigate the correlation between R and the ISI's IF rank, we obtain $\rho_{\text{IF}} = 0.6335$.

Figure 1 shows that there are journals for which $R_{G,j}$ is remarkably high and $R_{K,j}$ is slightly high. We refer to such a journal as "prestige journal." In order to select these jounals, we introduce the ratio; $F_j = R_{K,j}/R_{G,j}$ and define the conditional ranking of $R_{G,j}$ under F_j as follows:

$$R_j'(x) = R_{G,j}\left(F_j \geq x\right) \ , \qquad (2)$$

and regard this as a new measure of a journal impact. Figure 1 shows that there are papers of which $R_{G,j}$ is remarkably high and $R_{K,j}$ is slightly high. We refer to such a journal as an "extremely prestige journal." Table 1 shows the top 10 extremely prestige journals selected by the constraint $F_j > 3$. From the table, one will observe that there are more journals of information science.

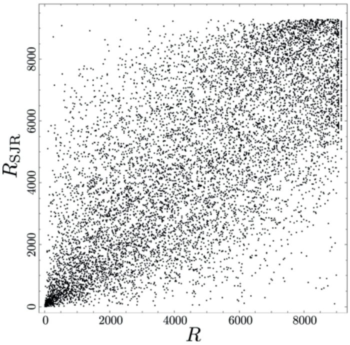


Figure 2. Scatter plots of R and R_{SJR}

Table 1. Top 10 prestige journals

Ranking	*Journal*
1	Advances in Physics
2	Annual Review of Fluid Mechanics
3	Computing Surveys
4	Advances in Organometallic Chemistry
5	Annual Review of Biophysics and Bioengineering
6	Advances in Parasitology
7	ACM Transactions on Computer Systems
8	Advances in Agronomy
8	IEEE Personal Communications
10	Advances in Cryptology - Eurocrypt 2000

Acknowledgments

This work is supported by Nihon University College of Science and Technology Grants-in Aid 2016.

References

Bollen, J., Rodriquez, M.A. & Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69, 669-687.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30, 107-117.

Chen, P., Xie, H., Maslov, S. & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1, 8-15.

Ma, N., Guan, J. & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing & Management*, 44, 800-810.

Souma, W. & Jibu, M. (2018). Progress of Studies of Citations and PageRank. In *Scientometrics* (pp. 213-231). IntechOpen.

Link Prediction of Knowledge Diffusion in Disciplinary Citation Networks based on Local Information

Zenghui Yue[1]   Haiyun Xu[2]   Guoting Yuan[3]   and   Qianfei Wang[4]

[1] *yzh66123@126.com*
School of Medical Information Engineering, Jining Medical University, Rizhao (China)

[2] *xuhy@clas.ac.cn*
Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu (China)

[3] *guotingyuan@hotmail.com*
School of Foreign Languages, Jining Medical University, Rizhao (China)

[4] *qianfeiwang@126.com*
School of Medical Information Engineering, Jining Medical University, Rizhao (China)

Introduction

With the rapid development of science and technology, the knowledge flow and exchanges among disciplines become more and more frequent with blurred boundaries. Knowledge flow based on citation could be considered an explicit process of knowledge diffusion. Link prediction and network science are mutually beneficial. An in-depth understanding of network structure can enhance the design of effective link prediction algorithms and the result and performance of link prediction algorithms can provide good estimation of network topological evolution and suggest guidelines for real network applications (Chen, Wang & Li, 2015). Link prediction of knowledge diffusion networks of disciplinary citation conducted in terms of complex network algorithms is an effective way to exploring structure changes and evolution trends of discipline knowledge flow and enhancing knowledge management and decision making.

Method

Based on structural information of knowledge diffusion network of disciplinary citation, nine similarity indices based on local information are employed for link prediction in both unweighted and weighted directed knowledge diffusion networks and a comparative analysis is conducted on prediction performance of each index. Finally, the disciplinary knowledge diffusion trend in the future is further predicted by employing RA in unweighted networks and CN in weighted networks. Link prediction of directed disciplinary citation networks falls into 2 categories, that is, link prediction of unweighted directed networks and link prediction of weighted directed networks. These similarity indices in unweighted directed networks and weighted directed networks are as shown in Table 1.
AUC is used to evaluate link prediction algorithms of knowledge diffusion in disciplinary citation networks (Lv & Zhou, 2010).

Table 1. Similarity indices based on local information in unweighted and weighted directed networks.

Similarity index	*Unweighted directed network*	*Weighted directed network*
CN	$S_{xy}=\lvert\Gamma_{out}(x)\cap\Gamma_{in}(y)\rvert$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)} w_{out}(x,z)+w_{in}(z,y)$
Salton	$S_{xy}=\frac{\lvert\Gamma_{out}(x)\cap\Gamma_{in}(y)\rvert}{\sqrt{k_{out}(x)\times k_{in}(y)}}$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{w_{out}(x,z)+w_{in}(z,y)}{\sqrt{s_{out}(z)\times s_{in}(z)}}$
Jaccard	$S_{xy}=\frac{\lvert\Gamma_{out}(x)\cap\Gamma_{in}(y)\rvert}{\lvert\Gamma_{out}(x)\cup\Gamma_{in}(y)\rvert}$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{w_{out}(x,z)+w_{in}(z,y)}{s_{out}(z)+s_{in}(z)-\sum(w_{out}(x,z)+w_{in}(z,y))}$
HPI	$S_{xy}=\frac{\lvert\Gamma_{out}(x)\cap\Gamma_{in}(y)\rvert}{\min\{k_{out}(x),k_{in}(y)\}}$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{w_{out}(x,z)+w_{in}(z,y)}{\min\{s_{out}(z),s_{in}(z)\}}$
HDI	$S_{xy}=\frac{\lvert\Gamma_{out}(x)\cap\Gamma_{in}(y)\rvert}{\max\{k_{out}(x),k_{in}(y)\}}$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{w_{out}(x,z)+w_{in}(z,y)}{\max\{s_{out}(z),s_{in}(z)\}}$
LHN-I	$S_{xy}=\frac{\lvert\Gamma_{out}(x)\cap\Gamma_{in}(y)\rvert}{k_{out}(x)\times k_{in}(y)}$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{w_{out}(x,z)+w_{in}(z,y)}{s_{out}(z)\times s_{in}(z)}$
PA	$S_{xy}=k_{out}(x)\times k_{in}(y)$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)} s_{out}(z)\times s_{in}(z)$
AA	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{1}{\lg k_{out}(z)}$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{w_{out}(x,z)+w_{in}(z,y)}{\lg(1+s_{out}(z))}$
RA	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{1}{k_{out}(z)}$	$S_{xy}=\sum_{z\in\Gamma_{out}(x)\cap\Gamma_{in}(y)}\frac{w_{out}(x,z)+w_{in}(z,y)}{s_{out}(z)}$

Data

By taking social network as an example, link prediction of knowledge diffusion among disciplines is conducted with nine well-known local similarity indices. The data was retrieved from the Web of Science. 28,168 primary pieces of literature were obtained. In terms of the comparison tables of discipline categories of journals in Journal Citation Reports (JCR), the annual knowledge diffusion networks of disciplinary citation are aggregated on the basis of journal citation data.

Results

Performance of nine similarity indices based on local information in disciplinary knowledge diffusion networks

From 1973 to 2016, AUC of nine similarity indices based on local information in unweighted and weighted disciplinary knowledge diffusion networks are shown in Figure 1 and Figure 2 respectively.

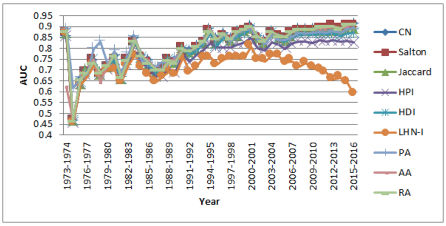


Figure 1. AUC of nine similarity indices based on local information in unweighted networks.

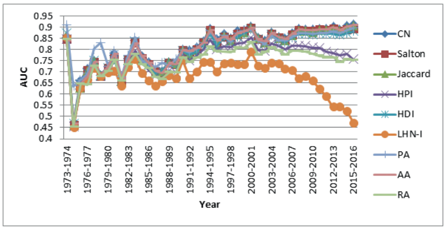


Figure 2. AUC of nine similarity indices based on local information in weighted networks.

Generally, concerning unweighted disciplinary knowledge diffusion networks, prediction performance of RA is the best with the AUC value reaching 0.89558, followed by that of AA and CN, and LHN-I shows the worst performance. Concerning weighted networks, prediction performance of CN is the best with the AUC value reaching 0.895602, that of AA is second to it, and that of LHN-I is still the worst.

Link prediction of knowledge diffusion in disciplinary citation networks

Knowledge diffusion network with newly added links predicted by RA based on unweighted disciplinary citations is shown in Figure 3. Edges in the network are newly added links among disciplines in prediction.

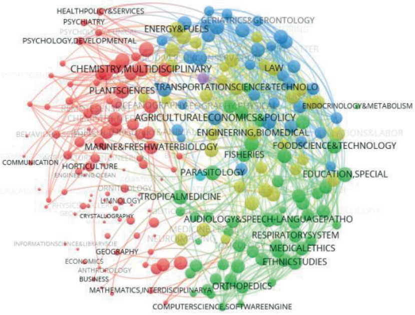


Figure 3. Knowledge diffusion network predicted by RA based on unweighted disciplinary citations.

Knowledge diffusion network with newly added links predicted by CN based on weighted disciplinary citations is shown in Figure 4.

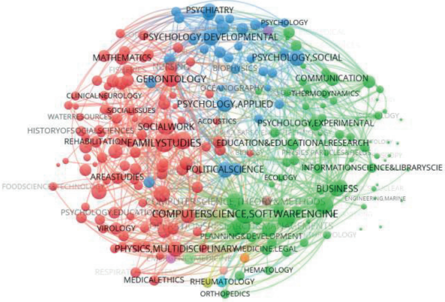


Figure 4. Knowledge diffusion network predicted by CN based on weighted disciplinary citations.

Discussion and Conclusions

Prediction accuracy of different indices shows dynamic changes in different time periods, which to some degree shall be related with the temporal dynamic structure and non-linear evolutionary topological characteristics of disciplinary knowledge diffusion networks.

In general, prediction performance of such indices as CN, Salton, Jaccard, HDI and PA in weighted networks is better than that in unweighted networks, while prediction performance of such predicators as HPI, LHN-I, AA and RA in unweighted networks is better than that in weighted networks. It means there exist certain degrees of weak-ties effect in disciplinary knowledge diffusion networks.

There are certain discrepancies of applicability of different link prediction indices in unweighted and weighted disciplinary knowledge diffusion networks.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 71704063), National Natural Science Foundation of China (Grant No. 71704170), Shandong Provincial Natural Science Foundation, China (Grant No. ZR2015GL015).

References

Chen, G.R., Wang, X.F. & Li, X. (Eds.). (2015). *Introduction to Complex Networks: Models, Structures and Dynamics (2nd)*. Beijing: Higher Education Press.

Lv, L.Y. & Zhou, T. (2010). Link prediction in complex networks: a survey. *Physical A*, 390, 1150-1170.

Dynamic Assessment of the Academic Influence of Scientific Literature from the Perspective of Altmetrics

WANG Feifei[1], JIA Chenran[2], LIU Jiayu[3] and LIU Junwan[4]

[1] *feifeiwang@bjut.edu.cn*
Research Base of Beijing Modern Manufacturing Development, College of Economics and Management,
Beijing University of Technology, 100124 Beijing (China)

[2] *18813085072@163.com*
Research Base of Beijing Modern Manufacturing Development, College of Economics and Management,
Beijing University of Technology, 100124 Beijing (China)

[3] *jyl420@163.com*
Research Base of Beijing Modern Manufacturing Development, College of Economics and Management,
Beijing University of Technology, 100124 Beijing (China)

[4] *liujunwan@bjut.edu.cn*
Research Base of Beijing Modern Manufacturing Development, College of Economics and Management,
Beijing University of Technology, 100124 Beijing (China)

Introduction

At present, with the deepening of the rapid development of open access to scientific literature has made academic communication and sharing a faster and more convenient way to communicate. The popularity of social networks in the Web2.0 environment enables metering at the paper level to produce massive, fast-growing, diverse, and widely covered data(Wang et al., 2015). Meanwhile, the dynamic development of altmetrics indicators has been growing quickly over time, so the dynamic assessment of the academic influence of scientific literature is an important direction of academic assessment and technology management.

In this paper, based on the comprehensive assessment system of the indicator weights and time weights of the sequential dynamic change, the traditional indicators and altmetrics indicators are used to assess the top tier scientific literature in genetic journals.

Data and methodology

Basic Data

The data retrieved from Web of Science with “PLOS Genetics” as the keywords, covering two types of documents (articles and reviews). The period covered is from 2006 to 2016. We obtained the citing literature of the published articles in the journal PLoS Genetics, including the full-record data of the reference. Then, this paper used the DOI information to obtain PDF downloads, HTML pageviews, and CiteULike storage through the PLoS ALMs tool. Finally, we used the Altmetric API to obtain the Tweets, altmetric score and other altmetrics indicators.

Methodologies

The assessment system is mainly composed of traditional indicators and altmetrics indicators, and uses objective weight to conduct research, which can avoid the uncertainty of subjective weight to a certain extent. The weights of the dynamic assessment indicators on different time windows were selected by the entropy weight method(EWM). This method has been applied to the study of university rankings (Jati & Dominic, 2017). The Grey Relational Analysis(GRA) method can be used to describe the degree of association on the timing sequence of multiple indicators of the assessment object. The method can realize the weight of time. Therefore, EWM and GRA method were selected to construct the assessment system. The comprehensive score expression for constructing the dynamic assessment system of the academic influence of scientific literature is as follows:

$$S = \sum_{k=1}^{n} W_{\gamma k}\left(\sum_{i=1}^{t} C_{ij} W_{ij}\right) \qquad (1)$$

where C_{ij} is the original data of the j_{th} indicator of the i_{th} month, W_{ij} is the time weight of the j_{th} indicator of the i_{th} month, t is the age of the single scientific literature, k is the k_{th} indicator, n is the number of indicators, γ is the age of the single scientific literature corresponding to the time window.

Results and discussion

To find the dynamic changes of timing sequence data, we used Python to perform the abnormality detection for citation frequency, PDF downloads, HTML pageviews, CiteULike storage and Tweet mentioned. To ensure the rationality of the time window for dynamic assessment and combined with the timing sequence of relevant research, a total of 10 time windows were divided with the month being the unit.

To establish a dynamic assessment system for scientific literature based on timing sequence data, EWM was used to obtain the specific indicator weights of five indicators in 10 time windows, as shown in Table 1. Considering the timeliness and diversity characteristics of the altmetrics indicators, the weight of the timing sequence first rising and then falling fully reflects the role of the altmetrics indicators in the dynamic assessment of scientific literature.

Table 1. Dynamic assessment indicator weight results on the time window

Time /month	Citation	PDF	HTML	CiteULike	Twitter
3	0.4567	0.0769	0.0696	0.1444	0.2524
6	0.2886	0.1001	0.0948	0.1838	0.3328
9	0.2171	0.1079	0.1037	0.2004	0.3708
12	0.1909	0.1094	0.1087	0.2056	0.3855
24	0.1702	0.1281	0.1096	0.2071	0.3850
36	0.1750	0.1318	0.1138	0.2048	0.3745
60	0.2056	0.1333	0.1404	0.1935	0.3272
72	0.2191	0.1362	0.1452	0.1899	0.3095
96	0.2061	0.1213	0.1264	0.3071	0.2392
132	0.2171	0.1294	0.1327	0.2973	0.2234

To scientifically explore the specific influence of time factors on the scientific literature in the timing sequence, GRA is performed by selecting the citation frequency, PDF downloads, HTML pageviews, CiteULike storage and Tweets mentioned in each month. The results are shown in Table 2. Due to the particularity of CiteULike storage and the Tweets mentioned data, the data of the evaluation object of these two indicators is highly discrete. Therefore, the range of the time importance of CiteULike storage is 96 months. The range of the time importance of the Tweet mentioned indicator is 84 months. The range of the importance order of the remaining indicators is 144 months. The impact of the citation frequency indicator on its academic influence is high in from one year to two years after the publication. The download volume and reading volume of the scientific literature from the beginning of the publication to the two-year time window shows a relatively high weight.

To achieve the dynamic assessment effect of the academic influence of scientific literature, we used the combination of time weight and indicator weight. After standardizing each indicator, the comprehensive assessment score was obtained using formula (1). The selected literature has a longer publication time, and the accumulated citation frequency, PDF downloads, and HTML pageviews value are higher, indicating that the publication age and the above three indicators have a greater impact on the influence of the literature. It can be seen that the Altmetric Score of the selected literature is not higher, which may be due to the fact that the algorithm strategy provided by Altmetric.com cannot verify the validity and accuracy of the data.

Table 2. GRA results based on the importance of timing sequence indicators

Time/ month	Citation	PDF	HTML	CiteULike	Tweet
1	0.0068	0.0076	0.0075	0.0288	0.0132
2	0.0067	0.0073	0.0073	0.0276	0.0114
…	…	…	…	…	…
…	0.0068	0.0067	0.0068	0.0100	0.0123

Conclusions

The rich assessment indicators of altmetrics have improved upon the lack of indicators in the traditional assessment system, and the characteristics of the participation of altmetrics have also broken the definition of peers in the traditional sense. Thus, dynamic academic influence assessment based on an altmetrics perspective has a higher theoretical and practical significance.

Based on traditional indicators, this paper introduces altmetrics indicators to evaluate the influence of the scientific literature. By analysing the dissemination process of the literature, a dynamic assessment model is constructed based on the timing sequence data, GRA is used to weight the time, and the weight of the indicator is given by EWM. By determining a reference score on the comprehensive assessment of the academic influence of scientific literature through double-weighted, the results are supposed to be more scientific.

References

Jati H.& Dominic D. D.(2017). A new approach of indonesian university webometrics ranking using entropy and PROMETHEE II. *Procedia Computer Science*, 124, 444-451.

Wang, X.W., Fang Z.C.& Wang H.Y.(2015). Study on the Continuous, Dynamic and Comprehensive Article-Level Evaluation System. *Science of Science and Management of S.& T.*, 36, 37-48.

Drawing the Conceptual Structure of Corporate Entrepreneurship using Co-Word Analysis

Manuel Castriotta[1], Michela Loi[1], Enrico Angioni[1] and Francesca Cabiddu[1]

[1] manuel.castriotta@unica.it
University of Cagliari, viale Sant'Ignazio, 74, Cagliari (Italy)

Introduction

Corporate entrepreneurship (CE) is a key construct for entrepreneurship (Kuratko et al. 2018). As a research topic, it has evolved dramatically over the last 40 years, during which definitions and research goals have multiplied (Nason et al. 2015). According to most of the studies, CE is a multidimensional construct, and many terms are used to refer to different aspects of the topic (Sakhdari 2016). For example, some authors associate conceptually and semantically the term CE with intrapreneurship, internal corporate entrepreneurship, corporate ventures, venture management, new ventures and, internal corporate venturing. Even though recent literature reviews show that the field is growing with several contributions, and progress has been made in its conceptualization (Sakhdari 2016), nevertheless there does not appear to be a universally accepted definition of CE (Nason et al. 2015). The main goal of this article, therefore, is to conceptually delimit the field by systematically describing its conceptual structure that we have drawn from the terms scholars were using to study CE over the last 26 years. The conceptual structure is here envisioned as a spatial representation of how terms are related to one another to form subgroups, in which distances among terms are also estimated (Cobo et al. 2014). To perform this study, we relied on bibliometrics. In particular, a co-word analysis is used to complement, improve and advance previous works on CE (Benavides-Velasco et al. 2013). To our knowledge, this is the first study in CE to be performed by the adoption of a quantitative method that considers words as a source of information. This systematic description helps identify the conceptual boundaries of the topic, showing its multiple meanings, and serves the scope of highlighting possible solutions to converge on a more shared conceptualization of the topic.

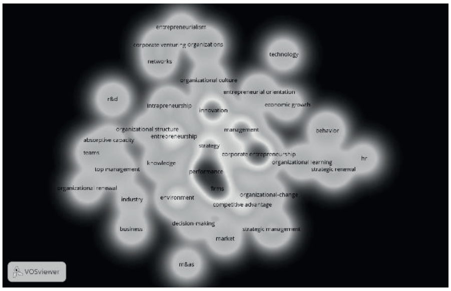
Method

The co-word analysis draws upon the assumption that a paper's keyword constitutes an adequate description of its content (Callon et al. 1983). Operationally, two words that co-occur within the same paper are an indication of a link between the themes they refer to. The presence of many co-occurrences around the same word or pair of words corresponds to a research theme (Ding et al. 2001). The data used for this paper were taken from the Web of Science (WoS) database. In our study, we cover 26 years of field history, as the chosen period ranges between the 1991 and 2017 years. This time frame is, then, segmented into three stages according to different spans: 1991 – 2006, 2007 – 2011, and 2012 – 2017, to investigate the field's evolution. Further, the document type selection was limited to scholarly journal articles and reviews written in the English language. In order to ensure that only relevant documents were considered in the final unit of analysis, a filtering procedure has been performed (Castriotta et al. 2019). We opted for a hybrid approach and merged the authors' keywords. This restriction ensures that our analysis reflected the intended emphasis of the authors (Zupic and Čater 2015). As a result of this merger between the author's keywords and keyword plus (hereafter "keyword"), we obtained a set of 2086 keywords. Keywords were then standardized by Apache NLP 1.5.3 vocabulary tool (Waltman et al. 2010). Through this preprocessing phase, the number of keywords decreased to 1828, among which 104 keywords occurred at least five times with which we built the row co-occurrence matrices required to perform the multivariate analysis.

Results

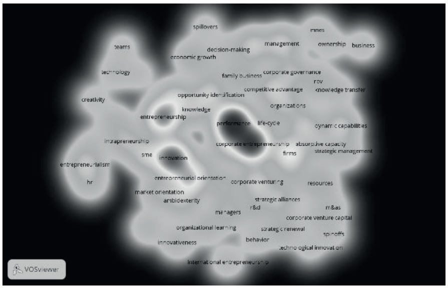
The distribution of the publications in the 1991-2017 period has a steady growth trend. 2016 is the most prolific year with 142 publications, followed by the 2015 (103 publications) and 2013 (94 publications). The top 10 keywords with the highest frequency are *performance, corporate entrepreneurship, innovation, entrepreneurship, entrepreneurial orientation, firms, management, SME, intrapreneurship* and *competitive advantage*. As regards the positioning of concepts in the first period, Figure 1 shows the terms having a polycentric structure, in which three main areas divide the center of the map. Terms such as *performance, corporate entrepreneurship* and *innovation* have a strong centrality. The second period (2007-2011), shows that *performance* and *corporate entrepreneurship* are still the central themes of the topic (see Fig. 2).

Figure 1 MDS density of the period 1991 - 2006



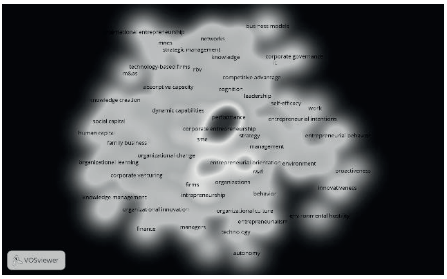
Terms such *as firms* and *management*, previously centrally located, moved to a semi-peripheral position.

Figure 2 MDS density of the period 2007 - 2011



The third period (2012- 2017) highlights deep changes in the conceptual structure of the topic (see Fig. 3). Neighboring concepts such as *corporate venturing, strategic renewal,* and *strategic entrepreneurship* are located close to each other on the left side of the map. *Intrapreneurship* is not too far at the bottom of the map.

Figure 3 MDS density of the period 2012 - 2017



Discussion and conclusion

Existing research has referred to CE by connecting this concept with various thematic areas such as strategy, innovation, management, organization, and finance. The coexistence of these different areas has contributed to expanding the CE field thematically, yielding the conceptual boundaries among the multiple components of CE to be less discriminable. By looking at our results, our analysis shows that the complex phenomenon of CE cannot be synthesized in only two sub-categories (corporate venturing and strategic entrepreneurship). From the MDS density maps, it emerges that *intrapreneurship, entrepreneurial orientation, entrepreneurship and entrepreneurialism* are also central concepts of the topic. Our findings reveal that words such as *strategic renewal* is still considered essential in describing CE and its activities as has been highlighted by Sharma and Chrisman in 1999. Furthermore, the thematic area of *corporate venturing*, divided into two different groups: *internal corporate venturing* and *external corporate venturing* appears to be more influenced by its external open innovation components and related to terms such as *R&D, spinoffs* and *M&As*.

References

Benavides-Velasco, C. A., Quintana-García, C., & Guzmán-Parra, V. F. (2013). Trends in family business research. *Small business economics*, 40(1), 41-57.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. Information (*International Social Science Council*), 22(2), 191-235.

Castriotta, M., Loi, M., Marku, E., & Naitana, L. (2019). What's in a name? Exploring the conceptual structure of emerging organizations. *Scientometrics*, 118(2), 407-437.

Cobo, M. J., Chiclana, F., Collop, A., de Ona, J., and Herrera-Viedma, E. (2014), A bibliometric analysis of the intelligent transportation systems research based on science mapping, *Intelligent Transportation Systems, IEEE Transactions on*, 15(2), 901-908.

Ding, Y., Chowdhury, G. G., and Foo, S. (2001), Bibliometric cartography of information retrieval research by using co-word analysis, *Information processing and management*, 37(6), 817-842.

Kuratko, D. F., and M. H. Morris. (2018). Corporate Entrepreneurship: A Critical Challenge for Educators and Researchers. *Entrepreneurship Education and Pedagogy*, 1(1): 42-60.

Nason, R. S., McKelvie, A., and Lumpkin, G. T. (2015), The role of organizational size in the heterogeneous nature of corporate entrepreneurship. *Small Business Economics*, 1-26.

Sakhdari, K. (2016). Corporate Entrepreneurship: A Review and Future Research Agenda. *Technology Innovation Management Review*, 6: 5-18.

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. Organizational Research Methods, 18(3), 429-472.

Current Status and Enhancement of Collaborative Research with ASEAN Countries: A Case Study of Osaka University

Shino Iwami[1], Toshihiko Shimizu[2], Melvin John F. Empizo[2], Jacque Lynn F. Gabayno[2], Nobuhiko Sarukura[2], Shota Fujii[2], and Yoshinari Sumimura[2]

[1] *iwami.research@outlook.com*
Osaka University, Suita (Japan)

[2] Osaka University, 1-1 Yamadaoka, Suita, Osaka (Japan)

Introduction

Osaka University (OU) has enacted Medium-term Plan (Osaka University, 2018a) every six year, and the third Medium-term Plan (April 2016 – March 2022) is being enforced as of 2018. OU has set the following numerical targets in order to promote globalization of its research.

- OU is running International Joint Research Promotion Program (Osaka University, 2018b), and build approximately 80 International Joint Labs until March 2022. [ID: 6-2]
- OU opens global campuses, and concludes at least 120 academic exchange contracts with abroad universities until March 2021. [ID: 10-2]

The shortage of labor force in super-aging society makes successions of sciences and technologies more difficult. Japanese universities cannot get adequate students only in Japan (Deguchi, 2018), and science and technology will be stagnant. If Japan wants to leave the technologies for the future generations somewhere in the world, the only solution is to delegate them overseas. Meanwhile, ASEAN countries are growing rapidly, and they are seeking the cutting-edge technologies for further economic development.

The purpose of this research is to overview current status of collaborative research between ASEAN countries and OU, and to identify: (1) strong academic collaborative countries for building global campuses, and (2) researchers performing international joint research to support in university strategy.

Methodology

This research was performed in the following steps:

(1) The bibliographic data related to OU in 2001-2018 were retrieved.
(2) Co-authoring relations of ASEAN – OU were counted, and transitions were illustrated.
(3) The bibliographic data related to the University of the Philippines (UP) in 2001-2018 were retrieved.
(4) Co-authoring relations of UP were counted to investigate the impacts of OU from the view of UP.

For (1) and (2), the bibliographic data of 105,772 academic papers were retrieved from the Web of Science; the data were provided by Clarivate Analytics (as of October 02, 2018 with the query, OG=("Osaka University"). For (3) and (4), the bibliographic data of 8,516 academic papers were retrieved from the Web of Science with the query, OG=("Univ Philippines*"). This research excluded academic papers with more than 1,000 authors.

OU made strategy every six years: 2004-2009, 2010-2015, and 2016-2021. Thus, co-authoring relations were counted every half of strategy period in addition to the previous three years (2001-2003). Integer counting (Park et al., 2016) was used.

Result and Discussion

Top Collaborative Researchers in OU

Figure 1 shows the breakdowns of world co-authoring partners with the top 10 OU researchers. They are limited to researchers who have written co-authored papers with researchers in ASEAN. The thickness of the ring becomes wider as the number of papers in that period increases.

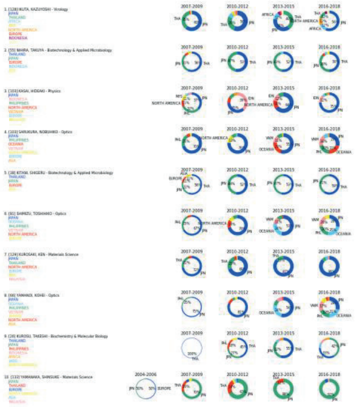


Figure 1. Share of Co-authoring Countries and Areas with ASEAN-related Top 10 Researchers in OU.

Table 1 shows the dependence rate for ASEAN in 2001-2018, and the high dependence rate will indicate the high potential of the collaboration about the strategy towards ASEAN. Within Table 1, researchers in biochemistry have focused on ASEAN. The dependence rate is defined as the below:

$$A_k = [number\ of\ ASEAN\ countries\ of\ coauthors]$$

$$B_k = [number\ of\ countries\ of\ coauthors]$$

$$Dependence\ rate = \frac{\sum A_k}{\sum B_k}$$

Table 1. Dependence Rate of ASEAN-related Top 10 Researchers in OU.

	Researcher	Dependence Rate
1	IKUTA, KAZUYOSHI	33.3%
2	NIHIRA, TAKUYA	54.8%
3	KASAI, HIDEAKI	33.1%
4	SARUKURA, NOBUHIKO	23.2%
5	KITANI, SHIGERU	54.7%
6	SHIMIZU, TOSHIHIKO	23.1%
7	KUROSAKI, KEN	17.1%
8	YAMANOI, KOHEI	23.7%
9	KUROSU, TAKESHI	60.9%
10	YAMANAKA, SHINSUKE	14.7%

Share of OU from the View of Collaborative Researchers: in the Case of UP

In Figure 2, the 5th Salvador and 10th Sarmago are high potential researchers that OU will be able to ask to collaborate on research for ASEAN campuses based on the past research collaborations. For them, the rate of OU is high, so they will cooperate with the OU more than other universities.

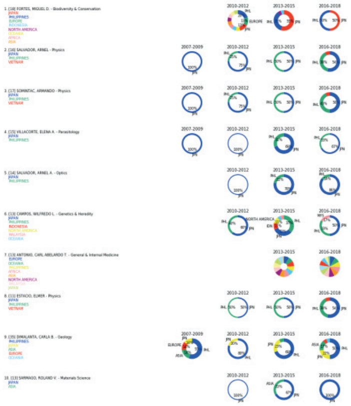


Figure 2. Share of Co-authoring Countries and Areas with ASEAN-and-Japan-related Top 10 Researchers in the UP.

Conclusion

This research was performed on the purpose of providing evidences for OU's strategy of internationalization from the view of evidence-based decision-making. The method is based on analyzing co-authorship relations, and it intends to let the strategy of internationalization succeed more securely than new exploration by expanding past research activities. Especially, the co-authoring relations between researchers are expected as deep and long-lasting relations, which will contribute to university's strategies.

In order to decide partners for OU's ASEAN campuses, the findings are:

- From the view of the strength on co-authoring countries, Thailand, Vietnam, Malaysia, and Indonesia should be selected from the view of collaborative achievements.
- From the view of the strength on co-authoring institutes, Mahidol University (Thailand) is suitable to tighten the collaborations.
- From the view of helps by currently active OU researchers, connections with the Philippines are expected.

The originalities of this research are time-series visualization of shares and the analysis from the view of partners. In the future, the analyses in this research will be expanded to the analyses in the other regions for the practical use of university strategy.

References

Deguchi, H. (2018, September 30). The direction of education in Japan. *The Japan Times*. Retrieved February 3, 2019, from https://www.japantimes.co.jp/opinion/2018/09/30/commentary/japan-commentary/direction-education-japan/

Osaka University. (2018a). Medium-term Plan (in Japanese). Retrieved September 5, 2018, from http://www.osaka-u.ac.jp/ja/guide/information/joho/keikaku.html

Osaka University. (2018b). International Joint Research Promotion Program. Retrieved September 5, 2018, from http://www.osaka-u.ac.jp/en/research/researcher_sp/international_joint

Park, H. W., Yoon, J., & Leydesdorff, L. (2016). The Normalization of Co-authorship Networks in the Bibliometric Evaluation: The Government Stimulation Programs of China and Korea. Retrieved from http://arxiv.org/abs/1605.03593

Accreditation of Graduate Courses in Brazil: Analysing the Evaluation of the First Proposals of Professional Doctorates in the Country[i]

André Brasil[1]

[1] *a.l.brasil@cwts.leidenuniv.nl*
Brazilian Agency for Support and Evaluation of Graduate Education (CAPES)
Centre for Science and Technology Studies (CWTS), Leiden University

Abstract

This paper investigates the assessment of the first-ever proposals for accreditation of professional doctorates in Brazil. This modality of course was implemented in the country in 1998 and was designed to bridge the gap between academia and the productive sector, further integrating scientific research and societal needs. At first, the modality was restricted to the master's level, and only in 2017 new legislation authorised institutions to submit professional doctorate proposals to CAPES: the Brazilian agency in charge of accrediting graduate education. By May 2019, 30 new courses were approved, and this research analysed the evaluation reports of all of the 135 proposals initially submitted, in order to identify the criteria used to either or not accredit the courses. From the coding of such reports, it was also possible to map what the agency expects to see in successful proposals, as well as to ascertain if the evaluation process has been conducted coherently and consistently across different fields. With that, this paper can also be seen as a contribution to Brazilian academia in the design of future professional doctorate proposals.

Introduction

Graduate education started in Brazil in the first decades of the 20th Century, at first as a reflection of the professional higher education model which dominated the country until the end of World War II. It was only from the 1950s, with the enhancement of the Brazilian development process, that universities began conducting research on top of teaching, and the activity found its place primarily within masters and doctorate programs (Sucupira, 1980). Balbachevsky and Schwartzman (2010) described this process as the graduate foundations of research in Brazil. The role of such programs today is impressive since they account for at least 80% of all research in science & technology conducted in the country (SBPC, 2018).

The numbers of the Brazilian System of Graduate Education are also notable. By January 2019, there were already 6592 active courses in the country, most of them in the academic modality: 2247 doctorates and 3557 masters (BRASIL. Ministério da Educação. CAPES, 2014a). The remaining 788 courses were professional masters: a modality first authorised in 1998 through legislation which allowed higher education institutions to develop courses to "articulate teaching with the professional application, in a differentiated and flexible way" (BRASIL. Ministério da Educação. CAPES, 1998). After two decades of experience at the masters level, in 2017 the Ministry of Education extended the authorisation of the modality for the doctoral level as well, and 135 proposals for the accreditation of such courses were presented since then (BRASIL. Ministério da Educação. CAPES, 2017a).

In Brazil, the accreditation of new graduate courses is compulsory, being one of the duties of the Brazilian Agency for Support and Evaluation of Graduate Education (CAPES). This organisation was founded in 1951 in order to strengthen the development of science, technology and innovation in the country: a task performed through the evaluation, accreditation and funding of graduate courses (Guimarães & de Almeida, 2012). For the agency, the assessment of the first proposals for professional doctorates was a challenge, as the criteria to be adopted for the evaluation was undefined. Legislation authorising new projects for these courses described only the need to strengthen the relations between research institutions and professional sectors, both public and private (BRASIL. Ministério da Educação. CAPES, 2017a).

As described by CAPES ordinance 161/2017[1], the evaluation process starts at the higher education institutions, where prospective graduate program directors (GPD) draw proposals to be submitted to CAPES. Such projects contain information about the course, including objectives, faculty involved, the institution's infrastructure for teaching and research, and more. With the approval of the institution's pro-rector for research and graduate education, the project is submitted to one of 49 possible evaluation fields at CAPES. Fields such as Economics, Philosophy, Chemistry, Education, exist at the agency in order not only to organise and manage graduate programs in Brazil but also to perform evaluations in a way that accounts for variations among such fields (BRASIL. Ministério da Educação. CAPES, 2017b).

The lack of clear guidelines for the design of the first professional doctorate proposals, both at the macro and at the field level, could be seen as a problem. However, according to CAPES, the idea behind the decision was that any criteria or indicators defined by the agency beforehand would influence and limit the ground-breaking potential of proposals. The idea was to allow Brazilian academia to present unrestricted proposals in this first round. From work conducted by CAPES and its scientific committees on their evaluation, a report would become available for the next cycle of evaluation. This report would contain expectations, guidelines and indicators for the accreditation of professional doctorates (Barata, 2017).

Based on that foundation, the evaluation of new professional doctorates took place from mid-2018 until May 2019, when the accreditation results for the last proposals were released (BRASIL. Ministério da Educação. CAPES, 2019). By analysing such results, this research aims to understand the performed assessment to identify the criteria adopted to either or not approve each new course. From that, the goal is to obtain a thorough understanding of what the agency and the scientific committees involved in the evaluation expect from such courses and then provide a guide to what higher education institutions should consider when designing new doctorate proposals in the professional modality.

Methods

At every evaluation cycle, which usually takes place yearly, CAPES receives hundreds of proposals for the accreditation of courses. The most recent cycle included proposals of 2017 and 2018: a record of 1.354 submissions, distributed as shown in Table 1.

Table 1. Proposals submitted to CAPES, in the evaluation cycle of 2017/2018, for the accreditation of new graduate courses in Brazil.

Level / Modality	*Masters*	*Doctorate*	*Masters and Doctorate*	***Total***
Academic	409	312	53	**774**
Professional	445	111	24	**580**
Total	**854**	**423**	**77**	**1.354**

Each proposal can consist of a single course, in the masters or doctorate level, or include the two levels. In this last case, results are independent, and accreditation can be given to both, either or neither proposed level. As stated before, the focus of this research is on the 135 proposals of professional doctorates submitted to CAPES (111 for doctoral courses only and 24 for both master's and doctorate levels). Such proposals were submitted to 31 distinct fields. Their distribution can be seen in Figure 1, where fields are arranged according to the broad groups to which they are associated.

[1] In February 2019, CAPES issued ordinance 33/2019 regulating the evaluation process of new graduate course proposals. Even though there is already updated legislation about the topic, the evaluation covered by this research was performed under ordinance 161/2017, so this is the one which was considered for the analysis.

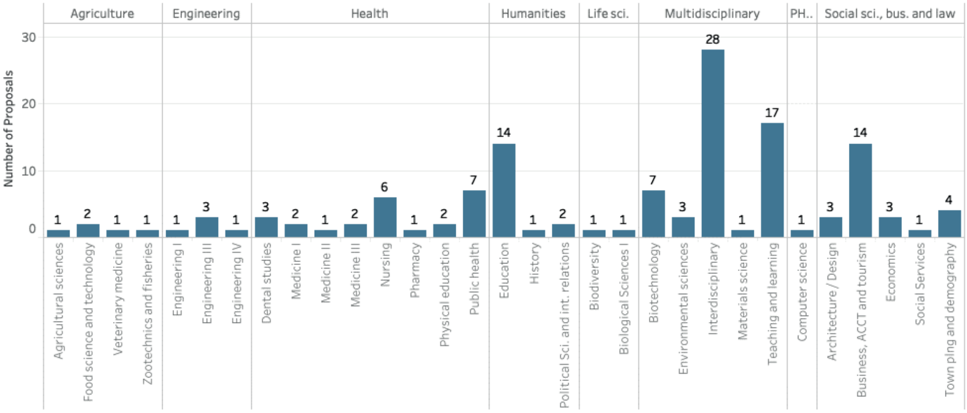


Figure 1. Distribution of submitted professional doctorate proposals in the 2017/2018 evaluation cycle, by evaluation field and broad group.

Figure 1 shows an uneven distribution. Several fields have received one or two proposals only. Three fields received more than five proposals (Nursing, Public Health and Biotechnology) and only four took in over ten (Education, Business, Teaching and Learning, Interdisciplinary). The high number of proposals in the Interdisciplinary field reflects a general tendency of the area. It received 141 submissions on all modalities, accounting for more than 10% of the 1.354 in total.

Even though there are differences in the evaluation process, every field appoints a scientific committee to assess their proposals. In order to judge the projects consistently and in a way that allows crossfield comparison, each committee conducts the analysis guided by a predetermined assessment form, shared by every field. In this, four dimensions are judged in merit, as described in Table 2:

Table 2. Shared dimensions analysed for the assessment of new graduate course proposals.

Dimension	*What is assessed by the committees?*
Conditions provided by the institution	Does the proposal provide indicators that the institution is committed to the implementation and success of the proposed course? Can the program count on an essential infrastructure to support its activities (physical structure, laboratories, library, computer resources, and more)?
Course proposal	Is the proposal adequately designed, with clearly defined and articulated objectives, concentration areas, research lines and curriculum structure?
Faculty size and workload	Is the number of professors, notably those full-time in the institution, enough to support the course activities, considering the concentration areas and the number of students expected?
Faculty productivity and research capacity	Does the program have, especially within its permanent professors, a group of researchers with scientific maturity confirmed by their production in the past five years? Are these researchers integrated in a way that allows the development of research projects as well as teaching and supervision activities?

From the assessment of each proposal concerning the four dimensions presented in Table 2, committees produce evaluation reports. These documents reflect a qualitative and quantitative analysis of the projects and carry the field recommendation of whether or not to approve the accreditation of the new course. After the assessment by field committees, reports are forwarded for examination by evaluators in distinct fields. For example, a proposal presented to the Economics committee might have its assessment reviewed by evaluators from the fields of Political Science and Architecture, in a process that could be described as an "evaluation of the evaluation". Even though such review must consider the criteria and characteristics of the field of the submission, its objective is to guarantee that the assessment is fair and coherent.

Once this analysis is performed and the results discussed between all parties, the report follows to a final evaluation by CAPES' Technical and Scientific Council for Higher Education (CTC-ES). This Council counts with 20 field coordinators representing all the nine broad groups dis-cussed earlier, as well as representatives from the Brazilian Association of Graduate Students (ANPG), the Brazilian Forum of Pro-Rectors for Research and Graduate Education (FOPROP), and from CAPES itself. At the CTC-ES meeting, the counsellors in charge of each proposal present their assessment of the whole evaluation process and, after the necessary debate and voting by all members of the Council, the final decision is included in the evaluation report and, from that, the results are made public.

This research is based on the examination of the reports of all the 135 proposals for accreditation of professional doctorates submitted in the 2017/2018 evaluation cycle, according to results released by May 2019[2] (BRASIL. Ministério da Educação. CAPES, 2019). From them, an analysis of how the four discussed dimensions were assessed for each proposal is performed in order to identify the most common strengths and fragilities of the evaluated projects.

Findings and discussion

To begin to understand the results of the professional doctorates assessment, it is essential to look at some descriptive statistics related to the evaluation results published by CAPES until May 2019 (BRASIL. Ministério da Educação. CAPES, 2018). For that, Figure 2 displays absolute and relative accreditation information on 1.404 courses. Courses of the 77 proposals that requested accreditation of masters and doctorate levels at once (see information on Table 1) are accounted separately on the graph. The reason is that, as stated earlier in the paper, the result of the analysis is independent for each level.

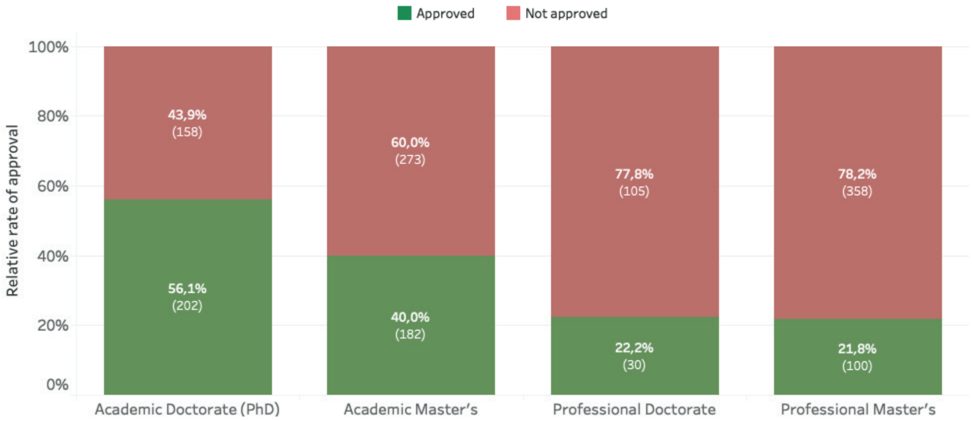


Figure 2. Relative approval rates, accompanied by absolute numbers, of new course proposals evaluated by CAPES by May 2019.

[2] This includes the first analysis for 100 proposals (subject to reconsideration requests) as well as results for such requests presented for 35 proposals, and which were already judged by May, 2019.

Figure 2 shows that 30 professional doctorates have been accredited, an approval rate of 22,2%. When compared to a success rate of 56,1% for academic doctorates and 40% for academic master's, this rate is strikingly low. Although, when compared to the approval of professional master's, at 21,8%, the low success rate seems to be related to professional proposals in general, rather than to a problem with the doctoral projects for the new modality. Thus, by investigating the reasons for not approving professional doctorates, it might be possible to understand why proposals in the modality have not been successful overall.

Another relevant analysis on the approval of professional doctorates comes by considering that a proposal can ask for the accreditation of a master's and a doctorate at once, or only of a doctorate. Figure 3 shows how the evaluation results relate to this aspect of the proposals.

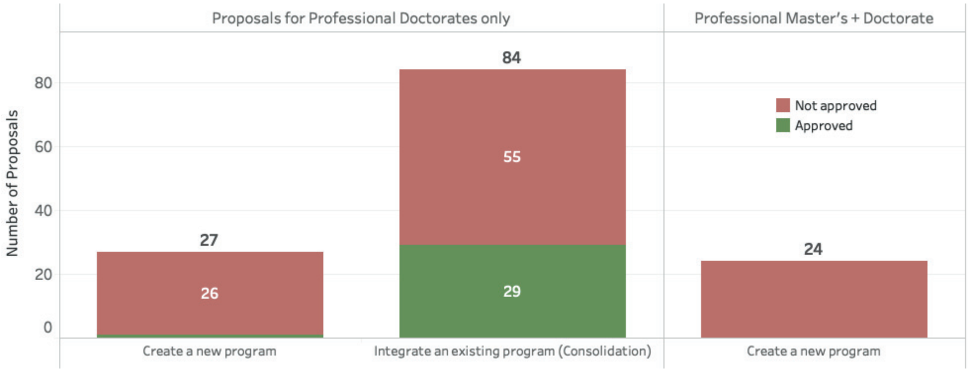


Figure 3. Panorama of the approval rate of professional doctorates, according to if is a joint proposal with a master's and if it will create a new program.

As it can be seen, even though the regulations for new course proposals do not restrict such submissions, joint professional master's and doctorate proposals – or even isolated doctorates without the previous experience of stablished professional masters – embodies little to no chance of accreditation. Actually, out of the 30 approved doctorates, 29 will integrate an existing program, building over the experience of a previous master's course already in place. The only exception is a single course establishing a new program, but its evaluation report mentions it is an associated endeavour of five distinct institutions, some counting with established master's individually. So, even this one course builds on previous experience.

After this initial understanding of the accreditation of professional doctorates, the next step would be the analysis of the strengths and weaknesses of the projects. In that sense, the first conclusion from the review of the 135 evaluation reports is that the ground-breaking proposals expected by CAPES did not materialise. What could be seen in every report, even for accredited courses, is that projects took few risks, mostly trying to replicate the traditional formula of previous professional master's and academic doctorate proposals. As a consequence, the evaluation process revolved around the feasibility of the presented projects.

An additional aspect of the evaluation reports is that favourable assessments regarding the four dimensions (Table 2) included recognition of quality, but did not provide detailed descriptions as to why quality was present. For example, a report might say that the course proposal is coherent for a professional doctorate, but would not explain the reasons why; or it might say the infrastructure is adequate for the proposal, but never describe what made it so. As a result, focusing on positive assessments did not generate enough information to map the criteria for the accreditation of new courses.

Fortunately, negative assessments in the reports were mostly followed by clear accounts of what should be improved to achieve the expected level of quality for accreditation. Thus, a decision was made to achieve the research's objective in a better way: instead of mapping the reasons why committees and the CTC-ES would accredit a new course, we focused on the more comprehensive descriptions to refuse accreditation and then extrapolated the conclusions to establish the main requirements to get a proposal approved. With that, Figure 4 focus on the non-approved proposals, offering some insight into how the four dimensions presented in Table 2 were evaluated.

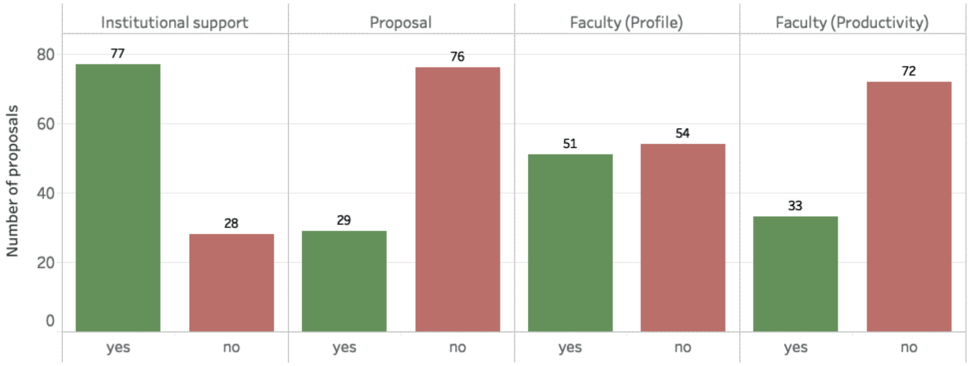


Figure 4. Panorama of the assessment of the four dimensions of analysis, considering only the 105 professional doctorate proposals which were not approved.

The first thing to notice in Figure 4 is that more than 70% of the non-approved proposals were evaluated positively regarding institutional support. This shows that most higher education institutions are not only committed to the implementation and success of their proposed courses but can also provide the necessary infrastructure to support these activities.

The examination of the reports for the 28 proposals which were not considered to be good enough in their institutional support shows that the main reason for a negative assessment is the lack of the necessary documentation either to show the commitment of the institution with the new course or to present the required course regulations. This problem appeared in 15 of the negatively evaluated proposals. The second most common reason for a poor evaluation in this dimension is a superficial, poorly detailed or imprecise description of the available infrastructure, which might indicate an absence of such foundations for teaching and research. This has been seen in 12 proposals and was followed by an actually inadequate infrastructure (observed in 5 proposals), or the lack of evidence that the institution would be able to maintain the presented infrastructure over time (two proposals).

While most institutions can provide adequate support for their courses in the first dimension, 54 out of the 105 non-approved proposals were assessed negatively in the dimension related to "faculty size and workload." The reasons for that are mostly regimental, and in most cases against general evaluation norms available at CAPES' website: 1. Reduced number of professors in relation to field expectations; 2. Professors are taking part in a larger number of courses than what it is allowed by the field or current regulations; 3. Professors do not have previous or adequate supervision experience; 4. Reduced number of faculty hours dedicated to the program. (BRASIL. Ministério da Educação. CAPES, 2019)

The dimension with the lowest evaluation scores is the "course proposal" itself. A total of 76 proposals received negative assessments in this dimension. The top five reasons for this, according to the reports were: 1. The proposal seems to have an academic approach, instead of a professional one; 2. The project lacks the depth or the level of innovation which is expected from a doctoral course; 3. The presentation is superficial, and either the objectives or the ways to accomplish them cannot be assessed; 4. The course proposed seems to be an unnecessary replica of another in existence at the institution; 5. The proposed syllabus, concentration areas or research projects are poorly designed or do not articulate with the course's objectives.

Regarding the fourth and last dimension, "faculty productivity and research capacity," the object of the evaluation is to assess the previous scientific production of the faculty in order to check their capability to develop the proposed research. To provide such information, proposals include a portfolio of up to five distinct products per researcher. However, in 72 out of 105 non-approved projects, proponents included mostly or only products with an academic orientation in the proposals, omitting the expected technical production. In some situations, it is evident that the information provided was selected based on journal rankings and usual scientific production metrics. As a consequence, some portfolios included products that had no direct relation to the objective of the new course, which would not help to measure the researchers' experience in the proposed field of work. Also, if portfolios lack technical or technological production, listing only papers published in indexed journals, evaluators could not verify the faculty's ability to run a professional, applied graduate course.

Even though the subject deserves a continuous and even more in-depth investigation, the main requirements for a proposal to obtain positive evaluations on any given dimension are now more evident: 1. Proposals should present proper institutional support for the new course, including the necessary infrastructure, adequately described and documented; 2. Proposals should provide evidence of their applied research approach, making clear that the new course is a professional one. The projects may not be superficial in this sense, and every aspect of the course's design should reflect that: from concentration areas to the syllabus; 3. Faculty size, profile and workload should respect field expectations, which are referenced in related legislation and also in field documents available at CAPES website; 4. Scientific production listed in researchers' portfolios should not only be a collection of their best work, indicator-wise. Such portfolios should display their ability to conduct high-quality research within the field of the proposed course, as well as their capacity to translate science into practice. Thus, technical production should be included.

The expectations for accreditation that were listed above appeared in most evaluation reports and no abnormalities were found across distinct fields. With consistent results, what is still unknown about the non-approved proposals relates to how far they are from being approved. To answer this question, Figure 5 displays how such proposals performed in terms of positively evaluated dimensions.

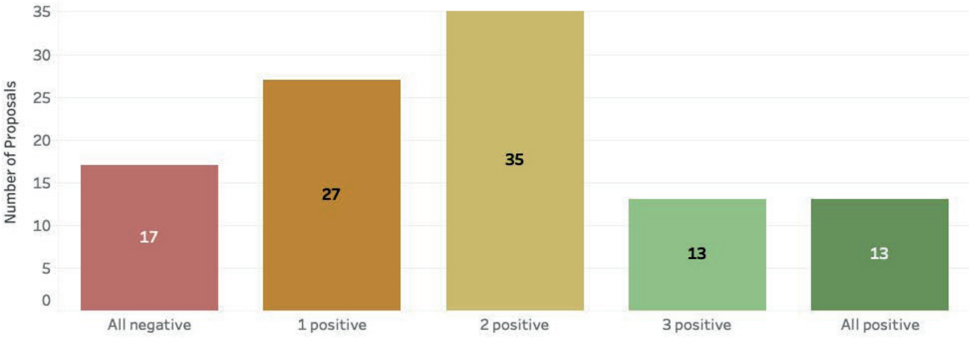


Figure 5. Distribution of proposals according to the number of dimensions that received a positive assessment (only non-approved proposals)

Out of the 105 non-approved proposals, only 17 received low evaluation scores across-the-board, showing to be very far from what it is expected from a graduate course at a doctoral level. Then there were 27 proposals with one positive dimension, 35 with two positive dimensions and 13 with only one negative dimension. These last ones would be "almost there," having to perfect just one single factor to be able to be accredited in a future evaluation cycle.

Figure 5 also shows that 13 proposals were evaluated positively in every dimension, but were not accredited nevertheless. Even though this might seem incongruous, the analysis of the reports confirms evaluations were conducted coherently in these cases, and two situations are present here. In the first one, seen in four joint proposals for master's and doctorates, the quality of the project was recognised in all four dimensions, but either the faculty or the project itself were not considered ready for the doctoral level. Thus, only the master's course was accredited.

In the second situation, the reports of nine proposals show that the evaluation field committees recommended the accreditation of the courses, but the CTC-ES disagreed with the assessment. In six of these cases, the Council requested proponents to provide additional documents or clarifications regarding the proposals, or even appointed a committee to visit the institutions to elucidate eventual doubts regarding the analysis. Ultimately, these proposals were not approved for three distinct reasons: 1. In one of the cases, the proponents could not make clear what was the difference for the new doctorate concerning the master's course already in place at the institution; 2. In four proposals the Council identified that a large portion of the faculty was already involved in other graduate programs, in numbers either against current regulations or in a percentage that would hurt the development of the proposed research; 3. Finally, in five cases, the CTC-ES considered that the proposal had an extremely academic profile, instead of the professional one that was necessary.

As discussed earlier in the paper, the evaluation process adopted by CAPES includes distinct phases of analysis and review, starting with scientific committes and going through external scrutiny and posterior examination in a multidisciplinary council. From the report analysis it became clear that such multilevel evaluation was relevant to guarantee that no proposal was wrongly assessed, as the review allowed an additional look at the projects and the eventual adjustment of the results.

Conclusion

The Brazilian Agency for Support and Evaluation of Graduate Education (CAPES) faced a challenge in designing the first-ever evaluation of professional doctorates in the country. One of the most critical choices in the process was not to set the assessment criteria and indicators in advance but to develop them throughout the evaluation process. According to Barata (2017), the reason for this decision was that the agency trusted the capabilities of Brazilian academia to "think outside the box," which could result in the submission of inspired proposals.

As argued, these ground-breaking proposals expected by CAPES did not materialise. The projects presented did not take the opportunity to innovate, avoiding risks by using established formulas previously seen for professional masters and academic doctorates. As a result, the evaluation process needed to focus on the viability of the proposed courses. That was an undesirable result, but there was also an additional threat, in the absence of predefined guidelines for the evaluation: the whole process could lack adequate coherence.

Fortunately, the analysis of the evaluation reports was able to show that CAPES' assessment was well conducted, and the expectations for the approval of new professional doctorates are consistent throughout the 31 fields that received proposals so far. By the time this paper is published, the official criteria used for the accreditation of professional doctorate proposals will probably be public. Despite that, it is already possible to make some statements. For example, even though there is not yet any rule against a professional master's/doctorate proposal or a standalone doctoral one, evaluators expect institutions to have prior experience at the professional master's level before they can venture into the doctoral level. Unless such institutions can present an exceptional case, they will be wasting their time and energy in such proposals.

Some other requirements for the approval of professional doctorate projects also became evident throughout this research, including the need for clear institutional support for new courses; the inclusion of an adequately sized and experienced faculty; as well as the presentation of a coherent course structure counting with suitable concentration areas, research projects and syllabus. Nevertheless, the most crucial element considered for the accreditation of these new professional doctorates was the applied research approach, indispensable for the success of any proposal.

At the time of writing, the results of the accreditation process are not yet final. This is due to the fact that institutions are allowed to request reconsideration of their proposals after a first negative assessment. However, what CAPES and the community of expert reviewers expect from professional doctorates became more evident from the conducted analysis. Thus, the hope for the evaluation process is that it will be improved and better documented from the first cycle of proposals.

References

Balbachevsky, E., & Schwartzman, S. (2010). The graduate Foundations of research in Brazil. *Higher Education Forum*, *7*(1), 85–101. Hiroshima.

Barata, R. de C. B. (2017). Avaliação Quadrienal: Balanço e Perspectivas. Presented at the XXXIII ENPROP, João Pessoa.

BRASIL. Ministério da Educação. CAPES. (1998). ***Portaria nº 80, de 16 de dezembro de 1998.*** *Dispõe sobre o reconhecimento dos mestrados profissionais*. *Diário Oficial da União*.

BRASIL. Ministério da Educação. CAPES. (2014a). Plataforma Sucupira. Brasília. Retrieved February, 2019, from http://www.capes.gov.br/avaliacao/plataforma-sucupira

BRASIL. Ministério da Educação. CAPES. (2017a). ***Portaria nº 131, de 28 de junho de 2017.*** *Dispõe sobre o mestrado e o doutorado profissionais. Diário Oficial da União*.

BRASIL. Ministério da Educação. CAPES. (2017b). ***Portaria nº 161, de 22 de agosto de 2017***. *Avaliação de propostas de cursos novos de pós-graduação stricto sensu. Diário Oficial da União*.

BRASIL. Ministério da Educação. CAPES. (2018). Sobre a Avaliação. *www.capes.gov.br*. Brasília. Retrieved May 2019, from http://capes.gov.br/avaliacao/sobre-a-avaliacao

BRASIL. Ministério da Educação. CAPES. (2019, May). Resultado de análises das propostas de cursos novos 2017/18. *www.capes.gov.br*. Brasília. Retrieved May 2019, from http://capes.gov.br/avaliacao/entrada-no-snpg-propostas/mestrado-profissional/resultados

Guimarães, J. A., & de Almeida, E. C. E. (2012). Quality assurance of Post-Graduate education: the case of CAPES, the Brazilian agency for support and evaluation of Graduate education. *Higher Learning Research Communications*, *2*(3), 3.

SBPC. (2018). 80% da pesquisa no Brasil está ligada a programas de pós-graduação. *portal.sbpcnet.org.br*. Retrieved February 2019, from http://portal.sbpcnet.org.br/noticias/80-da-pesquisa-no-brasil-esta-ligada-a-programas-de-pos-graduacao-2/

Sucupira, N. (1980). Antecedentes e primórdios da pós-graduação. *Forum Educacional*, *4*(4), 3–18. Retrieved from http://bibliotecadigital.fgv.br/ojs/index.php/fe/article/view/60545/58792

[i] This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

ISSI2019 Reframing the Absorptive Capacity's Mediating Effects on R&D Investment: Organizational Barrier and Quadruple-Helix Collaboration

Chang, Ching-Chun[1] and Liu, Tai-Ying[2]

[1] *jjzhang@narlabs.org.tw*
Science and Technology Policy Research and Information Center, National Applied Research Laboratories, Taipei, Taiwan

[2] *tyliu@narlabs.org.tw*
Science and Technology Policy Research and Information Center, National Applied Research Laboratories, Taipei, Taiwan

Introduction

This research is aim at reconceptualizing the organizational absorptive capacities, redesigning moderators and mediators in the translation processes of absorptive capacities's impact on innovation activities, and operationalizing the measurement strategy. Focused on the theories and rationales underpinning the individual-level and organization-level learning behaviors, Cohen & Levinthal (1990) conceptualized the "organizational absorptive capacity" as "the ability of a firm to recognize the value of new external information, assimilate it, and apply it to commercial ends." The firm's organizational absorptive capacity is determined by its "character and distribution of expertise" and the "internal structure of communication". It is presumed that knowledge sharing and knowledge diversity across individuals determine the organizational absorptive capacities.

Without directly measuring the the firm's absorptive capacity, Cohen & Levinthal (1990) measured the moderation effects of absorptive capacities on own R&D which is the primary sources of absorptive capacities by operationalizing the absorptive capacities through the concept of "ease of learning" affected by the "characteristics of knowledge" including:

- Degree of relevance between the external knowledge and firm's needs
- Accumulativeness and pace of advance
- Extent to be codified

It is assumed that lower "ease of learning" will require the firms to invest more on own R&D to be capable of identifying the technology opportunities, assimilating the new knowledge, and exploiting the benefits from the external knowledge. Based on the abovementioned operationalization approach, Cohen & Levinthal (1990) further assumed that:

- the cost of knowledge absorption depend on characteristics of the knowledge (ease of learning). Difficult learning requires more previous own R&D to be effective.
- Larger quantity of knowledge to be assimilated and exploited incentives the R&D investment

This paper tries to re-conceptualise the above measurement model regarding the absorptive capacity's impact on firms' own R&D investment.

Inadequcy of Cohen & Levinthal's measurement model

1.Not adequately operationalizing the ability to exploit: Even the firms with high knowledge diversity can recognize the technology opportunities, and successfully develope the product prototype, their ideas and innovation is still obstructed by the firm's business model and priorities (Uttal , 1983; Chesbrough, 2005). Their static view cannot characterize the dynamic mechanisms of absorptive capacities.

2.Oversimplified views on organizational communication structure: The technological capabilities to perceive, pursue, and develop can be limited by its culture and organizational structure even if the technology expertise per se are within the reach of the firms' human and financial resources (Christensen & Rosenbloom, 1995). The firms' communication structures is actually hindered by the information filter aligned with R&D organizational structure (Henderson & Clark, 1990).

3.Inadequate quantification of absorptive capacity: The magnitude of moderation effect by absorptive capacities cannot be adequately measured by their measurement model. Besides, firms' absorptive capacities to exploit the commercial potential of technologies might has less to do with technical knowledge and R&D investment. Besides, some firms try to collect the innovation ideas from mass fundraising which downplay the effects of absorptive capacities on firm's own R&D.

4. Underlying rationales of "Appropriability" work against the Open Innovation paradigm: The underlying presumption of the measurement model is that the exploitation of competitors' knowledge spillover relies on the interactions of the firms' absorptive capacities with competitors's spillovers. However, under Open Innovation paradigm allowing the outward flows of internal technologies to seek a path to market externally (Chesbrough, 2005), the

quadruple collaborations might be more relevant for the absorptive capacity.

Firms' performance assessment, absorptive capacity, and quadruple-helix interactions

Based on the organizational innovation theory proposed by March & Simon, it is assumed that firms' aspiration level is determined by:

- past performance
- performance of reference organization
- firms' absorptive capacities.

The failure to reach some aspiration level will instigate innovative activities. The absorptive capacity might distance the firms away from focus on performance to focus on perceived technology opportunities. I summarized the relationship among variables in the following Figure2.

The performance assessment mechanisms intensively condition the firms' organizational absorptive capacities. Many cases in researches (Christensen, 2008; Uttal, 1983; Chesbrough, 2005) illustrate how the performance assessment system such as ROI based on linear technology forecast rendered promising technology unattractive, and destructed the firms' ability to exploit the benefits of knowledge or technologies. Knott (2008) argued that the ROI concerns on R&D investment rather than absorptive capacity is the underlying drivers behind related R&D investment.

In my reframed measurement model (see Figure2), firm's R&D investment is impacted by the "performance of reference organization", "firms' past performance", and perceived technology opportunities". Furthermore, the absorptive capacity" will negatively moderated "firms' past performance" and Perceived Technology Opportunities. The more the quadruple-helix R&D collaborations, the larger organizational absorptive capacities.

Figures

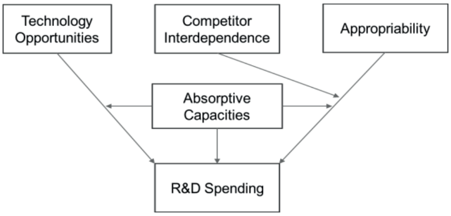


Figure 1. Measurement model of absorptive capacities' effects on firms' wwn R&D

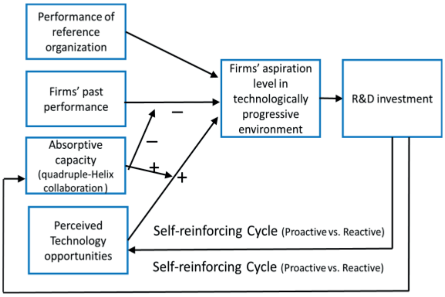

Performance of reference organization
Firms' past performance
Absorptive capacity (quadruple-Helix collaboration)
Perceived Technology opportunities
Firms' aspiration level in technologically progressive environment
R&D investment
−
−
+
+
Self-reinforcing Cycle (Proactive vs. Reactive)
Self-reinforcing Cycle (Proactive vs. Reactive)

Figure 2 Past performance, perceived technology opportunities, absorptive capacities (Triple-helix collaboration), and R&D investment

References

Christensen, Clayton M., Kaufman, Stephen P., Shih, Willy C. (2008). *Innovation Killers: How Financial Tools Destroy Your Capacity to Do New Things*. Harvard
Business Press, Boston, United States.

Christensen, Clayton, Musso, Christopher, & Anthony, Scott (n.d.). Capturing the Returns from Research. Retrieved from:
http://www.thefgi.net/wpcontent/uploads/2010/09/ Capturing-the-Returns-from-Research.pdf

Cohen, Wesley M. & Levinthal , Daniel A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*. 35(1): pp. 128-152.

Chesbrough, H. (2005). Open innovation: a new paradigm for understanding industrial innovation. In West, Joel (Eds.) (2006). *Open Innovation: Researching a New Paradigm.* Oxford University Press

Henderson, R. M., & Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly*, 9-30. Retrieved from:
http://dimetic.dime-eu.org/dimetic_files/HendersonClarkASQ1990.pdf

Knott, A. M. (2008). R&D/returns causality: Absorptive capacity or organizational IQ. *Management Science, 54*(12), 2054-2067.

Uttal, Bro (1983). The Lab That Ran Away from Xerox. In Burgleman, Robert A., Christensen, Clayton M., & Wheelwright, Steven C.(Eds.)(2009). *Strategic Management of Technology and Innovation (5th edition).* McGraw-Hill Irwin, Boston.

Co-occurrence of Cell Lines, Basal Media and Supplementation in the Biomedical Literature

Jessica Cox[1], Darin McBeath[1], Corey Harper[1] and Ron Daniel, Jr.[1]

[1] *[j.cox, d.mcbeath, c.harper, r.daniel]@elsevier.com*
Elsevier Labs, 230 Park Ave, New York, NY (USA) 10169

Introduction

Experimentation using *in vitro* cell culture is the backbone of biomedical research. Cell lines and primary cells are cultured and maintained in basal media, a specially formulated mixture of metabolites meant to support growth and proliferation. Basal medias are manufactured with the intention of minimizing variability between culture techniques. Typically, these basal medias are further supplemented to fit the needs of the specific cell line. The artificial environments in which cells are cultured has recently come into question (Yong, 2019) (Cantor, et al., 2017), as they do not simulate the physiologic conditions cells are derived from.

Despite these shortcomings, we found only one study that assessed use of media types in biomedical research (Arora, 2013). To understand the scope of the problem we have analyzed a corpus of 20,609 full length biomedical research articles containing mentions of basal media, published in ScienceDirect since 2000. Our contributions are to provide basic counts of the media types, cell lines, and supplement types; to provide information on the co-occurrences of those items, and to provide data on how the usage of those items has changed since the year 2000. We found an expected conformity amongst cell line and basal media co-occurrence, with significant differences in the kinds of supplementations used.

Data

Our primary resource is a corpus of all English-language full-length research articles published in 143 biomedical journals since 2000, for a total of 101,337 documents. The corpus was processed with Stanford Core NLP (Manning, et al., 2014) for sentence breaks and Part of Speech (POS) tags. We composed a list of queries around 27 unique basal medias (with abbreviations and names), sourced from two large commercial vendors, Thermo-Fisher and SigmaAldrich.

61,259 sentences mentioning basal media were extracted from 20,609 documents. Their text and POS tags were used in this study. For replication, the sentences, queries, and list of journals are available on Mendeley (Cox, et al., 2019).

Methodology and Results

To understand the basic characteristics of the reporting of culture media in the literature, we count the number of articles which mention one or more of the 27 basal media. The expected long-tailed distribution was observed. Table 1 provides the count of the 5 most commonly-mentioned media types, and their proportion of the corpus. Less than 5% of the articles mentioned one of the other 22 media. (Column 3 sums to more than 100% since some articles mention multiple media). Media are only mentioned in ~20% of the biomedical articles.

Table 1. Document level count of top 5 medias and their representation in biomedical research mentioning cell culture published since 2000.

Basal Media	Count	Proportion of literature mentioning media (n = 20,609)	Proportion of corpus (n = 101,337)
DMEM	12,620	61.2%	12.4%
RPMI	6,465	31.4%	6.4%
MEM	2,825	13.7%	2.8%
F12	1,049	5.1%	1%
IMDM	1,037	5%	1%

To see if the mentions of particular culture media are increasing or decreasing, counts were grouped by year. Figure 1 shows mentions of the top 5 media are increasing. For reference, the compound annual growth for the corpus since 2000 was 5.3%, while DMEM grew by 14%, RPMI by 8.6%, MEM by 10%, F12 by 9.2% and IMDM by 4.5%.

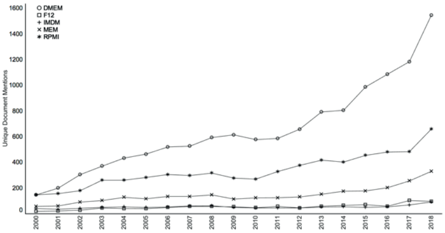


Figure 1 . Trends in top 5 basal media mentions since the year 2000.

We hypothesized that the cell line or tissue being cultured will typically be found in nouns which precede the mention of the basal media in the sentences, and the supplemental compounds will typically follow the mention of the media. Lists of the preceding and following nouns were collected and sorted by frequency. The top results were manually examined for mentions of cell lines and supplements, which were gathered into new lists.

Table 2: Mention types preceding and following media mention.

Token	Count	Position	Type
hela	1,177	preceding	cell line
hek293t	640	preceding	cell line
hek293	592	preceding	cell line
293t	510	preceding	cell line
hepg2	414	preceding	cell line
serum	20,404	following	supp
FBS	17,830	following	supp
Penicillin	9,054	following	supp
Streptomycin	9,035	following	supp
FCS	7,013	following	supp

Table 2 shows this to be an effective way of finding the cell lines and supplements mentioned. Note that there are many more mentions of supplements than cell line. This is expected since a media is typically supplemented with multiple compounds.

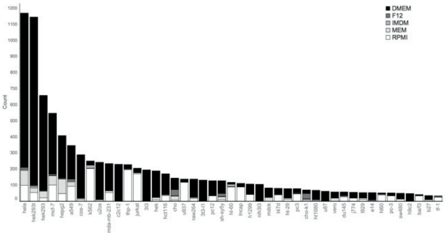


Figure 2: Co-occurrence of cell line and top 5 occurring basal medias

A more interesting research question is how the usage of media varies between the cell lines. We calculated co-occurrence of media types and cell lines in the sentences. Figure 2 shows that most cell lines are consistently cultured with one of the 5 top media types. There is little heterogeneity of the media within cell type. This suggests community norms govern how cell types are cultured, which is not surprising. Similarly, we are curious about how the use of supplements varies between the media. Figure 3 shows significant heterogeneity. This is unsurprising as media is typically supplemented with multiple compounds, as mentioned earlier.

Conclusions

This corpus analysis provides a first high-level look at how biomedical researchers are reporting their usage of culture media, cell lines, and supplements. Of our 27 unique media, we found the top 5 were mentioned in 97% of our corpus.

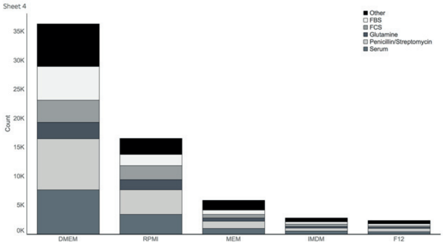


Figure 3: Co-occurrence of supplements and top 5 occurring basal medias

We found media was strongly correlated with the cell line; reflective of community standards and efforts to maintain reproducibility within a field. We also found supplements were weakly correlated with the media; reflecting considerable customization depending on experimental conditions and cell line being used. Taken together, these results strongly indicate that researchers are standardizing on certain media for reproducibility, but the media have limited physiological relevance. Going forward, we want to use a dictionary of cell lines from commercial sources, as well as look for novel lines mentioned in association with known media and supplements. We want to create a taxonomy that groups the cell lines and media by lineage, and the supplements by type. The main question in future work is to test the prediction that experiments will start to use more realistic and complex media, and to identify such articles where they exist.

References

Arora, M. (2013). Cell Culture Media: A Review. *Materials and Methods*.

Cantor, J. R., et al. (2017). Physiologic Medium Rewires Cellular Metabolism and Reveals Uric Acid as an Endogenous Inhibitor of EMP Synthase. *Cell*, 258-272.

Cox, J., & Groth, P. (2017). Indicators for the use of robotic labs in basic biomedical research: a literature analysis. *PeerJ*.

Cox, J., et al. (2019, February 8). Datasets for analysis of co-occurrence of cell clines, basal media, and supplementation in the biomedical literature. *Mendeley*. 10.17632/nvgy8pmkhk.1.

Manning, C. D., et al. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (pp. 55-60).

Yong, E. (2019, January 2). *Scientists Have Been Studying Cancers in a Very Strange Way for Decades*. Retrieved from The Atlantic: https://www.theatlantic.com/science/archive/2019/01/cancer-culture-media-plasmax/579283/

Lexical diversity as an indicator of journal scope

Philippe Mongeon[1], Maxime Sainte-Marie[2], and Marc-André Simard[3]

[1] *philippe.mongeon @ps.au.dk*
Aarhus University, Danish Centre for Studies in Research and Research Policy, Aarhus (Denmark)

[2] *maxime.sainte-marie@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, Montreal (Canada)

[2] *marc-andre.simard.1@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, Montreal (Canada)

Abstract

This work-in-progress proposes that the semantic similarity of articles published in a journal can be used to quantify the scope of the journal. Preliminary results of an analysis of a set of Library and Information Science (LIS) journals show that journals with a known broader scope have flatter, left-skewed similarity distributions, with lower averages, that can be used to identify the papers that best fit the journals in which they are published. Finally, we find different patterns regarding the evolution over time of the journals' scope.

Format

Journals play a role in the legitimization and development of scientific communities (Gingras, 2014; Mullins 1972) and help delineate scholarly communities (Milard 2008) with specific epistemic cultures that shape scholarly discourse (Wakeling et al. 2019). The scope of a journal crystalizes when researchers consider journal fit as a criterion for submitting their work (Tenopir et al., 2019; Solomon & Björk, 2012; Cronin & Younce, 2010; Jamali et al., 2014).

Editors generally also evaluate the fit of submitted manuscripts with the journal before sending it to reviewers who will also assess the relevance of the paper for the journal's readership. These mechanisms consolidate the cohesiveness of the journal for the benefit of its readership and, from a semantic perspective, we can suppose that they increase the average content similarity of papers published in a given journal. Within a discipline, some journals will cover a broad range of topics while others will focus more on a specific area of research. The premise of this work-in-progress is that the semantic similarity of articles published in a journal can be used as a proxy for its scope.

Several factors could influence the evolution over time of the journals' scope. The cognitive extent of science is expanding (Milojevic, 2015), and the lexical concentration of titles of scientific articles is increasing (Bérubé et al., 2018). The transition of journals from print to digital format and the pressure to publish could encourage journals to accept more publications and thus possibly expand its scope. Finally, new models of scholarly publishing such as open access may also be conducive to less lexical cohesiveness, as journal fit may have less weight in editorial decisions. In sum, many factors in the broad scientific environment could have observable effects on the lexical diversity of journals.

Research objectives

Using nine LIS journals as a case study, this work-in-progress proposes a measure of the lexical diversity of a journal as an indicator of journal scope, and More specifically, this work provides preliminary answers to the following research questions:

RQ1. How concentrated or diverse is the content of distinct LIS journals?
RQ2. How is the diversity of scientific discourses in specific journals evolving over time?

Below, we present the data collection steps and the methods used to measure the discursive diversity of journals and the fit of individual papers within these discourse distributions. We then present the answers to the two research questions and lay out an agenda for further research.

Using the Web of Science, we collected all articles, notes and reviews published between 1991 and 2017 in journals in the Library and Information Science category of the NSF classification. We only kept journals that publish in English and that were active over the whole 1991-2017 period for a resulting dataset of 9 journals and 12,549 publications. The title and abstract of were merged and segmented in vectors of 3-grams with TF-IDF-weighted dimensional values. This approach allows for semantically-related words to have non-zero similarity scores and offers comparable results to traditional word-based approaches. We then calculated the average cosine similarity between the text vectors of all articles published in the same journal in the same year.

Results

Table 1 shows the list of journals, the number of publications and the average similarity score between each pair or articles. A visual inspection of the distributions of similarity scores allowed us to confirm that they were approximately normal. In a broader-scope journal such as JASIST, articles will tend to be more diverse in content, as indicated by a lower average similarity score; inversely, a narrower-scope journal like Library Trends will have a higher average lexical similarity score.

Table 1. Average lexical similarity scores of nine LIS journals.

Journal	*N*	*Avg. sim.*
Info. Processing & Management	1,586	0.28
Info.Systems Journal	481	0.26
Info. Technology and Libraries	368	0.26
J of Documentation	877	0.32
J of Info. Science	893	0.25
J of the Society for Info. Sci. & Tech.	3,193	0.24
Lib. & Information Science Research	586	0.25
Lib. Trends	990	0.30
Scientometrics	3,575	0.27

Figure 1 shows the evolution of the average lexical similarity score of the nine journals over the 1991-2017 period. No general trend is followed by all journals. The average similarity score decreased for the Information Systems Journal, increased for the Journal of Documentation and Library Trends, and remained relatively stable for the other journals. This heterogeneity of trends between journals suggests the absence of a global trend in the lexical diversity of journals. Instead, the changes could result from editorial decisions, or changes in the research interests of the community of researchers that publish in a given journal.

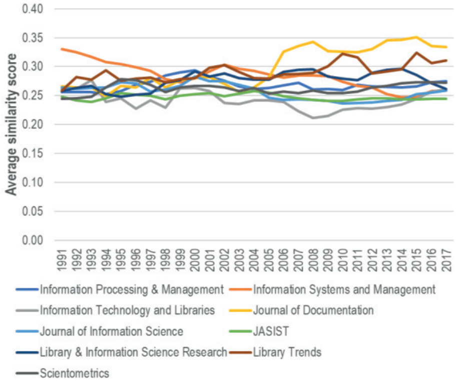


Figure 1. Average lexical similarity (1991-2017)

Discussion and conclusion

The methods and analyses presented in this work in progress lay the foundations for further in-depth analyses of the scope of journal and, more broadly, of the role that journals play in structuring knowledge dissemination. Furthermore, by enabling the identification of core, average, and peripheral papers in terms of similarity with the other papers in the journals, this method might be useful to investigate the publication practices of researchers and the relationship between journal fit and impact.

References

Bérubé, N., Sainte-Marie, M., Mongeon, P, & Larivière, V. (2017). Words by the tail: Assessing lexical diversity in scholarly titles using frequency-rank distribution tail fits. *PLoS ONE 13*(7): e0197775.

Cronin, B., & Younce, L. M. (2010). Publishing in open access education journals: The authors' perspectives. *Behavioral & Social Sciences Librarian, 29*(2), 118-132.

Gingras, Y. (2017). Sociologie des sciences. Paris cedex 14, France: Presses Universitaires de France.

Hagstrom, W. (1965). *The Scientific Community*. New York, basic books.

Jamali, H. R., Nicholas, D., Watkinson, A., Herman, E., Tenopir, C., Levine, K., ... & Nichols, F. (2014). How scholars implement trust in their reading, citing and publishing activities: Geographical differences. *Library & Information Science Research, 36*(3-4), 192-202.

Milard, B. (2008). La soumission d'un manuscrit à une revue: quelle place dans l'activité scientifique des chercheurs? Schedae (Presses Universitaires de Caen), 1, 1-12.

Milojevic, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics*, *9*(4): 962-973.

Mullins, N. C. (1972). The development of a scientific specialty: The phage group and the origins of molecular biology. *Minerva, 10*(1), 51-82.

Solomon, D. J., & Björk, B. C. (2012). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the American Society for Information Science and Technology, 63*(1), 98-107.

Tenopir, C., Dalton, E., Fish, A., Christian, L., Jones, M., & Smith, M. (2016). What motivates authors of scholarly articles? The importance of journal attributes and potential audience on publication choice. *Publications, 4*(3), 22.

Wakeling, S., Spezi, V., Fry, J., Creaser, C., Pinfield, S., & Willett, P. (2019). Academic communities: The role of journals and open-access mega-journals in scholarly communication. *Journal of Documentation, 75*(1), 120-139.

Behaviors and relationships among global universities on Twitter

Lili Miao[1], Cassidy R. Sugimoto[1] and Rodrigo Costas[2]

[1] *lilymiao08@gmail.com; sugimoto@indiana.edu*
Indiana University Bloomington, 901 E. 10th St, Bloomington, 47403 (USA)

[2] *rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (Netherlands)
Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Stellenbosch (South Africa)

Introduction

Social media platforms offer new venues to communicate and interact with large audiences. Among these, scientists have been fairly strong adopters. By 2014, nearly 40% of scientists reported that they used Twitter to share and comment on research articles (Van Noorden 2014). Early research suggested that Twitter could be a good tool for communication among educators and students ~~(~~(Rinaldo *et al.* 2011)~~)~~ and universities quickly adopted Twitter as tools for institutional news and outreach (Linvill et al., 2012). However, despite the key role of universities in knowledge production and dissemination, there is scant literature of the behaviors and relationships among universities on this social media platform.

Early research suggested that Twitter could be a good tool for communication among educators and students and universities quickly adopted Twitter as tools for institutional news and outreach (Linvill et al., 2012). Subsequent research focused on single country (e.g., Australia (Palmer, 2013) and Turkey (Yolcu, 2013)). There remains a need for a contemporary, large-scale analysis of university behaviour on Twitter. In particular, we focus on the ways in which universities use Twitter to build their online reputations and to spread their scientific knowledge. Following Shields (2016), we investigate the relationship between university rankings and network dominance. Our research draws on Twitter data and bibliometric ranking data from 682 global universities.

Methods and Data

The data used in this analysis are based on the CWTS Leiden Ranking (http://www.leidenranking.com/). We collected universities' homepage URLs for any institutions on the Leiden Ranking list. By visiting their official homepages, we automatically extracted the Twitter handles that appeared on the university homepages. This method yielded 686 university Twitter handles among the 938 universities in the Leiden Ranking list. After removing 4 universities with multiple Twitter accounts, the data set contains 682 identified university Twitter accounts. We used the Twitter API to extra additional account metadata, including the number of followers, the time when the university joined Twitter, and the list of followees.

The position of the universities in the Leiden Ranking by total number of citations (TCS) is used as an indicator to represent the overall prestige of a university. The numerical number assigned is in rank order, which means that lower scores are associated with higher ranked universities (e.g., Harvard is ranked in the number 1 position and therefore has a 1 associated with the rank).

Results

The Leiden Ranking covers 938 universities from 54 different countries. Among these 938 universities, the 682 universities for which Twitter accounts that were identified are distributed in 51 countries as showed in Figure 1(a). A large proportion of Twitter accounts come from North American and European universities. The universities without Twitter information mainly come from China and Iran where Twitter is banned, as showed in Figure 1(b).

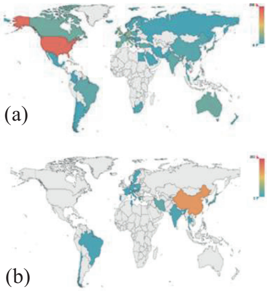


Figure 1. Leiden Ranking universities' distribution around the world (a): Countries are color-coded by the number of universities having Twitter accounts. (b): Countries are color-coded by the number of universities which do not have Twitter accounts.

As showed in Figure 2(a), universities in North America and Oceania joined Twitter earlier than universities in other areas. Universities in Asia joined Twitter more recently. Rank is also important: early joining universities are among those most highly ranked in terms of the Leiden Ranking (Figure 2(b)). Universities that joined Twitter after

2014 are, on average, ranked lower than those universities that joined Twitter earlier.

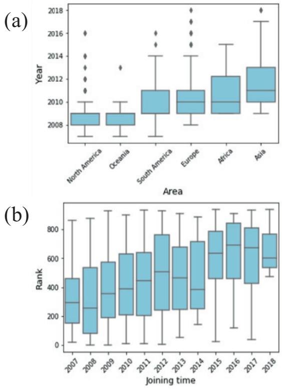


Figure 2 (a): Twitter joining time distribution within each world region. (b) University rank distribution in different Twitter join year.

A network analysis is implemented to see how universities are connected with each other, in terms of follower/followee relationships. The network visualization in Figure 3 shows how universities are grouped by geographical and linguistic factors. After a cluster extraction analysis, the result shows universities from North America are closely connected and constitute the largest group. The second largest group is formed by universities coming from Europe, although with several sub-networks. For instance: British universities formed one of the sub-clusters within the European cluster. South American universities and Spanish universities are also closely connected with each other.

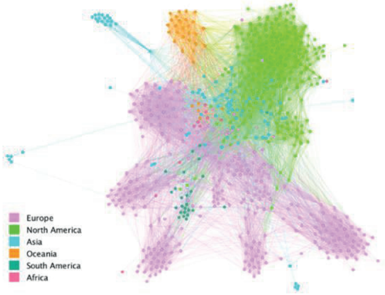


Figure 3: Network representation of Twitter following relationship among universities. Nodes are colored by their regions

In addition to the geographical coupling depicted above, we also studied whether the Leiden Ranking position also plays a role in building follower/followee connections. Universities are split into nine ranking categories based on their Leiden Ranking position. Universities ranked from 1 to 90 formed the top 10% ranking group, ranked from 90 to 180 formed the top 20% group and so on. As showed in Figure 4, within each group, there is a clear trend of most universities preferring to follow universities in the top 10% categories. This suggest that there is a clear hierarchical structure in the following relationship among universities. Thus, higher ranked universities tend to receive more followers from other lowly ranked universities and the relationship is not reciprocated.

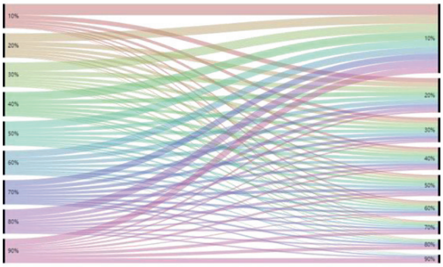


Figure 5: Following directions within each ranking group.

Conclusion

In this study, we have performed a first large-scale systematic analysis of the presence and activities of Leiden Ranking universities on Twitter. Our results show that North American and European universities are the most represented on Twitter, while Asian universities are underrepresented. Highly ranked universities and Western universities were early adopters of Twitter. The follower/followee relationships among universities are clearly dominated by both geographical and linguistic factors, and by the academic reputation of universities, as measured by their Leiden Ranking positions.

References

Rinaldo, S. B., Tapp, S. & Laverie, D. A. (2011). Learning by tweeting: Using Twitter as a pedagogical tool. Journal of Marketing Education, 33(2), 193-203.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. Nature news, 512(7513), 126.

Linvill, D. L., McGee, S. E., & Hicks, L. K. (2012). Colleges' and universities' use of Twitter: A content analysis. Public Relations Review, 38(4), 636-638.

Palmer, S. (2013). Characterisation of the use of Twitter by Australian Universities. Journal of Higher Education Policy and Management, 35(4), 333-344.

Yolcu, O. (2013). Twitter Usage of Universities in Turkey. Turkish Online Journal of Educational Technology-TOJET, 12(2), 360-371.

Shields, R. (2016). Following the leader? Network models of "world-class" universities on Twitter. Higher Education, 71(2), 253-268.

Can Crossref Citations Replace Web of Science for Research Evaluation? The Share of Open Citations

Tomáš Chudlarský[1] and Jan Dvořák[2]

[1] *tomas.chudlarsky@cvut.cz*
Czech Technical University in Prague, Computing and Information Centre,
Jugoslávských partyzánů 3, CZ-16000 Praha 6 (Czech Republic)

[2] *jan.dvorak@ff.cuni.cz*
Charles University, Institute of Information Studies and Librarianship,
U Kříže 8, CZ-15800 Praha 5 (Czech Republic)

Introduction

With the adoption of the Digital Object Identifiers (DOIs) by publishers of scholarly works advancing and with the recent availability of the partial citation network between scholarly works (Crossref) one can start to contemplate on "open scientometrics", where citation data need not be sourced from commercial providers. The prerequisite for that is that Crossref covers and openly provides sufficient part of citations of the today's de-faccto standard citation database at least for some fields of science. We studied this question in the context of Czech Technical University (CTU).

The proportion of open citations in Crossref is increasing. More than half of the citations in Crossref were classified as open (Shotton, 2017). Eck, Waltman, Larivière, Sugimoto (2018) show that while 77.1% of citations in WoS are present in Crossref, only 39.7% are classified as open.

We investigate the level of coverage of the established Web of Science citation database by the openly available citation links from the COCI project (OpenCitations, 2018) on the sample where the cited publications are those we track in our institution's Current Research Information System (CRIS). We provide a breakdown to individual faculties and fields.

Data Sources and Method

The November 2018 release of the Crossref Open Citations corpus (OpenCitations, 2018) was used. The "cited" side of the linking relationships is of very diverse quality. Some multiline values need to be straightened up. Some values seem to contain several DOIs concatenated, separated by spaces. To rectify these most severe errors we developed a script; its application made the data load possible and even slightly raised the number of citations to 449,843,367 (by 2,864 from the original 449,840,503). However, removing duplicate DOI pairs from the dataset leaves only 445,827,638 unique citation links (by 4,015,729 less). Some of the cited "DOIs" are still unsatisfactory: they contain internal spaces or illegal characters, end in an extra full stop, have superfluous parts in their contents or are incomplete. There clearly is room for further investigation and improvements which we are undertaking in a different thread of activity and plan to report on separately. Data quality problems on the side of Crossref citations clearly have a lowering effect on the recall of our study.

The Czech Technical University has a long tradition of running an in-house built institutional Current Research Information System. V3S, the research outputs tracking component, integrates our records and those harvested from the Web of Science web service interface, including the citations of our authors' works. Dvořák, Chudlarský, Špaček (2019) give a description of the CRIS and its many integrations. For our study we have used the WoS citation data from the end of January 2019.

We limit ourselves to publications from the period 2013–2017 which have both (1) a WoS accession number with a valid record in WoS, and (2) a DOI that is registered in Crossref. We exclude those tuples that have differing DOI values in the CRIS itself and in the WoS record. This gives the sample of 12,858 publications for which we look up the citations in both WoS and Crossref: the citing and the cited publication are both present in both WoS and Crossref.

Findings

We found than 53.7% of WoS are present in the COCI dump of the open citation network.

This is significantly more than the approximate 40% coverage measured by Eck, Waltman, Larivière, Sugimoto (2018) for 4 out of 5 broad main fields (in the CWTS Leiden Ranking classification). Note that the remaining main field of Social Sciences and Humanities is marginal in our sample, given the research profile of a technical university.

We found important differences in the coverage among faculties (ranging from 63% to 28%) – see Table 1.

Also, the coverage significantly differs among disciplines (ranging from 78% to 25%) – see Table 2. Only the disciplines (in the Czech national field of science classification) with more than one hundred publications are listed.

Conclusion

The open citations network in Crossref cannot replace the Web of Science citations at this time. The observed levels of coverage of citations are not yet sufficient for Crossref to be used as the source for citation analyses in research evaluation at the university and/or faculty levels. However, the coverage has an increasing tendency, and so has the percentage of open citations in Crossref; we therefore believe that at least for some disciplines this gap will be overcome in the upcoming decade.

References

Sandro La Bruzzo, Paolo Manghi & Andrea Mannocci (2019). OpenAIRE's DOIBoost - Boosting Crossref for Research. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science, vol 988. Springer, Cham, DOI 10.1007/978-3-030-11226-4_11

Jan Dvořák, Tomáš Chudlarský, Josef Špaček: Practical CRIS Interoperability In: 14th International Conference on Current Research Information Systems: FAIRness of Research Information. Amsterdam: Elsevier B.V., 2019. p. 256-264. Procedia Computer Science. vol. 146. ISSN 1877-0509. DOI 10.1016/j.procs.2019.01.077

Nees Jan van Eck, Ludo Waltman, Vincent Larivière, Cassidy Sugimoto (2018). Crossref as a new source of citation data: A comparison with Web of Science and Scopus. [blog] Retrieved 2019-02-06 from https://www.cwts.nl/blog?article=n-r2s234

OpenCitations (2018): Open Citation Indexes : COCI, the OpenCitations Index of Crossref open DOI-to-DOI references. [dataset] November 2018 Dump, the "Citation data (CSV)" file. DOI 10.6084/m9.figshare.6741422.v3

David Shotton (2017): Milestone for I4OC – open references at Crossref exceed 50%. Retrieved 2019-02-08 from https://opencitations.wordpress.com/2017/11/24/milestone-for-i4oc-open-references-at-crossref-exceed-50/

Table 1. Coverage of WoS citations in COCI by unit

Faculty/Institute	*Cover-age*
Institute of Exp. and Appl. Physics	62.5%
Faculty of Nuclear Sci. and Physical Eng.	59.5%
Faculty of Transportation Sciences	58.9%
Faculty of Mechanical Engineering	57.4%
Faculty of Electrical Engineering	46.4%
Faculty of Biomedical Engineering	46.3%
Czech Inst. of Informatics, Robotics and Cybernetics	41.7%
Faculty of Civil Engineering	35.6%
Univ. Ctr. of Energy Efficient Buildings	31.0%
Klokner Institute *(Building materials)*	30.6%
Faculty of Architecture	28.6%
Faculty of Information Technology	28.4%
Czech Technical University (whole)	**53.7%**

Table 2. Coverage of WoS citations in COCI by discipline

Field	*Cover-age*
Astronomy, Celestial Mechanics, Astrophysics	78.3%
Plasma and Gas Discharge Physics	69.9%
Theoretical Physics	69.1%
Elementary Particles and High Energy Physics	62.3%
Nuclear, Atomic and Molecular Physics, Colliders	60.0%
Nuclear & Quantum Chemistry	51.9%
Sensors, Measurement, Regulation	50.2%
Computer Applications, Robotics	49.1%
Solid Matter Physics & Magnetism	45.6%
Electronics & Optoelectronics, Electrical Engineering	44.9%
Computer Hardware & Software	44.7%
General Mathematics	40.2%
Other Materials	36.3%
Optics, Masers, Lasers	35.2%
Fluid Dynamics	35.1%
Control Systems Theory	34.6%
Informatics, Computer Science	34.1%
Non-nuclear Energetics, Energy Consumption & Use	32.6%
Composite Materials	32.0%
Civil Engineering	29.6%
Metallurgy	28.6%
Building Engineering	28.6%
Nuclear Energetics	25.0%

How Research Milestone Shape the Technology of Today - A Case Study of Highly Cited Researcher using Topic Model

Xiaoli Chen[1] Tao Han[1]

[1] *chenxl@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing, 100190, Peoples R China

Introduction

Each year Clarivate Analytic publishes highly cited researcher list based on citation metrics. These research elites or build milestones for their disciplinary, or has persistence impact on successors, or transformed the industry and shape the technology of today. We always wondering what ideas has these researchers bring to the world.

Recently researchers use topic modelling to find out above questions, which is far from perfection. This paper wants to conduct some preliminary experiment to find out what has these research milestones contribute to shape today's powerful technology.

In the model, we try to give our solution to find what is the influence highly cited researchers contribute to the scientific community by analysing their papers and citations by topic model. We call our topic model scientific influence model (SIM), to reveal what is the inheritance and what is the innovation.

Among scientific disciplines, we choose a highly cited researcher published by Clarivate Analytics in 2018 from Artificial Intelligence (AI) area. Geoffrey E. Hinton who is one of the 2018 highly cited researcher is one of the contributors who established AI milestones. In this paper, we use unsupervised topic model through researcher's perspective to find out what influence has the research elite has made.

Our Topic Model

Given a collection of D scientific articles and K topics expressed over V unique words. θ_d is the k dimensional document-topic distribution of citing document d. μ_d is the k dimensional document-topic distribution of citing document d. φ_k is the V dimensional topic-word distribution of citing document d. The whole generative process described:

- For each topic $k \in \{1,\ldots,K\}$,
    a) Drawφ_k from topic-word distribution $\varphi_k \sim \mathrm{Dir}(\beta)$.
- For each document $d \in \{1,\ldots,D\}$,
    a) Draw θ_d from document-topic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$ of citing document.
    b) Draw μ_d from document-topic distribution $\mu_d \sim \mathrm{Dir}(\varepsilon)$ of cited document.
- For each document $d \in \{1,\ldots,D\}$, citing document inherit topics from cited document is modelled with Beta distribution$\delta_d \sim \mathrm{Beta}(\lambda_c,\lambda_m)$
    a) For i*th* word in document d: $w_{d,i}$,
        i. Random variable s obeys Bernoulli distribution $s_{d,i} \sim \mathrm{Bernoulli}\ (\delta_d)$
        ii. If $s = 0$:
            1. Draw $z_{c,d,i}$ from multinomial distribution $z_{c,d,i} \sim \mathrm{Mult}(\mu_d)$,
            2. Draw $w_{c,d,i}$ from multinomial distribution $w_{c,d,i} \sim \mathrm{Mult}(\varphi_k)$.
        iii. If $s = 1$:
            1. Draw $z_{m,d,i}$ from multinomial distribution $z_{m,d,i} \sim \mathrm{Mult}(\theta_d)$,
            2. Draw $w_{m,d,i}$ from multinomial distribution $w_{m,d,i} \sim \mathrm{Mult}(\varphi_k)$.

Several works are similar to our model. (Dietz, Bickel, & Scheffer, 2007) propose copycat model and citation influence model. (He et al., 2009, n.d.) develop an inheritance topic model with the same assumption with (Dietz et al., 2007). (Kim, Kim, & Oh, 2017) present Latent Topical-Authority Indexing (LTAI) for jointly modelling the topics, citations, and topical authority to find topic authority. (Lu et al., 2014) develop a collective topic model on three types of objects: papers, authors and published venues, model any of these objects as bags of citations. (R. Nallapati & Cohen, n.d.; R. M. Nallapati, Ahmed, Xing, & Cohen, 2008; R. Nallapati, Mcfarland, & Manning, n.d.) proposed Link-LDA and Link-PLSA-LDA, which is similar to ccLDA (Paul, 2010), ccMix(Zhai, Velivelli, & Yu, 2004), and diffLDA(Thomas, Adams, Hassan, & Blostein, 2010), all these method are used for two text collection topic comparison. Our work extends cross collection comparison LDA model to find what topics are cited. What we do different is analyse from researcher's perspective. We use topic model to model their work and citations and find out what is the influence (Figure 1).

Result

We use Semantic Scholar Open Research Corpus(https://labs.semanticscholar.org/corpus/) from AI2's (Allen Institute for Artificial Intelligence) Semantic Scholar to collect Hinton's published paper and all the citations.

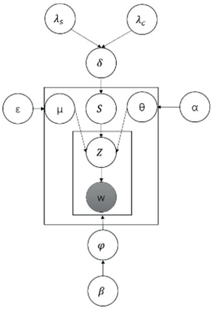


Figure 1. Plate diagram of scientific influence model

In order to find optimal number of topics for both cited collection and citing collection. We use two separate Latent Dirichlet Allocation instances on cited collection and citing collection. We use two evaluation metrics to determine to optimal topic number: perplexity and coherence.

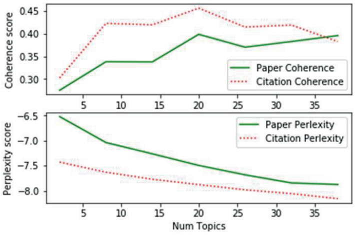


Figure 2. Perplexity and coherence of citing collection and cited collection

As coherence and perplexity metric in Figure 2 shows, the optimal topic number for paper collection and citation collection are both 20. Therefore, we will use 20 topics for both cited collection and citing collection in our Scientific Influence Model.

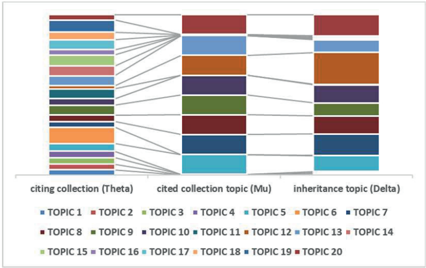


Figure 3. Perplexity and coherence of citing collection and cited collection

From topic influence from cited collection to citing collection in in Figure 3, though we choose 20 topic for both cited collection and citing collection. We still can see some of the topic only show up in the citing collection, which is the innovation topic of citing collection. The common topics are the citing collection's inheritance from cited collection. The goal of our paper is to find the influence of cited collection, which is also the inheritance of citing collection from cited collection.

Our Scientific Influence Model finds 8 inheritance topics and 12 innovation topics. From inheritance topics, we can see Hinton's influence are "boltzmann machine" related topic, "nonlinear system" related topic, "gradient descent" related topic, "deep learning" related topic, "perceptron" related topic, "connectionism" related topic, "neural networks" related topic and "machine learning" related topic. The innovation topics of Hinton's citations are "cluster analysis" related topic, "program optimization" related topic, "neural network simulation" related topic, "physical object tracing" related topic, "interpolation" related topic, "self-organization" related topic, "recognition application" related topic, "recurrent neural network" related topic, "convolutional neural network" related topic, "dimensionality reduction" related topic, "overfitting" related topic, "generative adversarial networks" related topic.

Conclusion

In this paper, we use topic model to study the influence of highly cited researcher. There are several improvements: Firstly, cited collection and citing collection do not always share the same number of topics. Secondly, slow convergence of our model due to vocabulary size of citing collection.

References

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. Proceedings of the 24th international conference on Machine learning - ICML '07 (pp. 233–240).

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, C. L. (n.d.). Detecting Topic Evolution in Scientific Literature: How Can Citations Help?

Kim, J., Kim, D., & Oh, A. (2017). Joint Modeling of Topics, Citations, and Topical Authority in Academic Corpora.

Lu, Z., Mamoulis, N., & Cheung, D. W. (2014). A collective topic model for milestone paper discovery. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14 (pp. 1019–1022). New York, New York, USA:

Nallapati, R., & Cohen, W. (n.d.). Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs.

Paul, M. (2010). Cross-Collection Topic Models: Automatically Comparing and Contrasting Text. Zhai, C., Velivelli, A., & Yu, B. (2004). A Cross-Collection Mixture Model for Comparative Text Mining.

Thomas, S. W., Adams, B., Hassan, A., & Blostein, D. (2010). DiffLDA: topic evolution in software projects. Computing.

Priorities for Social Sciences and Humanities Projects Based on Text Analysis

Ülle Must[1]

[1] ulle.must@archimedes.ee
Archimedes Foundation, Väike-Turu Str 8, 51004 Tartu (Estonia)

Introduction

In current paper we investigate the Social Sciences and Humanities (=SSH) thematic pattern through three funding instruments during the period 2007–2018: a) Societal challenges oriented top down European Union Framework Programmes (=FP) projects (FP7 Cooperation. Theme 8: Socioeconomic Sciences and Humanities and HORIZON 2020 Societal Challenge 6. Europe in a changing world – inclusive, innovative and reflective societies); b) Slovenian SSH projects funded by the Slovenian Research Agency, and c) Estonian SSH projects funded by the Estonian Research Council (until 2012 the Estonian Science Foundation).

Slovenia and Estonia have performed well in the FP, they also have a well-developed Research Information System, therefore we chose these countries as the best examples for the study. The aim of this paper is to monitor and analyse the evolution (or overlapping) of the SSH thematic pattern through those instruments since 2007. Also, in case of SSH we are used to talk about its local character. It was intriguing to see if and to what extent project writing practices affect this locality. Also, to what extent the style/keywords overlap in proposal writing. And whether it is possible to highlight hot topics from a given period with the help of text analysis.

Methods

We used publicly available tools to conduct the survey (https://www.onlineutility.org/text/analyzer.jsp; https://www.wordclouds.com/;http://bioinformatics.psb.ugent.be/webtools/Venn/).

Data collection

The Community Research and Development Information Service (CORDIS) is the European Commission's primary source of results from projects funded by the EU framework programmes for research and innovation (FP1 to Horizon 2020). From among collaborative projects we made a search by the thematic programmes: "FP 7 Cooperation. Theme 8: *Socio-economic Sciences and Humanities*", and "Horizon 2020 Societal Challenge 6. *Europe in a changing world – inclusive, innovative and reflective societies*". **The Estonian Research Information System (**ETIS) collects data about Estonian research institutions and researchers working in Estonia (CVs, publications, supervisions, patents, research projects and contracts). Data are available since 2006. All Estonian national research financing is processed via ETIS: researchers submit applications (and later financial or final reports) and the funding bodies process them. We performed a search by project field and by the Frascati Manual specialties "*Social Sciences*" and "*Humanities and the Arts*". From the collection we selected out research projects funded by the Estonian Research Council (until 2012 the Estonian Science Foundation).

The Slovenian Research Information System (SICRIS) collects data about research organisations, research groups, researchers, research projects, research and infrastructure programmes, and research equipment. The Projects database contains data on projects partly financed by the Slovenian Research Agency from 1998 onwards. The project classification scheme is based on CERIF. We performed searches using the science categories "*Social Sciences*" and "*Humanities*".

Data cleaning

It is said that data pre-processing, acquisition, and cleansing jointly represent up to 80% of the overall effort distribution from text analysis survey (Glenisson, et al, 2005). In this survey we used funded project titles and abstracts derived from the EU FP, Slovenia and Estonia RIS. As the texts used (especially in case of national funding) were not written in the majority of cases by English native speakers, we were aware of the risks involved. For some domains this information may not be enough to achieve high performance in text mining tasks (Gonçalves, et al, 2018). Also the bias risk caused by abstract's quality and completeness exists (Gómez-García, et al, 2017). The full text analysis has shown that within the categories, such as methodological or empirical research, substantial differences in profile and orientation can occur (Glenisson, et al, 2005-2).

Pre-processing steps include the removal of punctuation marks, numeric values, articles (a, the), prepositions (on, at, in), conjunctions (and, or, but) and auxiliary verbs, such as *"to be" (am, are, is, was, were, being)*, *"do" (did, does, doing)*, *"have" (had, has, having).* In the current survey we did not remove word suffixes, such as plurals, verb tenses and deflections during the first phase, because in English, nuances also play a role in differentiating the content. The final analysis and comparisons between different datasets were made on the basis of the 200 most frequent words.

Findings

After cleaning the data, 4854 unique words in ETIS, 4421 unique words in SICRIS, and 3950 unique words in FP were identified. Co-occurrence analyses were made on the basis of top 200 words. Across all funding instruments, about a quarter of the top words constitute a half of the word occurrences. Word frequency is an important measure in content analysis. This measure is used to identify the most important research topics or concepts in a field by focusing on the most frequently occurring words (Milojević, S., et al, 2011). As one aim of this paper was to examine to what extent FP affects national programs, then the results of this study showed, that in the majority of cases words do not overlap. There is more overlapping between words in case of SL and EE, and it is very marginal in case of SL-FP and EE-FP. As stated by Milojević, et al (Milojević, et al, 2011), all words are specific or nonspecific to some degree, depending on the context. In this case we divided words into three cognitive groups. The first group consists of so-called **project classics**. Since project preparation is subject to certain standards, there are terms in the text that are not related to the content but at the same time they are necessary to achieve the given criteria (*deliver, evaluate, engage, implement, network, develop, publish, area, present, findings, increase, process, platform,* etc.). In this case they make up the majority of the top 200 words. The most common pairs of words are formed from them (*research project, proposed project, proposed research, long term, project aims*). The second group is **content words** which form the core of the projects and follow the research trends in the given time frame. The following words overlap in all three datasets: cultur*, identity, countr*, econom*, innovat*, educat*. This shows that long-term trends are the same on both national and the European level. At the same time, the unique words in FPs are not among the national priorities (poverty, employment, inequality, citizen, ageing, security, etc.). In some cases, it may be due to using different vocabulary. The **geographical location** can easily be placed under the two previous groups. It looks like good practice of project writing favours mentioning the target area (Europe, Estonia, Slovenia). Because of this they should actually belong to the group *project classics.* On the other hand, geographic location shows the region that has the most interest for the country (Estonia – Russia, Baltic, German; Slovenia – Yugoslavia; FP – Mediterranean).

Conclusions

Our assumption that the FP affects national projects, was not confirmed. There is more overlapping between words in case of SL and EE. The same time the *project classics* form a large part of all three funding instrument. The unique words in FPs are not among the national priorities (poverty, employment, inequality, citizen, ageing, security, etc).

References

CORDIS. Retrieved October 12, 2018 from https://cordis.europa.eu/projects/en

ETIS. Retrieved October 12, 2018 https://www.etis.ee

Glenisson, P, Glanzel, W, Janssens, F., De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management*, 41, 1548–1572.

Gómez-García, F., Ruano, J., Aguilar-Luque, M., Alcalde-Mellado, P., Gay-Mimbrera, J., Hernández-Romero, J. L., Sanz-Cabanillas, J. L., Maestre-López, B., González-Padilla, M., Carmona-Fernández, P. J., Vélez García-Nieto, A., Isla-Tejera, B. (2018). Abstract analysis method facilitates filtering low-methodological quality and high-bias risk systematic reviews on psoriasis interventions. *BMC Medical Research Methodology series* – open, inclusive and trusted,1**7**:180. https://doi.org/10.1186/s12874-017-0460-z

Gonçalves C., Iglesias E.L., Borrajo L., Camacho R., Seara Vieira A., Gonçalves C.T. (2018) A Framework for Full Text Analysis. In: de Cos Juez F. et al. (eds) *Hybrid Artificial Intelligent Systems.* HAIS 2018. Lecture Notes in Computer Science, 10870. Springer, Cham.

Milojević, S., Sugimoto, C. R., Yan, E., Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology,* 62 (10), 1933-1953. SICRIS. Retrieved December, 20, 2018 https://www.sicris.si/

Why do researchers from Economics and Social Sciences cite online? Insights from an exploratory survey

Maryam Mehrazar[1], Hadas Shema[1], Steffen Lemke[1] and Isabella Peters[1,2]

[1] *[m.mehrazar| h.shema| s.lemke]@zbw.eu*
ZBW Leibniz Information Centre for Economics, Düsternbrooker Weg 120,
D 24105 Kiel (Germany)

[2] *i.peters@zbw.eu*
Kiel University, Department of Computer Science, Hermann-Rodewald-Str. 3,
D 24118 Kiel (Germany)

Introduction

A significant percentage of researchers interact with scholarly material using social media (Lemke et al., 2017). These interactions are measured using alternative metrics (often called 'altmetrics'). However, we have little knowledge about these interactions´ meaning and significance. In this poster, we present results of a survey on the possible motives of such interactions and their possible resemblance to citation motivations. We also examine which citer motivations are most relevant at which social media platform.

Data & Method

We conducted an exploratory online survey from July-September 2018 on researchers' motivations for interacting with 18 social media platforms. The survey was distributed via multiple mailing lists related to Economics and the Social Sciences. We contacted authors who published at least once after 2015 in Economics or the Social Sciences and had their email listed in the Web of Science or RePEc.

There were 16 different options provided as answers (Figure 1). The question´s formulation was built upon works by Weinstock (as cited by Cronin, (1984)), Harwood (2009) and Brooks (1986), to capture common motives of citation and adapt them to the social media era.

We analysed the stated frequencies and ranked the reasons based on their popularity among the users of all 18 social media platforms (Figure 1). For the most and the least popular reasons in the ranking, we then analysed whether the popularity varies for different social media platforms.

Survey Demographics

A total of 1,088 researchers participated in the survey (4% response rate). Most participants were from Germany (33.4%), the USA (14.2%) and the UK (5.9%) and they identified themselves as professors (58.0%), Postdocs/senior researchers (19.6%), PhD students/research assistants (14.8%) and others (7.5%). The gender distribution was male (68.2%), female (31.7%) and other (0.1%) of which 71.4% were Economists, 18.8% were Social Scientists and 9.8% were others.

Results & Discussion

Citing is a complex social process, which often has more than one motive (Brooks, 1986) and so do interactions of researchers with scholarly material in social media. Our survey shows that some scientists (about 15%) interact with scholarly material online, in particular from Wikipedia and code sharing platforms, because they rely on these products in their own work. Our findings are in line with Thompson and Hanley´s (2017) conclusion that "Science is shaped by Wikipedia".

Advertising, as Harwood (2009) named the phenomenon, was the second most popular reason for interacting with scholarly material online (about 14%). As can be seen in Figure 2, Twitter was the most popular platform for self-advertising of scholarly products, followed by scholarly blogs and reference management systems. On the other hand, only a small percentage of the researchers relied on scholarly products from Twitter in their work. Researchers interested in promoting their own work online, therefore, might do better by sharing code, answering Q & A, or editing Wikipedia so it would refer to their own work, rather than tweeting, blogging or writing posts in the social networks about it. We suggest seeing these activities as fulfilling some of the functions of self-citing in the scholarly literature. The study shows that few researchers criticize or correct their colleagues´ work online. This result echoes past research of citations (for a review of the literature, please see Bornmann & Daniel, 2008), which found a low percentage of negational references in the scholarly literature. Whether scientists simply ignore low-level work in social media, or make a conscious choice to avoid confrontation, remains to be seen. The findings contribute to our understanding of altmetric measurements and their relation to traditional impact indices.

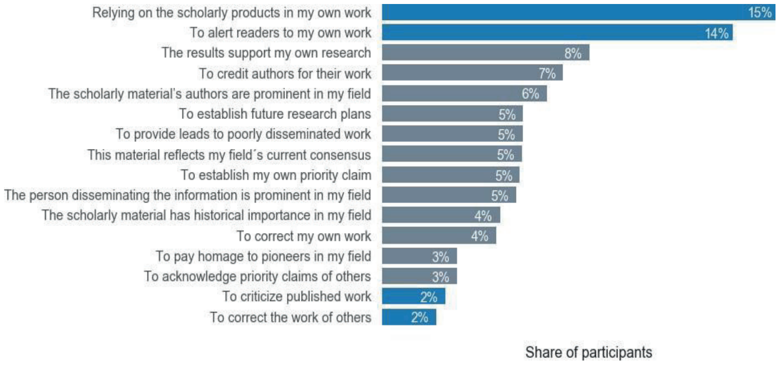


Figure 1. Why do researchers interact with scholarly material (e.g. like, share or post) in social media platforms? Top two most and least popular reasons are highlighted in blue.

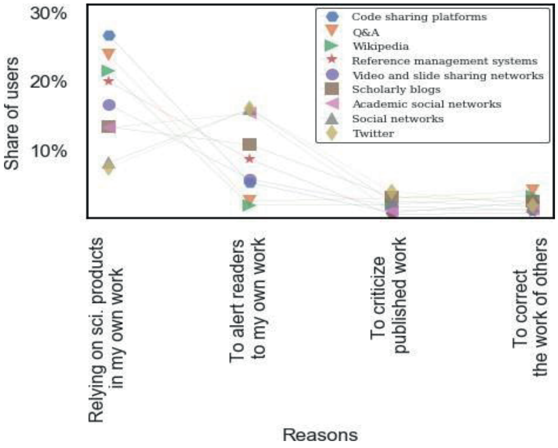


Figure 2. Differences between social media platforms and usage motivations.

As for disciplinary differences, we looked at Economists and Social Scientists and found that they both have very similar reasons for interacting with social media. We have observed two differences: More than half of Social Scientists rely on Q&A platforms for their own work, while only 28% of Economists do so. There was a slight difference in scholarly blogs usage: Social Scientists use scholarly blogs to inform readers about their own work more than Economists.

Acknowledgements

This work is part of the DFG-funded research project *metrics (project number: 314727790).

References

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45-80.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, *37*(1), 34–36.

Cronin, B. (1984). *The Citation Process: The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.

Harwood, N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, *41*(3), 497–518.

Lemke, S., Mehrazar, M., Peters, I., Beucke, D., Gottschling, M., Krausz, A., … Zagovora, O. (2017). *Exploring the Meaning and Perception of Altmetrics*. Poster at 4AM Conference. https://doi.org/10.5281/ZENODO.1037146

Thompson, N., & Hanley, D. (2017). Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial. *MIT Sloan Research, Paper No. 5238-17*. https://doi.org/10.2139/ssrn.3039505

The Comparison of Effectiveness between Direct and Indirect Support through the Meta-analysis: The Case of Korean R&D Policy for SMEs

Juil Kim[1]

[1] *juil@kistep.re.kr*
KISTEP (Korea Institute of S&T Evaluation and Planning), Center for National R&D Budget Strategy, 4F, 60, Mabang-ro, Seocho-gu, Seoul (Republic of Korea)

Introduction

As the real New Normal age arrives, the importance of SMEs (small and medium-sized enterprises) tends to be more strongly emphasized. South Korea also has reinforced the support for them, for economic growth and job creation. In particular, it searches for a variety of policy measures, including expansion of the tax incentive system and national R&D projects in order to promote R&D investment and technological innovation. Researchers' opinions on the efficiency of direct/indirect support for the complement of private R&D investment are differed, but there is little discussion on the effective combination between two policy measures. Now, it may be time to discuss policy mix of direct/indirect support as well as long-term directivity, where the ministry in charge of policies for SMEs is upgraded and the scale of government's R&D expenditure supporting them continues to be increased in South Korea. Over the past 50 years, South Korean exhibited rapid and successful economic prosperity by supporting for SMEs technological innovation. The process of exploring the optimal technological innovation supporting policy based on cases in South Korea can thus suggest an important example to various countries aiming for learning relevant policies.

The research questions of this study are as follows.

Q1: How large is the effect size of the government's direct and indirect support to promote R&D investment of enterprises?

Q2: With regard to Q1, what is the difference by firm size?

Method and Result

This study attempted a systematic literature review and a meta-analysis, by collecting 33 related studies which have been presented by Korean academic circles. This review used the descriptive statistics to analyze the effect of direct and indirect support in individual empirical studies by firm size. Such a process has an implication in that it summarizes related debates and empirically arranges them, beyond mere verification of government's policy support results in crowding-in of SMEs' own R&D investment. Findings from the systematic literature review were in disarray, yet indirect support showed more consistent complementary effects than direct support in terms of corporate R&D investment.

Descriptive statistics approach through the systematic review of literature can draw partial conclusions. The author attempted to conduct meta-analysis in order to draw quantified and empirical conclusion by overcoming these limitations. Meta-analysis is a statistical approach that integrates individual empirical analysis results to organize general knowledge in a certain field. In this study, this analysis was carried out to compare crowding-in effects of direct supports with indirect supports investigated in earlier studies.

There are 25 studies capable of producing effect size of correlation (r) among 33 studies used in the systematic literature review and these studies were set to subjects for meta-analysis. The meta-analysis found that crowding-in effects of indirect support (.192) were higher than direct support (.143) in the model containing all enterprises. The analysis on large enterprises showed that effect size of indirect support (.250) was more prevalent than direct support (.080). Direct support (.124) was more effective than indirect support (.098) to induce R&D investments by SMEs. This result shows that tax incentive is more effective for large enterprises, while subsidy is effective for SMEs.

However, Korean SMEs' R&D support currently has the excessively high proportion of direct support. Official statistics released by the South Korean government revealed that the current ratio of direct support to indirect support from government funds for SMEs turned out 72.7 : 27.3. Compared to 55.9: 44.1 relative ratio of effect size resulting from this meta-analysis, this ratio suggests that the proportion of direct support was excessive in terms of investing actual finance. Findings suggest that the proportion of direct subsidies should be diminished, while tax support should be progressively enlarged to promote the corporate technological innovation.

Table 1. Result of Meta-analysis

Firm type	Policy type	K	ES	95% CI	p
No Classification	Direct support	26	0.143	0.100~0.185	<0.001
	Indirect support	7	0.192	0.141~0.242	<0.001
	Over all	33	0.152	0.115~0.188	<0.001
Large Enterprises	Direct support	10	0.080	0.024~0.136	0.005
	Indirect support	2	0.250	0.021~0.454	0.032
	Over all	12	0.109	0.037~0.180	0.003
SMEs	Direct support	20	0.124	0.076~0.170	<0.001
	Indirect support	3	0.098	0.003~0.192	0.044
	Over all	23	0.120	0.078~0.161	<0.001

K=Number of research, ES=Effect Size, CI=Confidence interval, p=Significance level

Conclusion: Policy Suggestions

First, it is important to set up alternatives for supporting R&D tax to SMEs. The previous R&D tax system may serve as a barrier to give tax benefits to these enterprises. This defective factor should be identified and improved. In addition, incentives should be provided for enterprises to receive tax benefits on technological innovation. For instance, a solution can be considered to give extra points in supporting R&D tax when it comes to hiring new researchers.

Second, portfolio on long-term financial supports for R&D in SMEs should be established. Over the past years, the domestic technological innovation policies have focused on direct supports centered in national R&D project. On the contrary, indirect supports were slightly neglected. If financial portfolio encompassing both direct and indirect supports is regularly established, attempts to explore the optimal policy combination will be performed systematically in terms of supporting technological innovation and leads to boosting investment strategies.

Finally, official data for R&D taxation needs to be established and transparently opened among those who are involved. One of major causes of lack of systematic analysis on crowding-in effects of indirect supports was the lack of valid data in the academic field. As R&D tax support can play significant roles in promoting technological innovation of SMEs as much as R&D subsidies. Therefore, plenty of R&D taxation data needs to be established equivalent to national R&D project.

References

Becker, B. (2015), Public R&D Subsidies and Private R&D Investment: A Survey of the Empirical Evidence, *Journal of Economic Surveys*, 29(5), 917-942.

Correa, P., Andrés, L., & Borja-Vega, C. (2013), The Impact of Government Support on Firm R&D Investments: A Meta-Analysis, *Policy Research Working Paper 6532*, World Bank.

CPB Netherlands Bureau for Economic Policy Analysis (2014), *A Study on R&D Tax Incentives*, Working Paper N. 52-2014, Taxation Papers, Hague: CPB Netherlands Bureau for Economic Policy Analysis.

Cunningham, P., Gök, A., & Laredo, P. (2013), The Impact of Direct Support to R&D and Innovation in Firms, *Nesta Working Paper* 13/03.

David, P. A., Hall, B. H., & Toole, A. A. (2000), Is Public R&D a Complement or Substitute for Private R&D? A Review of the Econometric Evidence, *Research Policy*, 29(4-5), 497-529.

Dimos, C., & Pugh, G. (2016), The Effectiveness of R&D Subsidies: A Meta-regression Analysis of the Evaluation Literature, *Research Policy*, 45(4), 797-815.

Gaillard-Landinska, E., Non, M., & Straathof, B. (2015), More R&D with Tax Incentives? A Meta-analysis, *CPB Discussion Paper 309*, CPB Netherlands Bureau for Economic Policy Analysis.

Garcia-Quevedo, J. (2004), Do Public Subsidies Complement Business R&D? A Meta-Analysis of the Econometric Evidence, *KYKLOS*, 57(1), 87-102.

Negassi, S., & Sattin, J.-F. (2014), Evaluation of Public R&D Policy: A Meta-Regression Analysis, *Working Paper No. 2014-09*, Department of Economics, Alfred Lerner College of Business & Economics, University of Delaware.

Negassi, S., & Sattin, J.-F. (2016), Are Public R&D Subsidies Effective? Some Evidence from a Meta-Analysis of the Literature, *Proposition de Communication – Congrés AEI 2017*.

Petrin, T. (2018), A Literature Review on the Impact and Effectiveness of Government Support for R&D and Innovation, *Working Paper*, 5/2018 February, ISIGrowth.

Exploring Knowledge production in Europe. The KNOWMAK tool

Benedetto Lepori[1,2], Philippe Larédo[2], Thomas Scherngell[3], Diana Maynard[4], Massimiliano Guerini[5]

[1]*blepori@usi.ch.* Università della Svizzera italiana, Lugano, Switzerland.

[2]*philippe.laredo@enpc.fr.* University of Paris Est, France.

[1]*thomas.scherngell@ait.at.* Austrian Institute of Technology, Vienna, Austria.

[4]*d.maynard@sheffield.ac.uk.* University of Sheffield, Sheffield, United Kingdom.

[4]*massimiliano.guerini@polimi.it.* Polytechnic of Milan, Milan, Italy.

Introduction

The goal of the KNOWMAK project is to develop a web-based tool, which provides interactive visualisations and indicators on knowledge production in the European Research Area (ERA). Indicators are structured around three integrative elements:

- Research topics, by developing ontologies on Societal Grand Challenges (SGC) and Key Enabling Technologies (KET);
- Research actors (key actors in knowledge production);
- Geographical spaces and, more specifically, countries and (metropolitan) regions.

The tool will allow users to perform the following main tasks:

- Exploration: Allows the user to explore indicators in the tool in the form of tables and with different visualisations (maps, bar charts and line diagrams)
- Retrieval: Allows the user to download selected indicators in a suitable format for further statistical analysis or combination with other data
- Explanation: Supports the user in the navigation through the tool with online help features, user handbook, as well as some illustrative analytic examples (data stories)

Architecture

The overall architecture of the KNOWMAK tool is depicted in Figure 1. Primary data sources on knowledge production are structured and enriched in specific datasets (in strong relation to the EU funded RISIS infrastructure, risis2.eu). Harmonisation pertains to the three central integration dimensions, i.e. geographical space, topics, and actors. A core set of data is transferred to the KNOWMAK integrative database which, in turn, feeds indicator construction and the interactive online visualisation tool.

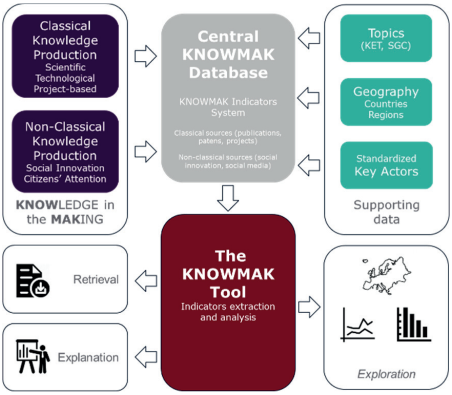


Figure 1. Architecture of the KNOWMAK tool

User involvement

A core part of the KNOWMAK project and web tool development is its co-creation approach. A participatory process from the beginning of the project has ensured that the tool is tailored to the needs of lead user groups. These groups are composed of members of all relevant stakeholder groups, i.e. policy-makers and research funders, managers of research organisations, regional actors and representatives of civil society, and the business sector.

Who we are

The KNOWMAK consortium is composed of eight partners. While all of them are involved in the joint development of the tool, their tasks are specialised, based on the competences and data they own. www.knowmak.eu

Acknowledgments

Funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992 is acknowledged.

Investigating knowledge production in Europe: Some illustrative examples

Small regions, big players in knowledge production

One main indicator of KNOWMAK gives an impression on the overall knowledge production volume and intensity, derived from three different types of knowledge production: Patents, publications and FP projects. In terms of intensity (normalised by population), it can be seen that some small to mid-sized regions (e.g. Eindhoven or Heidelberg) produce more knowledge per capita than their mega-city counterparts London, Paris and Berlin.

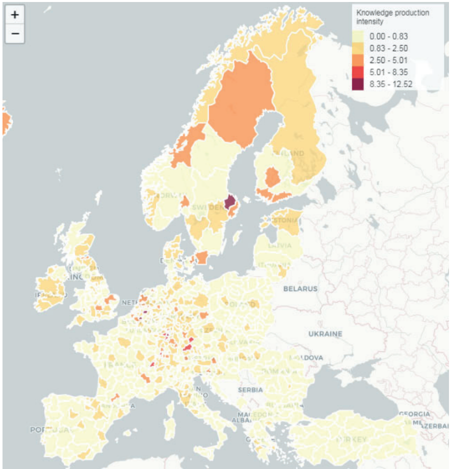


Figure 2. Knowledge production intensity by region

Clusters of knowledge production in genomics

The tool allows to analyse knowledge production at a detailed level of highly relevant topics, identified as crucial for the European socio-economic development. Figure 3 shows the example of Genomics research (one subclass of the KET Industrial Biotechnology) in publications. In 2013, The metropolitan region London turns out as leading hot spot, followed by Paris, Barcelona and East Anglia.

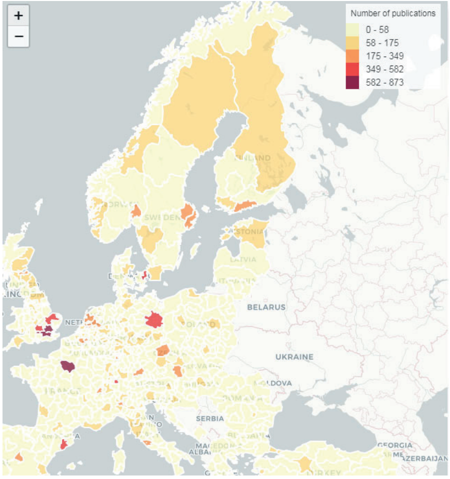


Figure 3. Publications in Genomics by region

Subtopics in nanoscience and technology

For the Nanotechnology KET, the KNOWMAK data allow disaggregating knowledge production by subtopic and type of output (Figure 4). They display systematic differences with nanoscale devices as being the most important subtopic for patents, nanoscale technology for publications and nanoscale materials for projects. This highlights differences in the science vs. technology orientation of domains within a KET, but also possible misalignments between EU funding policies and the European S&T basis.

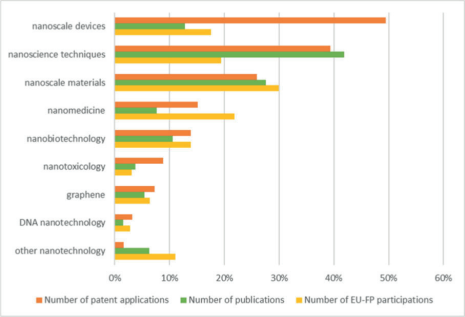


Figure 4. Knowledge production in Nanotechnology by subtopic

Investigating the Knowledge Spillover and Externality of Technology Standards

Pei-Chun Lee

pclee@nccu.edu.tw
Graduate Institute of Library, Information & Archival Studies, National Chengchi University, Taiwan

Introduction

The importance of technological standards has increased considerably over the past three decades. The increasing attention toward the standardization process is attributed largely to the growth of the information and communication technology industry. Before an industry standard is selected, there exist various attractive technologies. However, after industry participants select a standard and take steps to implement it, alternative technologies become less attractive. Similarly, the increasing emphasis on patenting by standard setting organizations (SSOs) reflects the strategy of an increasing number of firms to apply for patents for earning revenue from royalty payments for the use of their technology embedded in an industry standard. The key function of SSOs is to aggregate information from many different entities and coordinate efforts on relevant intellectual property claims before deciding on a standard (Lerner, Tabakovic, & Tirole, 2016). However, the knowledge influence of patented technologies derived from standard setting efforts and the externality of technology standards have not been analyzed in detail. Therefore, this paper proposes that analyzing the knowledge spillover and externality is a useful method of identifying the origin, direction, and magnitude of essential patents for supporting technology standards. This paper speculates that a technology standard facilitates a high degree of knowledge spillovers and externality in terms of a comprehensive technological influence (both forward and backward), wide geographical reach, and long-time span, especially when technological standardization is fulfilled. Despite the acknowledged importance of knowledge spillovers, there exist very few studies on the origin, direction, magnitude, and externality of knowledge spillovers, which influence the transmission of the spillovers effect across boundaries. Therefore, investigations are required to deepen our understanding of knowledge spillovers and the subsequent externalities. The novelty of this study is twofold. First, the knowledge spillovers of a technology standard is demonstrated and measured. Second, the significance of knowledge spillovers is identified in terms of the originality and externality indicators.

Theoretical background

Knowledge Spillovers and Externality

With the growth in the knowledge foundation over time, knowledge spillovers allow a large number of differentiated products to be introduced without a continual increase in research resources because the benefit of innovation accrues to the innovator and spills over to other organizations by raising the level of knowledge on which new innovations can be based. Thus, knowledge spillovers can serve as the engine of technological innovation to provide further access to new knowledge and increase the productivity of economic actors (Audretsch & Feldman, 2004). Furthermore, the significance, magnitude, and channels of international knowledge spillovers effects have been estimated by previous studies (Lichtenberg & De La Potterie, 1998)(Coe & Helpman, 1995)(Branstetter, 2000). For measuring the extent of knowledge externalities, models of endogenous economic growth have been tested by estimating the form of R&D spillovers across firms and/or from universities and public labs to firms (Branstetter, 2001). Patent citations allow researchers to quantify and measure knowledge spillovers and develop indicators of the significance of individual patents, which provides an alternative method of capturing the value of patents (Hall, Jaffe, & Trajtenberg, 2001).

Data and Methods

A patent that controls any part of the technology used in a standard is called a standard-essential patent (SEP). An SEP is a patent that claims an invention that must be used to comply with a technological standard. Patents supported by SSOs (SSO patents) can receive more citations than other patents from the same technological field and application year, which suggests that SSO patents have a high degree of economic and technological importance and monetary worth. Citations to SSO patents have limited distribution in the first few years after the patent is issued, which implies that SSO patents usually have a long life (Mehta, Rysman, & Simcoe, 2007). For capturing the dynamics of the knowledge spillovers that occurs during technology standardization, utility patents are downloaded from the USPTO patent database. Due to the selection and marginal effect, SSOs find either compelling technologies or technologies expected to become significant based on the consensus and open technologies built (Rysman & Simcoe, 2008). Patent information can be considered a type of technical problem addressed by engineers over time. When the flow of knowledge within the patent citation network is identified and the patents belonging to the trajectory can be scrutinized to obtain information regarding the engineering heuristics applied, the citations received in the heuristics enable the detection of the paradigmatic knowledge spillovers and externality. In this study, an attempt is made to explore whether the citations are isomorphic by analyzing three different patent datasets at the technical, organizational, and industrial standard levels.

Results and Conclusion

The significance of the knowledge spillover of technology standards is examined by analyzing SEPs, non-SEPs, SSO patents, and SEPs aligned with the 802.16 standard from 1976 to 2017. Two knowledge spillover properties are evaluated, namely the originality and externality of knowledge spillovers. As presented in Table 6, the number of inventor countries, CPC count, number of claims, originality index, country originality, industry originality, and assignee originality of SEPs are considerably higher than those of non-SEPs within the same technology fields. The sum of the forward citation distance, number of patent citations received, patent family size, generality index, country generality, industry generality, assignee generality, mean longevity, and maximum longevity of SEPs are significantly higher than those of non-SEPs. However, compared with non-SEPs, SEPs have fewer assignees, patent references, and non-patent references.

This research provides three contributions to existing research streams. First, it offers a comprehensive perspective of knowledge spillovers resulting from technology standardization. Second, this paper articulates the characteristic patterns of knowledge externality that occurs during and after technology standardization. With the growth in the knowledge foundation over time, knowledge spillovers allow a large number of differentiated products to be introduced without a continual increase in R&D resources. Knowledge that spills across organizational, industrial, and national borders may enable the application of the dominant design (standard) in the entire industry. Knowledge spillover can also cause a gradual shift toward a new generation or paradigm. Therefore, knowledge spillovers can serve as engines for creating a new technological trajectory or paradigm by providing access to new knowledge. To explore knowledge spillovers empirically, three datasets are used for ANOVA, which indicates the significance of SEPs, non-SEPs, and SSO patents. The results of this study indicate that on an average, SEPs have a low number of assignees, patent references, and non-patent references, which suggests that the patented technology standard usually originates from adjacent technological fields with few antecedents and a weak scientific basis. For instance, the knowledge spillovers of the SEPs supported by the 802.16 standard originates from fewer industries than that of the non-SEPs within the same technology fields. This finding is consistent with the characteristics of technology standards, which are often not the most advanced or cutting-edge technologies. Standards are generally adopted under pressure or due to their compatibility advantages, which benefit manufacturers, distributors, and consumers (Fritsch & Franke, 2004).

References

Audretsch, D. B., & Feldman, M. P. (2004). Knowledge spillovers and the geography of innovation. *Handbook of Regional and Urban Economics*, *4*, 2713–2739.

Branstetter, L. G. (2000). Looking for international knowledge spillovers a review of the literature with suggestions for new approaches. In *The economics and econometrics of innovation* (pp. 495–518). Springer.

Branstetter, L. G. (2001). Are knowledge spillovers international or intranational in scope?: Microeconometric evidence from the US and Japan. *Journal of International Economics*, *53*(1), 53–79.

Coe, D. T., & Helpman, E. (1995). International r&d spillovers. *European Economic Review*, *39*(5), 859–887.

Fritsch, M., & Franke, G. (2004). Innovation, regional knowledge spillovers and R&D cooperation. *Research Policy*, *33*(2), 245–255. https://doi.org/10.1016/S0048-7333(03)00123-9

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools* (Working Paper No. 8498). https://doi.org/10.3386/w8498

Lerner, J., Tabakovic, H., & Tirole, J. (2016). *Patent Disclosures and Standard-Setting*. National Bureau of Economic Research.

Lichtenberg, F. R., & De La Potterie, B. van P. (1998). International R&D spillovers: a comment. *European Economic Review*, *42*(8), 1483–1491.

Mehta, A., Rysman, M., & Simcoe, T. (2007). *Identifying the age profile of patent citations*. Boston University-Department of Economics.

Rysman, M., & Simcoe, T. (2008). Patents and the performance of voluntary standard-setting organizations. *Management Science*, *54*(11), 1920–1934.

Towards a multidimensional *valuation* model of scientists

Nicolas Robinson-Garcia[1], Rodrigo Costas[2], Thed N. van Leeuwen[2] and Tina Nane[1]

[2] *N.RobinsonGarcia@tudelft.nl; g.f.nane@tudelft.nl*
Delft Institute of Applied Mathematics (DIAM), TU Delft, Delft (Netherlands)

[2] *rcostas@cwts.leidenuniv.nl; leeuwen@cwts.leidenuniv.nl*
CWTS, Leiden University, Leiden (Netherlands)

Introduction

The use of scientometric indicators for individual research assessment has been severely criticized over the years due to their limited capacity to discriminate between different scientists and capture differences in a statistically reliable manner (Costas, van Leeuwen, & Bordons, 2010). Nevertheless, science managers and policy makers make use of these indicators for recruitment of scholars, promotion or allocation of funds. This has provoked strong reactions from the academic community, such as the San Francisco Declaration (DORA, 2014), a specific mention warning on the dangers of using bibliometrics for individual assessment (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015), or even a whole body of literature discussing the pros and cons of the H-index (Rousseau, García-Zorita, & Sanz-Casado, 2013), the most renown indicator for assessing individual research performance.

We argue that the greatest threat of the current use of bibliometric indicators for the assessment of scientists goes beyond technical or methodological decisions, and is more related to the irreflexive use of metrics at the individual level. We claim that this irreflexive use of metrics endangers the diversity of the scientific profiles researchers exhibit. This diversity is not only evident, but needed to ensure scientific progress (Milojević, Radicchi, & Walsh, 2018) and a breadth of societal and scientific outcomes (Woolley & Robinson-Garcia, 2017).

Some evaluation models for individual assessment have been proposed in the literature. But they have not been able to prevent the irreflexive use of bibliometric indicators. In our belief, there are three reasons behind this failure: 1) these models propose the introduction of a wide range of indicators, of which not all are necessarily operational; 2) they are framed in such terms that are difficult to operationalize; or 3) they deny the use of quantitative indicators without offering a viable and cost-efficient alternative.

By linking with the current literature and our own experience on conducting research evaluation, we here present a tentative *valuation* model which tries to balance between a conceptually-informed framework and a methodological viable operationalization. The model is designed so that it can be operationalized by making use of bibliometric indicators, although we acknowledge that it is sufficiently broad as to give room to non-bibliometric indicators.

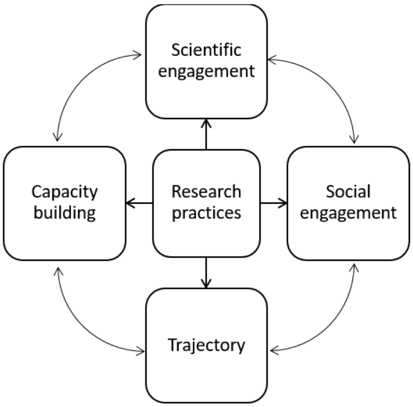


Figure 1. Evaluative dimensions of an individual

Main pillars of the valuation model

The model is structured into three distinct parts. The first and main one has to do with the actual performance of the individual in a set of five dimensions of the scientific practice. The second one addresses confounding effects derived from the individual's context, such as work environment, institutional logics or national policies shaping their performativity. The third pillar of the model relates to personal features of the individual. In principle, these characteristics hold little relation with researchers' performance, but can be of special interest for policy makers. For instance, science managers may be interested in promoting young researchers within a given programme, reduce gender inequality by encouraging the recruitment of women, or try to integrate and promote foreign born scholars.

Evaluative dimensions

We consider five dimensions as key factors to value the research performance of individuals. These are presented in Figure 1. Scientific engagement, social engagement, capacity building and trajectory look into diverse aspects of the individual's academic activities. However, the research practices dimension is represented as an overarching dimension which affects the other four. In the following, we describe each dimension.

Capacity building refers to the capacity of the individual to create new knowledge, train new scholars or develop novel applications. Some indicators operationalizing this dimension could be number of publications, normalized citation score, but also number of PhD students supervised or generation of patents.

Scientific engagement includes activities and actions reflecting a proactive engagement of the individual with the scientific community. This not only refers to scientific collaboration or division of labour, but also to reviewing papers, editing journals or organizing and participating in conferences and seminars.

Social engagement is conceived here as outreach and interaction with societal actors. For example, different modes of engagement would be considered (D'Este, Llopis, Rentocchini, & Yegros-Yegros, 2015) as well as social outreach for instance by written for non-academic audiences.

Trajectory reflects aspects related to the academic background of the individual such as geographical mobility, disciplinary changes or previous work experience.

Research practices are conceived here as an overlapping dimension which modulates each of the other four based on how open or closed these are. For instance, share of OA publications would reflect openness in capacity building, while diversity of stakeholders could apply in the case of social engagement.

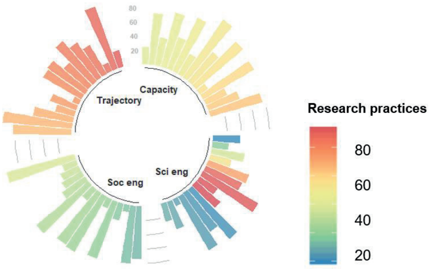


Figure 2. Profile of a fictitious researcher

Conclusions

This poster proposes a new valuation model of scientists which considers the wide variety of profiles and activities researchers perform. The model captures the heterogeneity of activities and roles researchers perform into five dimensions by which they can be profiled, also quantitatively. Figure 2 illustrates a potential visualization of such profiling. Furthermore, the model considers confounding effects mediating on individuals' performance as well as personal features which might be of relevance for science managers. The model is still under-development and still many caveats need to be solved as well as to the application of such a model on real case scenarios.

Acknowledgments

Nicolas Robinson-Garcia is a Marie Sklodowska-Curie Experienced Researcher in the LEaDing Fellows COFUND programme from the European Commission.

References

Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, *61*(8), 1564–1581.

D'Este, P., Llopis, O., Rentocchini, F., & Yegros-Yegros, A. (2015). Star vs. Interdisciplinary scientists? Exploring distinct patterns of engagement in university-industry interactions. Presentado en University-Industry Interactions Conference, Berlin.

DORA. (2014). San Francisco declaration on research assessment. Recuperado a partir de http://am.ascb.org/dora

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429-431.

Milojević, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences*, *115*(50), 12616-12623.

Rousseau, R., García-Zorita, C., & Sanz-Casado, E. (2013). The h-bubble. *Journal of Informetrics*, *7*(2), 294-300.

Woolley, R., & Robinson-Garcia, N. (2017). The 2014 REF results show only a very weak relationship between excellence in research and achieving societal impact. *Impact of Social Sciences Blog*. Recuperado a partir de https://blogs.lse.ac.uk/impactofsocialsciences/2017/07/19/what-do-the-2014-ref-results-tell-us-about-the-relationship-between-excellent-research-and-societal-impact/

Spanish scientific research in Psychology: an analysis of the differences in the production and scientific collaboration

Francisco González-Sala[1]; Julia Haba-Osca[2] and Julia Osca-Lluch[3]

[1] *francisco.gonzalez-sala@uv.es*
Dept. de Psicologia Evolutiva i de l'Educació, Facultat de Psicologia, Universitat de València, Av. Blasco Ibáñez nº 21, 46010, Valencia (Spain)

[2] *julia.haba@uv.es*
Dept. Filologia Anglesa i Alemanya, Facultat de Filologia, Traducció i Comunicació, Universitat de València, Av. Blasco Ibáñez nº 32, 46010 Valencia (Spain)

[3] *juosllu@ingenio.upv.es*
INGENIO (CSIC – UPV), Universitat Politècnica de València, Camino de Vera, s/n, 46022 Valencia (Spain)

Introduction

Scientific journals have a fundamental role in the different stages of the research activity, since they are the way in which researchers obtain recognition for their contributions to scientific progress. The publication of a work in a prestigious journal can help to increase the personal and social recognition of professors and researchers, also conditioning directly the progression in their academic career. One of the aspects that most differentiates some disciplines from others is related to the vehicle of dissemination of research results, but, in addition, in some disciplines, as in the case of Psychology, it is observed that there are also some differences in the publication habits and research dissemination. The objective of this work is to identify and characterize the publication habits of researchers working in Psychology at a Spanish institution.

Material and Methods

The data has been extracted from the Web of Science (WoS) database. The period of the study covers the period 2008-2017. We identified the works published in Psychology journals that are included in the Journal Citation Reports (SCI and SSCI) throughout the period examined. The research analyses all the aspects related to the Spanish scientific production in Psychology in the 11 psychological thematic areas existing in the databases used as sources of information and those related to scientific collaboration, in order to know and compare the differences between the diverse thematic areas dedicated to Psychology.

Results

15,563 Psychology works carried out in Spanish institutions were analysed. When the Spanish scientific production is analysed according to the thematic areas (see Tables 1 and 2) in which the journals where these works have been published are included, it is observed that it is the category *P. Multidisciplinary* (6,297 papers), where more works have been published throughout the period analysed; followed by the *Psychology* category of the JCR – SCI (3,875 works) and *P., Experimental* with 2,742 papers. On the contrary, the thematic areas with the lowest scientific output during the analysed period are *P. Psychoanalysis* (68 works), *P. Mathematical* (298 works) and *P. Educational* (795 papers).

A feature of current science is the increase of scientific collaboration, which is explained by several reasons, including the need to have access to more resources or interest in collaborating with prestigious authors or the increase of the visibility of the works. The results of the analysis of scientific collaboration in Spanish Psychology show that in this area collaborative works predominate, and that they increase over time, as shown in terms of percentage in Figure 1.

Table 1. Evolution of the number of papers per year and thematic area*.

Years	P	PA	PB	PC	PD
2008	365	54	62	147	59
2009	283	112	92	148	68
2010	317	78	77	175	99
2011	339	116	72	218	113
2012	399	116	82	185	94
2013	458	124	105	186	134
2014	495	128	90	264	111
2015	396	136	77	271	321
2016	447	203	90	280	100
2017	376	227	89	295	129
Total	**3875**	**1294**	**836**	**2169**	**1228**

*P=Psychology (SCI); PA=Psychology Applied; PB=Psychology Biological; PC=Psychology Clinical; PD=Psychology Developmental

Table 2. (Continuation) Evolution of the number of papers per year and thematic area.**

Years	PE	PEx	PM	PMu	PP	PS
2008	54	178	29	977	6	67
2009	71	261	36	389	3	68
2010	105	252	30	481	6	76
2011	77	250	24	508	3	77
2012	69	279	24	618	4	87
2013	86	304	29	567	15	82
2014	86	316	34	620	3	101
2015	88	296	30	642	15	82
2016	82	312	32	750	8	100
2017	77	294	30	745	5	104
Total	**795**	**2742**	**298**	**6297**	**68**	**844**

**PE=Psychology Educational; PEx=Psychology Experimental; PM=Psychology Mathematical; PMu=Psychology Multidisciplinary; PP=Psychology Analysis; PS= Psychology Social.

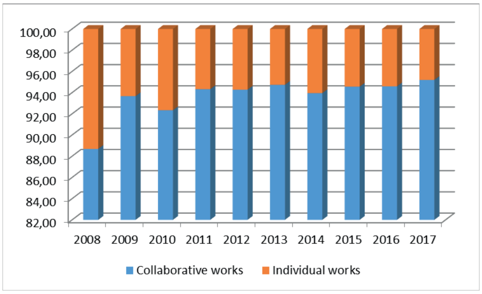


Figure 1. Evolution of scientific collaboration in Psychology.

However, there are significant different in the behaviour of researchers from different Psychology thematic areas. Table 3 shows the distribution in percentages of the number of works carried out by a single author and the number of works done in collaboration in the different thematic areas of Psychology. It is observed that the areas where a greater number of collaborative works are carried out correspond to *Psychology Biological* (96.41%), *Psychology Experimental* (95.55%) and *Psychology – SCIE* (95.41%), which confirms the idea that they have a behaviour much more similar to scientific and experimental areas. Whereas the area of *Psychology Psychoanalysis* (60.29%) stands out as it has the largest number of works done without collaboration, which corresponds to a behaviour much more common in the Humanities areas.

Table 3. Scientific collaboration per thematic areas (2008-2017).

Psychology Thematic areas	% Individual papers	% Collaborative papers
P. Applied	6,65%	93,35%
P. Biological	3,59%	96,41%
P. Clinical	4,79%	95,21%
P. Developmental	7,49%	92,51%
P. Educational	7,67%	92,33%
P. Experimental	4,45%	95,55%
P. Mathematical	15,10%	84,90%
P. Multidisciplinary	7,62%	92,38%
P. Psychoanalysis	42,65%	60,29%
P. Social	6,87%	93,13%
Psychology (SCIE)	4,59%	95,41%

Conclusions

This study shows the existence of different production and collaboration habits among the specialties within the Psychology field. Therefore, it is concluded that it is necessary to find an appropriate method that can serve to evaluate the activity of the different specialties in the most objective possible way by taking into account each of the different characteristics and peculiarities of the 11 psychological categories available in the JCR.

Acknowledgments

This paper is part of the research project UV-INV_AE18-785252 financed by the Universitat de València (Spain).

References

Haba-Osca, J., González-Sala, F. & Osca-Lluch, J. (2019). Education journals worldwide: an analysis of the publications included in the 2016 Journal Citation Reports (JCR). *Revista de Educación*, 383, 113-131.

Mayer, S. J. & Rathmann, J. M. K. (2018). How does research productivity relate to gender? Analyzing gender differences for multiple publication dimensions. *Scientometrics*, 117, 1663-1693

Osca-Lluch, J.& González-Sala, F. (2017). Evolución de las redes científicas y grupos de investigación. El caso de la psicología educativa en España durante los quinquenios 2004-2008 y 2009-2013. *Anales de Psicología*, 33, 356-364.

Uribe-Toril, J., Ruiz-Real, J.L., Haba-Osca, J. & Valenciano, J.D. (2019). Forests' First Decade: A Bibliometric Analysis Overview. *Forests,* 10(1), 72, https://doi.org/10.3390/f10010072

Co-citation in business translation research at Spanish centres: identifying topical similarities

Daniel Gallego-Hernández[1]

[1] *daniel.gallego@ua.es*
University of Alicante, Carretera Sant Vicent del Raspeig s/n, 03690, Sant Vicent del Raspeig - Alacant (Spain)

Introduction

As a professional activity in high demand in Spain and beyond, business translation in the broad sense (including economics, finance, etc.) has attracted researchers' attention in the first decades of the 21st century. Additionally, translator training centres have begun to include business translation in their curricula. However, business translation research has yet to be the object of a detailed bibliometric analysis aimed at mapping it in Spain.

The aim of this study is to start analysing business translation research performed by scholars affiliated to Spanish centres. This document describes the compilation of a bibliographic corpus, an initial bibliometrics-based study, and some results related to the distribution of publications by topics.

The main goals of the project are to map, heighten the visibility of and promote research on business translation.

Sample compilation and description

There is no specific bibliographic niche corresponding to business translation research. Publications on such research are scattered across journals, books and proceedings, and most are not indexed in mainstream databases. We thus had to manually retrieve a bibliographic corpus and record it in a spreadsheet, using various methods. Firstly, we consulted BITRA (Bibliography of Interpreting and Translation) (Franco, 2001-2018), the most comprehensive bibliographic database on translation and interpreting in the world, and TSB (Translation Studies Bibliography). We also consulted non-translation-specialised databases, namely WoS, Scopus, Google Scholar, BKCI-SSH and Mendeley Library. Secondly, we contacted scholars and business translation teachers by email and asked if they had published other works on business translation. We identified them by using the repertoire of subjects related to business translation in Spain compiled by Mateo (2014). Thirdly, we retrieved new publications while recording the references of works already included in the corpus on the spreadsheet.

The resulting bibliographic corpus contains a total of 539 publications by scholars affiliated to Spanish centres (206 articles, 256 book chapters, 24 books, two monographs, 34 PhD theses and 16 working papers), written in different languages (mainly Spanish and English). Figure 1 shows the growth in such publications over the years. The first publication dates from 1973. 2015 is the year with the highest productivity rate (53 publications), although this may be due to having not yet detected many publications from 2016-2018. In total, the works in the corpus cite more than 17,500 references (self-citations excluded).

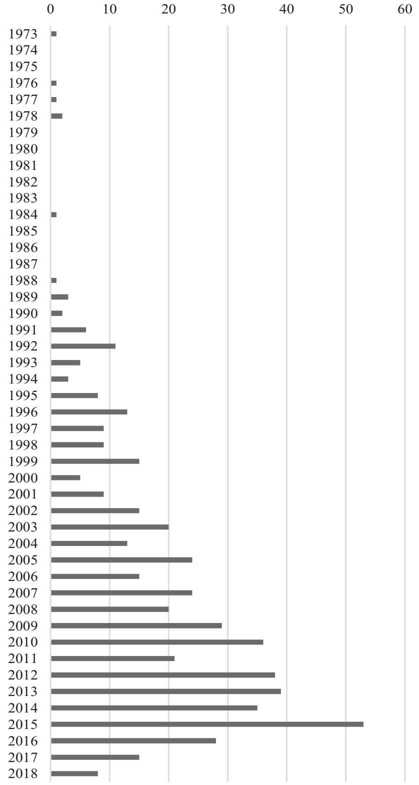


Figure 1. Publications by year

Analysis and results

We used co-citation analysis to establish subject similarities among business translation publications. Co-citation analysis was first introduced by Small (1973), as a better indicator of subject similarity than bibliographic coupling (Kessler, 1963). It can be defined as the frequency with which two documents are cited together. The more they are cited together, the stronger their relationship is.

We used BibExcel (Persson et al., 2009) and NodeXL Basic to detect co-occurrences from the spreadsheet containing the bibliographic corpus and create a network chart.

BibExcel extracted more than 50,000 pairs of co-cited references. We used 2,187 pairs of references cited together five or more times to create the co-citation network chart shown in Figure 2.

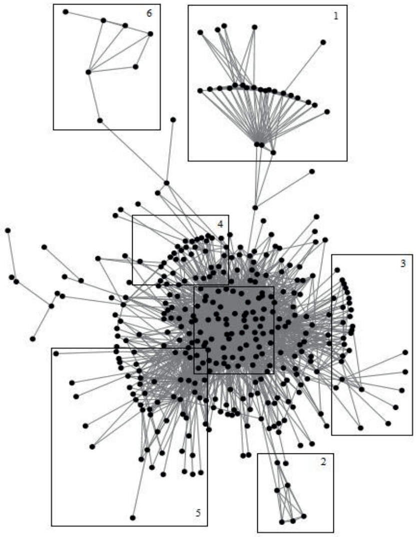


Figure 2. Co-citation network and subject areas

We have identified several subject areas. The core area (in the centre of the chart) represents theoretically-oriented translation studies, including the works of Newmark, Vázquez-Ayora, Hurtado, Nord, and Hatim and Mason.

Area 1 corresponds to lexicographical publications by authors such as Fuertes-Olivera, and Bergenholtz, Nielsen and Tarp from the Centre for Lexicography at Aarhus University in Denmark. Publications citing the studies in question apply the functional approach to lexicography to compiling/producing business dictionaries.

Area 2 represents the works of Bernardini, Zanettin, Corpas and Sánchez on corpus linguistics. Publications citing the works in question either use corpora to carry out research (corpus-based studies) or support the use of corpora as documentation resources for translation.

Area 3 comprises metaphor-related works citing: 1) main studies on metaphor, such as those of Lakoff and Johnson (general domain) or Henderson, Mccloskey and Charteris-Black (business domain); or 2) studies on metaphor translation, such as those of Samaniego or Deignan.

Area 4 includes the well-known works on genre theory of Swales and Bathia, as well as works by Gamero, García-Izquierdo and Borja from the GENTT (Textual Genres for Translation) research group of Universitat Jaume I (Spain). The publications related to business translation which cite them usually deal with a specific business genre and study its main features and translation.

Area 5 consists of Spanish works related to translation training. It includes publications by the PACTE (Translation Competence and the Acquisition of Translation Competence) research group from the Universitat Autònoma de Barcelona, Kelly's studies, the white paper on the Bachelor's Degree in Translation and Interpreting by ANECA (Spain's National Agency for Quality Assessment and Accreditation), and Roman's specific approaches to business translation training.

Finally, area 6 encompasses a small group of publications on Chinese and business translation.

Concluding remarks

We used co-citation to generate a graphical representation of the primary subject areas in business translation research in Spain. Despite the field being in its infancy, we identified various main areas. However, there are other main areas, such as history of business translation (only one co-occurrence), which were not retrieved using our technique. This may be due to the reduced number of publications on such topics, which should be enlarged in future research. Additionally, future research should identify minor subject areas by using other strategies, such as full text analysis or analysing scholarly documents from just one particular subject area. Such analyses will help improve the indexing of references in bibliographic databases such as BITRA.

References

Franco, J. (2001-2018). *BITRA (Bibliografía de Interpretación y Traducción).* Retrieved from: http://dti.ua.es/es/bitra/introduccion.html

Kessler, M. (1963). Bibliographic coupling between scientific papers. *American documentation,* 14(1), 10-25.

Mateo, J. (2014). Directorio de estudios de Traducción e Interpretación en España con materias de traducción de negocios. In D. Gallego (Ed.), *Traducción económica: entre profesión, formación y recursos* (pp. 201-208). Soria: Diputación.

Persson, O., Danell, R., Wiborg, J. (2009). How to use Bibexcel for various types of bibliometric analysis. In F. Åström, R. Danell, B. Larsen & J. Schneider (Eds.), *Celebrating scholarly communication studies* (pp. 9-24). Leuven: International Society for Scientometrics.

Small, H. (1973). Co-citation in the scientific literature. *Journal of the American Society for Information Science,* 24(4), 265-269.

The Character of the Tenure Track Professor Recruits at Aalto University

Leena Huiku[1], Anna-Kaisa Hyrkkänen[2], and Irma Pasanen[3]

corresponding author leena.huiku@aalto.fi
[1, 2, 3]Aalto University, P.O. Box 11000, 00076 AALTO, Finland

Introduction

Established in 2010, Aalto University is a new university with centuries of experience. It was created from the merger of three Finnish universities and today it consists of six schools with nearly 20 000 students and 4 700 employees, 390 of which are professors. Aalto University's research is concentrated around key areas combining four core competences in the fields of ICT, materials, arts, design and business together with three grand challenges related to energy, living environment, and health & wellbeing.

The university's ability in attracting and retaining the best professors is a key success element (Abramo, D'Angelo & Rosati, 2016). The tenure track system has clearly set evaluation criteria and selection process, which are based on the principles of predictability, transparency, and comparability with international standards. Since 2010, Aalto University has recruited over 300 professors on the tenure track.

Bibliometric analysis have been an integrated part of Aalto University recruiting processes since 2011 and the bibliometric service portfolio has evolved to support the different phases of the tenure track system. In the publication analysis the commonly used bibliometric indicators are used and the overall methodology follows the policies highlighted e.g. in the Leiden Manifesto and the Acumen Portfolio.

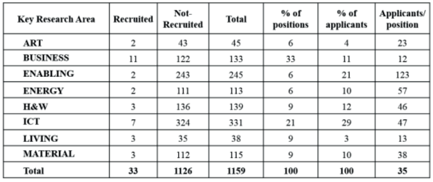
The research question in this analysis is: To which extent do the bibliometric indicators explain the choice of the recruited professor?

Data and methods

The statistical recruiting analysis is based on 33 tenure track positions where bibliometric analyses have been conducted. The level to which the tenure track professor will be appointed, is not always defined when the position is opened. The amount of the annual recruits is between 25 to 30. The total amount of applicants is 1159 splitted into eight key research areas. The mean value of the applicants per positions varies from 12 to 123 in the different areas.

The applicants are grouped to the Aalto's key research areas based on the respective areas of the recruited professors. All results are produced at the key research area level and at the university level as well.

Table 1. Basic information about the tenure track positions involved in the analysis.



Key Research Area	Recruited	Not-Recruited	Total	% of positions	% of applicants	Applicants/ position
ART	2	43	45	6	4	23
BUSINESS	11	122	133	33	11	12
ENABLING	2	243	245	6	21	123
ENERGY	2	111	113	6	10	57
H&W	3	136	139	9	12	46
ICT	7	324	331	21	29	47
LIVING	3	35	38	9	3	13
MATERIAL	3	112	115	9	10	38
Total	**33**	**1126**	**1159**	**100**	**100**	**35**

The data of bibliometric analyses have been gathered from the applicant's bibliometric reports. The bibliometric analyses are based on Web of Science and Scopus databases and on Google Scholar as well according to the choices by the recruiting committee. The journal rankings have also been utilized. These sources are commonly used in academic institutes e.g. Gorraiz & Gumpenberger (2015).

The data have been processed using analytical tools: SPSS, Power BI, RStudio, and Python.

The explanatory variables used in the analysis include the number of publications, the number of citations, H-index and the first year of publication in the respective database. The target variable is the binary variable recruited/not-recruited.

Results

The analysis indicates that Aalto University has recruited young, talented and highly cited researchers to the tenure track positions.

The recruited professors are slightly at the earlier phases of their scientific career compared to the not-recruited applicants in most key research areas (Figure 1.). The career length is estimated by substracting the year of the first publication in respective database from the year of the application. The method can be used as an approximation of career length according to the results by Nane, Larivière & Costas (2017).

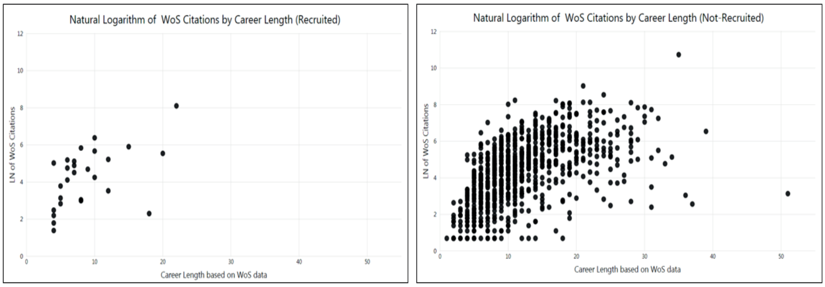


Figure 1. Career Length of the recruited/not-recruited based on Web of Science data.

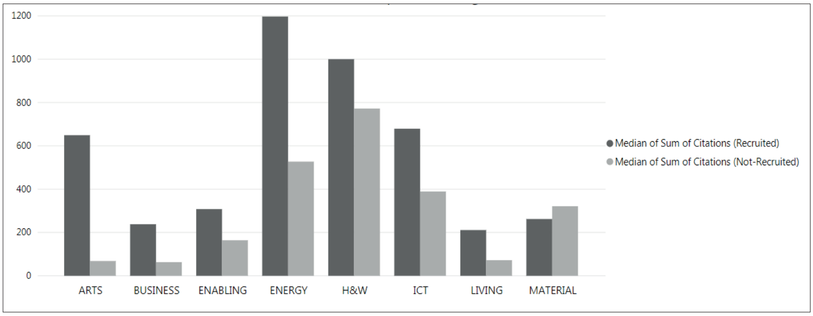


Figure 2. Median of the sum of citations based on publications in Scopus, Web of Science and Google Scholar.

The publications of the recruited professors have received more citations in the different databases than the publications of the not-recruited applicants in nearly all key research areas. The measure for this is the median of sum of citations, which is in the different databases higher among the recruited professors compared to the not-recruited applicants in nearly all key research areas (Figure 2.).

On the contrary, the number of publications per year doesn't differ significantly among the recruited and not-recruited. In some key research areas, the level of journals in which the applicants have managed to publish their articles, has been evaluated using different journal rankings. Here the results showcase that the recruited have published in more prestigious journals than not-recruited. The results can be interpreted that the same productivity has generated more attention in the scientific community among the recruited professors compered to not-recruited. (Kolesnikov, Fukumoto & Boseman, 2018)

The statistically significant difference between the amount of citations among recruited and not-recruited indicates that the bibliometric analyses conducted during the years have brought valuable information to the recruiting process.

Acknowledgments

Certain data included herein are derived from the SCI Expanded, SSC Index and A&HC Index © Copyright Clarivate Analytics® Philadelphia, PA, USA, 2018. All rights reserved.

Certain data included herein are derived from the Scopus® © Copyright ELSEVIER®, Amsterdam, Netherlands, 2018.

References

Abramo, G., D'Angelo, C.A. & Rosati, F. (2016). A methodology to measure the effectiveness of academic recruitment and turnover. *Journal of Informetrics, 10, 31-42.*

Gorraiz, J. & Gumpenberger, C. (2015). A flexible bibliometric approach for the assessment of professorial appointments. *Scientometrics*, 105, 1699-1719.

Kolesnikov, S., Fukumoto, E. & Boseman, B. (2018). Researchers' risk-smoothing publication strategies: Is productivity the enemy of impact? *Scientometrics*, 116, 1995-2017.

Nane, G.F., Larivière, V. & Costas, R. (2017). Predicting the age of researchers using bibliometric data. *Journal of Informetrics*, 11, 713-729.

The development of a new instrument to measure research agendas

Hugo Horta[1] and João M. Santos[2]

[1] *horta@hku.hk*
The University of Hong Kong, Pokfulam, Hong Kong, Hong Kong SAR, China

[2] *joao_marques_santos@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL), Avenida das Forças Armadas, Lisbon, Portugal

Introduction

The goal of this study is the development of an instrument capable of evaluating the dimensions which determine the research agendas of researchers across fields of science. It is inspired on a previous inventory which focused exclusively on the social sciences, and further improves it by introducing new dimensions and expanding its scope to include other fields of science.

The original instrument, as far as the authors know, is the first of its kind that is able to evaluate the dimensions that shape research agendas (Horta & Santos, 2016). The present study departs from the previous one by using a global sample of over 12,000 researchers who provided information on their own research agendas, allowing validation across fields of science as well as expanding the scope of the instrument and further improving its robustness.

Method

The analysis employed Exploratory and Confirmatory Factor Analysis (Kline, 2015) with the goal of conducting a full suite of validation exercises on the instrument, including Validity, Reliability, and Sensibility. Multi-group analysis was also employed with the goal of demonstrating measurement invariance.

Results

The validated iteration of the instrument contains 40 questions which cover 8 factors – Society Driven, Academia Driven, Discovery, Tolerance to Low Funding, Mentor Influence, Collaboration, Divergence, and Scientific Ambition. Model fit was adjudged as very good. Factorial and discriminant validity were both confirmed, while convergent validity exhibited minor issues. Furthermore, good reliability and sensitivity were also observed, and measurement invariance with full metric and scalar invariance, as well as partial construct invariance.

Discussion

The final instrument is comprised of 8 factors. The first one, Scientific Ambition, is measures the desire to acquire recognition through one's work and to acquire authority in the field (Latour and Woolgar, 2013). Collaboration, the second dimension, relates to the propensity to engage in collaborative ventures, as well as having the opportunity to do so. The third dimension, Tolerance to Low Funding, measures the degree of risk tolerance regarding engaging in scientific ventures where funding is considered limited or difficult to obtain. The fourth dimension, Mentor Influence, measures the degree to which the researcher's work is influenced by his or her PhD supervisor or mentor. The fifth dimension, Discovery, evaluates the researcher's propensity to engage in agendas which are considered riskier but also potentially more rewarding. The sixth dimension, Divergence, represents a researcher's willingness to engage in agendas which involve topics and knowledge from outside the researcher's own field.

The seventh dimension, and one of the new ones, is Academia Driven – this dimension represents the degree to which the researcher is aligned with the overarching agenda set by his field's community, as well as the demands of his or her institution. Finally, the eighth dimension, Society Driven, is a measure of alignment with the society at large – a researcher who scores high on this dimension is focused on tackling societal challenges and chooses his research agenda based on the needs of society.

Acknowledgments

This work was supported by doctoral grant PD/BD/113999/2015 from the Fundação para a Ciência e Tecnologia. This research was also funded by the Research Grants Council (Hong Kong) through project number: 27608516.

References

Horta, H., & Santos, J. M. (2016). An instrument to measure individuals' research agenda setting: the multi-dimensional research agendas inventory. Scientometrics, 108(3), 1243-1265.

Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

Kuhn, T. S. (2012). The structure of scientific revolutions. University of Chicago press.

Latour, B., & Woolgar, S. (2013). Laboratory life: The construction of scientific facts. Princeton University Press.

A Bibliometric Analysis of the #MeToo Movement in South Korea

Bitnari Yun[1,] Jinseo Park[2] and Sejung Ahn[3]

[1] *kisti0746@kisti.re.kr*
Future Technology Analysis Center, Korea Institute of Science and Technology Information, 66, Hoegi-ro, Dongdaemun-gu, Seoul, 02456 (Korea) · Data& High Performance Computing Science, University of Science and Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, 34113 (Korea)

[2] *jayoujin@kisti.re.kr*
Future Technology Analysis Center, Korea Institute of Science and Technology Information, 66, Hoegi-ro, Dongdaemun-gu, Seoul, 02456 (Korea)

[3] *sjahn@kisti.re.kr*
Future Technology Analysis Center, Korea Institute of Science and Technology Information, 66, Hoegi-ro, Dongdaemun-gu, Seoul, 02456 (Korea) · Data& High Performance Computing Science, University of Science and Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, 34113 (Korea)

Introduction

As big data analysis allows us to investigate the structure and value of large quantities of data, it has become possible to capture social issues in a different way from the fast. According to Zolli (2018), big data analysis has strengths in discovering and addressing social issues in terms of precision and resolution, frequency, scale and reach, predictive capacity, sophistication, and interoperability.

In this study, we identify social issues related to the MeToo movement in South Korea and examine its characteristics by analysing news web records. The MeToo movement is a movement against sexual violence that spreads through online social media, which began in the United States. News web records, one type of unstructured text data that provide insight on social issues, were explored to identify major social issues related to the MeToo and the characteristics of the issues.

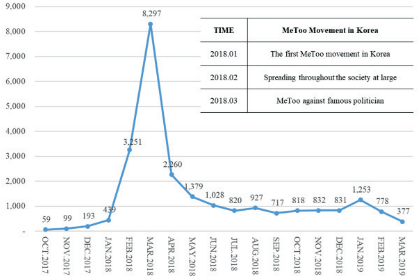
Data and Methods

To identify social issues of the MeToo movement in South Korea, we used the BIGKinds, a Korean analytical web service. It provides data for analysing social phenomena by converting unstructured text data such as daily newspapers, business magazine, regional daily newspapers, and broadcasters of Korea into structured text data developed at Korea Press Foundation (Korea Press Foundation, 2017). We retrieved the keyword 'MeToo' on the BIGKinds and composed a collection of 24,358 news from Oct. 2017 to Mar. 2019 for analysing.

To identify major issues related to Me Too movement, we used topic modelling algorithms using LDA (Latent Dirichlet Allocation) model. LDA is a generative probabilistic model of corpus, used to identify latent topics of a collection of documents (Blei, Ng & Jordan, 2007), topic modelling, which is fast algorithms for computing LDA, has been applied in many research domains (Blei, Carin & Dunson, 2010). In this study, we used topic modelling to reveal embedded structures in the collection of news data related to MeToo movement in South Korea.

Result and Discussion

Figure 1 shows the number of monthly news. In South Korea, a series of news about MeToo began to reported in October, 2017, when exposing sexual-abuse allegation against famous film producer in United States. In the early stage, a small number of news were reported, referring to overseas MeToo movement. However, on January 29, 2018, the number of news has increased sharply as it turned out that a female prosecutor was sexually harassed by a high-level prosecutor and disadvantaged in personnel transfers. It became a starting point of the Korea MeToo movement and the movement spread throughout the society at large as many women shared their experiences. The number of news surged in the March of 2018, as a prominent presidential contender and former province governor, Ahn H.J., was accused of committing sexual assault by his secretary, Kim J.E., and he resigned from the public service.



TIME	MeToo Movement in Korea
2018.01	The first MeToo movement in Korea
2018.02	Spreading throughout the society at large
2018.03	MeToo against famous politician

Figure 1. The number of news (inset) Timeline of MeToo movement in South Korea

To identify major issues of the MeToo movement, 10 major topics were drawn from the collection of news by using topic modelling algorithms and the keywords for each topic were extracted in order of importance. As shown in Table 1, various topics of politics, education, culture and art, social movement and so on were identified from the collection of news related to the MeToo movement in South Korea. In particular, the case of the governor mentioned in Figure 1 constituted Topic10, which had a huge influence on Korea MeTo Movement.

Table 1. Keyword distribution of topic related MeToo movement (10-Topic LDA Model)

Topic1	Topic2	Topic3	Topic4	Topic5
Parliament	Entertainment Industry	Education	Culture	Social Movement
Assembly Member Democratic Party Candidate Election Local	Actor MeToo Movie Sexual assault Broadcasting	Professor MeToo School Student Teacher	Lee Y.T. Paly Poet Sexual assault MeToo	Women MeToo Movement Society Organization
Topic6	**Topic7**	**Topic8**	**Topic9**	**Topic10**
Government	MeToo Movement	Investigation Process	MeToo Movement	Politics
Moon J.I. USA Time Local Representative	Sexual Violence MeToo Sexual Harassment Education Culture	Suspicion The Prosecution Seoul The Police Investigation	MeToo Movement Sexual Violence Exposure Damage	Ahn H.J. Chugnam Rape Secretory Kim J.E.

As a result of topic modelling, the following characteristics were found. First, a number of individual topics were formed centering on allegations against public figures in the various social fields (Topic2,3,4,10), the names of the perpetrators or the victims of relevant cases appeared among the topics frequently (Lee Y.T., Ahn H.J., Kim J.E.). Second, topics that examined the meaning, effects, and countermeasures of the MeToo movement in terms of social movements were observed (Topic5,7,9). Lastly, topics covering political responses to the MeToo movement (Topic1,6) and focusing on investigation process of the movement(Topic8) were observed also.

This relationship between topics is also appeared through the distance map of each topic. In Figure 2, the circles represent a single topic cluster, and the closer the distance is, the more relevant the topic is. The distance map shows that similar topics are located in close proximity.

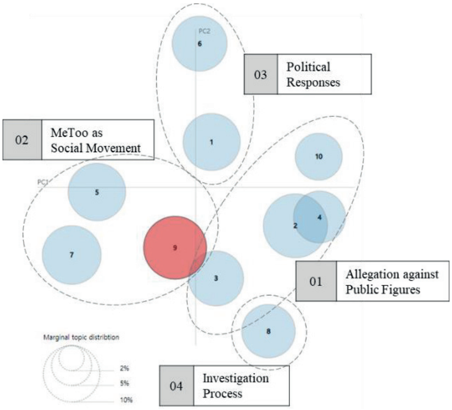


Figure 2 The intertopic distance map (via multidimensional scaling)

Conclusion

In this study, we identified 10 social issues regarding the MeToo movement in South Korea by using topic modelling algorithms on news. As a result, the allegation against public figures had the greatest influence on the issue formation, and keywords related to social movements, political responses, and the investigation process also affected issue formation of the MeToo movement. For more sophisticated study, it is required to conduct analysis such as a cross-national comparative study, sentiment analysis, and opinion mining in the future study.

Acknowledgments

This study was funded by the Korea Research Foundation (NRF-2017R1E1A1A03070975) and the Korea Institute of Science and Technology Information (K-19-L03-C03-S01).

References

Blei, D., Ng, A. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blei, D., Carin, L. & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6), 55-65.

Korea Press Foundation. (2017). *BIGKinds User Manual*. Seoul: Korea Press Foundation.

Zolli, A. (2018.01). After big data: The coming age of “big indicator”, *Stanford Social Innovation Review*. Retrieved March 20, 2019 from: https://ssir.org/articles/entry/after_big_data_the_coming_age_of_big_indicators#.

Study on open science: the general state of the play in Open Science principles and practices at European life sciences institutes

Pavla Foltynova[1] and Katerina Ornerova[2]

[1] *pavla.foltynova@ceitec.muni.cz*
CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 753/5
625 00 Brno (Czech Republic)

[2] *katerina.ornerova@ceitec.muni.cz*
CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 753/5
625 00 Brno (Czech Republic)

Introduction

Nowadays, open science is a hot topic on all levels and also is one of the priorities of the European Research Area. Components that are commonly associated with open science are open access, open data, open methodology, open source, open peer review, open science policies and citizen science. Open science may a great potential to connect and influence the practices of researchers, funding institutions and the public.
In this paper, we evaluate the level of openness based on public surveys at four European life sciences institutes.

Method

The level of openness was investigated through an online questionnaire disseminated to all the staff members of the examined organizations. It included 11 questions, including Likert Scale, Multiple-choice and Multiple-choice grid questions. The questionnaire was disseminated in February 2018 using a specific online form per institution, the target population included a total of 3517 people. A total of 334 participants completed the questionnaire. The profile of these participants: researchers (senior and PI) n = 87; junior researchers n = 130; technical support n = 25; outreach n = 22; management n = 59; and funding n = 11. Gender distribution: male n = 170; female n = 156. These data are part of WP2 of ORION project.
The institutes:

- Babraham Institute – UK (BI)
- Central European Institute of Technology - Czech Republic (CEITEC)
- Max Delbürck Center – Germany (MDC)
- Center for Genomic Regulation – Spain (CGR)

Results

Overall view on open science

The results showed that the most participants were declare that the Open Science is mostly positive for Science (all institutions more than 40%) and only minimum participants answered that Open Science is real threat to Science (2% at BI and 2.1% at CRG) or a worrying new perspective for Science (3% at CEITEC and 3% at MDC) as can be seen in graph 1. The opinion that Open Science is an exciting opportunity for Science had mostly participants from CRG (39%) and MDC (35%).

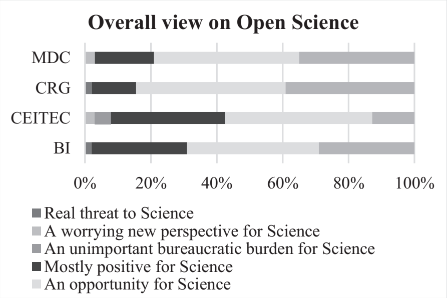


Figure 1: Question: Overall, if you had to summarise your view on Open Science, what would you say?

Openness to stakeholders

Generally, views on openness to different stakeholders are shifted towards a very open approach, especially for scientists from the same area/discipline (4.83 SD 0.26) and for scientists from other disciplines (4.72 SD 0.62). The lowest opening is observed for industry and companies (3.91 SD 1.29).
As can be seen in graph 2, the most opened institution for all groups is MDC (4.53 SD 0.31). Data for CEITEC showed low value (4.08 SD 0.44) in comparison with other institutions.

Reasons for and against open science

The most outstanding benefits for open science were marked Efficiency (97%), Equity (94%), Ethics (89%), Fairness (88%) and Rigour (87%) according to the participants.

On the other hand, the most important reasons against open science were Danger and potential misuse is the most relevant argument (49% of participants considered it an important or the most important reason), followed by Low quality (45%) and Unfairness (29%). The most relevant barriers that participants facing in relation with open science were Budget and funding (76% of participants), Lack of clear steps to follow (66%), Fears and uncertainties for career development (66%), Authentic public engagement (62%), Lack of proper infrastructure (61%). The item considered as a less relevant barrier is Time constraints, despite it is also being identified by 53% of participants as an important or the most important barrier.

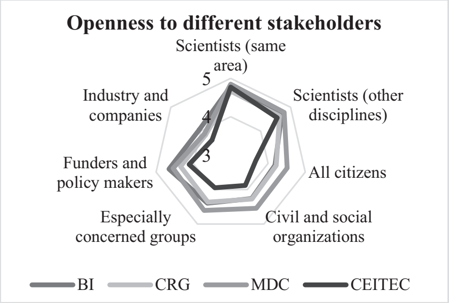


Figure 2: Question: In your opinion, to whom should science be opened?

Open Science activities

The action with the highest participation is Dissemination to scientists according to the survey, since 80% of respondent state participating regularly or sporadically. Next, is Collaboration with scientists; Dissemination to the public and Open access can be also considered as very common actions with 70, 66 and 64% of participants reporting their participation (whether regularly or sporadically). On another level, about 50% of participants declare their participation in three other activities or actions: Science education (54%), Open Data (50%) and Gender Equality (48%). Finally, three other tasks appear in the lower part of the ranking: Collaboration with industry (38%), Ethics and integrity (36%) and Collaboration with funders (30%).

Open access

Comparison of publications in the journals indexed in JCR® in open access (gold way) in shown in graph 3.

As can be seen, the percentage of open access publications increased by the time at all examined institutions. The highest percentage of open access publications was found at CRG (42%), followed by BI (32%) and MDC (29%), and the lowest value was found at CEITEC (23%).

Conclusion

The results showed that Open Science is mainly perceived as an opportunity for science, with the benefits outweighing the drawbacks. Science should be open to all the stakeholders, but especially to scientists themselves. Overall, responsibilities of science regarding Open Science are perceived as more relevant than benefits for science, and barriers to Science are perceived as more relevant than reasons against Open Science.

The results of the questionnaire analysis also show that dissemination and collaboration among scientists are the most frequent actions related with open science, and collaboration with funders and industry are the least frequent.

The highest level of openness from examined institutes was found at CRG, followed by MDC and BI, and the lowest level of open science was found at CEITEC.

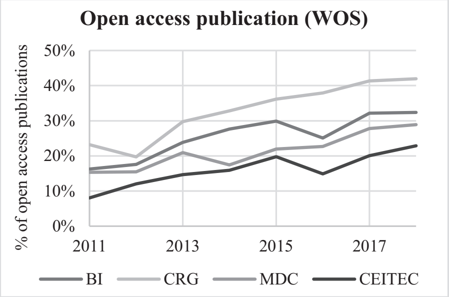


Figure 3: Percentage of open access publications in journals indexed in JCR (Clarivate Analytics®)

Acknowledgments

The data in this study are a part of project ORION. ORION has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 741527.

References

ORION. (2017). *Deliverable 2.1. Scheme for data collection* (Report D2.1) Retrieved March 28, 2019, from https://www.orion-openscience.eu/public/2017-09/D2.1.%20Scheme%20for%20data%20collection.pdf

ORION. (2018). *Deliverable 2.2 Analysis and Benchmarking: Self-assessment* (Report D2.2) https://www.orion-openscience.eu/publications/

Research Evaluation and Scientific Productivity: A bibliometric analysis of the Nigerian University of Calabar

Okon E. Ani

anioedet@yahoo.com

University of Calabar Library

University of Calabar, Calabar

Nigeria

Introduction

Research is critical for knowledge generation towards societal development and nation building. In view of this governments globally are investing on research as a veritable tool towards global transformation of the society. Dane (2011) describes scientific research as a vital process for making enquiry about the universe or society.It is globally asserted that scientific productivity is used as a measure of the research performance in universities (Aalfojarvi et al., 2008**).** The objectives of the paper are to determine the most productive scientists, cited scientists, and scientists with highest h-index.

Research method

Scopus database was used for data collection in the study. Scopus provides free access to "Author search" at its website. Three science-based faculties in the University of Calabar: Faculty of Physical Sciences, Faculty of Biological Sciences, and Faculty of Agriculture were used for the study. The lists of academic scientists indicating their names, ranks, and gender were obtained from all the Departments from the three faculties for the study. Graduate assistants and assistant lecturers were excluded from the study (since they are not active in research). Author search was conducted at the Scopus website (https: //www.scopus.com/home.uri) to obtain the number of publications, citations, and h-index for each academic scientist. These searches were carried out in November 2018.

Results

The results of the study are presented below:

Most productive scientists

Table 1: Most productive scientists

SN	Author	Rank	Sex	Faculty	Department	Article
1	A.E. Eneji	Professor	M	Agriculture	Soil Science	134
2	P. C. Okafor	Associate Professor	M	Physical Sciences	Chemistry	56
3	B. I. Ita	Associate Professor	M	Physical Sciences	Chemistry	52
4	A. E. Edet	Professor	M	Physical Sciences	Geology	41
5	O. E. Offiong	Professor	M	Physical Sciences	Chemistry	39
6	S.P. Antai	Professor	M	Biological Sciences	Microbiology	35
7	V. Ntui	Senior Lecturer	M	Biological Sciences	Genetics	31
8	E. I. Braide	Professor	F	Biological Sciences	Zoology	30
9	B. N. Ekwueme	Professor	M	Physical Sciences	Geology	25
9	A.E. Akpan	Professor	M	Physics	Physics	25
9	A. Brisibe	Professor	M	Biological Sciences	Genetics	25
10	F. I. Bassey	Associate Professor	M	Physical Sciences	Chemistry	24
10	E.V. Ikpeme	Senior Lecturer	F	Biological Sciences	Genetics	24

Most cited scientists

Table 2: Most cited scientists

SN	Author	Rank	Sex	Faculty	Department	Citations
1	P. C. Okafor	Associate Professor	M	Physical Sciences	Chemistry	2138
2	A.E. Eneji	Professor	M	Agriculture	Soil Science	1931
3	O. E. Offiong	Professor	M	Physical Sciences	Chemistry	885
4	B. I. Ita	Associate Professor	M	Physical Sciences	Chemistry	753
5	A.A. Ayi	Professor	M	Physical Sciences	Chemistry	708
6	A.E. Edet	Professor	M	Physical Sciences	Chemistry	582
7	S.P. Antai	Professor	M	Biological Sciences	Microbiology	494
8	S.W. Petters	Professor	M	Physical Sciences	Geology	402
9	M.E. Ikpi	Senior Lecturer	F	Physical Sciences	Chemistry	391
10	C.S. Okereke	Professor	M	Physical Sciences	Chemistry	378

Scientists with highest h-index

Table 3: Scientists with highest h-index

SN	Author	Rank	Sex	Faculty	Department	h-index
1	P. C. Okafor	Associate Professor	M	Physical Sciences	Chemistry	25
2	A.E. Eneji	Professor	M	Agriculture	Soil Science	24
3	A.E. Edet	Professor	M	Physical Sciences	Geology	16
3	O.E. Offiong	Professor	M	Physical Sciences	Chemistry	16
4	B.I. Ita	Professor	M	Physical Sciences	Chemistry	13
5	S. P. Antai	Professor	M	Biological Sciences	Microbiology	12
6	S.W. Petters	Professor	M	Physical Sciences	Geology	11
6	C.S. Okereke	Professor	M	Physical Sciences	Geology	11

Conclusion

Research is crucial for global development of the society. The findings of the study indicate that most productive and cited scientists are senior scientists at professorial cadre and from Chemistry Department, Faculty of Physical Sciences. It is recommended that the University Management should provide equitable research facilities/funding across all professional ranks, disciplines and gender to enhance quality of scientific research.

References

Aalfojarvi, I., Arminen, I., Auranen, O. and Pasanen, H. (2008). Scientific productivity, web visibility and citation patterns in sixteen nordic sociology departments. *ACTA Sociologica,* 51 (1): 5-22

Dane, F. C. (2011). *Evaluating Research Methodology for People Who Need to Read Research.* London: Sage Publications.

The Role of Research Collaborations for Academic Performance in Italy: An Empirical Analysis of Scopus Data

Luigi Aldieri[1], Gennaro Guida[2], Maxim Kotsemir[3] and Concetto Paolo Vinci[4]

[1] *laldieri@unisa.it*
Department of Economic and Statistical Sciences, University of Fisciano, Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Salerno (Italy)

[2] *gennaro.guida@uniparthenope.it*
Department of Economic and Legal Studies, Parthenope University of Naples, Via Generale Parisi, 13, 80132, Naples (Italy)

[3] *mkotsemir@hse.ru*
Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, 20 Myasnitskaya Street, 101000, Moscow (Russian Federation)

[4] *cpvinci@unisa.it*
Department of Economic and Statistical Sciences, University of Fisciano, Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Salerno (Italy)

Introduction

There are different streams in the literature to explore research collaboration (Acosta et al., 2010; Fantino et al., 2015; Bergé, 2017; Shashnov and Kotsemir, 2018; Kotsemir, 2019). The topic concerning the research collaboration assumes a key role to the development of an academic institution in Italy, where a higher share of fund-raising depends on scientific research output. As discussed in the literature (Aldieri et al., 2018), the knowledge flows arisen between researchers from different universities are relevant to enhancing the quality of research and at the same time Italian and Russian universities could improve their interactions with international institutional partners. Paying attention to the Italian case, we intend to investigate the distribution of collaboration activities for regions evidencing the explanatory variables able to produce such a result.

Following this approach, we explore the variables which are the most significant inside the research production function and which could be an important instrument to favour less opened institutions.

In order to achieve our objective, we consider the following set of variables: age, gender, academic institution, disciplinary field and position relative to authors; number of administrative staff employed; amount of study grants; foundation year; number of spin off, relative to universities. Thus, in order to evaluate the output of scientific research we use a mean of 5 bibliometric indexes as the IF5Y (the Impact Factor calculated on a period of the last 5 Years), the AIS (Article Influence Score), the IPP (Impact Per Publication) the SJR (SCImago Journal Rank), and the H Index for the period 2010-2014 (see detailed description of all these variables in Aldieri et al., 2019).

Methodological Approach

The aim of our investigation is to estimate the effect of single authorship (single), national external collaborations (nat_coll) and international collaborations (int_coll) on the quality indicator of Italian universities, measured by the new quality index (pc) that encompasses eight indicators (number of publications indicators of journal quality an citation indicators). PC variable is obtained from a principal component analysis (PCA) process (see more in Aldieri et al., 2019). In particular, we investigate the corpus 5002 publications of Italian researchers belonging to the statistical economic sector in Scopus for the period 2007 – 2016. Model that is estimated is the following (1):

$$PC_{i,k} = C\ (Coll,\ x_{i,k}\ ,\ z_i\ ,\ w_k\ )\ (1)$$

The university-specific characteristics (vector xi,k) include the number of students (stud) and Government transfers received (fund). The institution-specific characteristics that affect the quality of a unit’s publications (zi) consider the “age” of an academic institution (found), that is, the years elapsed from its establishment to 2017, the number of faculty staff (pstaff) and the number of spinoffs (SpOff).

The publications in scientific fields (wk in the model) are grouped into four sectors: 1) economics; 2) business management; 3) economic history; 4) statistics. This classification of scientific fields is used by the National Agency for Assessing University and Scientific Research (also known as ANVUR) for the assessment of scientific research in Italy. All variables used in our model are described in Table 1.

Table 1. Description of variables of the model

Variable	Description of Variable
PC	Dependent Variable of publication quality stemmed from the PCA process
Gend	Dummy variable assumes 1 if first author is male and 0 otherwise
Age	First author's age
Stud	Number of students enrolled in the University
Fund	Amount of funds received from University
pstaff	Number of administrative staff in the University
Found	Foundation year of University
SpOff	Dummy variable assumes the value of 1 if University has spin-off
Dist	Geographical distance in km between two Universities
Single	Dummy variable assumes the value of 1 if there is only 1 author in a specific publication
Int_coll	Dummy variable assumes the value of 1 if there is at least 1 foreign author (not from Italy) in a specific publication
Nat_coll	Dummy variable assumes the value of 1 if there is at least 1 author of other national (Italian) University in a specific publication

Results

In order to solve the potential endogeneity of collaborations, we use GMM (Generalized Method of Moments) techniques for instrumental variables, which allow endogenous variables to be instrumented by excluded instruments (See Table 2).

Table 2. GMM model results (dependent variable is PC)

Coeff. Est[d].	GMM	
	Coeff.	s.e.[a]
Gend	-32.05***	(11.25)
Age	0.81	(0.772)
Stud	-0.04***	(0.001)
Fund	0.01***	(0.001)
pstaff	-0.15***	(0.030)
Found	-0.26***	(0.028)
SpOff	-10.3***	(0.475)
Dist	-0.06***	(0.024)
Single	3.53***	(0.732)
Int_coll	13.47***	(0.406)
Nat_coll	2.53***	(0.377)
R2		

Notes. a: ***,**,* Coefficient significant at the 1% , 5% , 10%. b: Year, field and country dummies are included in the estimation procedure. c: Standard errors are corrected for heteroscedasticity. d: 5002 observations.

From the empirical findings, we can observe that both international and national external collaborations lead to a significant and positive effect on research evaluation process, but the magnitude of international collaborations is higher. Concerning the other control variables, our results show that male authorship and the amount of funds received from university increase the academic research quality. On the other hand, variables like number of students, number of administrative staff, university foundation year, the presence of spin-off and the average distance from another university lead to academic quality deterioration.

Conclusions

The full-article version of this research (Aldieri et al., 2019, available at https://rdcu.be/bqLxZ) provides more comprehensive analysis of the impact of research collaborations on the scientific performance of Italian universities, including geographical aspects (analysis of the effects of collaboration on the level of regions and provinces of Italy).

One the way of development of research is its replication for the case of several countries with different academic systems. The other way of development is the expansion of the sample (in terms of time span; number of universities, the corpus (and thematic coverage) of publications) to detect some effects due to trends and disciplinary aspects in the development of collaborations.

References

Acosta, M., Coronado, D., Ferrándiz, E., & León, M. D. (2010). Factors affecting inter-regional academic scientific collaboration within Europe: The role of economic distance. *Scientometrics*, *87*(1), 63-74.

Aldieri, L., Kotsemir, M., & Vinci, C. P. (2018). The impact of research collaboration on academic performance: An empirical analysis for some European countries. *Socio-Economic Planning Sciences*. 62, 13–30.

Aldieri L., Gennaro G., Kotsemir M., & Vinci C. P. (2019). An investigation of impact of research collaboration on academic performance in Italy. *Quality and Quantity*. 1-38, article in press.

Bergé, L. R. (2017). Network proximity in the geography of research collaboration. *Papers in Regional Science*, *96*(4), 785-815.

Fantino, D., Mori, A., & Scalise, D. (2015). Collaboration between firms and universities in Italy: The role of a firm's proximity to top-rated departments. *Italian Economic Journal*, *1*(2), 219-251.

Kotsemir M. (2019). Unmanned aerial vehicles research in Scopus: an analysis and visualization of publication activity and research collaboration at the country level // *Quality and Quantity*, 1-31. https://rdcu.be/btN2A

Shashnov, S., and Kotsemir, M. (2018). Research landscape of the BRICS countries: current trends in research output, thematic structures of publications, and the relative influence of partners. *Scientometrics*, 117(2), 1115–1155. https://rdcu.be/4rWv

The impacts of network mechanisms on scholars' perceptions and behaviours in research community

Chien Hsiang Liao

jeffen@gmail.com
Fu Jen Catholic University, Department of Information Management (Taiwan)

Introduction

Based on the existing literature, network mechanisms have been denoted to be related to scale-free network evolution and development. More specifically, these mechanisms reflect the people's tendencies to choose or connect a network, so that there are some signs of network development can be observed. According to the literature, this study concludes possible mechanisms (or tendencies) are preferential attachment, homophily, reciprocity, and triadic closure mechanisms. For instance, preferential attachment reflects that the existing nodes that are more attractive will have a higher chance to receive new connections, this mechanism leads to scale-free network's growth, formation and clustering (Spinellis and Louridas, 2008).

Nevertheless, this study believes that these mechanisms not only have an impact on the development of the network (from a macro perspective), but also on people's perceptions and social exchange behaviours (from a micro perspective). In the context of research community, scholar's research collaboration, perceived research quality, citation intention are always critical research issues in the past. In this vein, this study attempts to explore the impacts of network mechanisms on these behaviours and perceptions. Besides, performing citizenship behaviour toward the benefits of other members (courtesy behaviour) is positive feedback and beneficial for the development of research community (Chiu et al., 2015). Therefore, this study also incorporates it into the research model. The specific research purposes are - (1) Conceptualizing possible network mechanisms and developing the assessments; (2) Measuring and validating these network mechanisms by exploratory factor analysis; (3) Examining the associations between network mechanisms and scholars' perceptions and behaviours.

Theoretical background

Network mechanisms.

Social network is an intricate structure which contains complex interactions and relationships. Network theory (also called graph theory) provides a theoretic representation to interpret a complex structure via graph (Granovetter, 1983). As mentioned, the network mechanisms derived from network theory use a set of techniques to analyze its structure and explain network evolution and dynamic. Preferential attachment is a tendency of participants to link to the most popular participants (Barabási and Albert, 1999). However, past studies have made a slight difference in the definition of preference attachment. Omidi and Masoudi-Nejad (2010) suggest that preferential attachment attracts new entrants due to its 'popularity', while Ozmen et al. (2012) treat its effect as 'familiarity' and 'well-known' attribute. To carefully address preferential attachment, this study divides it into two attributes - popularity and familiarity mechanisms. Popularity mechanism refers to its high attraction to new members, while familiarity mechanism reflects its high visibility and well-known characteristic to members.

Although preferential attachment is a primary mechanism to explain new entrant's behaviour and tendency (Johnson et al., 2014), some possible network mechanisms should be noted. Homophily mechanism represents the tendency of people connect with others capturing similar features (Papadopoulos et al., 2012), e.g., similar research interests, expertise, and domain. Moreover, reciprocity mechanism is a motivation of user participation which new entrants want to obtain or share knowledge or resources with one another (Johnson et al., 2014). Finally, the triadic closure mechanism indicates that two individuals with mutual friends have a higher probability to establish a link (Romero and Kleinberg, 2010). Totally, five possible network mechanisms are included in this study.

The impacts on scholars' perceptions and behaviours.

Regardless of the mechanism that scholars are influenced by research community, this study attempts to examine whether these mechanisms have impact on scholars' perceptions and behaviours. Based on the research trends by prior studies, this study totally concludes four related research constructs, including willingness of *research collaboration*, *perceived research quality*, *citation intention*, and *citizenship behaviours toward the benefits of other individuals*. First of all, the nature of collaboration vary from one discipline to another, *research collaboration*, including its exogenous variables and outcome performance, has always been an interesting topics in bibliometrics. Hence, this study further investigates whether the network mechanisms stimulate the willingness of research collaboration.

Before the research collaboration, the quality of articles and citation intention are important reference indicators for the collaborators. In this study, *perceived research quality* is defined as the quality of research a person perceived from the works of target scholar, while *citation intention* is treated as the intention a person cite the works from target scholar. This study proposes both indicators might be affected by network mechanisms.

Performing citizenship behaviour or extra-role behaviour (e.g., voluntary behaviour) has been proven to be highly associated with effective user participation and success of an online community (Chiu et al., 2015), this study also includes community citizenship behaviour toward the individuals (CCBI) as the surrogate of extra-role behaviour in research community.

Methodology

As shown in Figure 1, there are nine constructs in the research model. Except the CCBI, the questionnaire items of other eight constructs are developed based on the definitions at the literature. More specifically, four items of CCBI are adopted from the measurement by Chiu et al. (2015), and each of other constructs is developed and measured by 2 to 4 items. Totally, there are 30 items for all constructs, and each item contains 6 anchor points, ranging from 1 "strongly disagree" to 6 "strongly agree". For instance, the respondents were asked to answer "when joining an academic community, I will first find friends who are familiar with everyone." for the measurement of 'familiarity' construct.

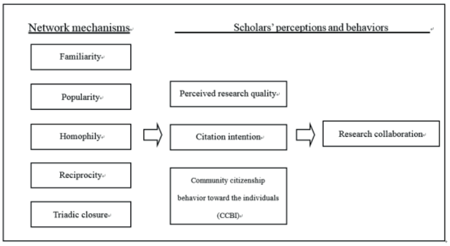

Network mechanisms
Scholars' perceptions and behaviors
Familiarity
Popularity
Homophily
Reciprocity
Triadic closure
Perceived research quality
Citation intention
Community citizenship behavior toward the individuals (CCBI)
Research collaboration

Figure 1. Research framework

Data source

Due to the research background, the sample collection target are faculty who have experiences in research community, no matter physical or virtual community. This study adopts online survey to collect data via invitation email. According to statistical requirements, this study has to collect more than five times of the questionnaire items (i.e., at least 30 items X 5 respondents) as sample size (Hair et al., 2006). But it was difficult to collect faculty samples, this study conducted a pilot test and only collected 73 respondents from April, 2018 to the end of 2018. There are only 34 valid respondents who have participated in research community. This study is currently in the process of data collection and will collect more data in the recent months.

Acknowledgments

This study was sponsored by the Ministry of Science and Technology, Taiwan (MOST-106-2410-H-030-018-MY3).

References

Barabási, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.

Chiu, C.M., Fang, Y.H., & Wang, E.T.G. (2015). Building community citizenship behaviours: The relative role of attachment and satisfaction. *Journal of the Association for Information Systems*, 16(11), 947.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, 201-233.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis (6th edition)*. Upper Saddle River, NJ: Pearson.

Johnson, S.L., Faraj, S., & Kudaravalli, S. (2014). Emergence of power laws in online communities: the role of social mechanisms and preferential attachment. *MIS Quarterly*, 38(3), 795-808.

Omidi, S., & Masoudi-Nejad, A. (2010). Network evolution: theory and mechanisms. *Computational Social Network Analysis*, Springer, London, 191-240.

Ozmen, O., Yilmaz, L., Smith, J., & Smith, A. E. (2012). A complex adaptive model of information foraging and preferential attachment dynamics in global participatory science. Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), *2012 IEEE International Multi-Disciplinary Conference*, 65-72.

Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M., & Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature*, 489(7417), 537.

Romero, D.M., & Kleinberg, J.M. (2010). The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. *the Fourth International Conference on Weblogs and Social Media (ICWSM)*.

Spinellis, D., & Louridas, P. (2008). The collaborative organization of knowledge. *Communications of the ACM*, 51(8), 68-73.

A Scientometric Analysis of the R&D Trends and National Research Activities in Organoid

Eunsoo Sohn[1] and Kyung-Ran Noh[2]

[1] *essohn@kisti.re.kr*
Future Technology Analysis Center, Korea Institute of Science and Technology Information, Seoul 02456 (Korea)

[2] *infor@kisti.re.kr*
R&D Investment Analysis Center, Korea Institute of Science and Technology Information, Seoul 02456 (Korea)

Introduction

Organoids can be defined as 3D constructs comprising tissue-specific cells with the invention of recapitulating the cellular microenvironment; organoids may also include ECM (extracellular matrix) components or biomaterials to achieve this aim (Lancaster MA et al., 2014). Organoids have been considered to be complex 3D structures that develop from stem cells or organ-specific progenitors through a self-organization process (Clevers H, 2016). Over the years, the term organoid has been used to define types of in vitro cultures, from tissue explants to organs-on-chips and human-on-chips (Zhang Y S et al, 2017).

Recent rapid advances in biomedical and tissue engineering technologies have driven widespread use of 3D cell culture platforms and organoids in various research fields including basic biological research, drug discovery and regenerative medicine.

Although scientometrics have provided researchers with valuable information and insights into hot topics, emerging trends, and the knowledge landscape in innovative technologies over the past few decades, there were no scientometric studies on organoid research to date.

The purpose of this study is to identify the global R&D trends and the national activities in this field from the collective knowledge of thousands of scientific publications using scientometric methods approach.

Data and Methods

As scientific publications on organoid research, 3207 articles and 2608 articles were retrieved and identified from the Scopus (1996~2019) and Web of Science (1986~2019) databases respectively. The query to collect the data for bibliometric analysis was as follows: TS=(organoid* OR tumoroid*)

The KnowledgeMetrix Plus and i*Metrics software developed at KISTI (Korea Institute of Science and Technology Information) were used for data processing and calculating various scientometric indicators. The VOSviewer of CWTS (Centre for Science and Technology Studies) was also used for science mapping and clustering.

Results and Discussion

Figure 1 shows R&D trends over time in major countries and CAGR (compound annual growth rate) of total datasets with three time intervals based on scientific publications regarding organoid. Over the past several decades there has been a significant growth in organoid research output with CAGR over 10%. In particular, in the last five years it has grown at a whopping 44%. This reflects that organoid research began to develop rapidly in the early 2010s and as an example organoid was selected as the most advanced scientific achievement in the magazine, *TheScientist* by 2013 (Kerry Grens, 2013).

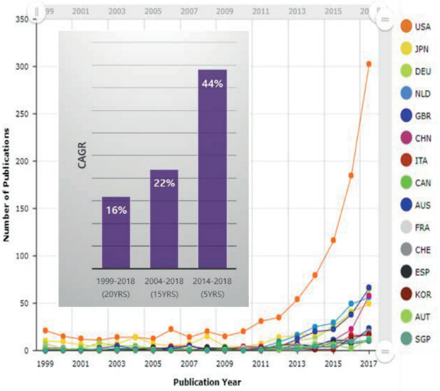


Figure 1. Annual publication output per country and CAGR in organoid research field

Figure 2 provides that several countries are showing strong development in organoid in many ways. As shown in Figure 1 and 2, the United States is the largest contributor to organoid research based on number of papers, but Singapore, the Netherlands and China in terms of growth rates, and the Netherlands, Austria and Canada in terms of MNCS (Mean Normalized Citation Score) which is scientometric indicator reflects the influence and excellence of research activities have made relatively high contributions to organoid research.

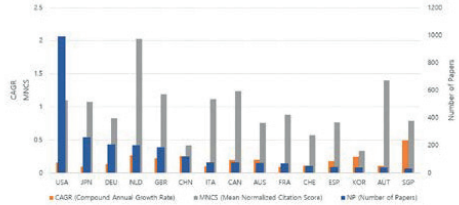


Figure 2. Comparison of scientometric indicator values for organoid research in major countries

To identify the technology convergence and the level of technological absorption and diffusion between organoid research and other research fields we adopted two indicators, DRO (Diffusion Rate to Other Fields) and ARO (Absorption Rate from Other Fields. As shown in Figure 3 (The node size represents number of publications.), the United States with the highest productivity in organoid research is located in the middle of the four quadrants, while the Netherlands and Austria which have influential research activities, have a relatively high DRO, and the technology traffickers China and Korea have relatively high ARO.

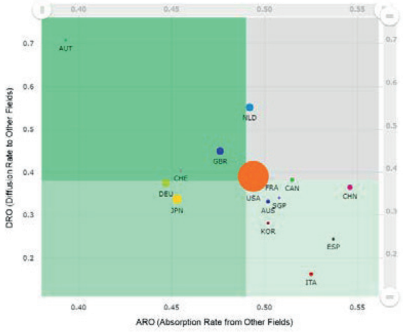


Figure 3. Indicator analysis for technology convergence and diffusion in organoid research

The influential research activities in the Netherlands and Austria can be also seen in Figure 4, showing higher AI (Activity Index) and AAI (Attractivity Index) indicator in the early 2010s when organoid research began to become active.

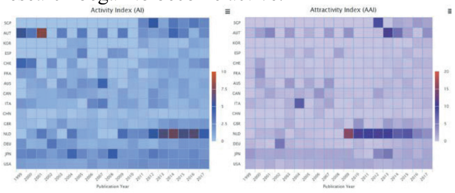


Figure 4. Heat maps with activity index (left) and attractivity index (right) indicators by country in organoid research

The knowledge map of keyword-network from co-word analysis is shown in Figure 5. Organoid has a wide range of applications such as disease modelling, drug discovery, regenerative medicine, precision medicine, and so on. Recently, studies on organogenesis of kidney, brain, and nerve as well as studies of intestines, liver, and cancer have been actively performed as shown in the color bar in the lower right corner of the visualization, which determined by the scores, the publication year in this case. Additionally, it can be seen that organs-on-chips and human-on-chips, organoid platform for more sophisticated organogenesis, bioprinting for personalized regenerative medicine have also been developed.

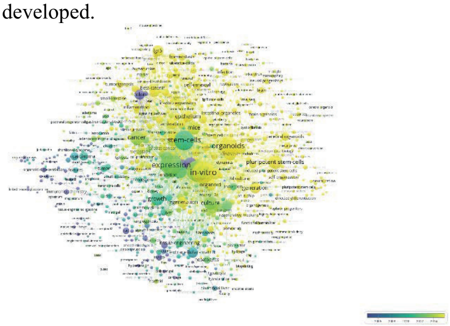


Figure 5. The overlay visualization of co-word map in organoid research field

In this study, we analysed global R&D trends and the national research activities in organoid field based on scientometric indicators and network analysis. Despite many obstacles and limitations, over the last decade, innovative organoid research has shown much progress around the world and has revealed that it is moving toward optimization studies for commercialization and practical use through the research performance analysis.

Acknowledgments

This research was supported by Korea Institute of Science and Technology Information (KISTI) under the program "Development of Future Technology Analysis System Based on Open Data" from the Ministry of Science and ICT in Korea.

References

Lancaster, M.A. & Knoblich, J.A. (2014). Organogenesis in a dish: modeling development and disease using organoid technologies. *Science*, 345(6194):1247125.

Clevers, H. (2016). Modelling development and disease with organoids. *Cell,* 165, 1586-1597.

Zhang, Y.S. et al. (2017). Multisensor-integrated organs-on-chips platform for automated and continual in situ monitoring of organoid behaviors. *PNAS Early Edition*. 114(12) E2293-E2302.

Kerry, G. (2013). 2013's Big Advances in Science. *TheScientist.* December 24, 2013.

Science at the Vatican

Ronald Rousseau[1,2]

[1]*ronald.rousseau@kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, Leuven, 3000 Leuven (Belgium)

[2]*ronald.rousseau@uantwerpen.be*
Faculty of Social Sciences, University of Antwerp, 2020 Antwerp (Belgium)

Introduction: The relation between the Vatican City State and the Holy See

The Vatican City State is an independent city-state enclaved within Rome, established in 1929 by the Lateran Treaty between the Holy See and Italy. The Vatican City is ruled by the Pope. The Holy See dates back to early Christianity, and is the primate episcopal see of the Catholic Church. The Holy See is the episcopal see of the Pope, and a sovereign entity of international law. Although the Holy See is closely associated with the Vatican City, the independent territory over which the Holy See is sovereign, the two entities are separate and distinct.

Scientific institutes in the Vatican

Historically the relation between the sciences and the Vatican has often been stressed: the Galilei case being the best known. Also nowadays the relation between what reproductive medicine can and wants to do and the official point of view of the Vatican do not correspond.

Yet, since the 19th century the Vatican has actively contributed to the natural sciences. Nowadays, the Vatican has several universities, academies and other institutes of higher learning. Besides universities, the Vatican has more than ten Academies and several Pontifical Institutes. Probably best known among these Academies is the Pontifical Academy of Sciences. Its goal is the promotion of the progress of the mathematical, physical, and natural sciences, and the study of related issues. As the Academy and its membership is not influenced by factors of a national, political, or religious character it represents a valuable source of information which is made available to the Holy See and to the international community. More than 45 Nobel Prize winners have been a member. The "father of the Big Bang", my country man Monseigneur Georges Lemaître was its president for a period of time. Besides the Pontifical Academy of Sciences there is also a Pontifical Academy of the Social Sciences.

Arts, humanities, social sciences and the Vatican

Considering the study of religion as a part of the humanities, it is clear that the Vatican is a huge contributor of original documents. In recent centuries some of these documents have had far-reaching social implications (Rerum Novarum; Humanae Vitae; Laudato Si). Moreover, cultural sites such as St. Peter's Basilica, the Sistine Chapel and the Vatican Museums play a leading role in the history of art.

A bibliometric study

We searched for "Vatican*" as a country (CU=Vatican*) or as a city (CI=Vatican*) in the Web of Science (WoS). To the results of this query we added documents with (Vatican* OR "Pontifical acad*) in the address field and Rome as city. Excluding documents dated 2019 this search yielded 733 documents. This is the primary data source (referred to further on as the Vatican Documents) we investigated. As a secondary data set we also collected the articles from the Vatican Observatory in Arizona (USA). This set contained 263 documents, with an overlap of 63 with the Vatican Documents. Although the large majority of these documents were related to astronomy, some dealt with religion, social issues and the history of the philosophy of science.

From now on we focus on the Vatican Documents, starting with the basic fact that they include 507 (normal) articles, 65 proceedings papers and 45 documents considered editorial material. There are further 41 meeting abstracts, 37 book reviews, 24 review papers, 17 letters and a few other items. The publication data are distributed as shown in Fig. 1.

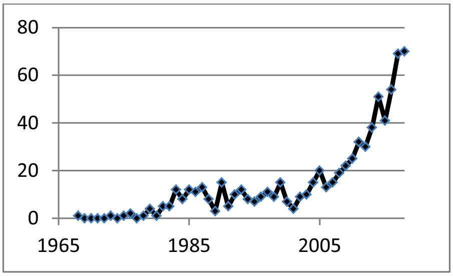


Figure 1. Yearly number of publications in the set of Vatican Documents

The h-index of the set is (only) 54. The large majority of these documents is written in English (659), followed by Italian (21 documents); none is written in Latin. As to the research areas covered (see Table 1), it is clear that Vatican scientists (at least as covered by the WoS) are mostly interested in astronomy and related physics, followed – at a large distance – by religion as a research area and medicine. We also mention philosophy, history and different aspects of arts. The interest in astronomy is also evident from the most used journals (not shown). Checking the institutes to which scientists contributing to the Vatican Documents belong we see (Table 2) that astronomical observatories are high on the list.

Table 1. Top research areas (WoS) covered by the Vatican Documents.

Area	Number of documents
ASTRONOMY ASTROPHYSICS	319
PHYSICS	72
RELIGION	68
GEOCHEMISTRY GEOPHYSICS	45
IMMUNOLOGY	35
ALLERGY	34

Table 2. Top Vatican institutes and those of collaborating scientists.

Institute	Number of documents
VATICAN OBSERV(ATORY)	311
PONTIFICIA UNIV. CATOLICA DE CHILE	139
SPECOLA VATICANA	114
UNIVERSIDAD ANDRES BELLO	100
EUROPEAN SOUTHERN OBSERV.	74
UNIVERSITY OF PADUA	63
BAMBINO GESU	52

The term "Specola Vaticana" refers to the astronomical observatory near Castel Gandolfo in Rome. Because of the light pollution in and near Rome, the Vatican Observatory established the Vatican Observatory Research Group, with offices at the Steward Observatory of the University of Arizona in Tucson. Headquarters remained in Rome. Hence the terms "VATICAN OBSERV", SPECOLA VATICANA and VATICAN ASTRON OBSERV, essentially refer to the same scientific institute. Recall that to those publications with a "Vatican" address we should add the 200 other ones with only an address in Arizona. Vatican scientists, moreover, make use of the European Observatory in Chile. Finally, we mention the Bambino Gesu Hospital in Vatican City where scientists in medicine (allergy, immunology) perform their research.

Scientists in these Vatican Documents belong to the following countries: Vatican (539), Italy (309), USA (218), Chile (180), and many others. Clearly many scientists write their address as Vatican City, Italy, explaining why not all documents have the Vatican City-state as country.

Among the Vatican Documents we found a single-author paper with 11 reprint addresses (Zichichi, 2017). This case provides an example of "address inflation". Finally, we mention a special item (classified as editorial material) co-authored by Pope Francis and Patriach Kiril (2016).

Discussion and Conclusion

It is well-known that WoS (and Scopus), as international databases, have a good representation of the natural sciences and medicine, yet they do not reflect the real situation in the humanities. As such the contribution of the Vatican in medicine, astronomy and related fields is probably correctly represented in this study. However, this statement does not hold for the humanities. Most pontifical academies have their own series of acta, scripta, studia selecta, etc., but these are not included in the WoS (or Scopus). As the Vatican scientific institutes have a double purpose, namely providing a means for a dialogue between science (including the humanities) and the Church, on the one hand, and acting as sources of objective scientific information for the Pope, on the other, it seems to us that more could be done to reach this purpose. Nowadays, many countries develop current research information systems which include all peer-reviewed publications in which a resident of one of the country's scientific institutes has contributed. This might be a way to make Vatican scientific results comprehensively visible.

Although the Vatican State cannot be considered a major force in science, it does have its own niche and plays a preeminent role in the field of astronomy. Adding to this its unique position in religion and arts makes a bibliometric study of this small state quite fascinating.

References

Francis, Pope & Kirill, Patriarch (2016). The Divided Mind of the Black Church: Theology, Piety, and Public Witness. Ecumenical Review, 68(1), 139-146.

Zichichi, A. (2017). Abdus Salam, the electroweak forces, ICTP and beyond. International Journal of Modern Physics A, 32(8), Article Number: 1741004

recerTIC UPC: a new approach to bibliometric analysis for a research university

Rubèn Pocull[1], Miquel Codina-Vila[1], Ruth Inigo[1], Sara Matheu[1], Andrés Pérez[1], Javier Clavero[2]

[1]*ruben. pocull@upc. edu; miquel. codina@upc. edu; ruth. inigo@upc. edu; sara. matheu@upc. edu; andres. perez@upc. edu*
Polytechnic University of Catalonia. Rector Gabriel Ferraté Library, C / Jordi Girona 1-3, 08034, Barcelona (Spain)

[2]*javier. clavero@upc. edu*
Polytechnic University of Catalonia. Libraries, Publications and Archives Service, C / Jordi Girona 1-3, 08034, Barcelona (Spain)

Introduction

The Biblioteca Rector Gabriel Ferraté (Rector Gabriel Ferraté Library, BRGF) of the Universitat Politècnica de Catalunya (Polytechnic University of Catalonia, UPC) presents *recerTIC UPC* <https://tinyurl.com/y4os625c>: a set of ten bibliometric works on the same number of subjects, which is designed to generate a meaningful and easily understandable map of the UPC's scientific publications on topics relevant to the field of Information and Communications Technologies (ICT). These studies place emphasis on showing the links between internal and external collaborations of researchers who author these publications, as well as the dynamics of transversality between the subject areas of the analyzed research publications.

Methodology

The studies are based on journal articles and conference publications published by UPC professors between 2007 and 2017.

The process of creating *recerTIC UPC* began with the directors of ICT learning centers petitioning for bibliometric studies on specific subject areas that had previously been left out of bibliometric studies prepared by the BRGF.

These subjects were determined in close collaboration with UPC researchers specialized in the corresponding areas. Consideration was given to aspects such as the relevance and emergence of each technology, its social impact, its future potential, and its strategic relevance for the UPC. Ultimately, the areas chosen were: 5G, Computer Security, Embedded Systems, Machine Learning, Smart Sensors, Bioinformatics, Data Science and Engineering, IoT, Robotics and Vehicle-to-Everything.

The database chosen as the primary source of information was Web of Science Core Collection (WoS), by Clarivate Analytics. In choosing this, the definition, scope and level of granularity for the WoS *Subject Categories* were taken into account. This is appropriate for the subsequent analysis not only of the relationships between subject nodes, but also of the inclusion of keywords and subject descriptors in most registries.

The references of the WoS publications were extracted by using non-controlled vocabulary in the WoS field *TS=Topic*. This field tag retrieves terms within the fields *Title, Abstract, Author Keywords* and *Keywords Plus®*.

Choosing keywords was the most critical part for obtaining reliable results. Their selection relied on the collaboration of relevant researchers who identified concepts and terms that define the contents of each area of study. These terms may be of a generic scope and designate various disciplines, techniques and methodologies that are closely related to the research field being considered.

Paradigmatically for the fields that form the theoretical basis of many *recerTIC UPC* subjects (for example, Statistics, Mathematics and Physics), the difficulty lies in recovering publications relevant to the study subject when this subject is often tacit. In order to minimize this problem, we have attempted to ensure that the terms related to the most basic and methodological aspects of each study do not distort or add noise to the results.

However, it was necessary on various occasions to iterate the algorithms so that they would introduce concepts that were not included in the first approaches, which was done in order to obtain balanced and representative results.

Generating maps with GenMap

The interactive node maps were created with GenMap, an application that was custom-designed for the ITC staff at the Libraries, Publications and Archives Service. We used the library cytoscape.js, which specializes in representing graphs of data, nodes and edges – all of which are extracted from various information sources (currently: WoS, Scopus, and DBLP, among others) – with the aim of facilitating visualization while making it more interactive.

Once the map is created, the next step is to import the extracted file from the database corresponding to the information we want to view. Parameterizing the concepts for representation allows generating a graph in which the nodes indicate the analyzed concept, its identifying title, the size according to a chosen criterion and a color representing the characteristic intended to be emphasized. Edges connect the nodes and represent the relationships between them.

Once the data entry process has been completed, the visualization module allows all users to view the knowledge map and interact with it.

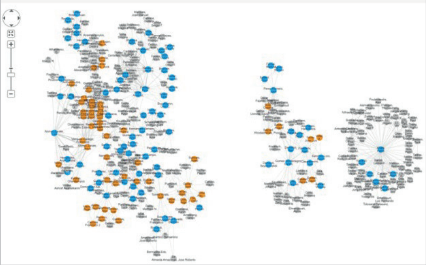


Figure 1. Interactive node map created with GenMap

The user can adjust the visualization by means of the following: zoom; moving the nodes; searching nodes on the map with a drop-down menu; selecting and highlighting nodes and/or edges; viewing node and edge information; modifying the map's graphical settings; exporting the map in various formats (with the possibility of inserting it into a web page); and, finally, sharing the map via social networks.

Results

The results obtained by the *recerTIC UPC* studies show data and significant information regarding: the location and analysis of clusters and communities of collaboration and co-authorship; relationships between subject areas; and other aspects pertaining to the evolution of UPC scientific production in the areas analyzed.

The co-author maps show different clusters of authors in each subject area. Diverse clusters exist because various research groups take different approaches to working in this area in separate stages of development. Information is also collected on production that forms a part of the methodology or the practical application of the research in another field (for example, machine learning applied to fish farm management or 5G applied to traffic management).

Distinguishing the nodes by color reveals several dynamics regarding internal co-authorship, co-authorship with related research centers, and co-authorship that is external to the UPC.

Regarding the node maps corresponding to the WoS categories assigned to the publications, these indicate the weight of each category and how they relate to each other within the framework of the subject area studied.

Conclusions

The aim of *recerTIC UPC* is to analyze and highlight the power of our university's scientific production in certain areas covering the most relevant research on current technology. To achieve this, we generated a "satellite image" of research at the UPC, which can be browsed as a set and then "downloaded" in order to discover details and new niches for investigation, as well as for applying and/or transferring the research.

In addition, *recerTIC UPC* constitutes a step forward in our work with bibliometric data. Some of the characteristics that define this new generation of our bibliometric productions are: the use of new technologies; a presentation that facilitates analyzing results; and options for interacting with the data that we have processed through our searches.

References

Codina-Vila, M., & Íñigo, R. (2015). De la investigación al investigador. Adaptando servicios en la Biblioteca Rector Gabriel Ferraté. *El profesional de la información, 24* (5), 648-655. doi:10.3145/epi.2015.sep.13

Falagas, M. E., Pitsouni, E. I, Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal 22*(2), 338-342. doi:10.1096/fj.07-9492LSF

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In Henk F. Moed, Wolfgang Glänzel, & Ulrich Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 257-276). Dordrecht: Springer.

Jansen, B. J., & Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology 61*(8), 1517-1534. doi:10.1002/asi.21358

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science 34* (4), 605-621. doi:10.1177/0165551507087238

Scientific collaboration in institutes of chemical engineering in Taiwan during the declining research manpower

Tung-Wen Cheng[1] and Yu-Wei Chang[2]

[1] *twcheng@mail.tku.edu.tw*
Tamkang University, New Taipei City 25137 (Taiwan)

[2] *yuweichang2013@ntu.edu.tw*
National Taiwan University, Taipei 10617 (Taiwan)

Introduction

Natural sciences research relies on more scientific collaboration than does social sciences and humanities research based on the prevalence of co-authored articles. The culture of scientific collaboration in natural sciences justifies that graduate students are the essential research resources for faculty to improve their research productivity (Know et al., 2015; Miller, Coble, & Lusk, 2013). With the continuing expansion in the number of universities over the period of 1986-2015 in Taiwan, the increase was observed in the quantity of research articles in numerous fields including the field of chemical engineering.

Responding to changes in education of chemical engineering in Taiwan, Chang and Cheng (2012) tracked the characteristics and trends of research articles authored by researchers affiliated with institutes of chemical engineering (CE researchers) in Taiwan before 2010. Major changes in authorship patterns, types of collaboration, and research interests were identified. However, the continuous decline in the numbers of master and doctorate students has been observed in fields of science and technology since 2011 (Ministry of Education, 2019). This phenomenon is expected to directly affect faculty's research productivity. Moreover, the percentage of education expenditures to GDP has reduced since 2013 (Ministry of Education, 2018). This study aimed to track the research productivity contributed to by CE researchers in Taiwan during the shortage in research manpower and budget. In particularly, collaboration is regarded as a practical approach to face the problem of research resource shortage. The characteristics and trends of research articles during 2011-2017 were focused on. Research focuses include changes in the numbers of graduate student, faculty, and articles, and in authorship pattern, types of collaboration, and research interests before and after 2010.

Methodology

Data collection

Articles by CE researchers in universities in Taiwan were retrieved from the Web of Science (WOS) database, which is a multidisciplinary citation index database with author affiliation information. Articles indexed by WOS are valued in Taiwan. In particularly, nature sciences researchers make efforts to increase WOS articles. Therefore, WOS can help us understand the academic efforts of CE researchers in Taiwan. One author with chemical engineering expertise in the paper identified the institutes of chemical engineering in Taiwan because not all institute names contain keywords "chemical engineering." The list of institutes of chemical engineering assisted us to obtain bibliographic records of articles by CE researchers in Taiwan. The latest year of publication of articles was 2017. The country was limited to "Taiwan," the document type was limited to "Article" referring to research articles. In addition, the abbreviation form of keywords referring to specific institutes such as "chem engn" or "chem & mat engn" or "app chem & mat sci " were included in address data of authors. Some authors provided another addresses outside of Taiwan. When one address provided is in Taiwan, the author was classified as one affiliated with institutes in Taiwan. After excluding 21 disqualified articles using manual examination, a total of 28,801 articles were analyzed.

Data processing

Coauthored articles were divided into articles from domestic collaboration and from international collaboration based on address information. In addition, sample articles were divided by subject category. The source journal of each article indexed by WOS was assigned at least one subject category according to journal categorization system of Journal of Citation Report. If an article was assigned with n subject categories, each subject category was assumed to have 1/n article. The number of articles belonging to a specific subject category was calculated.

Results

Declining research productivity after 2014

Figure 1 shows that the continuing decreasing trend in total number of faculty and graduate students in institutes of chemical engineering in Taiwan during 2014-2017. Meanwhile, the decreasing trend in annual number of articles was observed in the same period. Overall, a similar change appeared in both numbers of research manpower and research productivity. The declining number of graduate students is one of possible factor causing the decreasing research productivity of faculty.

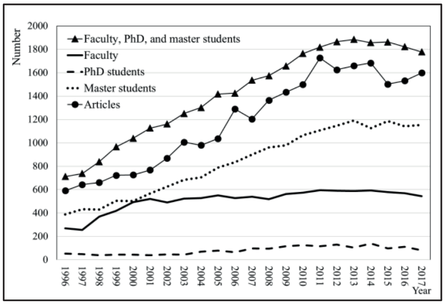


Figure 1. Numbers of faculty, graduate students, and articles by year.

Rapid growth in international collaboration

It was rare for CE researchers to publish a single-authored articles (764, 2.7%). Although two-authored articles accounted for the largest part (24.9% of 28,801 articles), followed by three-authored articles (23.4%), their notable declining trend have appeared over one decade. With the continue increase in the average number of authors per article, the inter-institutional collaboration has replaced the intra-departmental collaboration. Although the percentage of international collaboration articles was expected to be lower than that of domestic collaboration articles due to several barriers to scientific collaboration, a considerable growth was observed after 2010 and reached the peak in 2017 (39.7%). The average number of authors per international-collaboration (IC) article was higher than that per domestic-collaboration (DC) article (5.85 vs. 3.17). Obviously, international collaboration has been highly emphasized since 2010. CE researchers in Taiwan tended to collaborate with more researchers in Taiwan and other countries during the declining domestic research manpower.

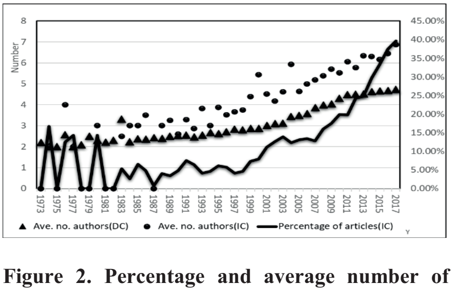


Figure 2. Percentage and average number of authors of articles from international collaboration by year.

Articles distribution by subject category

The 28,801 articles published in 1,652 journal titles belonging to 163 subject categories, indicating the diverse research interests. *Journal of the Taiwan Institute of Chemical Engineers* (JTICE) ranked the top 1 journal with the highest number of articles (4.5% of 28,801 articles), followed by *Journal of Applied Polymer Science* (JAPS) (3.6%). A considerable increasing trend was identified in JTICE, whereas a decreasing trend appeared in JAPS after 2000. Polymer science (16.28% of articles) and chemical engineering (16.05%) were main subject categories. In addition, the number of subject categories continued to increase with year, revealing that CE researchers expanded their research to other fields. With the expansion of subject categories, the percentage of articles in two main subject categories has been considerably decreasing since 2000. An emerging dominate subject category will be expected.

Acknowledgments

The authors are grateful for the financial support provided by the Ministry of Science and Technology of Taiwan (Grant numbers: MOST 107-2622-8-032-002-TE4; MOST 107-2221-E-032-026).

References

Chang, Y. W., & Cheng, T. W.（2012）. Characteristics and trends of research articles authored by researchers affiliated with institute of chemical engineering in Taiwan. *Journal of the Taiwan Institute of Chemical Engineers, 43*(3), 331-338.

Kwon, K.S., Kim, S. H., Park, T.S., Kim, E. K., & Jang, D. (2015). The impact of graduate students on research productivity in Korea. *Journal of Open Innovation: Technology, Market, and Complexity, 1*(21), https://jopeninnovation.springeropen.com/track/pdf/10.1186/s40852-015-0024-6

Miller, J. C., Coble, K. H., & Lusk, J. L. (2013). Evaluating top faculty researchers and the incentives that motivate them. *Scientometrics, 97*(3), 519-533.

Minister of Education (2019). Number of graduate students in universities and colleges by year. Retrieved March 21, 2019 from http://stats.moe.gov.tw/files/important/OVERVIEW_U02.pdf

Ministry of Education (2018). Education Statistics 2018: The Republic of China. Retrieved March 21, 2019 from Ehttp://stats.moe.gov.tw/files/ebook/Education_Statistics/107/107edu.pdf

Construction of Knowledge Map by Co-Citation Analysis: A Case Study on Information Behavior

Ming-Yueh Tsay[1], Yu-Wei Tseng[1], and Chien-Hui Lai[1]

[1] *mytsay@nccu.edu.tw*, *wendy.tseng@gmail.com*, *105155501@nccu.edu.tw*
National Chengchi University, Graduate Institute of Library, Information and Archival Studies
Taipei City 11605 (Taiwan, R.O.C.)

Introduction

Since the paradigm shift from early system-base to information needs, information seeking, information searching, information behavior, and user-centered information use, the research paradigm of information retrieval induces large amount of relevant research in past decades. (In this study, the term information behavior will be used as a broader term covering the aforementioned topics). Relevant empirical research was founded on various theories, and new information behavior theories and models has emerged. Most researches on information behavior theories focused on qualitative method. For instance, Julien et al. (2011) continuously discussed the research characteristics of information behaviors at different periods with content analysis of the relevant literature. Jamali (2013) studied the relationship between information behavior modules and theories with bibliographic coupling. However, bibliographic coupling could merely measure the growth point of subject development trend, while co-citation analysis could understand disciplinary evolution, reflect the semantic relationship among cited literatures, and further grasp the historical dynamic of discipline structure and the link between disciplines or literatures. Accordingly, this study aims to investigate the citation, co-citation as well as the co-citation context analysis in the field of information behavior through informetrics and visualize the relationship by social network analysis tools. It aims to outline the theoretical content, to understand the mutual relationship among researchers on information behavior as well as the citation relationship with other disciplines and further to construct the knowledge context and the knowledge map of this field.

Research method

Research sample selection

In the present study, the references of the prominent book of research on information behavior, *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (hereafter referred to as *Looking for Information*) 4th edition in 2016, co-authored by Donald O. Case and Lisa M. Given, is regarded as the research object. The book maintains the existing characteristics of Best Information Science Book Award of 2003 Association for Information Science and Technology (ASIS&T), is resupplied the successive ideas and theories, expands the chapter of research method, stresses on qualitative research, and deepens context study and relevant issues to enrich the core subject and grasp the research trend in recent research on information behavior.

Data collection and search

The representative multidisciplinary database, Web of Science (WoS), is utilized for retrieving the publication information as well as the cited and co-cited data of the research sample for successive statistical analyses. The data collection and search in this study is preceded with three stages. First, the bibliographic references of *Looking for Information* are categorized by the document type of journal, book, book chapter, and proceedings. Second, the "basic search" function of WoS database, made by Clarivate Analytics, is used for checking the journals in the reference of the book being collected in the database and organizing the number of citation times, disciplinary area, and keyword provided by the database. The number of citation times is ordered for top cited literatures. Third, senior researchers of this study strictly select representative literatures on information behavior from the top cited literatures, i.e. representative studies and the relevant research of important researchers on information behavior, as the samples for co-citation analysis. The paired cited reference search in WoS database is conducted for the literature co-citation data. They are further organized and stated according to the correlation coefficients transformed from co-citation data and the numerical values required for cluster analyses.

Research result: Citation, co-citation analysis and the knowledge map

Highly cited literatures

Eight hundred and twenty five articles in *Looking for Information* that collected in WoS database were searched in February, 2017 for citation data. The results show, total 117 articles, merely one-tenth of total journals (14.2%), are cited up to a hundred times, as highly cited literatures. Eight literatures being cited for more than a thousand times, covered in the fields of sociology, psychology, business and economics, general medicine, and operation research and management science, which are the classical articles in the research fields.

Representative literatures on information behavior studies and the co-citation analysis

Aiming at 117 highly cited literatures, researchers on information behavior check the literatures one by one to pick out 53 representative ones as the co-citation analysis samples. The 53 representative literatures on information behavior are paired for total 1,378 sets of data. The cited reference search in WoS database is used for confirming the co-citation time to further make a symmetric group pair matrix; the search date is June 22-July 5, 2017. The co-citation value of the 1,378 sets of paired literatures appears in 0-139 times, where 323 sets of paired literatures show zero co-citation, about a quarter of total paired sets (23.4%). The proportion is rather high, revealing extremely high segmentation of important literatures. Three paired sets show more than a hundred co-citation times. The first set, Kuhlthau (1991) and Wilson, T. D. (1999), presents up to 139 times; the second set, Belkin, Oddy & Brooks (1982) and Taylor (1968), shows 111 times; and, the third set, Kuhlthau (1991) and Ellis (1989), reveal 104 times.

To analyze the correlation among literatures, SPSS is applied in this study to transform the original value from co-citation into Pearson product-moment correlation coefficient to calculate the standardized data. Hierarchical cluster analysis (with inter-group link) is further used for distinguishing various relevant clusters of representative literatures on information behavior. The classification result reveals that the 53 representative literatures could be divided into three major clusters. As the example of cluster 1, 22 literatures are covered, numbered 1, 3, 9, 7, 2, 4, 6, 5, 8, 19, 21, 23, 15, 11, 29, 28, 38, 27, 32, 16, 17, and 30.

Social network of representative literatures on information behavior

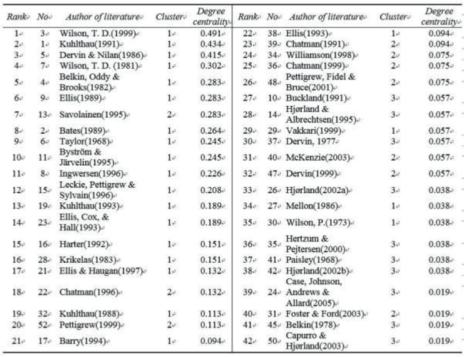
The UCINET is further utilized for drawing the social network map to explain the strength of co-cited literatures which could not be presented on hierarchical cluster analysis and multidimensional scaling analysis. As shown in Table 1, literatures of Wilson, T. D. (numbered 3), Kuhlthau (1), and Dervin & Nilan (5) in cluster 1 appear in the core of co-citation network and show tight citation relationship with other representative literatures on information behavior, while Savolainen (13) is the most influential literature in cluster 2. To observe the evolution of the knowledge map of the representative literatures, the search period, 1968-February 2017, is divided into 6 phases, with 10 years as an interval for further distinguishing the degree of core.

Conclusion and discussion

The results show that Dervin & Nilan (1986), Bate (1989), and Wilson, T. D. (1999) keep on top three since the publication. Dervin & Nilan (1986) is the milestone to change system-oriented research paradigm into user-oriented. Bates (1989) describes the process of people searching for information as picking strawberries, and the suggestions are broadly applied to the design of information systems. Wilson, T. D. (1999) reviews important information behavior theories and models in past years and interprets the content and coverage of research on information behavior. Such three become the classical literatures for research on information behavior. Comprehensively reviewing the evolution of literatures with top ten degree centrality after 1980s, the representative literatures on information behavior are classified into following categories.

1. Being stably in the core: As mentioned above, Dervin & Nilan (1986) remains in top three.
2. Important ideas or model: Taylor (1968), Wilson, T. D. (1981), Kuhlthau (1991, 1993), and Ellis (1989) keep in top ten.
3. Gradully retreating to the secondary core: Belkin, Oddy & Brooks (1982) is declining in past years.
4. Fluctuated concerns: Paisley (1968), Buckland (1991), and Ingwersen (1996) attract the attention when being published, are still for some time, and then are concerned again.
5. Other literatrues are in top ten in the time of publication or the next decade, but retreat to the secondary core afterwards, e.g. Dervin (1977), Belkin (1978), Harter (1992), Savolainen (1995), and Leckie, Pettigrew, & Sylvain (1996).

Table 1. Degree centrality of literature with more than 20 co-citation times.



Rank	*No*	*Author of literature*	*Cluster*	*Degree centrality*	*Rank*	*No*	*Author of literature*	*Cluster*	*Degree centrality*
1	3	Wilson, T. D.(1999)	1	0.491	22	38	Ellis(1993)	1	0.094
2	1	Kuhlthau(1991)	1	0.434	23	39	Chatman(1991)	2	0.094
3	5	Dervin & Nilan(1986)	1	0.415	24	34	Williamson(1998)	2	0.075
4	7	Wilson, T. D. (1981)	1	0.302	25	36	Chatman(1999)	2	0.075
5	4	Belkin, Oddy & Brooks(1982)	1	0.283	26	48	Pettigrew, Fidel & Bruce(2001)	2	0.075
6	9	Ellis(1989)	1	0.283	27	10	Buckland(1991)	3	0.057
7	13	Savolainen(1995)	2	0.283	28	14	Hjørland & Albrechtsen(1995)	3	0.057
8	2	Bates(1989)	1	0.264	29	29	Vakkari(1999)	1	0.057
9	6	Taylor(1968)	1	0.245	30	37	Dervin, 1977	3	0.057
10	11	Byström & Järvelin(1995)	1	0.245	31	40	McKenzie(2003)	2	0.057
11	8	Ingwersen(1996)	1	0.226	32	47	Dervin(1999)	2	0.057
12	15	Leckie, Pettigrew & Sylvain(1996)	1	0.208	33	26	Hjørland(2002a)	3	0.038
13	19	Kuhlthau(1993)	1	0.189	34	27	Mellon(1986)	1	0.038
14	23	Ellis, Cox, & Hall(1993)	1	0.189	35	30	Wilson, P.(1973)	1	0.038
15	16	Harter(1992)	1	0.151	36	35	Hertzum & Pejtersen(2000)	3	0.038
16	28	Krikelas(1983)	1	0.151	37	41	Paisley(1968)	3	0.038
17	21	Ellis & Haugan(1997)	1	0.132	38	42	Hjørland(2002b)	3	0.038
18	22	Chatman(1996)	2	0.132	39	24	Case, Johnson, Andrews & Allard(2005)	3	0.019
19	32	Kuhlthau(1988)	1	0.113	40	31	Foster & Ford(2003)	2	0.019
20	52	Pettigrew(1999)	2	0.113	41	45	Belkin(1978)	3	0.019
21	17	Barry(1994)	1	0.094	42	50	Capurro & Hjørland(2003)	3	0.019

Acknowledgments

This work was supported by grant MOST 104-2410-H-004-175-MY2 from the Ministry of Science and Technology, Taiwan, R.O.C.

References

Jamali, H. R. (2013). Citation relations of theories of human information behaviour. *Webology*, 10(1), 1-16.

Julien, H., Pecoskie, J., & Reed, K. (2011). Trends in information behavior research, 1999-2008: A content analysis. *Library & Information Science Research*, 33(1), 19-34.

Improve the Reliability of Short Term Citation Impact Indicators by Taking into Account the Correlation between Short and Long Term Citation Impact

Xing Wang[1] and Zhihui Zhang[2]

[1] *wangxing830914@gmail.com*
Shanxi University of Finance & Economics, School of Information, 696 Wucheng Road, 030006 Taiyuan (China)

[2]*zhangzh0501@yeah.net*
CNKI, Dongsheng Science Park, 66 Xixiaokou Ave, 100192 Beijing (China)

Introduction

The normalized citation indicators are not free of limitations. One concern is the reliability of normalized indicators when a short citation time window is used. The normalized citation indicator may not be reliable enough when a short citation time window (e.g. 2 years) is used, since the citations of papers published recently are not as reliable as citations of papers published many years ago. In a limited time of period, the publications usually don't have enough time to accumulate their citations to reach a stable state (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). In other words, the reliability of citations is related to the length of the citation time window. However, normalization methods themselves can't solve this problem (Wang, 2013).

To solve this problem mentioned above, we introduced a weighting factor to the traditional mean-based field normalized citation indicators (i.e. CNCI). The weighting factor reflects the reliability of the citation window and the reliable degree of the normalized indicator. Taking 500 universities as a sample, we made a comparison of the performance and ranking of the universities before and after introducing the weighting factor in the indicator CNCI.

Materials and Methods

Determination of weight

According to the results of Wang (2013), the reliability factor can be calculated as the correlation coefficient between the citation counts in the short time window and that in a fixed reliable long citation time window (e.g. 31 years). For example, if the correlation coefficient between the citation counts of papers published 2 years later and 31 years later in Chemistry is 0.55, the normalized citation count of a Chemistry paper with a citation window of 2 years should be multiplied by 0.55 to obtain its reliable scientific impact. The shorter (longer) the time window after publication is, the lower (higher) the correlation coefficient and degree of reliability is.

Wang (2013) calculated the correlation coefficients between the citation counts in each time window of 1-10 years and those in the citation time window of 31 years in 14 subjects. We first plotted Figure 1 according to these correlation coefficients.

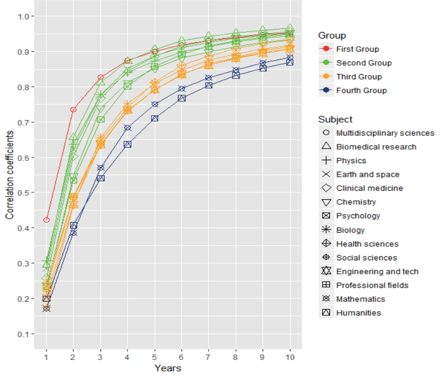


Figure 1. The correlation coefficients between the citation counts in each time window of 1-10 years and those in citation time window of 31 years in 14 subjects.

In order to simplify the calculation of our weightings later, we classify the 14 subjects into four groups according to their shapes, where subjects whose curves are close to each other are classified into the same group. The subjects in the same group share the same weighting factors. The weighting factor of a group for a specific time window was calculated by averaging the correlation coefficients of all subjects in the same group for the given time window. Next, we mapped the 226 WOS subject categories to the 14 subjects according to the description of Wang (2013). Once a paper is identified by its publication year and WOS category, the weight of the paper with the given citation window is then determined.

Data collection

We downloaded the information of papers published by top 500 universities in Shanghai Ranking 2017 between 2007 and 2016 from InCites. The information includes the CNCI value (with citation time windows of 1-10 years), the publication year, and the WOS subject categories of each paper.

Weighted and unweighted CNCI

The CNCI value of a paper is defined as *c*/*e*, where *c* is the raw citation count of the paper and *e* is the expected citation count of the paper. The CNCI value of a university is defined as the average CNCI values of all papers authored by the university, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{e_i}$$

where *n* is the total number of papers published by the university. The indicator is called WCNCI in this paper when the weighting factor is introduced. Similarly, the WCNCI value of a university is defined as follow, where w_i is the weighting factor of the paper *i*:

$$\frac{1}{n}\sum_{i=1}^{n}w_i\frac{c_i}{e_i}$$

Results

Figure 2 shows the scatter plots of universities' WCNCI scores against CNCI scores, and Figure 3 shows the comparison of rankings under these two indicators. The results showed that there was a strong positive correlation between the WCNCI and CNCI scores, where the correlation coefficient was 0.987 (*p=0.000*); there was also a strong positive correlation between the rankings under WCNCI and CNCI, where the correlation coefficient was 0.985 (*p=0.000*).

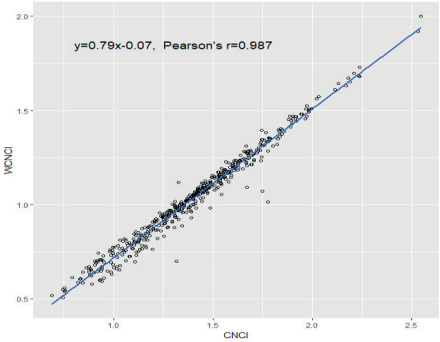


Figure 2. Correlation of CNCI and WCNCI scores.

Although Figures 2 and 3 show strong correlations, some universities' rankings changed substantially after the weighting factor is introduced. Table 1 shows some universities rising most or dropping most after the weighting factor is introduced. (A complete list of the changes in rankings for all the 500 universities is not given due to the limited length of this paper.)

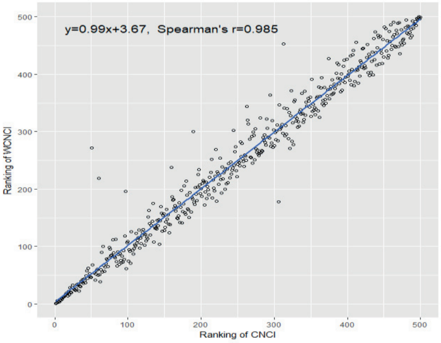


Figure 3. Correlation of the university rankings under CNCI and WCNCI.

Table 1. Some universities rising or dropping most in the ranking after introducing the weighting factor.

University	Ranking changes
Universite de Bordeaux	+128
University of Regensburg	+50
UT Medical Branch at Galveston	+48
Saint Louis University	+42
Hebrew University of Jerusalem	+41
King Abdulaziz University	-221
King Abdullah Univ of Sci &Tech	-159
Univ of Paris Dauphine - Paris IX	-140
Aalborg University	-111
Ulsan Natl Inst of Sci &Tech	-99
Western Sydney University	-81
University of Adelaide	-79
University of Technology Sydney	-70

Conclusion

The results showed that although there is a strong positive correlation before and after introducing the weighting factor, some universities' performance and ranking changed dramatically. This demonstrates that the weighting factor which reflects the reliability of citation impact indicators is essential and should not be ignored in the actual research evaluation practices.

References

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics, 87*(3), 467-481.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851-872.

An Analysis of the Relative Citation Ratio in NIH-Funded Articles

Christopher W. Belter[1]

[1] *christopher.belter@nih.gov*
NIH Library, Office of Research Services, NIH, Bldg 10, Rm 1L09G, MSC 1150, Bethesda, MD 20892 (USA)

Introduction

The Relative Citation Ratio (RCR) is a new field-normalized citation impact indicator developed by analysts at the NIH Office of Portfolio Analysis (Hutchins, 2016) and frequently used at the NIH and other funding agencies. The RCR differs from other field-normalized indicators in that it uses an article's co-citation network to define the article's field rather than the journal in which the article was published. The indicator is also freely available through a website (icite.od.nih.gov) and an API.

Although it is being rapidly adopted at a number of institutions, it has also been the subject of debate. Waltman (2015) and Janssens et al (2017) published theoretical critiques of the RCR's calculation. A formal response to Janssens et al (2017) was published in Hutchins (2017). Bornmann and Haunschild (2017) provided an empirical analysis and found that the RCR was strongly correlated with other bibliometric indicators but that all indicators were weakly correlated with peer review scores. Purkayastha et al (2019) also found that the RCR was moderately to strongly correlated with other bibliometric indicators.

This poster provides a large-scale analysis of the RCR as compared to a similar metric, the Category-Normalized Citation Impact (CNCI) indicator, in all NIH-funded articles from 2011 through 2015 at both the individual paper level and at the subject category level using the journal-based subject categories from Web of Science (WOS). It attempts to discover similarities and differences between both the metrics themselves and the numerators and denominators of both metrics in this data set.

Methods

Publications and citation metrics were obtained from PubMed, iCite, and InCites. NIH-funded articles were identified in PubMed using the search string "nih[gr] AND 2011:2015[dp]" and the results were downloaded as a list of PMID numbers. I then uploaded the PMIDs into InCites and downloaded the InCites citation metrics for the matched articles. I then used the iCite API to obtain iCite citation metrics for these articles. Finally, I merged the two sets of metrics using the PMID as a matching key. All data were retrieved on 31 December 2018.

I then analysed the resulting data at the individual article level and at the WOS subject category level. Articles assigned to multiple subject categories were counted as a whole article for each subject category, in line with the InCites method of calculating CNCI values for these articles. For stability, only subject categories with 100 or more publications were retained. Since both the CNCI and RCR are calculated by dividing an article's citation count, or rate, by an expected citation count, or rate, I not only compared the final RCR and CNCI values of these articles, but also the times cited counts and expected citation rates obtained from the two databases.

Results

A total of 558,449 articles were successfully matched in both InCites and iCite. At the individual paper level, the RCR and CNCI were strongly correlated with each other (Spearman's rho = 0.87), but both the RCR and CNCI were also strongly correlated with the articles' original times cited counts (rho = 0.91 for the RCR and rho = 0.85 for the CNCI). A scatterplot of the CNCI and RCR values for these articles is given in Figure 1.

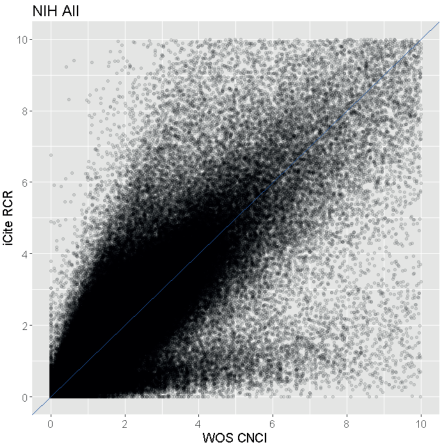


Figure 1. Scatterplot of paper-level CNCI and RCR values. Note that the axes are scaled from 0 to 10 to exclude outliers.

Despite the strong correlation between the metrics, Figure 1 shows a number of papers with diverging metric scores. Approximately 44% of these papers had an absolute difference of less than 0.2 between the indicators, while 56% had an absolute difference of 0.2 or greater. Approximately 27% of the papers had an absolute difference of 0.5 or greater between the metrics.

At the subject category level, there was virtually no correlation between the median RCR and CNCI values for articles in each category (rho = 0.08). The median RCR by category was highly correlated with each category's median citation count (rho = 0.86), but their median CNCI was not at all correlated with median citation counts (rho = -0.05).

Figure 2 indicates that a number of categories had a substantially higher median CNCI than RCR. Nearly all of these categories were in the fields of computer science, chemistry, engineering, and the social sciences. Analyses of the median citation counts and median expected citation rates by subject category indicate that many of these fields had some combination of lower citation counts and higher expected citation rates in iCite than in InCites.

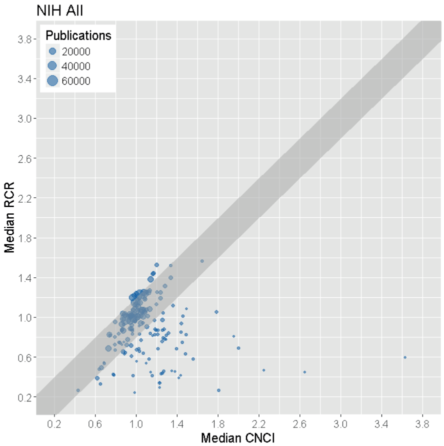


Figure 2. Scatterplot of median CNCI and RCR values by WOS subject category for categories with 100 or more papers.

Discussion

These results raise questions about the appropriateness of the RCR for citation impact measurement. The strong correlation between the RCR and times cited counts at both the individual paper level and at the subject category level raises the possibility that the RCR may not be adequately adjusting for citation differences across fields. While it is logical that both the RCR and CNCI should be correlated with citation counts at the paper level, the fact that the RCR is also strongly correlated with citation counts at the category level, while the CNCI is not, should be cause for concern.

The strong correlation between the CNCI and RCR found here, which agrees with previous studies, seems to mask the divergence (absolute difference > 0.2) of the metrics in the majority of these papers and the disagreement (absolute difference > 0.5) between the metrics in over a quarter of the articles.

Substantial differences in RCR and CNCI values by subject category in most non-biomedical subject categories suggest that the RCR as currently implemented in iCite may be systematically undervaluing articles published in non-biomedical fields as compared to the CNCI. Use of either the RCR or the CNCI for evaluation purposes might therefore lead evaluators to different conclusions about the citation impact of same articles depending on the indicator selected.

Additional research is needed to replicate these results and to further validate the use of the RCR for citation impact measurement.

References

Bornmann, L., & Haunschild, R. (2017). Relative Citation Ratio (RCR): An Empirical Attempt to Study a New Field-Normalized Bibliometric Indicator. *Journal of the Association for Information Science and Technology, 68*(4), 1064-1067. doi:10.1002/asi.23729

Hutchins, B. I., Hoppe, T. A., Meseroll, R. A., Anderson, J. M., & Santangelo, G. M. (2017). Additional support for RCR: A validated article-level measure of scientific influence. *Plos Biology, 15*(10), e2003552. doi:10.1371/journal.pbio.2003552

Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. (2016). Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *Plos Biology, 14*(9), e1002541. doi:10.1371/journal.pbio.1002541

Janssens, A. C. J. W., Goodman, M., Powell, K. R., & Gwinn, M. (2017). A critical evaluation of the algorithm behind the Relative Citation Ratio (RCR). *Plos Biology, 15*(10), e2002536. doi:10.1371/journal.pbio.2002536

Purkayastha, A., Palmaro, E., Falk-Krzesinski, H. J., & Baas, J. (2019). Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). *Journal of Informetrics, submitted.*

Waltman, L. (2015). NIH's new citation metric: A step forward in quantifying scientific impact? Retrieved from http://www.cwts.nl/blog?article=n-q2u294&title=nihs-new-citation-metric-a-step-forward-in-quantifying-scientific-impact

Two indicators rule them all: Mean and standard deviation used to calculate other journal indicators based on lognormal distribution of citation counts

Zhesi Shen[1] , Liying Yang[1], Jinshan Wu[2]

[1] shenzhs@mail.las.ac.cn, yangly@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences, Beijing 100190 (China)

[2] jinshanw@bnu.edu.cn
School of Systems Science, Beijing Normal University, Beijing 100875 (China)

Introduction

On one hand, it has been noted that often citation counts of papers in a journal in certain period of time window follow more or less the lognomral distribution (Radicchi et al. 2008; Stringer et al. 2010). On the other hand, often journal indicators, such as journal impact factor (JIF), journal h index, journal one-by-one-sample comparison citation success index (Stringer et al., 2008, Milojević et al., 2017), journal multiple-sample comparison success rate (Shen et al. 2019), and minimum representative sizes (Shen et al. 2019) are calcualted from raw data of citation counts of journal papers. In this work, we want to brige the two: starting from a lognomral distribution with given parameters (*m* and *v as* defined later*)*, it is possible to derive and estimate other indicators, and check accuracy of such an estimation.

We will show that those estimated indicators are consistent with those calculated directly from the raw citation data. Besides its theoretical value, in practice, being able to estimate other indicators from core indicators allows those who do not have access to the raw citation counts data to work with those other indicators simply using the core indicators, which are often readily available.

Data and Methods

To implement and test our idea, papers published in 2015 and 2016 and their corresponding citation counts in 2017 of the top 30 journals with the highest JIF listed in Web of Science-MEDICINE, GENERAL & INTERNAL category from the Journal Citation Report 2017 are extracted.

Results

For a log-normal distribution, the parameter μ and σ are related to the basic statistics *m* and *v* by

$$\mu^i = \ln\left(\frac{m^i}{\sqrt{1+\left(\frac{v^i}{m^i}\right)^2}}\right), \sigma^i = \sqrt{\ln\left(1+\left(\frac{v^i}{m^i}\right)^2\right)}. \quad (1)$$

H-index

H-index is one of the most widely used indicators which takes both quantity and quality into consideration. H-index is defined as the largest value *h* that there are *h* papers are cited at least *h* times. Under the log-normal assumption, the H-index *h* can be estimated via solving the following equation numerically

$$\hat{h} = N \times \int_{\hat{h}+1}^{\infty} \rho(x, \mu, \sigma) dx \quad (2)$$

where $\rho(x, \mu, \sigma)$ is the log-normal probability density function. By combing Eq.1 and Eq.2, we can estimate H-index from *m* and *v*, as shown in Fig.1. From Eq.2, we also can see how journal size *N* is related to journal H-index theoretically.

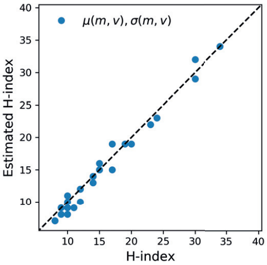


Figure 2. Comparison of real H-index and estimated H-index.

Citation Success Index (CSI)

The CSI, S_r^t, of the journal *t* compared with the journal *r* is defined as the probability of the citation count of an randomly selected paper from journal *t* being larger than that of a random paper from journal *r* (Stringer et al., 2008, Milojević et al., 2017). Under the log-normal assumption, S_r^t can be estimated (Shen et al. 2018) as

$$S_r^t = \int_{-\frac{(\mu^t - \mu^r)}{\sqrt{(\sigma^r)^2 + (\sigma^t)^2}}}^{\infty} N(x; 0, 1)\, dx, \quad (3)$$

where $\mathrm{N}(x;0,1)$ is the standard Normal distribution. Figure 3 shows the estimation of CSI based on *m* and *v* using Eq.1 and Eq.3.

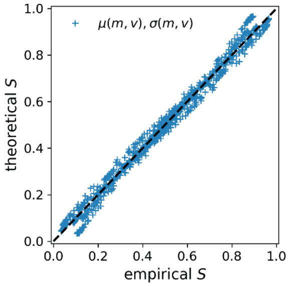


Figure 3. Comparison of real CSI and estimated CSI.

CSI can be further extended to a *group-group comparison* case, S_{r,k_r}^{t,k_t}, defined as the probability of the average citation of k_t random samples from journal *t* being larger than the average citation of k_r random samples from journal *r* (Shen et al. 2019). Such situation is often faced in organization evaluation using journal-level indictors. As the average of random samples from a log-normal distribution can also be approximated as log-normal distribution with following parameter μ_k and σ_k,

$$\mu_k = \mu + \frac{\sigma^2}{2} - \frac{\sigma_k^2}{2} + \ln k, \sigma_k^2 = \ln\left[\frac{e^{\sigma^2} - 1}{k} + 1\right]. \qquad (4)$$

we can efficiently estimate S_{r,k_r}^{t,k_t} by combing Eq.4 and Eq.3.

Minimum representative size

Minimum representative size is a recent proposed indicator when comparing two journals according to their means (Shen et al. 2018) and is closely related to the group-group comparison CSI. It asks how large k_t and k_r should be so that the group-group comparison is reliable, i.e., S_{r,k_r}^{t,k_t}>0.9.

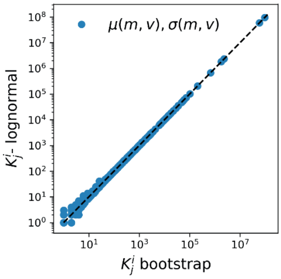


Figure 4. Comparison of real *K* from bootstrap sampling and estimated *K*.

Conclusion

We have illustrated the high consitency between the journal indicators estimated from lognormal distribution and the ones calculated directly from raw data. This firstly partially confirms the validity of lognormal distribution: even if it is not exactly true mathematically, practically when calculating many journal indicator, we can still use this lognormal distribution assumption. Secondly, this provides researchers a way to get journal indicators without direct access to the raw data of citation counts of individual papers. Conceptually, this also shows how these indicators are related to each other.

References

Milojevićv S., Radicchi F. , Bar-Ilan J. (2017). Citation success index – An intuitive pair-wise journal comparison metric. *Journal of Informetrics*, 11, 223-231.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 17268–17272.

Shen, Z., Yang, L., Di, Z. and Wu, J. (2019). Large enough sample size to rank two groups of data reliably according to their means, *Scientometrics*, 118: 653-671.

Stringer M.J. , Sales-Pardo M. , Amaral L.A.N. (2008), Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, 3, e1683.

Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *Journal of the American Society for Information Science, 61,* 1377-1385.

Understanding Roles of Collaborators from Their Byline Orders and Affiliations

Chao Lu[1], Chengwei Zhang[2], Ying Ding[2], Dandan Ma[3], Yingyi Zhang[4],

[1] luchaoink@gmail.com, [2] zhang334@indiana.edu, [3]dingying@indiana.edu, [3].jane.ma1030@gmail.com,
[4]yingyizhang@njust.edu.cn

[1] Hohai University, Business School, 8 West Focheng Road, Nanjing, Jiangsu, 210098 (China)

[2] Indiana University, School of Informatics, Computing, and Engineering, 700 N Woodlawn Ave, Bloomington, IN 47408 (USA)

[3] Nanjing University of Finance & Economics, School of Information Engineering, 3 Wenyuan Rd, Nanjing, Jiangsu, 210046 (China)

[4] Nanjing University of Science & Technology, School of Management and Economics, 200 Xiaolingwei St, Nanjing, Jiangsu, 210094 (China)

Introduction

Collaborations are prevailing in science currently (Wuchty, Jones, & Uzzi, 2007). It is believed to have incomparable advantages such as bringing diverse ideas to breed innovations and sharing various facilities and equipment to enrich scientific practice. Collaborations have been encouraged in many disciplines. However, little is known how a team really functions from the detailed division of labor within the team. Here, we continue our study on scientific collaboration and division of labor within individual scholarly articles (Lu et al, 2018) by analyzing the relationship between collaborators' roles and their byline orders and affiliations.

Data and Method

Author Information Parsing

Nearly 170,000 full-text articles published in *PLoS*[i] from 2006 to 2015 are collected in XML formats with their metadata, including author information. First, the author contribution statements of these papers are extracted and parsed using natural language processing techniques assisted by necessary manual correction as exemplified in Table 1. Only those statements that are completely and correctly parsed are kept, leaving us 138,787 articles correctly parsed. Then, authors' full names, byline orders, and affiliation information are extracted from the authors' full names to match the author name abbreviations in contribution statements, byline orders and affiliations (Here, we removed 33,595 articles where author names cannot be completely matched). Then, we remove 1,331 single-authored articles, which leads to articles excluded from our initial data set. So our final data set contains 103,861 articles with their author contribution statements parsed to identify their roles in collaboration and their author byline orders and affiliations to assist us understand their roles in collaborations.

Table 1. An author contribution parsed sample from our dataset[ii].

Id	Authors	Task
1	EG; ES; JD	Conceived and designed the experiments
2	ES; JD; MH; JP; MS	Performed the experiments
3	EG; ES; FC; JD; JP; MS	Analyzed the data
4	ES; JD; MH; JP; MS	Contributed reagents
5	ES; JD; MH; JP; MS	Contributed materials
6	ES; JD; MH; JP; MS	Contributed analysis tools
7	EG; ES	Wrote the paper

Types of collaborators

Network Construction

Weighted undirected network model is adopted here to construct an author co-contributorship network for every study using the parsed author contribution statements. One *node* in the network denotes a collaborator. Every e*dge* in the network represents task(s) shared by co-author(s). A self-looped edge indicates task(s) performed independently. The *weight* of an edge is the number of tasks performed by author(s). Then we can identify three types of collaborators from the network: Specialists, Versatiles, and Team-players (Lu et al, 2018).

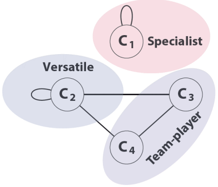


Figure 1. Types of Collaborators edited from (Lu *et al*, 2018).

Byline orders

Byline order is the order where authors' names are assigned in their publications, usually demonstrating some degree of their contributions to their work or teams (Corrêa Jr, Silva, Costa, & Amancio, 2017). We use Formula (1) to calculate each author's normalized byline order in every single article and mitigate the effect caused by different numbers of authors among articles:

$$NB_i^j = \frac{B_i^j - 1}{N_i - 1}, N_i \geq 2, 0 \leq NB_i^j \leq 1 \quad (1)$$

Where B_i^j is one author's byline order in his/her collaborated article *i* and N_i denotes the total number of authors in the article i and NB_i^j represents the normalized byline order of the author, which is in the range of [0,1].

Affiliation index

we proposed AFI (Affiliation Index) to depict the disparity between one author's affiliations and the affiliations of the whole team. We use Formula (2) to calculate AFI index of each author:

$$\text{AFI(k)} = \frac{N_c^k}{N-1}, N \geq 2 \quad (2)$$

In the formula, N_c^k stands for the number of author k's colleagues within the team (sharing same affiliations); *N* denotes the total number of authors in the study. For instance, one paper is collaborated by three authors (i.e., A, B, and C): A is affiliated with aff_1 and aff_2; B belongs to aff_2 and aff_3; and C's affiliations is aff_3. The AFI for author A is $\frac{1+0}{2} =$ 0.5. Intuitively, when AFI is 1, it means the author is a colleague of the rest of authors; when AFI is 0, it means the author is affiliated with a different organization from other collaborators.

Result

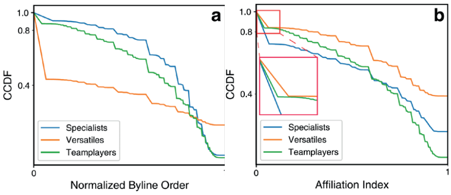


Figure 2. CCDFs for Byline orders (a) and AFIs (b) of collaborators (p<0.0001 in both Kolmogorov–Smirnov test between groups).

Review of our former study

Given the co-contributorship network of a paper, we defined three types of contributors: Specialists, Team-players, and Versatiles (in Figure 1). Specialists are those who contribute to all their tasks alone; team-players are those who contribute to every task with other collaborators; and versatiles are those who do both. We found that team-players are the majority and tend to contribute to the five most common tasks as expected, such as "data analysis" and "performing experiments". The specialists and versatiles are more prevalent than expected from random-graph null models. Versatiles tend to be senior authors associated with funding and supervisions. Specialists are associated with two contrasting roles: the supervising role as team leaders or marginal and specialized contributions.

Byline Order

Figure 2(a) plots the CCDF (complementary cumulative distribution function) for the three types of collaborators. In the plot, versatiles usually demonstrate their leading positions in collaborations among authors, which takes accords with our observations; while specialists usually sign their names at the end of their bylines, suggesting their more marginal contributions to teams. In between lie the team-players, who usually perform the common tasks within a team; their names are more frequently placed in the middle. However, versatiles can also occasionally appear at the end of bylines, indicating their authorities in research as corresponding authors.

Affiliation index

A larger affiliation index value of an author usually indicates one collaborates with his/her colleagues in a single study. Versatiles demonstrate their much stronger connections with other collaborators than those of the other two types of collaborators, confirming their core role in communication and coordination. Team-players, as the main labor source, tend to have a larger affiliation index value than specialists. However, specialists take over the leading position when affiliation value exceeds 0.5. It might suggest that specialists can also partake the role of communication and coordination as versatiles within teams and thus, confirm our former findings.

Conclusion

In this study, we use authors' byline orders and affiliations to understand different types of collaborators. The results extend our former findings about different types of collaborators and their roles in study and also imply the usefulness of affiliation index to identify author roles in scientific collaborations.

References

Corrêa Jr, E. A., Silva, F. N., Costa, L. D. F., & Amancio, D. R. (2017). Patterns of authors contribution in scientific manuscripts. *Journal of Informetrics*, *11*(2), 498-510.

Lu, C., Ding, Y., Zhang, Y., Bu, Y., & Zhang, C. (2018). Types of Scientific Collaborators: A Perspective of Author Contribution Network. *iConference 2018 Proceedings*.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036-1039.

[i] https://www.plos.org/

[ii] See original text: http://t.cn/EMSVcGp

Citation2vec: A New Method for Citation Recommendation Based on Semantic Representation of Citation Context

Jinzhu Zhang[1,2], Yue Wang[1], Duanwu Yan[1,2], Jingjie Liu[1] and Wenqian Yu[1]

zhangjinzhu@njust.edu.cn

[1] Nanjing University of Science and Technology, Dept of Information Management, Xiaolinwei Str 200, Nanjing (China)

[2] Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Xiaolinwei Str 200, Nanjing (China)

Introduction

Tens of millions of research papers have been published and still increasing rapidly with time. Thus finding relevant research works for citation from the gigantic number of published papers has become a nontrivial problem.

In order to solve this problem, researchers have explored several citation recommendation approaches to list multiple recommendation candidates for citation. The methods could be classified into three classes, including search engine based, citation text based and citation relation based methods (Beel, Gipp, Langer, & Breitinger, 2015). They play an important role in different research process.

This paper focuses on citation relation based recommendation. It is often been done in a co-citation network formed by reference sequences after the content of each paper, e.g., a reference sequence shown in Table 1. Then multiple network metrics, e.g., common neighbours and its improvements, are used to find recommendation candidates in the co-citation network (Beel et al., 2015). Moreover, some network embedding methods, e.g., deepwalk(Perozzi, Alrfou, & Skiena, 2014), node2vec(Aditya Grover, 2016) and LINE(Large-scale Information Network Embedding) (Tang et al., 2015), are applied to learn semantic representation of citations in the co-citation network for citation recommendation.

However, these methods have not considered the order of citation, i.e., all citations in a paper do the same impact on each other. Furthermore, the reference sequence after the content is not totally the same as the citation sequence occurring in the content of a paper sometimes, e.g., *r7* and *r8* in Table 1. More important, some references could be cited several times in a paper, e.g., *r1* and *r2* in Table 1, which often be simply denoted by frequency. This may lead that rich citation context information surrounding the reference have been lost, e.g., *r1* and *r2* have more citation context in citation sequence than reference sequence in Table 1. At the same time, the context of citations that occurred only one time has been enriched, e.g., the citation context of *r5* in citation sequence is more than in reference sequence in Table 1.

Table 1. An example of reference sequence and citation sequence in a paper.

Reference Sequence	r1 r2 r3 r4 r5 r6 r7 r8
Citation Sequence (CS)	r1 r2 r3 r4 r2 r5 r6 r1 r8 r7
CS without repetition	r1 r2 r3 r4 r5 r6 r8 r7

Therefore, this paper extracts and defines the citation context in the citation sequence and uses deep learning method to learn the semantic representation of each citation by its citation context. Then the semantic similarity among citations is computed for citation recommendation which evaluated by link prediction.

Data and method

Firstly, the citation sequence in the content of each paper is extracted. Then they are transformed into the input of deep learning method such as word2vec for capturing the context information and generate the semantic vectors of each citation as the output. Finally, the semantic similarities among citations are calculated through vector similarity indicator, which are used for citation recommendation and quantitative compared by link prediction.

Data description

We collected a research dataset in PLOS ONE under the subject area of Artificial Intelligence (AI) in Nov 25, 2018. We downloaded 1,675 full text papers with XML format and eliminate 5 papers because these are correction papers. Thus, there are 1,670 papers retained with 60,097 distinct citations. The details of divided datasets are shown in Table 2. In testing set, there are 13,228 of 581,100 co-citation relations occurred in non-exist relations obtained from training co-citation network, so we use 13,228 co-citation relations as the true testing set.

Table 2. The details of divided datasets.

	Training set (2007-2016)	**Testing set (2017-2018)**
Num of papers	1,150	520
Num of co-citations	1,134,398	581,100

Definition of citation context

This paper uses xml.etree.ElementTree to parse the XML full text papers. According to the occurring sequence of citations in the content, the citation sequence is formed just like in Table 1, where one citation can appear several times. In this dataset, there are 1,150 citation sequences in the training set. The citation context is defined as the 5 citations surrounding current citation, e.g., the citation context of *r5* is "*r1 r2 r3 r4 r2*" and "*r6 r1 r8 r7*" in the citation sequence of Table 1.

Semantic representation of citation context based on deep learning

Deep learning method can capture the semantic information as a semantic vector. Word2vec is one of the word embedding methods that reconstructs linguistic contexts of each word with the input of words sequence. Corresponding to our work, the citation sequences can be seen as the word sequences as the input, and each citation's context can be reconstructed and represented as a semantic vector as the output for denoting this citation.

Semantic similarity calculation based on vector similarity indices

The semantic similarity among citations forms the predictor for citation recommendation. It can be calculated by multiple vector similarity indices. This paper selects cosine similarity index for semantic similarity computation.

Quantitatively evaluation based on link prediction

Link prediction method is used for quantitatively evaluating the effectiveness of indicators. This paper uses the citation sequences in 2007-2016 as the training set and 2017-2018 as the testing set. We choose AUC（Area Under Curve）as the indicator for citation recommendation.

Result

We generate three types of citation sequences for comparison. The first one is in accordance with the order of citations appeared in papers completely where citation can occur multiple times (CS1). The second one is the same as the first one but without repeated citations (CS2). The third one is in accordance with the order of reference after the content of a paper (CS3). Furthermore, we use the training set of citation sequence to form a co-citation network for comparison (CCN).

For three types of citation sequences, the word2vec is used for learning semantic vectors of citations, which are called as C2V1, C2V2 and C2V3 respectively. For corresponding network, the deepwalk is used for learning semantic vectors of citations, which is called as N2V. Deepwalk is an network representation learning method and performs better than traditional similarity indicators such as common neighbours (Perozzi et al., 2014).

The efficiency and effectiveness of different methods are shown in Table 3. The result shows that C2V1 performs best with highest AUC. It means that CS1 contains more citation context than others. N2V performs next but takes the most time. C2V2 performs a little better than C2V3 means that the CS2 is more accurate than CS3.

Table 3. The efficiency and effectiveness of different methods.

	C2V1	**C2V2**	**C2V3**	**N2V**
AUC	*75.4%*	62.3%	60.0%	66.6%
Time(s)	21.2	19.6	*19.4*	967.8

Conclusion

This paper enriches the citation context and applies the deep learning method to learn the semantic representation of each citation with its citation context. This method performs best when compared with others. In the next step, we would like to apply other deep learning methods that may do better in semantic representation of citation context. In addition, the dataset's size will be increased because the bigger the data size the more citation context it has.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 71503125), the Social Science Fund Youth Project of Jiangsu Province (No. 17TQC003), the Fundamental Research Funds for the Central Universities (No. 30918013108), the National Key Research and Development Program of China (2017YFB1401903) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Aditya Grover, J. L. (2016). *node2vec: Scalable Feature Learning for Networks.* Paper presented at the Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2015). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries, 17*(4), 1-34.

Perozzi, B., Alrfou, R., & Skiena, S. (2014). *DeepWalk: Online Learning of Social Representations.* Paper presented at the Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). *Line: Large-scale information network embedding.* Paper presented at the Proceedings of the 24th International Conference on World Wide Web.

Representation of Libraries in Funding Acknowledgments

David E. Hubbard[1] and Sierra Laddusaw[2]

[1] *hubbardd@library.tamu.edu*
Texas A&M University, University Libraries, 5000 TAMU, College Station, Texas (USA)

[2] *sladdusaw@library.tamu.edu*
Texas A&M University, University Libraries, 5000 TAMU, College Station, Texas (USA)

Introduction

Acknowledgments are an important aspect of the scholarly communication process. The importance of these paratextual components were pioneered by Cronin in the 1990s and is now an established field of study (e.g., Cronin, 1995). While an acknowledgment is a small part of a publication, it highlights the contributions of others to the research and scholarship. More recently, linguistic analysis has found recognition beyond funding in Web of Science (WoS) full-text funding acknowledgments (FAs) (Paul-Hus et al., 2017). Some researchers have begun to use library acknowledgments (LAs) in publications as a way to examine the impact of academic libraries (Finnell 2014; Hubbard et al., 2018; Scrivener, 2009). This study fills a gap by quantifying and characterizing the representation of libraries in WoS FAs. More specifically, (1) Are libraries acknowledged in journal article FAs and to what extent and context? (2) How do LAs differ across disciplines and time? and (3) How do FAs mentioning libraries differ among peer universities?

Methods

This study examined FAs of six universities, Texas A&M University (TAMU) and five peer universities (P1-P5), for 2008-2018. Acknowledgments were obtained from WoS by searching for publications using Organization-Enhanced and Funding Text fields. Acknowledgments were further refined to those associated with libraries using a truncated term (librar*). The LAs were then categorized by the following: facilities, people, resources, services, and general. Inter-rater reliability was determined for TAMU; the other five were then divided among the two co-authors. The LAs were also examined over time and by WoS categories.

Results/Discussion

Articles at all six universities had LAs, though the numbers and percentages are low (Table 1). The values in the last column of Table 1 include those with and without local open access (OA) funding. All subsequent analyses are performed on publications/acknowledgments that exclude local OA funding since its inclusion distorts comparisons for those universities that do not offer OA funding.

Table 1. Summary of articles and acknowledgments (2008-2018)

Univ.	Total Article Count	Total FAs # (%)	Librar* FAs # (%)	Relevant Library FAs # (%)
TAMU	45,066	28,785 (63.9)	182 (0.63)	107 (0.37) 19 (0.07)[a]
P1	62,820	41,216 (65.6)	126 (0.31)	35 (0.08) 33 (0.08)[a]
P2	49,983	34,906 (69.8)	167 (0.48)	36 (0.10) 36 (0.10)[a]
P3	59,079	38,306 (64.8)	136 (0.35)	30 (0.08) 24 (0.06)[a]
P4	42,663	29,046 (68.1)	91 (0.31)	19 (0.07) 19 (0.07)[a]
P5	57,792	39,336 (68.1)	186 (0.47)	67 (0.17) 44 (0.11)[a]
Average	52,901	35,266 (66.7)	148 (0.42)	49 (0.14) 29 (0.08)[a]

[a] Excludes OA funding from home university.

Many LAs were false hits (e.g., DNA library), while a smaller number were deemed relevant. The inter-rater reliability, Cohen's kappa, was 0.92 for TAMU indicating almost near perfect agreement with respect to categorization. Figure 1 summarizes the types of LAs found within journal articles of TAMU and five peer universities. It should be noted that each WoS FA may contain more than one library acknowledgment (e.g., one FA may acknowledge use of a library collection, thank a librarian for assistance, and express indebtedness for internet access at another library). The Resources category, which includes funding from libraries, was one of the larger categories across all six universities even without OA funding. People and Services also figured prominently. Facilities were seldom, if ever, mentioned. Selected examples of LAs include: (1) "Maps were generated with help from the Map and

GIS Collections and Services at [TAMU] Libraries…and bathymetry data are from Tobin Global Planner…" [Resources, Services]; (2) "Archival research was facilitated by…Herbarium Library of the [P1] Museum of Natural History." [General]; and (3) "[P3] Library Data Learning Centre for the statistical analysis and interpretation." [Services].

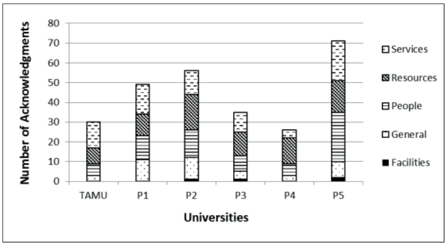


Figure 1. Categories of acknowledgments by university.

The cumulative number of LAs annually for the six universities increased approximately 10-fold from 2008 to 2018, though some of the increase may be due to more thorough full-text FA entries within WoS. It is beyond the scope of this study to explore why there is an increase in LAs, but is certainly worth exploring in a future study.

Table 2 shows the WoS categories assigned to the journals associated with the LAs presented in Figure 1. Of the 252 WoS categories, 97 were among the 175 articles containing LAs. Table 2 is limited to those with 5 or more WoS category counts. The natural sciences were well represented among the journals with LAs (Zoology, Environmental Sciences, Ecology, Plant Sciences, and Water Resources). These five subject areas correspond to a number of LAs with flora/fauna libraries, in addition to interlibrary loan services, technical assistance, and library funding found throughout many of the LAs.

Limitations

WoS covers certain disciplines more thoroughly than others, plus some disciplines are more journal-centric (e.g., sciences) compared to others (e.g., humanities). Collectively this may result in some scholarship and therefore LAs being missed. It should also be noted that only a single truncated English-language search term was used. Additional analogous terms (e.g., archives, bibliothèque, etc.) may yield more acknowledgments.

Conclusion

The number and percentage of acknowledgments to libraries and/or librarians were found to be low, but all six universities had similar percentages and types. This approach may offer an additional means for libraries to demonstrate their impact on research at their respective universities and the larger scholarly community, beyond those typically used for library assessment and therefore provide data for a richer qualitative narrative.

Table 2. Web of Science categories of library acknowledgments

WoS Category	TAMU	P1	P2	P3	P4	P5	TOTAL
Zoology	5	5	5	–	–	1	16
Environmental Sciences	2	1	2	3	–	7	15
Information Science & Library Science	–	–	8	1	2	2	13
Multidisciplinary Sciences	–	2	2	4	–	5	13
Ecology	4	3	2	1	–	1	11
Plant Sciences	1	3	1	1	3	–	9
Public, Environmental & Occupational Hlth	2	–	1	1	–	3	7
Genetics & Heredity	–	–	3	–	–	3	6
Health Care Sciences & Services	1	–	–	–	1	3	5
Nutrition & Dietetics	2	–	1	1	1	–	5
Toxicology	1	–	1	1	1	1	5
Water Resources	1	–	1	–	3	–	5

Acknowledgments

We would like to thank Bruce Neville for providing feedback and helpful comments on the manuscript.

References

Cronin, B. (1995). *The Scholar's Courtesy: The Role of Acknowledgment in the Primary Communication Process.* London: Taylor Graham.

Finnell, J. (2014). Much obliged: Analyzing the importance and impact of acknowledgements in scholarly communication. *Library Philosophy and Practice.* Retrieved from http://digitalcommons.unl.edu/libphilprac/1229/

Hubbard, D. Laddusaw, S. Kitchens, J. & Kimball, R. (2018). Demonstrating library impact through acknowledgment: An examination of acknowledgments in theses and dissertations. *The Journal of Academic Librarianship,* 44(3), 404-411.

Paul-Hus, A., Diaz-Faes, A., Sainte-Marie, M., Desrochers, N., Costas, R., & Lariviere, V. (2017). Beyond funding: Acknowledgement patterns in biomedical, natural and social sciences. *PloS ONE,* 12(10): e0185578.

Scrivener, L. (2009). An exploratory analysis of history students' dissertation acknowledgments. *The Journal of Academic Librarianship,* 35(3), 241-251.

How does author ethnic diversity affect scientific impact? A study of nanoscience and nanotechnology

Jielan Ding[1], Zhesi Shen[1], Per Ahlgren[2], Tobias Jeppsson[3], David Minguillo[3]

[1] *dingjielan@mail.las.ac.cn, shenzhs@ mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing 100190 (China)

[2] *per.ahlgren@uadm.uu.se*
Uppsala University, Department of Statistics, Box 256, 751 05 Uppsala (Sweden)

[3] *tjep@kth.se, davidmin@kth.se*
KTH Library, KTH Royal Institute of Technology, Stockholm 100 44 (Sweden)

Introduction

Earlier bibliometric research on the relationship between author ethnic diversity and scientific impact has found a positive relationship between these two variables (Freeman & Huang, 2014; Freeman & Huang, 2015; AlShebli, Rahwan & Woon, 2018). However, more research is needed to understand the reasons for this effect. To further explore how ethnic diversity affects scientific impact, this study put forward a model to connect ethnic diversity with scientific impact, assuming novelty and audience diversity as mediators. Our research hypotheses are as follows:

H1 Ethnic diversity would have positive effects on scientific impact.

AlShebli, Rahwan and Woon (2018) found that group and individual ethnic diversity can have positive effects on scientific impact.

H2 Ethnic diversity would have positive effects on novelty (of ideas).

Peterson (2001) showed that multi-national collaboration, where the collaborators have different cultural (and educational) backgrounds, tend to stimulate new ideas and develop new approaches to theoretical or practical problems.

H3 Ethnic diversity would have positive effects on audience diversity.

Freeman and Huang (2014) suggested that a publication generated by a more diverse research group could tap into different networks and thus attract greater diversity with respect to citing authors.

H4 Novelty would have positive effects on scientific impact.

Wang, Veugelers and Stephan (2017) found that highly novel publications deliver high gains to science: they are more likely to be highly cited papers (top 1%) in the long run.

H5 Audience diversity would have positive effects on scientific impact.

Kerr (2008) showed that ethnic technology transfers are particularly strong in high-tech industries and among Chinese economies. The strong Chinese outcomes of technology transfer may be due to unique qualities of ethnicity's network (for example, size and network effects).

Data and Methods

We use 97,983 publications, of the type *Article*, from the Web of Science (WoS) subject category *Nanoscience & Nanotechnology*. The data was collected from Bibmet, the bibliometric version of WoS at KTH Royal Institute of Technology (Sweden). The publication period is 2008-2012.

We use *NamePrism* to automatically identify the ethnicity of authors. *NamePrism* is the most accurate, fine-grained nationality classifier available (Ye et al., 2017). *NamePrism* gives ethnicities for any name given as input to it. Each name occurring in the 97,983 publications has been assigned an ethnic group. In our analysis, we use structural equation modelling (SEM), which was performed with the AMOS software. Variables and measures used are reported in Table 1.

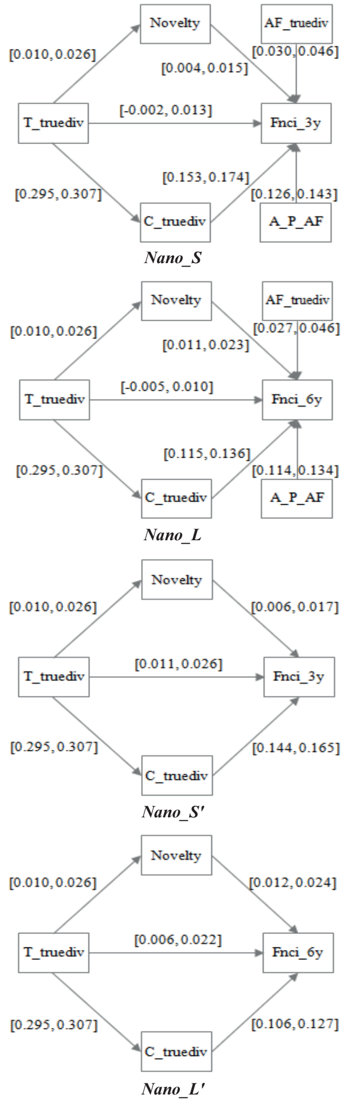
Table 1 Variables and measures

	Variables	*Measures*
Independent variable	Ethnic diversity	***T_truediv***—true diversity measure (Zhang, Rousseau & Glänzel,2016) for author ethnic diversity of target publications
Mediate variables	Novelty	***Novelty***—new combination of referenced journals (Wang, Veugelers & Stephan, 2017)
	Audience diversity	***C_ truediv***—true diversity measure (Zhang, Rousseau & Glänzel,2016) for author ethnic diversity of citing publications
Response variables	Scientific impact	***Fnci_3y***—field normalized citation rate with 3 years citation window ***Fnci_6y***— field normalized citation rate with 6 years citation window
Control variables	Affiliated country diversity	***AF_truediv***—true diversity measure (Zhang, Rousseau & Glänzel,2016) of affiliated countries for a target publication
	Affiliated country size	***A_P_AF***—Average number of publication fractions of the affiliated countries for a target publication

Note: Ethnicity similarity, used in *T_truediv* and *C_truediv*, is calculated from the ethnicity identification results of our data from *NamePrism*, as *NamePrism* gives 3 candidate ethnicities for each name. The similarity between affiliated countries is their geographic distance.

Results

RMSEA and *GFI* are two commonly used measures of global fit for SEM models. The *RMSEA* values of the models of Figure 1–*Nano_S*, *Nano_*L, *Nano_S*' and *Nano_L'*–are 0.044, 0.044, 0.067 and 0.067, while the *GIF* values are 0.997, 0997, 0998 and 0.998. Those values satisfy fit thresholds for the two measures (*RMSEA*<0.08, *GFI*>0.9). The four models can be regarded as valid.



Nano_S

Nano_L

Nano_S'

Nano_L'

Note: 95% confidence intervals for std. regression coefficients.

Fig 1 Model results

Figure 1 displays the path models with regressions. Regression coefficients were standardized for all paths to allow comparisons between relationships. In the reduced models (Model *Nano_S'* and *Model Nano_L'*), which do not take control variables into consideration, *T_truediv* has a weak positive impact on *Fnci_3y* and *Fnci_6y*. But in the models *Nano_S* and Model *Nano_L*, where control variables are used, *T_truediv* does not have any positive impact on *Fnci_3y* or *Fnci_6y*. Without control variables, *T_truediv* has more impact on *Fnci_3y* than on *Fnci_6y*. In all four models, *T_truediv* has a positive relationship with *C_truediv*, but it has a much weaker relationship with novelty.

Discussion and conclusions

In this work, we try to explain the mechanism of how ethnic diversity affects scientific impact by studying the two-layer relationship between ethnic diversity and scientific impact with novelty and audience diversity as mediators.

We find that ethnic diversity can increase scientific impact when not considering affiliated country diversity and affiliated country size (in terms of publication output), but that ethnic diversity does not have any effect on scientific impact when taking the two variables into account. The affiliated country diversity and affiliated country size are related to international collaboration to some extent. Earlier studies have found that ethnic diversity can increase scientific impact, which might be due to the fact that these studies did not control for affiliated country variables. When not considering affiliated country variables, ethnic diversity seems to promote short-term scientific impact rather than long-term. Our research further suggests that ethnic diversity can generate scientific impact via audience diversity. For future research, we intend to extend our analysis to other scientific fields.

References

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communication -s*, 9, 1-10.

Freeman, R. B., & Huang, W. (2015). Collaborating with people like me: Ethnic coauthorship within the United States. *Journal of Labor Economics*, 33(S1), S289-S318.

Peterson, M. F. (2001). International collaboration in organizational behavior research. *Journal of Organizational Behavior*, 22(1), 59–81.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8), 1416-1436.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67(5), 1257–1265.

Improving RA-index by Using the Weighting Mechanism Number of Citations to Filter "Spike" Signal of the Citation Data of Indonesian Authors

Adian Fatchur Rochim[1] and Riri Fitri Sari[2]

[1] *adian@undip.ac.id*
Departement of Computer Engineering, Diponegoro University, Semarang, 50275 (Indonesia)

[2] *riri@ui.ac.id*
Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok 16424 (Indonesia)

Introduction

Number of citations and the number of papers were combines as H-index (Hirsch, 2005). H-index is an index to figure the profile of the authors. H-index is a well-known index that is used by the database indexers such as Clarivate Analytics, Scopus and Google Scholar. From the literature we found some weaknesses of H-index, including 1) the productive and perfectionist researcher were not accommodated by the H-index (Mesiar, et.al, 2016), 2) self-citation was calculated, 3) the citation weight of the main researcher is considered equal to other researchers, and the frequency of citation in a paper has not been considered (Bai et al., 2018) (Mesiar, et.al, 2016) (Gagolewski, et.al, 2009) (Zhu, et.al, 2015). Many H-index improvement proposals have been made. This includes the proposal of Egghe in 2006, which accommodates the impact value of perfectionist researchers (Egghe, 2006). Improvement and new indicator to measure the impact of researchers was needed for a better evaluation. Rochim, et.al. in 2018, proposed the RA-index as an alternative indicator of fairer-based bibliometrics to measure the impact of researchers (Rochim, et.al., 2018). Glanzel in 2016, stated that it is important to consider some methods and models to accommodate the needs (Glänzel, et.al, 2016).

This poster proposed an initial work to weighting mechanism and to filter the "spike" of citation. Subsequently, the filter is applied, and the result of citation data is calculated by the RA-index. RA-index is a more fairness-concerned variant of H-index (Rochim et al., 2018) (Rochim, et.al, 2017).

This work to measure and to differentiate of two authors with the same H-index value using the weighting citations and RA-index method. We investigate the phenomenon of the "spike" of the number of citation, and the initial solution to prevent/filter impact of the cartels/citation circle. The "spike" of citation phenomenon is the raise of the number of sudden citations within a short period of time, which is obtained from co-authors of multiple papers. Cartels/citation circle can be defined as follows: 1) The activity of an author that act as also a reviewer for multiple papers at the same time and a joint-work among friends in a peer review ring to increase the record of papers and citation numbers (Gamboa, 2014), and 2) The activity of an author cite his/her friend's papers, and at the same time these friends also cite the author's papers (Witold Kienc, 2015). Tscharntke in 2007 classified the weighting for each author in a publication text into four weighting methods groups. The four groups are: 1) Sequence-determining-credit (SDC), 2) Equal Contribution (EQ), 3) First-author-emphasis (FLAE) and 4) Percent-contributed-percentage (PCI) (Tscharntke, 2007). In 2018, we have identified that a small number of Indonesian researchers conducted some activities of "*citation circles*" to increase their H-index values. "*citation circle*" is an activity in which someone cites the work of his friends, and will get a citation for the same way (Witold Kienc, 2015). This is a part of the "black hat" technique. The technique is not accepted or illegal for academics.

Methodology

In order to prevent the activity of "creating citation circle", we recommend the weighting mechanism for the citation data. The citation data is weighted before it will be calculated by the RA-index method. This weighting mechanism is proposed to give an appreciate the first author and the corresponding authors. The corresponding author is normally the supervisor of the author. The proposed method accommodates the regulations of the Indonesian Government in granting credits for scientific publications. The method of the weighting mechanism is based on the combination of PCI and EC methods. For example, one paper has ten citations, and written by four authors i.e. main author (1), corresponding author (1) and other authors (2). The citation calculation obtained by each author is different and based on the following proportions as follows. The main author and correspondent get the maximum publication index value of 100% of the publication index value.

author's publication index value = ma x 100 % (1)
ma value = the number of total citations of a paper.

others author's citation value = (ma x 50%)/n (2)
n = total number of the others author.

Others author get a value of 50% of the maximum value divided by the number of other authors. We called the combination of the PCI and EC methods as the *maximization* or *ma* method for a weighting number of citations of author. After weighting, is done, the data were calculated by the RA-index method. The combination method of the *ma* method and the RA-index method is then called the RAMa-index method. In the previous work, a model for calculating the impact of researchers has been done using the RA-index method (Rochim et al., 2018). The following is the RA-index equation. The following is the example of the calculation process: Two authors are A and B. Author A has an H-index value of 10 and Author B of 12. Figure 1. shows the number of citations of the author each years.

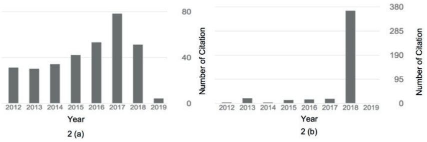


Figure 1. Comparison of the two authors.

Based on the visual observations, it appears that the number of papers and the number of citations of researcher A is greater than that of researcher B. However, the H-index value of researcher A is smaller than that of researcher B. The area of the number of citations of the researcher A is greater than researcher B in the period 2012 to 2019. Researcher B only had a large citation in 2018. For this reason, we need to filter the "spike" citation data of the author.

Table 1. Comparison the number of papers, citation and number of citation after weighted.

Researcher A				
No. of Papers	No. of Citation	Number of Authors	Author Position	Number of Citation
1	151	3	*	25
2	74	3	1	74
3	33	5	*	3
4	23	3	*	4
5	13	6	1	13
6	12	6	*	1
7	12	2	*	3
8	10	5	*	1
9	10	5	1	10
10	10	5	*	1

Researcher B				
No. of Papers	No. of Citation	Number of Authors	Author Position	Number of Citation after Weighted
1	40	10	*	2
2	32	14	*	1
3	25	10	*	1
4	25	24	*	1
5	22	5	*	2
6	18	6	*	2
7	16	148	*	0
8	15	10	*	1
9	14	5	*	1
10	14	2	1	14

*: author is a co-author.

Discussion

Table 1. shows the comparison of the authors based on number of citations, number of papers and authors position. In the initial investigation we used 10 highest citation of the authors' papers.

Table 2. shows the results of the H-index calculation and RAMa-index calculation. Researcher A gets the RAMa-index 11 and Researcher B gets the RAMa index value 1. The calculation results show that the RAMa-index method is able to filter out citation data that is not actually the impact of the researcher as a main author, so that it is not included in the calculation of the researchers' impact.

Table 2. Comparison of the H-index and RAMa-index value of the authors.

Researcher A			Researcher B	
H-index	RAMa-index		H-index	RAMa-index
10	11		10	1

Conclusion

This work is a research in progress. The model was proposed to give weighting value of Indonesian researcher citation. The citation data was filtered by the combination of PCI and EC methods. The initial investigation shows that the weighting mechanism of the *maximization* can be used to filter the "spike" citation of the authors. We will continue to optimize the weighted values of the co-authors. In the future, it is planned to test the model by the real data of the Indonesian researchers will be conducted.

References

BiHui, J. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, *52*(6), 855–863. https://doi.org/10.1007/s11434-007-0145-9

Egghe, L. (2006). Theory and practise of the g -index. *Scientometrics*, *69*(1), 131–152. https://doi.org/10.1007/s11192-006-0144-7

Gagolewski, M., & Grzegorzewski, P. (2009). A geometric approach to the construction of scientific impact indices. *Scientometrics*, *81*(3), 617–634. https://doi.org/10.1007/s11192-008-2253-y

Gamboa, C. (2014). SAGE statement on Journal of Vibration and Control. Retrieved February 6, 2019, from https://uk.sagepub.com/en-gb/asi/press/sage-statement-on-journal-of-vibration-and-control

Glänzel, W., Thijs, B., & Debackere, K. (2016). Productivity, performance, efficiency, impact—What do we measure anyway? *Journal of Informetrics*, *10*(2), 658–660. https://doi.org/10.1016/j.joi.2016.04.008

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Mesiar, R., & Gagolewski, M. (2016). H-index and Other Sugeno Integrals: Some Defects and Their Compensation. *IEEE Transactions on Fuzzy Systems*, *6706*(c), 1–1. https://doi.org/10.1109/TFUZZ.2016.2516579

Rochim, A. F., Muis, A., & Sari, R. F. (2017). Discrimination Measurement Method on H-index and G-index Using Jain ' s Fairness Index. *ISSI 2017 - 16th International Conference Scientometrics and Informetrics*, 446–447.

Rochim, A. F., Muis, A., & Sari, R. F. (2018). Improving Fairness of H-index : RA-index. *Journal of Library and Information Technology*, *38*(6), 378–386.

Witold Kienc. (2015). Should you care about your h-index, and if so, how to improve it? | Open Science. Retrieved October 11, 2018, from https://openscience.com/should-you-care-about-your-h-index-and-if-so-how-to-improve-it/

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427. https://doi.org/10.1002/asi.23179

Research on the Development Trend of Ships Diesel Engine Based on Patentometrics

Rongying Zhao [1], Danyang Li [2] and Xinlai Li [3]

[1] *zhaorongying@126.com*
Research Center for Chinese Science Evaluation, Wuhan University, Wuhan 430072 (China)

[2] *whusimldy@163.com*
School of Information Management, Wuhan University, Wuhan 430072 (China)

[3] *lixinlai_whu@163.com*
School of Information Management, Wuhan University, Wuhan 430072 (China)

Introduction

This paper demonstrates the development trend from the perspectives of technology development life cycle and direction of research and development based on the research about the patents of ships diesel engine, and the direction of research and development is illustrated from three aspects including technology concentration, industry concentration and regional diffusion. We use the methodologies and tools of social network clustering, technology life cycle S curve, visual analysis and so on.

It is found that the technology of ships diesel engine is more competitive and the technology tends to be saturated. The coverage is quite complete from the hull and the ships diesel propulsion system design to the internal combustion engine design. Germany, Japan and South Korea have strong competitiveness in this field and established technological advantages in this field.

Data and Methods

The research object of this paper needs to obtain patent data of many countries around the world, especially in the leading countries of shipbuilding industry (the United States, Japan, South Korea, etc.). So we choose to use Derwent Innovations Index (DII) to search.

In this paper, an exhaustive search strategy is adopted to improve the comprehensiveness of the search results based on the English search keywords related to the topic, and the search characteristics of the database are adjusted. As of April 12, 2018, a total of 386 records had been retrieved, and the result after removing the duplication was 313.

Although the exhaustive strategy was adopted to ensure the completion rate of retrieval, the accuracy rate was not improved. In this regard, we select all DC classification Numbers to formulate co-occurrence matrix, conducts aggregation subgroup analysis, and obtains classification number clustering related to the topic, as shown in figure 1.

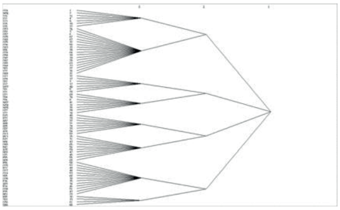


Figure 1. Aggregation Subgroup Analysis of Patent DC Classification Number of Ships Diesel Engine.

61 DC classification Numbers were divided into 8 clusters based on the co-occurrence matrix of DC classification Numbers. DC classification Numbers with frequency greater than 50 were selected for further analysis, mostly concentrated in three regions. We think that the classification Numbers of these three regions are related to the research topic of this paper, with a total of 32 DC classification Numbers. In order to improve the pertinently of patent technical analysis, the deduplication data were screened according to the above classification number. Finally, 276 records with strong correlation classification number were obtained.

Results and Discussion

Life cycle analysis of patent technology development of ships diesel engine

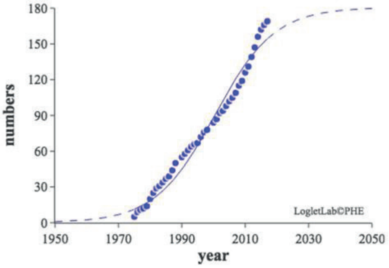


Figure 2. Fitting Diagram of Technology Development Life Cycle

Analysis of patent technology research and development direction of ships diesel engine

Technology concentration analysis.

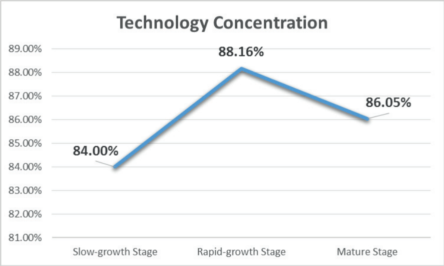


Figure 3. Technology Concentration Analysis

Analysis of concentration degree of patentee.

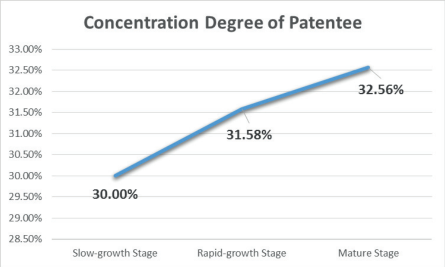


Figure 4. Analysis of Concentration Degree of Patentee

Analysis on concentration degree of dominant patentee.

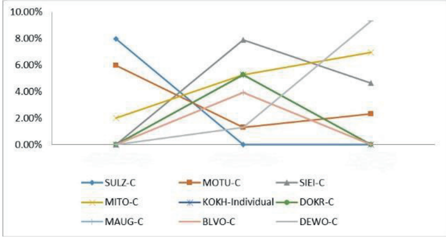


Figure 5. Concentration of Representative Patentees in Each Stage

Regional diffusion analysis.

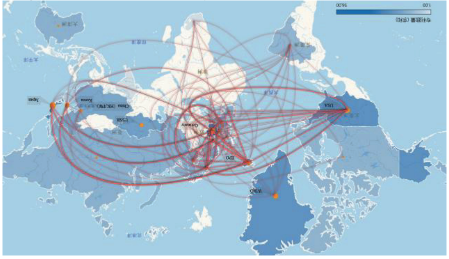


Figure 6. Regional diffusion in the slow-growth stage

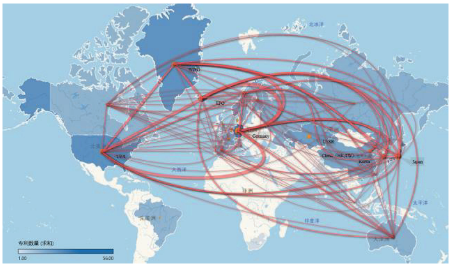


Figure 7. Regional diffusion in the rapid-growth stage

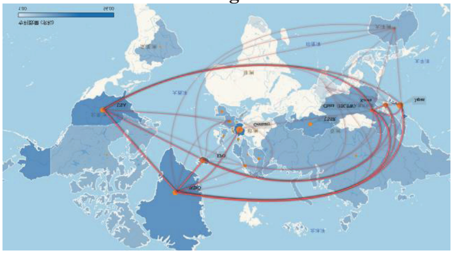


Figure 8. Regional diffusion in the mature stage

Acknowledgments

This research is funded by the National Social Science Found Major Project of China (18ZDA325) , the National Social Science Found Project of China (16BTQ055) and the Fundamental Research Funds for the Central Universities (Independent Research Projects of Wuhan University (Humanities and Social Sciences)).

References

Lijun Fu, Lufeng Liu, Gang Wang, Fan Ma, Zhihao Ye, Feng Ji, Luhui Liu. (2016). The research progress of the medium voltage DC integrated power system in China. Chinese Journal of Ship Research 11(01), 72-79.

Weiming Ma. (2015). Electromechanical Power Conversion Technologies in Vessel Integrated Pow-er System. Journal of Electrical Engineering 10(04), 3-10.

Project Napier sees twin-track plan adopted to resolve Type 45 problems, Warship Technol-ogy, https://www.rina.org.uk/2016_Editions3.html, 2016(7/8):13-16[2018-5-28].

Brockhoff K. (1992). Instruments for Patent Data Analysis in Business Firms. Technovation (12),41-58.

Dejing Kong, Kun Wang. (2018). To Assess a Technology Investment based on Patent Data——From an Empirical Analysis on Express Logist. Journal of Technical Economics & Man-agement (8), 14-20.

Idea Diffusion Patterns: SNA on Knowledge Meme Cascade Network

Zhentao Liang, Jin Mao*, Yujie Cao and Gang Li

*liangzht5@whu.edu.cn, * maojin@whu.edu.cn, cathy0021@163.com, imiswhu@aliyun.com*
Wuhan University, Center for the Studies of Information Resources, Bayi Rd. #299, Wuhan, Hubei(China)

Introduction

A meme in culture plays a fundamental role as carrying cultural ideas, behaviours, or practices that spread among people (Dawkins, 1976). Knowledge system can be modelled as a set of ideas in an idea space, where the diffusion of ideas is crucial to knowledge creation and growth (Olsson, 2000). Recently, the concept of *scientific meme* is proposed in the study of scientific information. Analogy to genes acting as a significant inheritance in evolution, scientific memes play a similar role in shaping the evolution of science through the spread of knowledge (Kuhn, Perc & Helbing, 2014).

Citations among papers symbolize the spread of scientific knowledge, which is the basis of many current studies on scientific knowledge diffusion. The pioneering work was Price (1965). One well-applied approach is to construct a citation network for the papers in a domain or multiple fields to quantify knowledge flow therein. To measure the impact and diffusion pattern of a single article, some recent studies focus on the citation network originated from an individual paper (Huang et al., 2018; Min et al., 2018).

The observed objects of the above studies are at the level of scientific papers. However, to our best knowledge, how ideas spread in the scientific system has not been fully exploited. We attempt to investigate on the patterns of idea diffusion through citation networks with respect to knowledge memes by applying social network analysis (SNA). This study helps us understand the diffusion process of ideas in science.

Methodology

Dataset

As a case study, we choose the emerging field of medical informatics. We downloaded 37,650 records published in the 24 journals under the category of Medical Informatics (2016 version) from Web of Science. The type of records was set as article and the publish year was restricted to the range from 1900 to 2016. The fields of title, abstract, author keyword, and Keyword Plus were parsed.

Knowledge meme detection

The method in Kuhn et al. (2014) was used to identify knowledge memes of medical informatics. The fundamental assumption is that if a term occurs in both citing paper and cited paper, the citation between the two papers indicates implicit diffusion of the term. A meme should frequently occur in papers citing meme-carrying papers, but appear rarely in papers that do not cite a meme-carrying paper. Thus, *Meme Score* is defined to rank the importance of memes. Single words, 2-grams, and 3-grams from titles and abstracts, as well as author keywords and Keyword Plus terms, constitute the vocabulary of potential memes. The *Meme Score* of all items in the vocabulary were calculated and 18,995 memes with Meme Score greater than 0 were determined as memes to be analysed in this research.

Knowledge meme cascade network construction

A citation cascade of a paper is a citation tree originated from the paper and consisting of its following offspring citing papers (Mazloumian et al., 2011). Similarly, a cascade network of a knowledge meme is defined in this study as the citation tree only composed of publications containing the specific meme. For example, Figure 1 shows the cascade network of the meme of *RFID*. By this method, cascade networks were constructed for all memes. Isolated papers that even contain the meme but have no citation relation were excluded since they do not lie in any diffusion path of the meme.

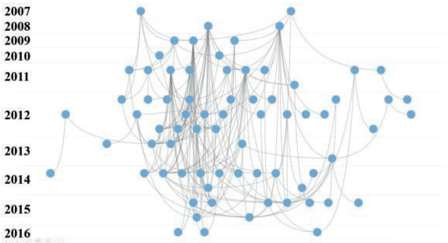


Figure 1. Cascade network of *RFID*

Results and Discussion

The basic metrics were obtained for all knowledge meme cascade networks and listed in Table 1. It is shown that on average, knowledge meme cascade networks have a small cascade depth (aka. network diameter, 1.89) and a low clustering coefficient (CC, 0.09). The structure is different from small-world networks that typically have a small diameter but a relatively large clustering coefficient. A possible reason is that cascade networks are treelike where branches interact less frequently.

Table 1. Basic metrics averaged over all knowledge meme cascade networks

Nodes	Edges	Depth	Density	CC	Comp.
10.01	11.05	1.89	0.71	0.09	2.34

The averaged component number (Comp.) is 2.34 as shown in Table 1. It indicates that multiple separated subgraphs exist in knowledge meme cascade networks. This is a major difference between knowledge meme cascade networks and paper citation cascades that have only one subgraph. A citation cascade starts from a publication, while a knowledge meme cascade network may have multiple origins that do not cite each other. This is further justified by the number of origins in cascade networks as presented in Figure 2.

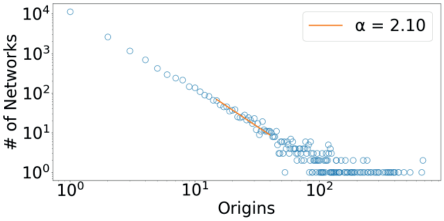


Figure 2. The distribution of the number of origins in cascade networks.

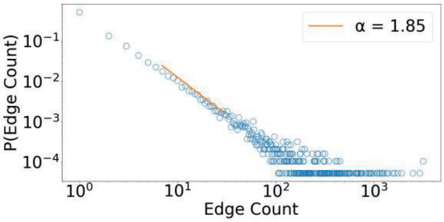


Figure 3. Edge count distribution.

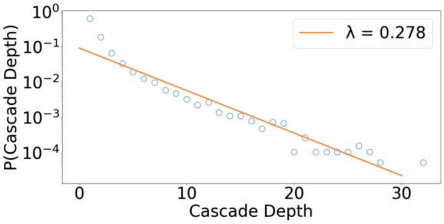


Figure 4. Cascade depth distribution.

Figure 3 shows that the edge counts of cascade networks follow a fat-tailed power law distribution. Many networks have more edges than expected by the fitted distribution, which reveals that some knowledge memes spread broadly and form extreme large cascade networks.

Cascade depth, i.e., the diameter of the cascade network, reflects the depth of influence. From Figure 4, a large portion of memes have a cascade depth smaller than 10. It is due to the fact that many papers with a meme only receive a few citations from papers carrying the meme. Cascade depth demonstrates an exponential distribution with λ=0.278.

Figure 5 presents that both in-degree and out-degree fit power-law distributions, showing a scale-free property of the diffusion. However, the out-degree curve decays faster than the in-degree curve. This is because cited paper cumulate citations from citing papers, thus between a citation pair, the in-degree of the cited paper is basically greater than the out-degree of the citing paper.

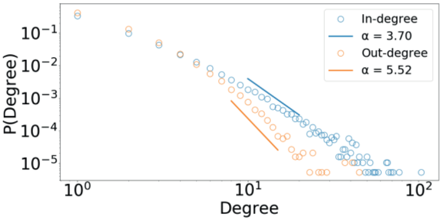


Figure 5. In-degree distribution and out-degree distribution

Conclusion

The case study shows preliminary results about diffusion patterns of knowledge memes by applying SNA on knowledge meme cascade networks. A knowledge meme cascade network may have multiple origins. Edge counts, in-degree and out-degree of meme cascade network show power-law distributions, while cascade depth fits exponential distribution. More diffusion patterns could be identified and the diffusion differences between memes and papers are to be compared among multiple disciplines in future.

References

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2018). Number versus structure: towards citing cascades. *Scientometrics*, 117(3), 2177-2193.

Kuhn, T., Perc, M., & Helbing, D. (2014). Inheritance patterns in citation networks reveal scientific memes. *Physical Review X*, 4(4), 041036.

Mazloumian, A., Eom, Y. H., Helbing, D., Lozano, S., & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and Nobel prizes. *PloS one*, 6(5), e18975.

Min, C., Ding, Y., Li, J., Bu, Y., Pei, L., & Sun, J. (2018). Innovation or imitation: The diffusion of citations. *Journal of the Association for Information Science and Technology*, 69(10), 1271-1282.

Olsson, O. (2000). Knowledge as a set in idea space: An epistemological view on growth. *Journal of Economic Growth*, 5(3), 253-275.

Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 510-515.

Article-level matching of Web of Science to a local database in a comparative context

Linda Sīle[1] and Raf Guns[1]

[1]*linda.sile@uantwerpen.be, raf.guns@uantwerpen.be*
University of Antwerp, Faculty of Social Sciences, Centre for R&D Monitoring (ECOOM), Middelheimlaan 1, 2020 Antwerp (Belgium)

Introduction

The low coverage of social sciences and humanities (SSH) journals in Web of Science (WoS) is well known (Kulczycki et al., 2018; Ossenblok, Engels, & Sivertsen, 2012). Over years the coverage, however, has been increasing and more journals are indexed. At the same time, these developments highlight the need for a continued monitoring of coverage.

To monitor coverage, one requires comprehensive bibliographic data on research output as reference data and a sound technique to identify which articles in this reference dataset are indexed in Web of Science. The challenge is to find an approach where one would have reasonable balance between accuracy and the time required for article matching. Here we describe an article-level approach.

Context

The search for an approach suitable for the use in a comparative context emerged in the context of bibliometric analyses based on data from two different national bibliographic databases (VABB-SHW in Flanders, Belgium and Cristin in Norway). That study, although not focused on WoS coverage, required information on WoS indexation (for further details see Sīle et al. 2019).

Our goal is to identify which articles can be matched to a record in data retrieved from WoS. In this matching we strive for maximum accuracy and speed, and minimum number of metadata categories. The latter is especially crucial when working in a comparative context, where different sources do not always have the same metadata.

Article-level approach to be used in comparative settings

Data

The proposed matching procedure is applied to two datasets derived from two national bibliographic databases (VABB-SHW in Flanders, Belgium and Cristin in Norway). The datasets are limited to journal articles (2006-2015) in social sciences and humanities (SSH) by authors affiliated to universities (n_{Flanders} = 31,550; n_{Norway} = 26,007). These datasets are referred to as the reference datasets.

For WoS, we use datasets retrieved from the ECOOM-Leuven in-house WoS database. We delineate the data by country (Belgium or Norway), year (2006-2015) and indices (SCIE, SSCI, and AHCI). These datasets henceforth are referred to as the WoS datasets.

Our approach combines algorithmic and manual steps. In brief, we match bibliographic data from VABB-SHW and Cristin with the WoS-datasets. This matching is done in three steps: we identify records automatically, first, with identical metadata, and, second, with approximately identical metadata. Finally, we identify matching records (semi-) manually. For the overview of results see Table 1.

Table 1. Results from article-level identification of indexation in Web of Science

	Flanders		Norway	
	#	%	#	%
Step 1	8533	63	7476	79
Step 2	3904	29	1577	17
Step 3	1111	8	400	4
Total	13548	100	9453	100

Step 1. Identical matches

First, we identify matching records using the following rule: (1) identical title of the article (punctuation removed, case ignored), AND (2) identical page numbers, AND (3) identical ISSN, AND (4) identical publication year.

Step 2. Approximate matches: LSH

Occasionally identical records are not identified due to discrepancies in bibliographic control practices or simply due to inaccuracies in records. For instance, titles, especially if reported by authors themselves, sometimes do not exactly match the title as it appears on the published version. The same applies for ISSNs, page numbers, titles of journals, etc. While approximate string matching by e.g. edit distance can theoretically offer a solution, the number of comparisons quickly grows too large to be feasible in practice. Following Abdulhayoglu and Thijs (2018), we use a solution based on Locality Sensitive

Hashing (LSH). More specifically, we use an LSH Forest (Bawa et al., 2005), which allows to retrieve the top-n best matches by estimated Jaccard similarity, as implemented in DataSketch (Zhu & Markovtsev, 2017).

We compiled 'reference' strings for all records in both the reference and WoS datasets using the title of an article, first author, page numbers, and the journal title (e.g., 'jacobs, s 2015 consumers' health risk-benefit perception of seafood and attitude toward the marine environment: insights from five european countries environmental research 11 19'). Afterwards, each reference was converted into a set of 3-grams. We then identified the three most likely matching reference records for each WoS record, using the LSH Forest.

For each pair of potential matches, we computed six similarity scores: Jaccard similarity for titles, arithmetic difference for publication year, arithmetic difference for start and end page numbers, Levenshtein distance for ISSN, and Levenshtein distance for authors. Combinations of these measures were explored in relation to the reference dataset (the VABB-SHW). First, we explored the distributions of these difference measures for the reference dataset: correct matches/incorrect matches/all. The aim here was to find a minimum set of rules that, on the one hand, identifies the largest number of correct matches but, on the other hand, includes as few false matches as possible (<10). This iterative exercise led to a set of rules, which were applied sequentially. This step led to the identification of 3904 additional records for Flanders and 1577 records for Norway.

Step 3. Semi-manually matched records

Finally, matching records were searched manually. This was necessary to identify articles the titles of which, for example, are published in non-English languages (e.g. in Dutch or Norwegian)in WoS are indexed in English. This was done only for records that had an ISSN in the WoS datasets. Using the ISSN, pairs of potential matches were generated. For each pair, we calculated similarity measures for each pair: arithmetic difference for publication year, arithmetic difference for start and end page numbers, Levenshtein distance for the first author, and Levenshtein distance for titles. All record pairs that had the following difference (or smaller) in any of these categories were checked manually: publication year +-2 OR one exact page number match OR first author match OR Levenshein difference for titles <20. This step led to the further identification of 1111 records for Flanders and 400 records for Norway.

Conclusion

This approach allow to identify a substantial number of records relatively quickly. If we put aside the time that was spent to develop this approach, the identification of WoS-indexed records in document sets consisting of 31,550 or 26,007 records required approximately 10 hours (most of the time was required to carry out the manual matching in the step 3). Overall, this approach requires considerably less time than attempting to match the same amount of records manually or by 'naïve' approximate string matching.

In this approach we used basic bibliographic metadata categories that are not database-specific and are included in most data sources. This means that the same approach can easily be used to match bibliographic data from other data sources.

For future development, it is possible to improve this method by exploring possibilities to use more bibliographic data categories. Also, further work can be done in the use of similarity measures, which, in combination with more fine-tuned algorithms could lead to improved accuracy.

References

Abdulhayoglu, M. A., & Thijs, B. (2018). Use of locality sensitive hashing (LSH) algorithm to match Web of Science and Scopus. *Scientometrics*, 116(2), 1229–1245. https://doi.org/10.1007/s11192-017-2569-6

Bawa, M., Condie, T., & Ganesan, P. (2005). LSH Forest: Self-tuning Indexes for Similarity Search. *Proceedings of the 14th International Conference on World Wide Web*, 651–660. https://doi.org/10.1145/1060745.1060840

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., … Zuccala, A. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*. https://doi.org/10.1007/s11192-018-2711-0

Ossenblok, T. L. B., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science--a comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation*, *21*(4), 280–290. https://doi.org/10.1093/reseval/rvs019

Sīle, L., Guns, R., Vandermoere, F., & Engels, T.C.E. (2019). Comparison of classification-related differences in the distribution of journal articles across academic disciplines: the case of social sciences and humanities in Flanders and Norway (2006-2015). *Proceedings of ISSI 2019*.

Zhu, E. & Markovtsev, V. (2017).DataSketch. *Zenodo*. http://doi.org/10.5281/zenodo.290602

Public Administration and Social Media: An analysis of the journal literature

Alessandra Ordinelli[1], Barbara Colonna[1] and Carla De Iuliis[1]

[1]a.ordinelli@izs.it
Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM), 64100, Teramo, Italy.

[1]b.colonna@izs.it
Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM), 64100, Teramo, Italy.

[1]c.deiuliis@izs.it
Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM), 64100, Teramo, Italy.

Introduction

The recent years have been characterized by an increase in Social Media applications in Public Administrations (PA), giving rise to a new form of Institutional Communication.

This new type of communication, incorporates traditional communication channels (one-to-many) (e.g. Newspaper, Radio, Television), with Social Media communication (many-to-many) (e.g., Facebook, LinkedIn, wikis, YouTube).

Governments are adopting Social Media to provide complementary information dissemination, communication, and participation channels whereby citizens can access government and government officials and therefore make informed decisions (Song, Ch. & Lee, J. 2015).

The aim of this study is to present an overview of the scientific production (publications) concerning the relation between PA and Social Media by using a scientometric analysis.

Materials and methods

The data set was obtained from Advanced Search Function of Web of Science Database (WoS) (Reuters, T. 2014), that uses field tags, Boolean operators, and query sets to create specific queries. Then we analysed the data using Biblioshiny, a shiny app providing a web-interface of the Bibliometrix R-package.

Bibliometrix R-package is a tool for quantitative research in scientometrics and bibliometrics. It provides various routines for importing bibliographic data from Scopus, Web of Science, PubMed and Cochrane databases, performing bibliometric analysis and building data matrices for co-citation, coupling, scientific collaboration analysis and co-word analysis (Aria, M. & Cuccurullo, C. 2017).

Results

We have obtained that 1469 authors have written a total number of 611 documents (as Article, Book, Review, Proceedings Paper), of which 272 articles from 2000 to 2018 years. The number of publications shows that researches have grown exponentially since 2007 and that the trend has continued at relatively stable rates with a peak in 2015 (Fig. 1.).

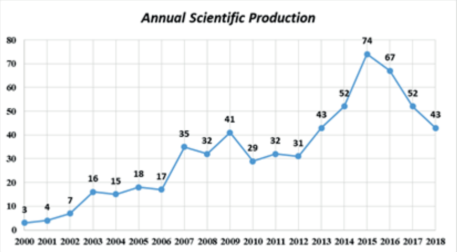


Figure 1. The WoS publication from 2000 to 2018

In addition, to the study on publication growth, we carried out an analysis about the Word Dynamic Graph (Fig. 2.) which helps to understand the keyword dynamics over time.

The results of Figure 2. show the five keyword dynamics: the two keywords “e-government" (61 occurrences) and “management” (44) are the most dynamic between 2014 and 2018.

In particular, “e-government” represents the digital administration that uses information and communication technologies (ICT) (including social media) to ensure PA efficiency, improving the quality of services for citizens and decreasing costs for the community.

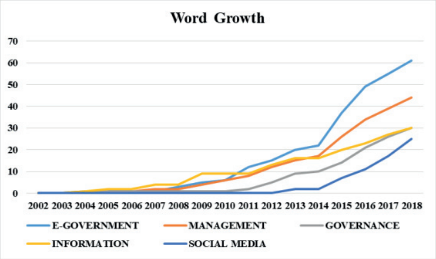


Figure 2. Word Dynamics Graph

Furthermore, we analysed the contributions of countries and the interaction between authors. In particular, the Countries' Scientific Productions are shown in Table 1. As we can observe from the results, the researchers from the USA (102 papers), Italy (93 papers), and Spain (80) have had a major role in scientific publications about the topics we selected.

Table 1. First Ten Country Scientific Production

Country	N° Production
USA	102
Italy	93
Spain	80
China	58
Russia	45
Germany	40
Romania	37
Portugal	36
Greece	35
Brazil	30

At the same time, the analysis of the Country Collaboration Map (Fig. 3.) shows that there are not strong collaborations between the researchers of different countries: the most significant link is between the USA and China.

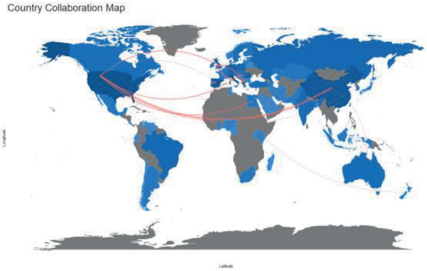


Figure 3. Country Collaboration Map

Discussion

We have applied an analysis of the journal literature to understand the current intellectual core between PAs and Social Media compared to the research productivity of individual authors and countries.

Finally, we reviewed and analysed 276 articles from 66 different countries.

The data show how the study on PAs and Social Media have evolved over time and the constant flow of scientific productions on these topics has created continuous change in keywords. In fact, it is possible to observe (Fig. 2) that some words are present every year, such as "governance"; others disappear or emerge over time such as "social media", used the first time in 2013. In addition, the analysis of Word Growth allows to observe the growing or declining trend of the keywords; it can help us to make a selection for the specific topic or using the keywords can attract researchers and consequently potential bibliometric citations.

The studies about Scientific Productions and Country Collaboration suggests that the PA production in Social Media is not distributed equally among nations and there is a lack of communication among the researchers. This may be due to the fact that there is lack of legislation at global and European level regarding Social Media in PA. In addition, PAs may use different Social Media implementation strategies to interact with citizens or study specific cases regarding their own countries, such as the studies by Pr. David Špaček about "Social Media Use in Public Administration: The Case of Facebook Use by Czech Regions".

Conclusions

This study has tried to provide an analysis of the journal literature related to PAs and Social Media.

The data show that the scientific production related to PAs and Social Media is growing, with marked differences in terms of quantity, quality and international collaboration. Specifically, the analysis of Country Collaboration Map shows that the United States and China are pioneers on this topic.

Future studies could improve the link in the international research network and contribute to set the basis for adequate legislation at all levels, and a proper use and efficiency of Social Media in PAs.

References

Aria, M. & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), pp 959-975, Elsevier.

Reuters, T. (2014) Web of Science Core Collection. Available from: http://thomsonreuters.com/web-of-science-core-collection/

Song, Ch. & Lee. J. (2015). Citizens' Use of Social Media in Government, Perceived Transparency, and Trust in Government. *Public Performance & Management Review* 39, 430 – 453.

Špaček, D. (2018). Social Media Use in Public Administration: The Case of Facebook Use by Czech Regions. *NISPAcee Journal of Public Administration and Policy* 11(2):199-218.

Developing a rule-based method for identifying researchers on Twitter: The case of vaccine discussions

Björn Ekström[1]

[1] bjorn.ekstrom@hb.se
University of Borås, The Swedish School of Library and Information Science, Allégatan 1, 503 32, Borås, (Sweden)

Introduction

This study seeks to develop a method for identifying the occurrences and proportions of researchers, media and other professionals active in Twitter discussions. As a case example, a dataset from Twitter vaccine discussions is used. The study proposes a method of using keywords as strings within lists to identify classes from user biographies. This provides a way to apply multiple classification principles to a set of Twitter biographies using semantic rules through the Python programming language.

Theory

The theoretical outline is based on rule-based text classification. As described by Glushko (2013, 374), a rule-based system can serve to separate words in terms of tokenization, where textual components are divided using spaces, and stemming, where terms are derived to their word stems. While the rule-based process provides domain-based classification, issues may occur with regards to how punctuation complicates tokenization and how semantic ambivalence can occur from incorrect stemming.

Method

9 647 plain text biographies from Twitter profiles engaged in discussions related to vaccines are studied as a prominent case example. The case dataset is provided through the research project Data for Impact. The method includes a qualitative content rule-based analysis process using the Python programming language and data wrangling software OpenRefine where patterns within the biographies are set to correspond to predefined classes. A set of keywords as strings within lists are represented by variables. Each variable is then matched against the biographies as plain text and returns one of the predefined classes if any of the strings are present.

Strings used to identify biographies are influenced by and partially reused from previous studies (Côté and Darling 2018; Vainio and Holmberg 2017), although amended in order to suit the nature of the biographies used as a dataset in this study. As discussed by Patton (2015), the identification process is performed by working back and forth between the classes and the data in order to verify accuracy. Eleven types of classes are used, as described in Table 1, corresponding with a set of keywords. The class *General public* is used when the biographies does not match any class. Twitter profiles lacking biographies are classed as *Unknown*. Users can also belong to more than one class. Spelling variations are used where needed.

Table 1. Classes, keywords and biography extracts.

Class	Keyword example	Biography extract example
Science student	student, phd student, phd candidate	[City] University [discipline] Student
Graduated	MS, MA, graduate	[…] Engineering graduate. […]
University faculty	lectur, prof., professor	Professor of [discipline], teaches [subjects].
Other scientist or science-associated group	technician, lab manager, ologist	[…] biologist […]
Education and outreach professionals	curator, teacher, librarian	Language teacher [subject]
Applied science organization	nonprofit, policy officer	[…], nonprofit board member […]
Other professional	recruiter, entrepreneur, manager	Entrepreneur, marketer […]
Media professional	journalis, corresponden, publisher	correspondent for [media outlet]
Policy/decision maker	congressman, senator, parliament	District […] Congressman [year span]
General public		
Unknown		-

Findings

The findings of the classification and their occurrences are presented in Table 2 below.

Table 2. Occurrences and proportions of classes.

Class	**No.**	**% (out of 10255 classes)**
Science student	165	1.61 %
Graduated	58	0.57 %
University faculty	191	1.86 %
Other scientist or science-associated group	394	3.84 %
Education and outreach professional	283	2.76 %
Applied science organization	56	0.55 %
Other professional	704	6.86 %
Media professional	1127	10.99 %
Policy/decision maker	23	0.22 %
General public	7188	70.09 %
Unknown	66	0.64 %
Total	**10255**	

As per this case example, academic professionals, organizations and students are engaged to the following extent and order in relation to the total number of classes identified (10 255): *Other scientist or science-associated group* (3.84 %), *Education and outreach* professional (2.76 %), *University faculty* (1.86 %), *Science student* (1.61 %), *Graduated* (0.57 %), *Applied science organization* (0.55 %). The proportion of *media professionals* amounts to approximately a tenth of all classes (10.99 %) while the class of *other professionals* such as recruiters, entrepreneurs and managers amounts slightly lower (6.86 %). A substantial share of the Twitter profiles engaged (70.09 %) does not belong to any of the professionally related classes but rather belongs to the class of *General public.* The classes of *Policy/decision makers* and *Unknown* relates to small proportions (0.22 % and 0.64 % respectively).

The method does provide a certain error margin when examining the outcome through close-reading. For instance, a biography simply mentioning the word "scientist" may be classed as *University faculty.* Although tendencies can be examined on the occurrences and proportions of academic and media voices in the Twitter vaccine discussion, the biographies' free form provides some classification noise.

Conclusion

The rule-based classification process presented provides a method of identifying the occurrences and proportions of researchers and other professionals engaged in discussions related to vaccines based on a set of predefined rules. Keywords as strings within lists are matched to user biographies collected from Twitter. The study has proven to give the project an indication of the relevant share of the collected data. Of these, 7.88 % are academic (class 1 - 4), 3.31 % are academically related (class 5 - 6) and 10.99 % are media related (class 8). 7.08 % consist of other classes (class 7 + 9). 70.09 % are classed as the *General public* (class 10) and 0.64 % are classed as *Unknown* (class 11).

While prior studies have used search-and-replace methods through regular expressions, the method proposed provides a way to apply multiple classification principles to a set of Twitter profile biographies using the Python programming language and data wrangling software OpenRefine. This enables a better understanding of the the occurrences and proportions of researchers as well as other professionals being present in Twitter discussions. Future studies on new classification methods with regards also to natural language processing are needed in order to further develop such methods.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 770531. Special thanks to Ettore Rizza – PhD Student in Information Sciences & Technologies at Université libre de Bruxelles, Belgium – for programming advice.

References

Côté, I. M., & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops? *FACETS*. https://doi.org/10.1139/facets-2018-0002

Glushko, R. J. (2013). *The Discipline of Organizing*. Cambridge, Massachusetts: The MIT Press.

Patton, M. Q. (2015). *Qualitative research & evaluation methods : integrating theory and practice*. Thousand Oaks, California: SAGE Publications, Inc.

Vainio, J., & Holmberg, K. (2017). Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions. *Scientometrics*, 112(1), 345–366. https://doi.org/10.1007/s11192-017-2368-0

Research on Identification and Selection on Key Fields of Science and Technology

Hui Wang[1] and Xiaowei Yang[2]

[1] *wanghui@mail.las.ac.cn*
National Science Library，Chinese Academy of Sciences，33 Beisihuan Xilu, Zhongguancun , Beijing (P.R.China)
University of Chinese Academy of Sciences，No.19(A) Yuquan Road, Shijingshan District, Beijing (P.R.China)

[2] *yangxiaowei@caas.cn*
Agricultural Information Institute of Chinese Academy of Agricultural Science, No.12 Zhongguancun South St.,Haidian District, Beijing (P.R.China)

Introduction

The selection of key areas of scientific and technological development reflects the national S&T development goals and strategies. It plays an important role in obtaining major opportunities for scientific innovation, finding the combination of scientific development, national economic and social development goals, guiding and supporting basic and strategic S&T research.

In this study, we take the United States, Britain, Japan, South Korea, the European Union and China as the research objects, explore the method of combining quantitative analysis and qualitative research with multiple sources, design the multi-dimensional comprehensive analysis indicators and weights based on expert advices, select and identify key areas of S&T development in the target countries. The main methods of information collection and analysis used in this study are expert consultation, bibliometric analysis (Alberto M., Elisa B. & Geraldine J., 2018), text mining (Xin L, Qianqian X & Tugrul D. , 2019).

The results provide a reference basis for the planning and layout of key areas of S&T development in China objectively and systematically from a global perspective.

Data Sources and Methods

The analysis indicators and weights are determined from three dimensions of the past, present and future. The S&T achievements, S&T input and S&T strategic plans are corresponding to three first-level indicators, each first-level indicators included 2-3 secondary indicators. Weights are determined by experts' ranking of the importance of three dimensions and indicators. The Indicators framework and weights detailed in Table 1.

Data Sources of second-level Indicators

The articles in the last five years from the Web of Science Core Collection database and Incites journal citation reports are used to estimate the past dimension. The secondary indicators are the discipline scale, discipline influence and expert judgement. The S&T projects funded by the selected countries are used to estimate the present dimension. The secondary indicators are awards count and award amount. The national and international S&T strategic plan in these countries and region are used to estimate the future dimension. The secondary indicators are research fronts and expert forecast.

Analysis method

The data sources of the three dimensions are different in domain classification. This study built a 3-level mapping table (Table 2） that incorporated different topic descriptions into the unified fields based on subjects classified by the National Natural Science Foundation of China, Research Areas in NSF and expert consultation.

Articles published in the past 5 years are retrieved and divided by countries/region and categories. the articles count in various disciplines are calculated and normalized. The analysis method of the discipline scale indicator was shown in Table 3 and the formula.

$$Y_n = weight * (X_n - \mathrm{MIN}(X_n))/(\mathrm{MAX}(X_n) - \mathrm{MIN}(X_n))$$

The S&T funding programs statistics information is used as input value and calculated using the formula above to output the awards count and award amount indicators.

The experts take the outline and the topic frequency which are text mined from the full-text of national and international S&T strategic plans for reference, combine their professional knowledge to estimate the research fronts and forecast indicator values.

According to the indicators and weights in table 1, The sum of three dimensions of the key fields is calculated and ranked in the major countries/ region. We use the Min-max standardized extremum scaling method to map all sums into the range of [0, 1]. The closer the value is to 1, the more important the field is, the national/regional top 10 S&T key fields are identified and ranked by the standardized value.

Results

The Top10 key S&T fields in countries/region are selected and identified. There are 25 key S&T fields involved by comparing the Top10 key S&T fields in different countries/region.

The key fields are Material Genome, Artificial Intelligence and Robot and Intelligent Manufacturing, Biomedical Engineering and Biomaterials, Safe and Effective Medicine and Medical Health, Renewable Energy and Effective Utilization and Green Energy, Global Climate and Environmental Change, Transgenic Biotechnology and Gene Breeding, Large Data Analysis and High Performance Computing and Application, Nanomaterials and Nanotechnology, Gene Medicine, Space Technology, Automobile Technology, Intelligent Transportation System, Prevention and Treatment Of Major Non-Communicable Diseases, New Quantum Devices and Quantum Information Security Technology, Information and Communication Technology, High Temperature Superconducting Technology and Materials, Brain and Cognitive Science, Health Information Technology, Food Safety, Synthetic Biology and Biomanufacturing, Digital Network, Energy Saving Technology,3D Printing, Disaster Prediction Technology.

Discussion and Conclusions

The results are based on an investigation and analysis conducted in 2015. In the 2019 study, this study will expand the coverage of funding and strategic information. The indicators of the third dimension depend on the subjective judgment of the expert team, the representativeness and academic reputation of the expert team will be improved. With the rapid development of science and technology, mapping tables need to be further supplemented and improved.

Table 1. Indicators and weights for comprehensive analysis of key fields

First-level Indicators	**weight**	**Second-level Indicators**	**weight**
Research achievements (Past)	25%	Discipline scale	10%
		Discipline influence	12.5%
		Expert judgement	2.5%
S&T input (Present)	35%	Awards count	15%
		Award amount	20%
S&T Plans (Future)	40%	Research fronts	25%
		Expert forecast	15%

Table 2. 3-level mapping table

Level 1-Discipline	*Level 2-Research field*	*Level 3-Topic*
Discipline 1	Research field 1	Topic 1, ……, Topic n
	……	Topic 1, ……, Topic n
	Research field n	Topic 1, ……, Topic n
……	Research field 1	Topic 1, ……, Topic n
	……	Topic 1, ……, Topic n
	Research field n	Topic 1, ……, Topic n
Discipline n	Research field 1	Topic 1, ……, Topic n
	……	Topic 1, ……, Topic n
	Research field n	Topic 1, ……, Topic n

Table 3. Analysis method of the discipline scale indicator

Discipline	*Article Count*	*Discipline scale indicator*
Discipline 1	X_1	Y_1
……	……	……
Discipline n	X_n	Y_n

References

Alberto M., Elisa B. & Geraldine J. (2018). A bibliometric-based technique to identify emerging photovoltaic technologies in a comparative assessment with expert review. *Renewable Energy,* 123, 407-416.

Xin L, Qianqian X & Tugrul D. (2019), Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology, *Technological Forecasting and Social Change*(In Press)

Bibliometric differences between funding and non-funding papers on substance abuse scientific research

Juan Carlos Valderrama-Zurián[1], Lourdes Castelló-Cogollos[2], David Melero-Fuentes[3], Rafael Aleixandre-Benavent[4], Francisco Jesús Bueno-Cañigral[5]

[1] juan.valderrama@uv.es
Departament d'Història de la Ciència i Documentació. Universitat de València, Spain. UISYS, Mixed Research Unit, CSIC-University of Valencia, Palau de Cerveró, Plaza Cisneros, 4, 46003 València. Spain.

[2] lourdes.castello@uv.es
Departament de Sociologia i Antropologia Social. Universitat de València, Spain. Palau de Cerveró, Plaza Cisneros, 4, 46003 València. Spain

[3] david.melero@ucv.es
Universidad Católica de Valencia "San Vicente Mártir", Instituto de Documentación y Tecnologías de Información (INDOTEI). Carrer de Quevedo 2, 46001 València (Spain)

[4] rafael.aleixandre@uv.es
UISYS, Mixed Research Unit, CSIC-University of Valencia. Ingenio (CSIC-Universitat Politècnica de València), Spain. Palau de Cerveró, Plaza Cisneros, 4, 46003 València. Spain

[5] fjbueno@valencia.es
Pla Municipal de Drogodependències-UPCCA València, Ajuntament de València, Spain

Introduction

Substance abuse is considered a disease that creates serious social problems and significant health expenditure for research, prevention and treatment (Green et al, 2004). There are several worldwide research initiatives aimed at better understanding the problem and seeking solutions to mitigate it (Savic et al, 2017; Moulahoum et al, 2019).

The economic support of research projects is essential for the proper performance of the scientific system and especially in areas related to the health and welfare of the population, as it facilitates discoveries and the advancement of science (Fortin and Currie, 2013).

However, the funding of the drugs of abuse research and their impact are unknown. The aim of this work is to identify the relationship between the funding resources allocated to the study of four most used drugs of abuse (cannabis, cocaine, opioids and psychostimulants), and several other variables such as the annual trend, number of authors, international collaboration, subject area and the relation between funding and citation.

Methods

A search strategy validated in a previous work (Khalili et al, 2018) and improved that combines general terms related with the abuse of substances or papers included in the research area "Substance abuse" and terms related to four specific drugs was used for the study. The search was performed on the Web of Science Core Collection (WoS) on 25 March 2019 and was restricted to the decade 2009-2018, as WoS only systematically records information on funding since 2008. Documents written in English and classified as article, review, letter and proceeding paper, were selected.

Several indices were determined to measure the relationship between the financing work and the type of drug, patterns of collaboration, subject research areas and citation. The number of citations per article has been calculated by dividing the number of citations by the number of years since their publication. We have used the "citation per paper" and the "citation difference" indices, defined by Zhao et al (2018). A bivariate analysis (Chi square tests and tests of means) was used to determine whether there were statistically significant differences ($p<0.05$) between funded papers, international collaboration and number of citations received.

Results

The total number of documents retrieved was 47,981, of which 24,589 (51.2%) dealt with opiate abuse, 13,255 (27.6%) with cocaine, 12,566 (26.2%) with psychostimulants and 11,578 (24.1%) with cannabis. 65.4% of the total number of papers was funded, with 2016 being the year with the highest percentage of funded papers.

Cocaine abuse has attracted a higher percentage of funded papers, followed by psychostimulants abuse (figure 1). The percentage of financed papers declined in recent years for all drugs except cannabis, which rose slightly.

Statistically significant differences were observed with respect to the number of authors, the existence of international collaboration and the number of citations received between funded and non-funded papers. The funded papers showed a higher mean number of authors, a higher mean of international collaboration and a higher number of citations in the articles ($p<0.001$).

Funding was higher for research on cocaine abuse and psychostimulants abuse. The subject areas most funded (with at least 100 published documents) were Virology (90.6%), Multidisciplinary Sciences (88.6%), Neuroscience (86.3%), Behavioral Sciences (84.9%) and Psychology Biological (84.3%), while those that receive the lowest funding were Law (16.2%), Criminology and Penology (17%), Emergency Medicine (26.1%), Pathology (28.9%) and Legal Medicine (29.9%).

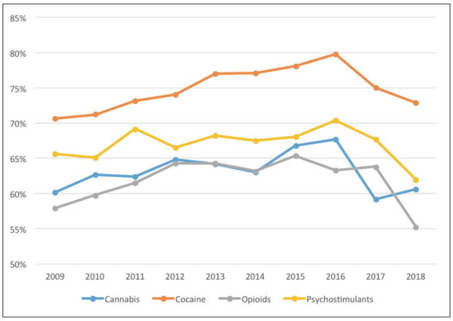


Figure 1. Annual evolution of funded papers according to type of drug of abuse

Figure 2 shows Citation per paper overall (CPP), Citation per paper in funded papers (CPPF), Citation in non-funded papers (CPPN) by type of substance. The CPP is higher in the funded papers cannabis and cocaine. The CD citation difference (CPPF/CPPN) is greater than 2.3 in all substances which means that funding increases citation.

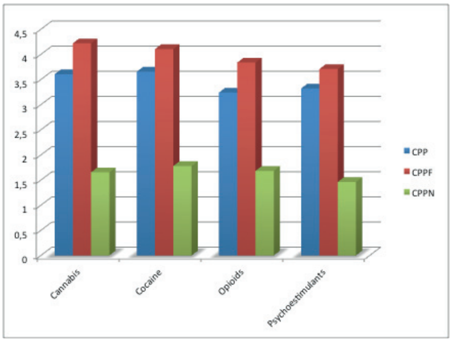


Figure 2. Citation per paper indexes by drug of abuse

Conclusions

We have identified differences in funding of most drugs of abuse that affect the annual percentage of funded papers, number of authors, international collaboration and number of citations received between funded and non-funded papers.

One limitation is that the probability of receiving citations is higher in older papers, which has been corrected by dividing citations by the years since the papers were published. Future work must deepen the analysis of these indicators according to subject areas because substance abuse research is multidisciplinary.

Acknowledgments

1. Plan Municipal de Drogodependencias; UPCCA; Ajuntament de València; Spain. 2. Plan Nacional sobre Drogas (Project n. 2016-028), Ministerio de Sanidad, Consumo y Bienestar Social, Spain.

References

Fortin, J.M., Currie, D.J. (2013). Big Science vs. Little Science: How Scientific Impact Scales with Funding. *PLoS ONE, 8(6)*, e65263.

Green, C.A., Lynch, F.L., Polen, M.R., McFarland, B.H., Dickinson D,M., Freeborn, D.K. (2004). Talk about costs: health care professionals' views about expenses related to substance abuse treatment. *Administration and Policy in Mental Health and Mental Health Services Research, 31(6)*, 483-94.

Khalili, M., Rahimi-Movaghar, A.; Shadloo B.; Mojtabai R.; Mann K; Amin-Esmaeili M.b (2018). Global scientific production on illicit drug addiction: A two-decade analysis". *European Addiction Research; 24 (2):* 60-70.

Moulahoum, H., Zihnioglu, F., Timur, S., Coskunol, H. (2019). Novel technologies in detection, treatment and prevention of substance use disorders. *Journal of Food and Drug Analysis;27(1):*22-31.

Savic, M., Best, D., Manning, V., Lubman, D.I. (2017). Strategies to facilitate integrated care for people with alcohol and other drug problems: a systematic review. *Substance Abuse Treatment, Prevention, and Policy, 12(1),*19.

Zhao, S.X., Lou, W., Tan, A.M., Yu, S. (2018). Do funded papers attract more usage? Scientometrics, 115, 153-168.

Observatory for the Scientific Evaluation of Catholic Universities in Spain, Latin America and the Caribbean

Juan-Carlos Valderrama-Zurián[1], Remedios Aguilar-Moya[2], David Melero-Fuentes[3], Rafael Aleixandre-Benavent[4] & Francisco Jesús Bueno-Cañigral[5]

[1] jc.valderrama@ucv.es
Departament d'Història de la Ciència i Documentació. Universitat de València, Spain. UISYS, Mixed Research Unit, CSIC-University of Valencia. Spain. Palau de Cerveró, Plaza Cisneros, 4, 46003 València. Spain
[2] remedios.aguilar@ucv.es
Universidad Católica de Valencia "San Vicente Mártir", Departamento de Ciencias de la Educación, Calle Sagrado Corazón 5, 46110 Godella (Spain)
[3] david.melero@ucv.es
Universidad Católica de Valencia "San Vicente Mártir", Instituto de Documentación y Tecnologías de Información (INDOTEI). Carrer de Quevedo 2, 46001 València (Spain)
[4] rafael.aleixandre@uv.es
UISYS, Mixed Research Unit, CSIC-University of Valencia. Ingenio (CSIC-Universitat Politècnica de València), Spain. Palau de Cerveró, Plaza Cisneros, 4, 46003 València. Spain
[5] fjbueno@valencia.es
Pla Municipal de Drogodependències-UPCCA València, Ajuntament de València

Introduction

Bibliometric indicators allow, among other types of analysis, to position the different scientific agents through indicators weighted in rankings, such as the Journal Citation Reports (Clarivate Analytics, n.d.) or the Scimago Journal & Country Rank (www.scimagojr.com).

At present, there are numerous rankings that position universities on the basis of various criteria, such as those derived from research, teaching, infrastructures and websites. However, the presence of Catholic universities from Latin America, the Caribbean and Spain in these rankings is low.

Catholic universities, like other universities, present research, teaching and common university services, but they also provide the inspiration and light of the Christian message (Laghi, 1995).

Based on our knowledge, there is not observatory that evaluates the scientific activity of universities with a Catholic identity, giving visibility and positioning through bibliometric indicators. For this reason, the general objective of this pilot study is to evaluate the research of Catholic universities in Spain and Latin America and the Caribbean with classic bibliometric indicators and to compare and position their scientific production within the framework of the development of an Observatory of Catholic Universities that will contemplate all Catholic universities in the world.

Methods

An exhaustive process of collection, treatment and analysis of bibliographic data was carried out.

Data collected: We used the papers collected in the scientific bibliography databases that include the most important and quality journals at international level: Web of Science Main Collection (WoS) (specifically the indexes of citations Science Citation Index Expanded, Social Science Citation Index, Arts & Humanities Citation Index), belonging to Clarivate Analytics and; Scopus (includes Medline), belonging to Elsevier.

Bibliographic data were retrieved and downloaded on 3 May 2018 from the international databases described for all documents published in scientific journals during the period 2007-2016 (10 years) in which at least one author was affiliated to one of the universities associated with the Organization of Catholic Universities in Latin America and the Caribbean (ODUCAL) (http://www.oducal.com) or to the Spanish Catholic universities belonging to the International Federation of Catholic Universities (IFCU) (http://www.fiuc.org).

Data processing: Data processing consisted of the following steps: (1) Creation of a relational database with the bibliographic information of the 117,030 records recovered; (2) elimination of 44,234 overlapping records from the downloads of the databases used; (3) normalization of the institutional affiliations of the universities under study. After this process, 4,490 false positives were eliminated, i.e., records in which none of the universities to be analyzed signed; and (4) standardization of disciplines from scientific journals to the Essential Science Indicators Research Areas (see http://ipscience-help.thomsonreuters.com/inCites2Live/filterValuesGroup/researchAreaSchema/esiDetail.html).

Analysis: An analysis of the productivity, productivity by discipline, collaboration, visibility, impact and indexes of each University was carried out through the following bibliometric indicators:

- Productivity measures (WoS and Scopus data): number of documents and number of documents per year.

- Productivity measures per discipline (WoS and Scopus data): number of documents per discipline.

- Collaboration measures (WoS and Scopus data): number of institutional signatures, signatures per document, percentage and number of documents in collaboration and without collaboration, and percentage and number of documents in internal collaboration.

- Visibility measures (WoS data): number of documents in WoS, percentage and number of documents in first, second, third and fourth quartile of Journal Citation Reports (JCR).

- Impact measures (WoS data): citations received, citations by paper, citations received in the most cited papers, percentage and number of documents cited, percentage and number of documents not cited, Essential Science Indicators top 0.01, 0.1, 1 and 10 percent.

- Indexes (WoS data): h-index (Hirsch, 2005), Institutional Field Quantitative-Qualitative Analysis Index (IFQ2A-Index), QNIF of IFQ2A-Index and QLIF of IFQ2A-Index (Torres-Salinas et al., 2011).

Results

All the results of the study can be viewed in http://www.oecuc.com or http://www.observatoriouniversidadescatolicas.com

Table 1 shows how, in general, the results of the signature/paper and citation/paper/WoS indices are not related.

Table 1. Number of papers, signature/paper and citations/paper/WoS rates in the 10 most productive universities

University	n papers	Signature /paper	Citation /paper/WoS
Pontificia Universidad Católica de Chile	18,095	13.22	14.24
Pontificia Universidad Católica do Río Grande do Sul	5,143	3.73	11.63
Pontificia Universidad Católica do Río de Janeiro	4,140	9.83	11.30
Pontificia Universidad Javeriana	3,549	6.16	13.70
Pontificia Universidad Católica de Valparaíso	3,409	2.76	7.31
Pontificia Universidad Católica do Paraná	2,624	3.51	10.72
Universidad Católica del Norte	2,562	3.70	8.83
Universidad Ramón Llull	2,110	13.35	13.32
Pontificia Universidad Católica de Minas Gerais	1,969	11.29	18.10
Universidad Católica de Brasilia	1,735	3.52	11.68

Table 2 shows how, except in the 3 most productive universities, the remaining ones maintain differences in their position in the ranking considering different indicators. The main disciplines where Catholic universities publish are clinical medicine (19.4%) and social sciences (11.1%), and those that publish less are immunology (1%), and pharmacology and toxicology (1.4%).

Table 2. H-index, IFQ²A and ESI Top 1 in the 10 most productive universities

University	Rank H-index	Rank IFQ²A	Rank ESI Top 1
Pontificia Universidad Católica de Chile	1	1	1
Pontificia Universidad Católica do Río Grande do Sul	2	2	2
Pontificia Universidad Católica do Río de Janeiro	3	3	3
Pontificia Universidad Javeriana	7	5	7
Pontificia Universidad Católica de Valparaíso	13	10	19
Pontificia Universidad Católica do Paraná	8	7	9
Universidad Católica del Norte	10	9	14
Universidad Ramón Llull	5	4	6
Pontificia Universidad Católica de Minas Gerais	10	8	8
Universidad Católica de Brasilia	9	13	14

* (Hirsch, 2005). ** (Torres-Salinas et al., 2011).

Conclusion

The present study has made it possible to observe the situation of these universities within their own typology of universities with the same identity and, in future editions, to construct indicators specific to universities with a Catholic identity. Likewise, this study makes it possible to promote the visibility of Catholic universities with a great research tradition, as well as to make known other Catholic universities that do not appear in other recognized international rankings.

Acknowledgments: Funded by a grant of the Catholic University of Valencia "San Vicente Mártir" (UCV) for the development of research projects 2017-2018 (Project 2018-258-001).

References

Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, *102(46),* 16569.

Laghi, P. (1995). The Catholic University as University and as Catholic. *Seminarium*, *35*, 369-376.

Torres-Salinas, D., Moreno-Torres, J. G., Delgado-López-Cózar, E., & Herrera, F. (2011). A methodology for Institution-Field ranking based on a bidimensional analysis: the IFQ 2 A index. *Scientometrics*, *88(3),* 771-786.

Towards Leiden Manifesto version 2.0

Lorna Wildgaard (lowi@kb.dk)[1,2] and Marianne Gauffriau (mgau@kb.dk)[2]

[1] Institute of Information Studies, Faculty of Humanities, University of Copenhagen, Copenhagen, Denmark.

[2] KUB Research Support, Copenhagen University Library, Royal Danish Library, Copenhagen, Denmark

Introduction

In Leiden Manifesto (LM) (Hicks et al 2015) bibliometric evaluation is explained as a combination of quantitative and qualitative methods, allowing the use of different metrics, disciplinary knowledge and research performance strategies. Both bibliometricians and consumers of bibliometrics are encouraged to communicate and use the LM principles to acknowledge what they know and do not know, what is measured and what is not measured, thus legitimizing the use of metrics applied in an evaluation. However, we have previously observed that it is unclear how the LM principles should be interpreted in a concrete evaluation and that evaluations may differ in their interpretations of the LM principles (Madsen, Wildgaard, and Gauffriau 2017a, 2017b). Our concern is that interpretations randomly differ from case to case. The present study investigates how the LM principles have been interpreted in concrete evaluations. Based on the investigation we suggest possible future developments of the LM.

Method

Creating a corpus

To create a corpus for the investigation we searched Web of Science, Scopus, Google Scholar and the Leiden Manifesto Blog (Leiden Manifesto for Research Metrics n.d.) on the 10th of July 2018 for published articles and reports in English or Danish that implement all 10 LM principles on a single evaluation case. The publication had to be available in full-text. Local interpretations and translations of the LM where excluded, as were our own previous publications on the LM as these were considered conflicts of interest.

Of 653 potentially relevant publications, three publications were identified as applying all 10 principles on an evaluation case. Thus our corpus consists of the following publications: de Oliveira and Amaral 2017, (29 pages); Marzolla 2016, (31 pages); Orduna-Malea et al. 2017, (18 pages).

Coding the 10 LM principles

Two raters independently annotated sentences and words describing the definition and general application of each LM principle in all corpus-documents and in the LM. Specific examples and anecdotes were coded as "other". There was strong agreement (*Cohen's* kappa coefficient, κ) between the two raters identification of how each principle was defined, κ= .999, and applied, κ= .844.

Thereafter all phrases describing the application of each LM principle from the three corpus publications were matched to the LM's formulation of the principles. Phrases that could not be matched to the LM were further analyzed to investigate if they are extending the boundaries of the LM.

Results

Size of the formulation of each principle

Word count is used to indicate the authors' attention to each principle.

Figure 1. Number of words describing each principle

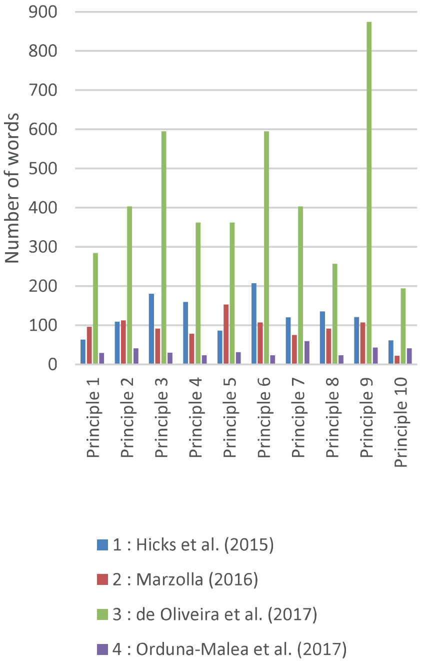


Figure 1 shows the size of the formulation for each principle. Longer formulations indicate the principle is discussed more and has engaged the author, perhaps with difficulties or challenges in the application of the indicator whereas short formulation could be an indication of the transparency of the indicator and ease of application.

We do not observe any pattern in the formulation size of the principles across the corpus documents.

Analysis of application of LM principles

The comparison of the phrases and terminology describing the application of the LM principles between the corpus documents and the LM, exemplified how the LM is being extended in practice beyond its current formulation. Below we report our main observations and also suggested improvements of the LM in practical applications:

Principles 1-10 State which actors are involved in and who is responsible for which deliverables throughout the bibliometric evaluation – bibliometrician, administrator, assessor, panel, national committee, area coordinators, ad hoc consultants, evaluand, etc.

Principle 2: State clearly the profiles of who or what is being evaluated (colleagues, applicants, researchers, teachers, research areas, etc) and provide accordingly a rationale for which materials are and are not included in the evaluation.

Principle 3: Present a wider view of what can be included in the evaluation (documents and activities) and how these contribute to different forms of impact. Otherwise, clearly limit the LM as a distillation of best practice for the evaluation of *publications* only.

Principles 4 & 5: All persons collecting the data, conducting the bibliometric analysis and interpreting the data, as well as the evaluation system itself should work as open as possible and abide by recommendations for good data practices. Ensure access to the data underlying the evaluation and consider how long the data should be available, for whom and where the data should be stored; ensure the data has a Read Me file, appropriate anonymity of evaluands and appropriate metadata; demonstrate how indicators are calculated and validated; give stakeholders the opportunity to validate the final report before publication.

Principles 8 & 10: State how bibliometric indicators will be summarized and interpreted in a final report, and consequently how bibliometric scores correspond to interpretive benchmarks, e.g. adequate, good, excellent. Those designing the evaluation should refer to best practice guidelines for statistics.

Principle 9: When applying a suite of indicators, ensure they measure different aspects of for example productivity and impact.

Conclusions and Further work

The phrase analysis illustrates how the boundaries of the LM are being extended and the challenges we face in responsible evaluation practices. As the LM has captured the attention of the evaluation and administrative community, we encourage follow-up studies of the implementation of the LM to ensure its robustness and continued use in the long term. The current need to update the LM is critical, ”so researchers can continue to hold evaluators to account, and evaluators can continue to hold their indicators to account.” (Hicks et al 2015, p. 430)

References

de Oliveira, T.M., and Amaral, L. 2017. Public Policies in Science and Technology in Brazil: challenges and proposals for the use of indicators in evaluation. In R. Mugnaini, A. Fujino and N.Y. Kobashi (Eds.), *Bibliometria e Cientometria No Brasil: Infraestrutura Para Avaliação Da Pesquisa Científica Na Era Do Big Data* . Universidade de São Paulo. Escola de Comunicações e Artes. https://doi.org/10.11606/9788572051705.

Hicks, D, P Wouters, Ludo Waltman, S. de Rijcke, and I. Rafols. 2015. ‘Bibliometrics: The Leiden Manifesto for Research Metrics’. *Nature*, 2015. ‘Leiden Manifesto for Research Metrics’. [n.d.]. Retrieved from: http://www.leidenmanifesto.org/.

Madsen, H, L Wildgaard, and M Gauffriau. 2017a. ‘Bottom-up Implementation of Leiden Manifesto’. In *Nordic Workshop on Bibliometrics and Research Policy*.

Madsen, H, L Wildgaard, and M Gauffriau. 2017b. ‘Consumer Labels for Bibliometric Analyses Based on Leiden Manifesto’. *Leiden Manifesto for Research Metrics*. http://www.leidenmanifesto.org/2/post/2017/11/consumer-labels-for-bibliometric-analyses-based-on-leiden-manifesto.html.

Marzolla, Moreno. 2016. ‘Assessing Evaluation Procedures for Individual Researchers: The Case of the Italian National Scientific Qualification’. *Journal of Informetrics* 10 (2): 408–38. https://doi.org/10.1016/j.joi.2016.01.009.

Orduna-Malea, Enrique, Alberto Martín-Martín, Mike Thelwall, and Emilio Delgado López-Cózar. 2017. ‘Do ResearchGate Scores Create Ghost Academic Reputations?’ *Scientometrics* 112 (1): 443–60. https://doi.org/10.1007/s11192-017-2396-9.

Technology Foresight Study of Human Phenomics

Xu Li[1], Yao Chiyuan[2], Wang Yue[3], Xu Ping[4]*

[1] *xuli@sibs.ac.cn* Shanghai Institutes for Biological Sciences, CAS, Shanghai, (China)

[2] *cyyao@sibs.ac.cn* Shanghai Institutes for Biological Sciences, CAS, Shanghai, (China)

[3] *wangyue@sibs.ac.cn* Shanghai Institutes for Biological Sciences, CAS, Shanghai, (China)

[4]*xuping@sibs.ac.cn* Shanghai Institutes for Biological Sciences, CAS, Shanghai, (China)

Introduction

Although human phenomics (the study of the interaction between human genes and the environment) has developed as an engrossing hot spot of research in the recent five years, it demands technological advancements. This paper probes the current R&D status of human phenomics, through data investigations and literature analysis, combined with expert consultations and field interviews. We also analyzed the realization possibility, time and developmental approaches of key technologies in the field of human phenomics.

Methods

Investigation data obtained through strategic planning, prudent measures, reports and periodicals was collected from Web of Science database, while information on patents was acquired from Derwent Innovation and Innography database. The search keywords used in this study were "Phenome" and "Phenomics". Consultants were chosen on the grounds of their professionalism, authority and comprehensive caliber. The methods are as follows:

- To analyze key points, hotspots and cutting-edge technologies in the field, in order to further develop a full-fledged system of technology.
- To prepare a list of alternative substantial technologies through literature/patent analysis, hotspot clustering and expert consultation.
- To design Delphi Questionnaire and establish analysis model of investigation results.
- To take a city as an example.

Results

Selection of the key technologies in the field of human phenomics

Bibliometric analysis was performed based on the investigation completed at the initial stage, and the VOS Viewer software was applied to induce keyword clustering analysis, in order to explore hotspots in the field of human phenomics.

The topics of human phenomics technology prediction are divided into three technological groups (Figure 1), which are further subdivided into 23 alternative key technologies (Table 1).

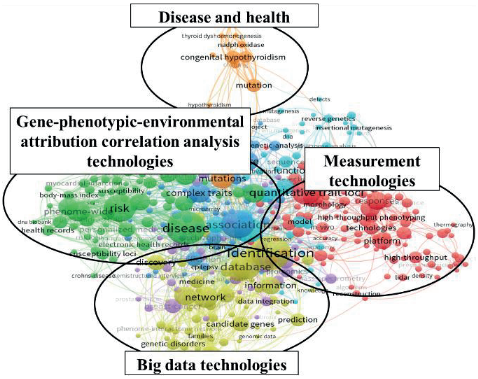


Figure 1. Knowledge mapping of VOS viewer-based investigation hotspots in the field of human phenomics

Table 1. List of alternative key technologies for technology predictions

Technology fields	Alternative key technologies
1 Measurement technologies of human phenomics group	1.1 Whole body and organ magnetic resonance imaging technology
	1.2 High resolution three-dimensional dynamic imaging of human body
	1.3 Artificial intelligence-assisted imaging technology
	1.4 Precision skeletal system detection technology
	1.5 Human bioelectric impedance measurement technology
	1.6 Flow coding technology for single cell mass spectrometry
	1.7 Precision proteomics detection technology
	1.8 Precision metabonomics detection technology
2 Big data technologies of human phenomics group	2.1 Human phenomics data standard
	2.2 Individual and group feature recognition
	2.3 Data network and Internet of Things (IoT) system for the acquisition of human phenomics data
	2.4 Big Data blockchain technology for healthcare
	2.5 Annotation of AI Training Data
	2.6 Cross-scale phenomics integration and retrieval technology
	2.7 Phenomics data visualization technology
	2.8 Phenomics prediction model
3 Gene-phenotypic-environmental attribution correlation analysis technologies	3.1 Detection of molecular and cellular markers
	3.2 Screening of drug targets
	3.3 Evolution of human health path
	3.4 Reference mapping of human phenomics
	3.5 Risk prediction and assessment of major diseases
	3.6 Early diagnosis and warning of major diseases
	3.7 Human health control and guidance system

Design of Delphi questionnaire and establishment of analysis model of investigation results

The indexes for the questionnaire design are mainly divided into five categories (Table 2).

Table 2. Design of main indexes for technology evaluation of Delphi questionnaire

Index types	Indexes	Options
Indexes of technology importance	Importance to the competitiveness of health industry	①High; ② Relatively high; ③ Moderate; ④ Relatively low
	Importance to economic growth of a city	①High; ② Relatively high;③ Moderate; ④ Relatively low
Indexes of technology realization possibility	Current development level of a city	①International level; ② 1–3 years lagging behind; ③ 3–5 years lagging behind; ④ over 5 years lagging behind
	Development stage	①Laboratory research; ② Preliminary application; ③ Extensive application
	Leading cities in China	①Beijing; ② Shanghai; ③ Guangzhou; ④ Shenzhen; ⑤ Other cities
Indexes of realization time	Time for extensive application	① 1–3 years; ② 3-5 years; ③ 5–10years; ④ over 10 years
Indexes of realization approaches	Technology development model of a city	①Independent research and development; ② Joint development; ③ Imitation ④ Introduction
Indexes of basic conditions	Leading countries/regions	①US; ② Europe; ③ Japan; ④ China; ⑤ Other countries
	Factors limiting development	①Technological bottleneck;② Commercial application; ③ Policy standard; ④Human resources; ⑤ R&D input; ⑥ Infrastructure

Considering the statistical analysis of Delphi questionnaire survey results, experts were scored for weight quantification using five-point and four-point methods based on specific indexes, on the scale of 0 to 100 points. The number of experts matching "high", "relatively high", "moderate", "relatively low" and "low" for an index were N1, N2, N3, N4 and N5, respectively.

$$f(x)=\frac{\sum_{n=1}^{5} N_{ij}}{N}$$

The importance of a technology framework mainly depends on two indexes, i.e., I_1, which refers to the importance of the competitiveness of health industry, and I_2, which denotes the importance of the economic growth of a city.

$$f(x)=\frac{\sum_{j=1}^{3}\sum_{i=1}^{n}\alpha_{ij}I_{iJ}}{3N}$$

Indexes of technology realization possibility seem to be in direct proportion to the current development level of a city (R), but inversely proportional to its developmental stage (C).

$$f(x)=\frac{\sum_{n=1}^{n} R_{ij}}{\sum_{n=1}^{n} C_{ij}}$$

Technology prediction study performed by taking a city as an example

A city in East China was considered as an exemplar to conduct the technology foresight study. A total of 150 Delphi questionnaire sets were distributed, and amounting to a recovery rate of about 63%. It is to be noted that these selected experts, in general, were highly proficient in the field, thus increasing the reliability level of the predicted results.

The results are shown in Table 3:

- The city shows a leading advantage in Big Data technology of human phenomics and gene-phenotype-environment attribution correlation analysis technologies.
- AI-assisted image technology is highly crucial for the development of this city.
- The human phenomics data standards demonstrate efficient performances in terms of realization possibility of the city's development.

Table 3. Predicted technology results of a city in the field of human phenomics

Alternative key technologies*	Importance	Realization possibility	Expected realization time (year)	Technology development path**
1.1	75.00	1.22	1-3	a
1.2	58.24	1.02	3-5	a
1.3	94.44	1.52	3-5	a
1.4	64.06	1.17	1-3	a
1.5	68.75	1.25	1-3	a
1.6	79.17	0.92	3-5	a, b
1.7	58.33	0.66	3-5	a
1.8	70.09	0.74	3-5	a
2.1	68.18	2.02	1-3	b
2.2	83.33	1.07	3-5	b
2.3	83.33	0.73	3-5	b
2.4	64.02	1.62	1-3	b
2.5	71.50	1.36	1-3	b
2.6	75.00	1.17	3-5	b
2.7	68.18	1.41	1-3	b
2.8	69.32	1.21	3-5	b
3.1	78.13	1.53	1-3	b
3.2	89.29	0.94	3-5	a, b
3.3	75.00	0.67	3-5	a, b
3.4	84.38	0.77	3-5	a
3.5	63.18	1.22	3-5	b
3.6	83.33	0.60	3-5	b
3.7	75.00	0.45	3-5	b

* See Table 1. ** The letter "a" represents "Joint development", and "b" represents "Independent development and research".

Acknowledgments

This study was supported by the Shanghai Soft Science Research Program (Grant No. 18692109000).

References

Baker, M. (2013). Big biology: The 'omes puzzle. Nature. *Nature*, 494(7438), 416-419.

Freimer, N. & Sabatti, C. (2003) The Human Phenome Project. *Nature Genetics*, 34(1), 15-21.

Kendall, P. (2018). Technology to watch in 2018. *Nature*, 553(7689), 531-534.

Analysis of disaster-related research trend in South Korea using topic modeling

YuCheong Chon[1] and Geonwook Hwang[2]

[1] ycchon@kistep.re.kr
Office of National R&D Coordination, Korea Institute of S&T Evaluation and Planning (KISTEP),
60 Mabang-ro, Seocho-gu, Seoul 06775 (South Korea)

[2] geunouk@kistep.re.kr
Office of National R&D Coordination, Korea Institute of S&T Evaluation and Planning (KISTEP),
60 Mabang-ro, Seocho-gu, Seoul 06775 (South Korea)

Introduction

In recent years, disasters and safety accidents were increased in South Korea due to the emergence of new infectious diseases such as Mers, the occurrence of earthquakes and fire. The Korean government has invested 1.258 trillion Korean Won (KRW) in the government's budget for disaster and safety R&D in 2019, which is about 5% of the total government R&D (20.4 trillion Korean Won (KRW)). As the budget for disaster and safety R&D increases, it is necessary to analyze research trends of that field. The main purpose of this study was to investigate sub technology fields of disaster-related R&D by using national R&D data.

Literature Review

Topic modeling

Topic modeling is a statistical method in which subjects are inferred by analyzing vocabulary used in a vast amount of literature (Beli, 2012). Assuming that there are probabilistic topics in the literature, researchers deduce the variables hidden in the literature. As a result, topics and words that exist stochastically in each document could be investigated. Using clues related to context, researchers deduce a topic by clustering words with similar meanings, analyzing how topics are connected and how they change with time. Researchers used topic modeling in trend analysis and inferred topics in particular fields. Griffiths and Stevyer (2004) analyzed the topics that were highlighted and decimated by time, after extracting them from the green of the papers published in PNAS from 1991 to 2001.

Co-occurrence analysis

Co-occurrence analysis used to understand trends and time changes in various subject fields by using the frequency of simultaneous appearance of the keywords or classification codes in the document sets. It also used to investigate trends and temporal changes in various subject areas. Co-occurrence means that two keywords are found in the same range as the same document, paragraph or sentence. When two keywords appear at the same time in the literature, it could be judged that the research topics represented by the two keywords were related to each other. The more frequently the two words were found together, the higher the relevance of the two keywords.

Methods

Data

Data were collected from National science & technology information service (NTIS) database which provides all national R&D information for the last 2 years (2016-2017). We analyzed literature which defined disaster and safety R&D and obtained candidates for technical keywords in the precision medical field. Expert review was performed to select keywords which could define precision medicine, and as a result, 68 keywords were derived. NTIS database were used to search patents that contained the 68 keywords in the R&D project's title or abstract. As a result, 13,201 R&D projects were used for analysis.

Method 1: Latent Dirlet Allocation (LDA)

In this study, the Latent Dirlet Allocation (LDA) algorithm were used. LDA was a probabilistic model that identified what topics exist in a particular set of documents. LDA has been one of the popular methods for summarizing and extracting topics from text documents (Lee and Sohn, 2017). It assumed that the actual literature was being prepared, and models each parameter for what topics to include in each document to produce the literature, and which words to select and place in which subject. It was widely used in text mining analyses because it helped to reduce the dimension of data along with the characteristics of simplicity among many topic modeling techniques. Also, it produced topics which were meaningful and consistent.

Method 2 : Network analysis

Network analysis is composed to relative data, correlation matrix, and networks. First of all, correlation data could represent the relationship between pair of elements that make up the network. Correlations and interactions of data were used to obtain a correlation matrix, which had a property of value and direction and determined the type of correlation and network. The correlation matrix could be a binary matrix, expressed as 0 or 1, or could be a valued matrix. Then, researchers could use this correlation matrix to find networks.

Results

Topic discovery with LDA

The results of analyzing the topics were shown in the following table 1. The table lists 10 topics and topic-specific keywords, with the order of keywords being in the order in which they are most likely to be in the topic. The topics were related to technologies such as infectious diseases, vulnerable groups, and environmental pollution.

Table 1. Topics and keywords of disaster-related R&D in 2016-2017

Topic No.	Keywords
1	micro, compound, test, by-product, mask, multi-functional, extract, simulation, CO2
2	system, monitoring, real-time, smart, data, network, ecosystem, structure, radioactivity , radiation
3	system, plasma , smart, sludge, solution, reservoir, ultrasonic waves, cable, pedestrian, laser
4	program, adolescent, cyber, victim, protocol , depression, sexual violence, nuclear energy, manual, long-term
5	structure, concrete, system, scenario, nuclear fuel, cement, facility, nuclear reactor, complex, risk
6	system, smart, monitoring, medicine, platform, facility, traffic accident, driver, semiconductor, simulation
7	virus, protein, system, bio, network, stress, next generation, tubercular bacillus, risk, sensitivity
8	system, car, multipurpose, guideline, medical appliance, simulator, waste, plastic, tire, platform
9	system, service, smart, infrastructure, design, senior citizen, safety accident, downtown area, safety net, disabled person
10	microorganism, heavy metal, bio, food poisoning, farmland, agricultural products, insecticide, waste, antibiotic, system

Keyword network with co-occurrence analysis

The frequency of simultaneous occurrence of the keywords included in the abstract was analyzed. The words with high in-degree of centrality were safety (0.373) environment (0.239), and analysis (0.239). Also, the words with out-degree of centrality were analyzed in terms of system (03657), infrastructure (0.448) and assessment (0.358). The network by keyword was as follows.

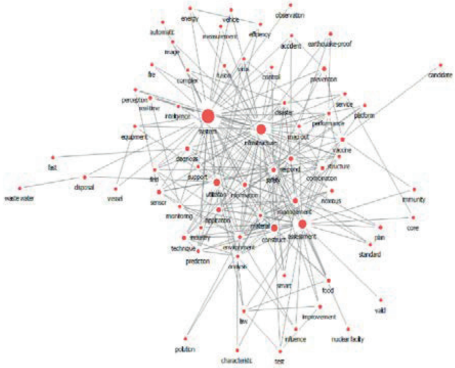


Figure 1. Keyword networks of disaster-related R&D

Conclusion

This study quantitatively investigated the trends of disaster related research in South Korea. The topic modeling was used to define the disaster-related technologies from the disaster related governmental R&D projects in which various areas of science and technology were utilized. It is anticipated that it would be effectively used in establishing R&D policies and research in the future.

Acknowledgments

This work was supported by Korea Institute of S&T Evaluation and Planning (KISTEP) grant funded by the Ministry of Science and ICT (MSIT). The title of project was "A study on Disaster and Safety R&D Investment Analysis System".

References

Blei, D. M. (2012, April). Probabilistic topic models. *Communications of the ACM, 55(4)*, 77-84.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Lee, W., & Sohn, S. (2017). Identifying emerging trends of financial business method patents. *Sustainability*, *9*(9), 1670-1691.

One research field, multiple subjects integrated: Subfield differences and correlations in “computer science, artificial intelligence” in WoS

Jiajun Cao[1], Yuefen Wang[2], Shengzhi Chen[3] and Bentao Zou[4]

[1] *jscjj95@126.com*
School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094(China)

[2] *yuefen163@163.com*
School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094(China)

[3] *908892765@qq.com*
School of Computer Science & Engineering, Nanjing University of Science & Technology, Nanjing 210094(China)

[4] *zoubentao@njust.edu.cn*
School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094(China)

Introduction

Scientific research is gradually showing the characteristics of highly integrated subjects (Wagner, C. S., et al, 2011). Current relevant research measured with different methods (Lillquist, E., et al, 2010; Wei, J., et al, 2012), but few scholars pay attention to a given topic and explore the similarities between the subjects who had the same research topic. Artificial intelligence (AI) is a typical research field supported by multiple related subjects. This study aims to perform an analysis based on papers in “computer science, artificial intelligence (CS,AI)” in Web of Science (WoS). It takes AI as an example to find the attributes. The analysis focused on the following: (1) the paper quantities and distribution in related subjects; (2) co-occurrence between these subjects; (3) research similarity between these subjects and that between “CS,AI” and others.

Data and Methodology

The overall research framework is shown in Figure 1. WoS core collection divides subjects into 252 categories. Each journal and book included in it belongs to at least one subject category, expressed as “WC”. We searched ‘WC = "Computer Science, Artificial Intelligence"’, time-span was up to 2017, and retrieval time was May 11, 2018. There were total of 771375 records. We removed data whose DE was missing and there were 389358 data left. These data include conference papers and journal articles.

After pre-processing, *Bag-of-words model*, *TF-IDF* and *Cosine similarity* were used. Keywords in DE that appear at the same time with the subject formed a dictionary. The frequencies of keywords were regarded as the feature values of subject vectors. The subject is equivalent to the “document” in *Bag-of-word*, and then the related keywords and word frequencies of each subject are counted, following the TF-IDF values of keywords in each subject are calculated and the similarities between two subjects are measured by cosine similarity of subject vectors. In addition, some Python packages were also used.

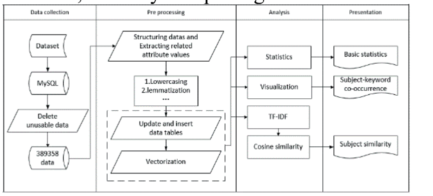


Figure 1. Flow chart of Research procedures.

Results Analysis

Basic Statistic

Top-ranking subjects include “Engineering, Electrical & Electronic (En,E&E)”, “Automation & Control Systems”, “Robotics”, “Imaging Science & Photographic Technology”, “Telecommunications”, “Operations Research & Management Science”, etc, except “CS,AI”. There are no articles in “Mathematical & Computational Biology” before 2006 and in “Management” before 2001. There are no papers about “Engineering, Industrial”, between 2010 and 2013, and it is the same as “Acoustics” from 2011 to 2015.

Some subjects attach to same areas, such as computer science (CS) and engineering (En). Figure 2 shows their paper quantities. In CS, AI research distributes in six subjects. In En, it mainly distributed in “Electrical & Electronic”.

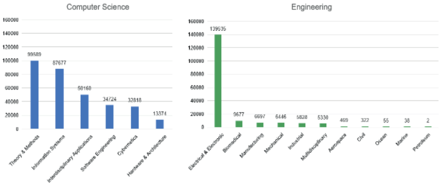


Figure 2. The paper quantities of subjects in the field of CS (left) and En (right).

Research of AI Related Subjects Based on Co-occurrence Analysis

In Figure 3, co-occurrence degree is quite large in the field of CS, En and automation. Except them, the degree between "Robotics" and "Automation & Control System" is relatively larger than the rest. There are still many links to "Imaging Science & Photographic Technology", and among them, the degree between it and "En, E&E" is the largest.

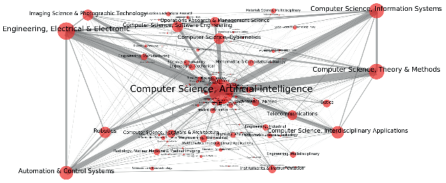
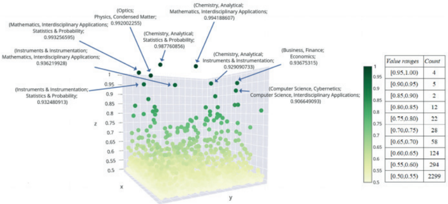

Imaging Science & Photographic Technology
Computer Science, Information Systems
Operations Research & Management Science
Computer Science, Software Engineering
Engineering, Electrical & Electronic
Computer Science, Cybernetics
Engineering, Manufacturing
Statistics & Probability
Engineering, Mechanical
Mathematical & Computational Biology
Computer Science, Theory & Methods
Computer Science, Artificial Intelligence
Mathematics, Applied
Optics
Telecommunications
Engineering, Industrial
Robotics
Computer Science, Hardware & Architecture
Engineering, Biomedical
Computer Science, Interdisciplinary Applications
Mathematics, Interdisciplinary Applications
Radiology, Nuclear Medicine & Medical Imaging
Engineering, Multidisciplinary
Automation & Control Systems
Instruments & Instrumentation

Figure 3. The Co-occurrence network among subjects in AI research.

Research of AI Related Subjects and its Evolution Based on Cosine Similarity

Table 1. Part of subjects (Top13) and their sequences and similarities in different intervals.

Subjects	*Sim*	*1996-2000*		*2001-2005*		*2006-2010*		*2010-2015*		*2016-2017*	
		Seq	*Sim*	*Seq*	*Sim*	*Seq*	*Sim*	*Seq*	*Sim*	*Seq*	*Sim*
CS, Interdisciplinary Applications	0.72	2	0.92	6	0.90	4	0.92	4	0.93	4	0.91
CS, Cybernetics	0.71	8	0.79	2	0.94	3	0.93	7	0.85	10	0.80
CS, Hardware & Architecture	0.68	7	0.80	15	0.77	12	0.82	10	0.82	8	0.84
Telecommunications	0.67	26	0.60	10	0.82	8	0.86	12	0.78	6	0.85
Engineering, Multidisciplinary	0.67	12	0.76	16	0.76	14	0.82	11	0.80	13	0.73
Imaging Science & Photographic Technology	0.66	23	0.66	8	0.86	10	0.84	14	0.73	12	0.76
Engineering, Industrial	0.64	15	0.75	18	0.74	20	0.70	29	0.61	21	0.62
Mathematics, Applied	0.64	21	0.67	21	0.71	14	0.82	16	0.73	18	0.66
Engineering, Manufacturing	0.63	11	0.76	17	0.75	17	0.74	18	0.69	26	0.61
Materials Science, Multidisciplinary	0.61	43	0.52	54	0.53	72	0.52	17	0.70	48	0.53
Remote Sensing	0.61	24	0.64	13	0.79	27	0.64	33	0.59	25	0.61
Medical Informatics	0.61	10	0.76	24	0.67	26	0.65	19	0.68	19	0.64
Instruments & Instrumentation	0.60	20	0.67	11	0.80	22	0.68	27	0.62	33	0.59



Value ranges	*Count*
[0.95,1.00)	4
[0.90,0.95)	5
[0.85,0.90)	2
[0.80,0.85)	12
[0.75,0.80)	22
[0.70,0.75)	28
[0.65,0.70)	58
[0.60,0.65)	124
[0.55,0.60)	294
[0.50,0.55)	2299

Figure 4. Cosine Similarity among subjects in AI and its count distribution in different ranges.

Table 1 gives part of subjects related to "CS,AI", and their sequences and similarities in different intervals. Their similarities are all over 0.60. Except subjects attached to CS, "Telecommunications" has the highest degree of similarity. Most similarities did not change much in each interval. However, the similarity of "Materials Science, Multidisciplinary" in 2010-2015 is quite larger than in other intervals, the research of this subject in this interval improves the similarity of its overall research.

In figure 4, z-axis represents the value of similarity, *x* and *y* axis represent subjects. We labeled top 9 values with their subjects. The similarity between "Chemistry, Analytical (C,A)" and "Mathematics, Interdisciplinary Applications (M,IA)" is the largest. The similarity between "Statistics & Probability" and "M,IA" is next to it. Mathematics plays an important role. The left part of top 9 points are the similarities between subjects, which represent basic theory and method. "C,A" has high similarities with "Statistics & Probability", "Mathematics, Interdisciplinary Applications" and "Instruments & Instrumentation", it represent that the application of AI in chemistry is highly related to its own characteristics. It is the same as the similarity between "Business, Finance" and "Economics". While these subjects don't have high similarities with "CS, AI", they have high similarities in AI research with subjects within their self-fields.

Conclusions

The most detected subjects are with little span of AI including methods or applications and they play a key role in the interdisciplinary of AI. AI research has gradually expanded from technology to others, and more and more AI research relevant to subjects within their self-fields. AI is gradually integrating with management. Gradual changes are more conducive to establishing links with new subjects.

This study did not consider the hierarchy of subjects. We will supplement it in the future and cluster analysis of subjects will also be carried out.

Acknowledgments

Major bidding project of National Social Science Fund of China (16ZDA224); Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX18_0344).

References

Hu, J., & Zhang, Y.. (2017). Discovering the interdisciplinary nature of big data research through social network analysis and visualization. *Scientometrics*, 112(1), 91-109.

Lillquist, E., & Green, S.. (2010). The discipline dependence of citation statistics. *Scientometrics*, 84(3), 749-762.

Wagner, C. S., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of informetrics*, 5(1), 14-26.

Wei, J., Sun, Y., & Su, X. (2012). Research on Interdisciplinary Knowledge Mining Model. *Information Studies: Theory & Application*, 35(04), 76-80.

Development of a user-friendly app for exploring and analyzing research topics in psychology

André Bittermann

abi@leibniz-psychology.org
ZPID – Leibniz Institute for Psychology Information, Universitätsring 15, D-54296 Trier (Germany)

Background

Keeping track of the developments in a scientific field can be challenging. Regarding increasing numbers of publications, summarizing the contents of hundreds of thousands of scientific publications on specific topics is necessary to gain insights into the processes of a scientific field.

Many databases offer classifications, i.e., broad subject headings for categorizing the publications' contents. Past research has treated these categories as research topics (e.g., Krampen, 2016), but regarding the level of detail, topicality, and flexibility this approach has been criticized (e.g., Bittermann & Fischer, 2018).

Techniques for automated content analysis represent a promising approach for getting insight into large text corpora. Topic modeling (e.g., Blei, Ng, & Jordan, 2003), in particular, is gaining in popularity in scientometrics. In their well-known paper, Griffiths and Steyvers (2004) demonstrated how to find scientific topics by applying topic models to a corpus of scientific abstracts.

A topic-guided and user-friendly interface for databases of scientific literature can open publication trends to a broader audience with various user scenarios: exploring the current "hot topics," investigating the ups and downs of topic popularity over time, or comparing publication trends concerning societal processes (e.g., the increasing trend of a topic referring to refugees and emotional trauma in psychological publications from the German-speaking countries after 2015).

Aim

The goal of this project was to develop a user-friendly web-based application for exploring and analyzing research topics in psychology. This app is considered as an entry point to further research of scientific literature by informing the user about past and current developments of publication topics. To this end, the topics are directly linked to search queries in a database for psychological literature.

Method

Data

The psychological research topics were derived from PSYNDEX – the comprehensive database containing references for German- and English-language publications in psychology and closely related disciplines from the German-speaking countries. It is developed and hosted by the Leibniz Institute for Psychology Information (ZPID; Trier, Germany). In April 2019, there were more than 350,000 psychological articles, book chapters, reports, and dissertations indexed in PSYNDEX. In the development of the app, documents published between 1980 and 2017 were included (N = 329,240 in early 2019).

The PSYNDEX editorial staff assigns controlled terms from the *Thesaurus of Psychological Index Terms* published by the American Psychological Association (Tuleya, 2007). This standardized vocabulary of keywords the input for topic modeling. Main advantages compared to abstract texts are, inter alia, the direct usability for efficient literature search for this topic, the avoidance of stemming, stop words, and synonyms, as well as faster computation time (Bittermann & Fischer, 2018).

Software

All analyses were conducted in R version 3.5.1 (R Core Team, 2018). For inference of research topics, the package *topicmodels* 0.2-8 (Grün & Hornik, 2011) was employed. The user interface was built as a *Shiny app* using *shiny* 1.2.0 (Chang, Cheng, Allaire, Xie, & McPherson).

Topic Modeling

Topic modeling based on *latent Dirichlet allocation* (Blei et al., 2003) was applied. Following the best-practice recommendations by Maier et al. (2018), several candidates for the alpha hyperparameter (0.0001, 0.0005, 0.001) and the number of topics k (250–550) were examined. Delta was fixed to 0.01. The final model (alpha = 0.0005 and k = 325) was selected regarding interpretability and document–topic assignments. Finally, only reliable topics with stability across multiple inference runs were included to increase the robustness of the results.

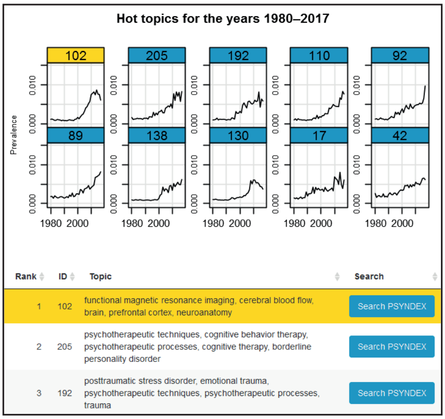
Results

Topic Model

The final model comprised 213 topics. The five terms with highest probabilities were included in the app. The topic with the highest prevalence overall was “psychoanalysis, psychotherapeutic processes, psychotherapeutic transference, countertransference, psychoanalytic theory.” The most strongly increasing trend over the whole range of years was shown by the topic “functional magnetic resonance imaging, cerebral blood flow, brain, prefrontal cortex, neuroanatomy.” The “hottest” topic during the last three years was “posttraumatic stress disorder, emotional trauma, refugees, trauma, war.”

Features of the App

Users can set the range of years from which popularity and trends are dynamically calculated. For exploring the topics, they can switch between “popular topics,” “hot topics” (see Fig. 1), “cold topics,” and an overview of “all topics.” Each topic entry has a search button that forwards a search query to PSYNDEX for literature relevant to this topic. For optimizing results, the terms in this query are weighted according to the term probability in the topic model.



Rank	ID	Topic	Search
1	102	functional magnetic resonance imaging, cerebral blood flow, brain, prefrontal cortex, neuroanatomy	Search PSYNDEX
2	205	psychotherapeutic techniques, cognitive behavior therapy, psychotherapeutic processes, cognitive therapy, borderline personality disorder	Search PSYNDEX
3	192	posttraumatic stress disorder, emotional trauma, psychotherapeutic techniques, psychotherapeutic processes, trauma	Search PSYNDEX

Figure 1. Hot topic view showing topics with most strongly increasing trends from 1980–2017. A demo version of the app can be accessed for free via https://abitter.shinyapps.io/psychtopics/

Conclusions and Future Developments

Initial user experiences confirm the app’s ease of use. The implemented search queries help to clarify the topics’ contents and offer a low-threshold starting point to literature search. Topic inference is data-driven and independent from prior knowledge about a database’s contents. Since the standardized vocabulary used in PSYNDEX is updated on a regular basis, the topic model can be updated as well and kept up to date with low maintenance efforts.

Future features will include current developments from our research group. For instance, topics with a high/low degree of empirical evidence can be shown, which may be of interest for research synthesis or explorative research. Using forecasting techniques, the observed trends can be compared to expected courses over time and help to quantify sudden increases and decreases in publication numbers. Author information could be included for investigating topical author networks.

The app can be applied to other databases with only few modifications necessary. The code is available for free on PsychArchives (http://dx.doi.org/10.23668/psycharchives.2410).

References

Bittermann, A., & Fischer, A. (2018). How to identify hot topics in psychology using topic modeling. *Zeitschrift für Psychologie, 226*(1), 3–13. https://doi.org/10.1027/2151-2604/a000318

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2018). shiny: Web Application Framework for R. R package version 1.2.0 [Computer software].

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13). https://doi.org/10.18637/jss.v040.i13

Krampen, G. (2016). Scientometric trend analyses of publications on the history of psychology: Is psychology becoming an unhistorical science? *Scientometrics, 106*, 1217–1238. https://doi.org/10.1007/s11192-016-1834-4

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Schmid-Petri, H. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. [Computer software].

Tuleya, L. G. (Ed.). (2007). *Thesaurus of psychological index terms* (11th ed.). Washington, DC: American Psychological Association.

Enriching Bibliographic Data by Combining String Matching and the Wikidata Knowledge Graph to Improve the Measurement of International Research Collaboration

Ba Xuan Nguyen[1], Jesse David Dinneen[1] and Markus Luczak-Roesch[1]

[1] *ba.nguyen @vuw.ac.nz, jesse.dinneen@vuw.ac.nz, markus.luczak-roesch@vuw.ac.nz*
School of Information Management, Victoria University of Wellington, Wellington (New Zealand)

Introduction

As publications with international research collaborations receive, on average, a higher number of citations (Glänzel et al., 2001), researchers are incentivised to collaborate. As a result, both the number and the ratio of international co-authored papers have risen (Fortunato, 2018). Research managers and policy makers are interested in measuring international research collaboration (IRC; Luukkonen, 1993), for example to determine if relevant policies are effective.

Problem statement

Currently, the common measure of IRC is a count of co-authored research publications (Chen, Zhan, & Fu, 2018) reported in various data sets (e.g., Web of Science, Google Scholar). The most commonly used data sets in IRC research are SCI/Web of Science and Scopus (Guerrero Bote et al., 2013; Luukkonen, 1993). Each set entails considerable practical challenges for researchers; for example, only 500 records can be downloaded from WoS at a time, or 2,000 from Scopus. Further, these sets (and Google Scholar) do not have as comprehensive general coverage as, for example, Microsoft Academic Graph (MAG; Paszcza, 2016; Sinha et al., 2015), and may not have as complete domain-specific coverage as, for example, ACM Digital Library (ACM DL) and IEEE Xplore provide for computer science. All sources have in common one considerable practical challenge, however: measuring IRC requires mapping the affiliation data from each publication to the relevant countries, and no method for doing this has been previously (e.g., in prior work). The task is non-trivial because, for example, there are many records with varying affiliation formats (e.g., ending with country, or with state/province, or just an institution like "McGill University") or dirty data (e.g., ending in "#TAB#"), and there is no standard method for associating such values with the parent country. In short, measuring IRC is desirable, but currently difficult. Here we describe a method to address this difficulty, and evaluate it using both general and domain-specific data sets.

Preparation of data sets

In this paper we test our method on MAG, a general scholarly bibliographic data set, and ACM DL, a scholarly bibliographic data set containing works published by the Association for Computing Machinery and primarily related to computer science. To make the results of our evaluations comparable across the two sets we filtered out records from MAG that were not relevant to computer science: a list of fields of study (FOS) was compiled from records present in both ACM and MAG, and the 38 top FOS terms (94% of papers in the overlap) were used to filter out irrelevant works. Overlapping papers were also filtered from the ACM set to make the sets distinct. Finally, single-author records were filtered out to identify only co-authored papers. Table 1 summarises the results.

Table 1. Summary of data sets used

Features	ACM DL	MAG
Total works	182,791	212,689,976
Date range	1951-2017	1965-2017
Unique, co-authored, CS works	121,672	594,036

Contributed method

Two steps were implemented to identify the collaborating countries using the authors' affiliation data in co-authored papers. First, for records with author affiliation data (i.e., in the *org* field in MAG and *affiliation* in ACM) containing names or abbreviations of countries or their component parts (e.g., US states), substrings of the location names were extracted and matched to a list of countries. The UK was considered in this study as a whole entity for all its component parts. Second, for records having no country names or state information, we then used the remaining information (e.g., university name) to query the SPARQL endpoint of Wikidata[1] executing the following query[2]:

```
PREFIX schema: <http://schema.org/>
PREFIX wdt:
<http://www.wikidata.org/prop/direct/>
SELECT ?countryLabel WHERE
{<https://en.wikipedia.org/wiki/[AFFILIATION]>
schema:about ?datalink. ?datalink wdt:P17
?country.SERVICE wikibase:label
{bd:serviceParam wikibase:language "en".}}
```

This query returns English names of countries associated with the location data if there is a matching Wikidata item. For example, querying "McGill University" returns Canada. We implemented both steps in *R* and have made the source code freely available for use in future work.[3]

Method evaluation

Our method identifies countries for approximately 70%-80% records in each data set (details in Table 2), with the remaining records being either unidentified (~15%) or unidentifiable (8-14%) because of empty affiliation values (e.g., NA). Specifically, while the substring matching approach identifies countries for 60-70% of records, the Wikidata querying approach adds an additional 11% in ACM and 8.41% in MAG. In other words, the method provided identifies approximately 85% of the possible records. These results suggest our approach succeeds in matching the majority of bibliographic records, in both general and domain-specific data sets, to the relevant countries.

Table 2. Results of country identification

Results	ACM DL	MAG
Affiliations	384,672	831,888
NA, Null, etc values	52,454 (13.66%)	65,674 (7.89%)
Country names identified	136,671 (35.60%)	549,992 (66.11%)
Component parts identified	98,622 (25.69%)	31,738 (3.82%)
Identified by Wikidata	42,050 (10.94%)	69,985 (8.41%)
Not identified (Other values)	54,875 (14.29%)	116,982 (14.06%)

Conclusion

A current problem in IRC research is that it is difficult to identify countries by the affiliation information that bibliographic records provide. Previously, no method was available to overcome this, so methods were *ad hoc,* impractical, and likely inconsistent with each other, potentially resulting in varying results across even studies using the same data sets, or worse, preventing IRC measurement altogether. Here we provided and evaluated a novel method for addressing the problem, using substring matching and the SPARQL endpoint of the Wikidata knowledge graph. The method appears promising for use with other data sets as well, especially given that Wikidata will continue to grow and thus improve in matching affiliations to countries.

References

Chen, K., Zhang, Y., & Fu, X. (2018). International research collaboration: An emerging domain of innovation studies? *Research Policy*. doi:10.1016/j.respol.2018.08.005

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Vespignani, A. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics, 51*(1), 69-115. doi:10.1023/A:1010512628145

Guerrero Bote, V. P., Olmeda-Gómez, C., & de Moya-Anegón, F. (2013). Quantifying the benefits of international scientific collaboration. *Journal of the American Society for Information Science and Technology*, *64*(2), 392-404.

Luukkonen, T., Tijssen, R., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. Scientometrics, 28(1), 15-36.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. P., & Wang, K. (2015, May). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243-246). ACM.

[1] https://query.wikidata.org/sparql

[2] "[AFFILIATION]" in this query gets replaced with the remaining information extracted

[3] https://github.com/baxuan/IRCM/tree/master/Comparing-data-sets

How Grant Reviewers Evaluate Impact Statements: Two Cases from Science Foundation Ireland (SFI)

Lai Ma[1,] Junwen Luo[1], Thomas Feliciani[2] and Kalpana Shankar[1]

[1] *lai.ma@ucd.ie, junwen.luo@ucd.ie, kalpana.shankar@ucd.ie*
School of Information and Communication Studies, University College Dublin, Dublin 4, Ireland

[2] *thomas.feliciani@ucd.ie*
School of Sociology and Geary Institute for Public Policy, University College Dublin, Dublin 4, Ireland

Introduction

'Impact' is a contested concept in research assessment and science policy. Different frameworks and assessment strategies provide various definitions based on the purposes and objectives of research policy (Donovan, 2011; Langfeldt, Bloch & Sivertsen 2015; Penfield, 2014). Research impact is generally traced using citations, whereas economic and societal impact is generally evaluated using different frameworks (see, for example, Muhonen, Benneworth & Olmos-Peñuela, 2019; Spaapen & van Drooge, 2011; Bornmann, 2014). However, these measures and frameworks are mainly designed for *ex post* assessment.

For the peer review of grant proposals, *ex ante* assessment of economic and societal impacts are often required. Research councils and funding agencies often provide their own definitions and criteria for such assessments in lieu of using *ex post* assessment frameworks. Studies have shown that there are often discrepancies between the notion of impact defined by funding agencies and how it is understood by reviewers (de Jong et al., 2016), and that applicants and reviewers are reluctant to address the societal impact of a proposal as they feel more competent in assessing its intellectual merits (Holbrook & Frodeman, 2011).

In this work-in-progress, we examine two funding programmes at Science Foundation Ireland (SFI) to understand how reviewers interpret peer review criteria when evaluating proposals. Specifically, while "economic and societal impact" has been emphasised in both programmes, the term is defined differently for each call and in the guidelines for each set of reviewers. We use content analysis of reviewers' anonymised comments to understand how peer reviewers evaluate and assess economic and social impact. The main objectives of the two programmes as follows (Science Foundation Ireland, 2016, 2017):

- Industry Fellowship Programme (IF): to maximise the *economic and societal impact* of Irish state-funded research by developing and deepening effective industry-academia collaboration through research
- Investigators Programme (IvP): to support excellent scientific research that has potential *economic and societal impact* aligned with Ireland's research and innovation strategies

In future studies, we will be conducting semi-structured interviews with applicants and reviewers to further our understanding.

Data and Method

The researchers obtained a corpus of postal reviews of IF (2013-2017) and IvP (2012-2016) reviews which were redacted by SFI to ensure the anonymity of applicants and reviewers. Content analysis was conducted by two independent researchers. One researcher conducted this analysis in Nvivo 12 with an evolving coding scheme containing thirteen categories derived from the review criteria suggested by the SFI Guidelines. Another researcher conducted textual analysis and created an inductive coding scheme using a bottom-up approach. The two coding schemes will be compared and consolidated in a subsequent round of coding.

Preliminary Findings and Conclusion

Preliminary analyses show lack of specificity in reviewers' comments on impact statements and a tendency for reviewers to reiterate (i.e., confirm/reject) the review criteria prescribed by the funding agency. There is also a tendency to comment on scientific and economic impact rather than societal impact in their evaluations. For instance, when evaluating impact statements, reviewers tend to translate given criteria in such a way that their technical knowledge is used to comment on the feasibility of the proposal's contribution to economy and society. Also, reviewers often do not recommend collaborative applications with an unbalanced benefits and costs (in their estimation) between academic and industrial partners.

In sum, this study examines how grant reviewers evaluate impact statements with a broader objective of understanding the challenges of *ex ante* impact assessment. In the poster, we will present the results of the content analysis of the two programmes and compare them with relevant studies if appropriate (e.g., Reinhart, 2010).

Acknowledgments

This study is supported by the Science Foundation Ireland under Grant No.17/SPR/5319.

References

de Jong, S., Smit, J., & van Drooge, L. (2016). Scientists' response to societal impact policies: A policy paradox. *Science and Public Policy*, *43*(1), 102-114.

Donovan, C. (2011). State of the art in assessing research impact: introduction to a special issue. *Research Evaluation*, *20*(3), 175-179.Henshall, C. (2011). The impact of Payback research: Developing and using evidence in policy. *Research Evaluation*, *20*(3), 257-258.

Holbrook, J. B., & Frodeman, R. (2011). Peer review and the *ex ante* assessment of societal impacts. *Research Evaluation*, *20*(3), 239-246.

Langfeldt, L., Bloch, C. W., & Sivertsen, G. (2015). Options and limitations in measuring the impact of research grants—evidence from Denmark and Norway. *Research Evaluation*, *24,* 256-270.

Muhonen, R., Benneworth, P., & Olmos-Peñuela, J. (2019). From productive interactions to impact pathways: Understanding the key dimensions in developing SSH research societal impact. *Research Evaluation*. https://doi.org/10.1093/reseval/rvz003

Penfield, T., et al. (2014). Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation*, *23*(1), 21-32.

Reinhart, M. (2010). Peer review practices: A content analysis of external reviews in science funding. *Research Evaluation*, *19*(5), 317-331. DOI: 10.3152/095820210X12809191250843

Science Foundation Ireland (2016). SFI Investigators Programme 2016. Available at https://www.sfi.ie/resources/SFI-Investigators-Programme-2016-Call-Document.pdf

Science Foundation Ireland. (2017). Industry Fellowship Programme 2017. Call for Submission of Proposals. Available at https://www.sfi.ie/resources/SFI-Fellowship-Programme-07.07.17.pdf

Spaapen, J., & van Drooge, L. (2011). Introducing 'productive interactions' in social impact assessment. *Research Evaluation*, *20*(3), 211-218.

The Prospect of Chemistry Research in India

Swapan Deoghuria[1*] and Gayatri Paul[2]

[1]ccsd@iacs.res.in * *Corresponding Author*
System Manager, Indian Association for the Cultivation of Science, Jadavpur, Kolkata – 700032 (India)

[2]*libgp@iacs.res.in*
Documentation Superintendent, Indian Association for the Cultivation of Science, Jadavpur, Kolkata – 700032 (India)

Introduction

We see that all countries have their own strength and weaknesses in different subject domains in terms of research output. Our objective is to find out in which subject India is doing well and why. Data show that India is steadily improving in the field of chemistry research. Contribution by Indian researchers covered in Web of Science database is compared with other most productive countries. We have analysed the research activity of Indian scientists in terms of total number of publication, global share, share of international collaborative publications and visibility & citation impact for the period 2009-2014. The trend of research output in chemistry clearly indicates that India is steadily emerging as a potential contender in chemistry research.

Garg, Dutt & Kumar (2006), Glänzel & Gupta (2008), Gunasekaran, Batcha & Sivaraman (2006) and Gupta & Dhawan (2009) have studied the different aspects of trend in S&T research in India. Refining the data according to "Research Areas" as defined in Web of Science we see that India is doing very well in "Research Area" chemistry and stands at 3rd position since 2014.

Methodology

Data sources and processing

Bibliometric data have been extracted from WoS Core Collection during 2009-2014 in the research area "Chemistry".

Results and Discussions

In chemistry a total 1,085,080 number of papers has been published during the period 2009-2014. USA and China are leaders in this field in terms of number of publications with global share of 22.51% and 20.80% respectively. India is at 5th position with global share of 5.76%. Chemistry research output of ten most productive countries excluding USA and China in terms of global share has been shown in Figure 1. India stands on a firm position during this period and acquired 3rd position in 2014 followed by USA and China, with global share of 6.46%. India has published maximum number of research papers in Chemistry compared to other research areas and its global share in chemistry research has been increased steadily during 2009 to 2014. It is evident from Figure 1 that global share of Japan has been decreased from 2009 to 2014 and its positions in global ranking have been fallen from 3rd position (7.28%) in 2009 to 5th position (5.82%) in 2014. Although global share of Germany in Chemistry research has been decreased slightly during this period but Germany has managed to keep its position at 4th during the entire period. South Korea has made notable progress in Chemistry research during this period and it has improved its position from 10th (3.34%) in 2009 to 6th (4.33%) in 2014 in respect of global share. Iran has increased its research output in chemistry steadily in terms of global share during this period and in 2014 Iran is just behind Russia at 12th position. Research output of other countries (France, England, Spain, Italy) shown in this Figure are comparable to each other in chemistry and they are placed in between 7th to 10th positions during this period.

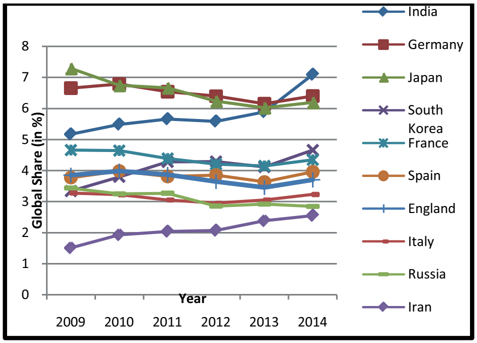


Figure 1. Global share of countries in chemistry

Table 1 shows India's ranking in major research areas covered in WoS during 2009-2014. We see that growth rate of India in terms of number of publications is more than that of global growth rate for chemistry. In terms of number of publications and global share, India's performance is the best in Chemistry.

In Table 2 we have shown the average citation per article of Indian publications in chemistry during 2009-2016. We see that average citations per article are comparable with that of productive countries like USA and China.

Conclusions

This study clearly indicates the trends in chemistry research during 2009-2014 for different countries in terms of number of publications and global share.

Table 1. **Position of India in major research areas in terms of global share**

Research Areas	2009	2010	2011	2012	2013	2014
Physics	10	9	8	8	7	**6**
Chemistry	5	5	5	5	5	**3**
Materials Science	7	6	6	6	**5**	6
Engineering	11	12	11	6	**4**	6
Computer Science	12	12	9	**3**	4	11
Biochemistry Molecular Biology	12	11	11	11	10	**9**
Neuroscience Neurology	18	17	17	16	**16**	17

Table 2. Citation of Indian publications in chemistry during 2009-2016

Year	2009	2010	2011	2012	2013	2014	2015	2016
Average Citation	20.6	19.7	18.9	17.2	15.86	14.2	11.6	8.75

Some of the possible answers have been listed below for success in Chemistry research in India. India was good in pure science research even before its independence and that legacy is still continuing. We find that most of the key persons in science policy makers in India are having chemistry background. Indian scientists working in the field of chemistry are more focused and recognized worldwide as many of them have been awarded TWAS prize and fellowship, FRS, and other distinguished international fellowships and medals. Many of them are members of various reputed international journals. Strong collaboration between India and other countries in chemistry research is worth mentioning as 11,424 numbers of papers out of total 62,196 are published in collaboration. Research scholars working in the field of Chemistry have a better chance to work in the best laboratories of the world as their research guides have a strong collaboration with those laboratories or they are alumni of those laboratories. As a traditional subject, most of the Indian universities teach chemistry and around 40% of total publications is contributed by the universities. Research laboratories also get a steady flow of trained students with chemistry background from universities. Looking at the distribution of the publications to the institutes we see that CSIR laboratories publish most (10,830) followed by IITs (8,470) and DST (3,397). Some of the most productive laboratories in chemistry research in India are BARC (2,472), IICT (2,633), IISc (2,207), IACS (1676) and NCL (1,580). Prominent universities in chemistry research are JU (1,298), DU (1,211) and BHU (1,184). We see that there is almost no role of industries as per the funding of research is concerned in the field of chemistry in India. CSIR, DST and UGC are the major sponsors in chemistry research in India. As per the topic or subject category is concerned where Indian scientists publish more, we see Physical chemistry is the most focused (29%) followed by Organic (19%), Inorganic (10%), Analytical (9%), Applied (8%), Nanoscience (5%) and Atomic-Molecular (5%) respectively. The bright side of chemistry research in India is also reflected in the number of patents granted in this subject area. From Derwent Innovations Index of WoS, we see that out of total 462 numbers of patents granted to Indian innovators during 2009-2014, 330 numbers i.e. 71% are in the field of chemistry. Interestingly, DRDO, India holds most (79%) of the patents. The picture is not much different in Indian patent database (http://ipindiaservices.gov.in/publicsearch/) where we see 4,801 numbers of patents (i.e. 37%) have been granted in chemistry research area out of total 12,982 patents granted in all fields during 2009-14. India has a large consumer base. As a result chemical industries in different sectors like fertilizer, pesticide, plastic, paint, petro-chemical, medicine, cosmetics and health care products are thriving in India. So career as research scientist in chemistry is attractive for better placement in the R&D labs of those industries. India's contribution in chemistry research has been recognized by ACS and designated IACS, Kolkata on 15/12/1998 as International Historic Chemical Landmark for C V Raman and the Raman Effect. In a study by Arunan, Brakaspathy, Desiraju, & Sivaram (2012) it is shown that India is still not using its full potential in Chemistry research and India can do much better in near future if suitable policy is implemented. The reasons behind subject preference of other countries may be studied separately.

References

Arunan E, Brakaspathy R, Desiraju G R & Sivaram S (2012). Chemistry in India – unlocking the Potential. Angew. Chem. Int. Ed. (51), 2–6.

Garg, K C, Dutt, B & Kumar, S. (2006). Scientometric profile of Indian sciences as seen through Science Citation Index. Annals of Library and Information Studies, 53, 114-125.

Glänzel, W & Gupta, B M (2008). Science in India. A bibliometric study of national and institutional research performance in 1991-2006. Proceedings of WIS 2008, Berlin.

Gunasekaran, S., Batcha, M. S. & Sivaraman, P. (2006). Mapping chemical science research in India: A bibliometric study. Annals of Library and Information Studies, 53, 83-95.

Gupta, B. M. & Dhawan, S. M. (2009). Status of India in science and technology as reflected in its publication output in scopus international databases, 1996 – 2006. Scientometrics, 80(2), 473-490.

Scientometric Implosion of Armenian Journals

Shushanik Sargsjan[1] Aram Mirzoyan[2] Viktor Blaginin[3]

[1] *shushaniksargsyan8@gmail.com*
Center for Scientific Information Analysis and Monitoring (CSIAM), Institute for Informatics and Automation
Problems of the National Academy of Sciences
1 Paruyr Sevak str., 0014 Yerevan, Armenia
Medical Physics Department, Yerevan State Medical University after Mkhitar Heratsi
2 Koryun str., 0025 Yerevan, Armenia

[2] *mirzoyan.aram@gmail.com*
Center for Scientific Information Analysis and Monitoring (CSIAM), Institute for Informatics and Automation
Problems of the National Academy of Sciences
1Paruyr Sevak str., 0014 Yerevan, Armenia

[3] *v.a.blaginin@usue.ru*
Center on Scientometrics and Ranking Researches, Ural State University of Economics
8th of March str., 620014 Ekaterinburg, Russia

Introduction

Evaluation of scientific journals has almost a century long history. Most probably the first and already a classic study in the field was published by Gross and Gross in 1927 titled “College libraries and chemical education” (Archambault & Larivière, 2009). One of the main issues of this study was to find out which periodicals to purchase for the libraries, especially of the small colleges with relatively small budget with a view to find out the most important and influential periodicals for the given scientific field (Archambault & Larivière, 2009). The idea of impact factor (IF) was presented by Eugene Garfield in 1955 (Garfield, 2006). Later on, in 1961, the Science Citation Index (SCI) was published. Shortly after that (in 1963) E. Garfield and Irving Sher presented the journal impact factor by re-sorting the author citation index into the journal citation index (Garfield, 2006). It was done to assist in selecting journals for the SCI.

There was and still is a criticism on the journal IF. However as Hoeffel (1998) notes “IF is not a perfect tool to measure the quality of articles, but there is nothing better...”. The issue of evaluation of the scientific journals has received importance in Armenia since 1991 when the country regained its independence. One of the main challenges for Armenia is integration into the international scientific community as an independent unit. For that end the preservation and further development of a strong scientific community has become vital. The establishment of a high-quality network of Armenian scientific journals is one of the steps towards this goal. Until 2010 there were only two tools for measuring Armenian scientific journals: a) international indexing platforms and b) the list of recommended journals by the Supreme Certifying Commission (the special state agency granting academic degrees and statuses). Since 2010 the Center for Scientific Information Analysis and Monitoring has imported the third tool – Armenian Journal IF (ArmJIF) providing the scientific community both in Armenia and abroad with another indicator for Armenian scientific journals.

Current state of Armenian Scholarly Journals

Nowadays there are nearly 120 scholarly journals in Armenia, but only about 100 of them are included in the ArmJIF database (Figure 1) due to different reasons (absence of archive, irregular periodicity, large number of non-scientific articles, etc.).

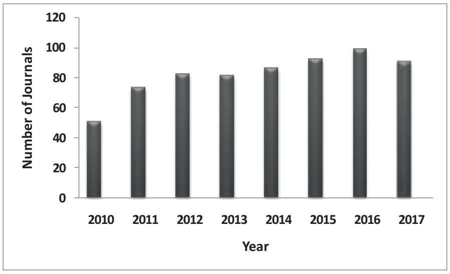


Figure 1. The number of Armenian journals in the ArmJIF database (2010-2017)

The use of IF in the assessment of Armenian scholarly journals has contributed to the raise of competition, maintenance of publication ethics, online access of journals, trilingual bibliographic data, etc.

Unfortunately, only 5% of Armenian scholarly journals are indexed in international scientific databases (ISD) such as Web of Science (WOS), Scopus, Russian Index of Scientific Citation (RISC) (Table 1).

Table 1. The list of Armenian journals indexed in WOS, Scopus, RISC Core Collection

№	**WOS**	**Scopus**	**RISC Core**
1	Astrophysics	Astrophysics	Astrophysics
2	J CONTEMP PHYS-ARME	J CONTEMP PHYS-ARME	J CONTEMP PHYS-ARME
3	J CONTEMP MATH ANAL	J CONTEMP MATH ANAL	J CONTEMP MATH ANAL
4	NAMJ **(ESCI)**	NAMJ	-
5	WISDOM **(ESCI)**	WISDOM	-
6	Armen.J. Math. **(ESCI)**	Armen.J. Math.	Armen.J. Math.

Taking into consideration the above mentioned, it is interesting to see the visibility of local journals not indexed in the ISD.

Table 1 demonstrates the citations from the WOS (to cover international field) and the citations from the RICS (to cover the so-called Russian field).

The Table also reveals that the subject coverage of journals is extremely limited and their citation rate is quite modest.

Table 2. Total Citations of journals indexed in WOS, RISC Core (2013-2017)

№	Journal Name	ISSN	Times cited (WOS)	Times cited (RISC Core)
1	ASTROPHYSICS	0571-7256	2277	1073
2	J CONTEMP PHYS-ARME	1068-3372	548	185
3	J CONTEMP MATH ANAL	1068-3623	337	132
4	NAMJ	1820-0254	23	-
5	Armen.J.Math.	1829-1163	36	0
6	WISDOM	1829-3824	8	11

Meanwhile, this testifies about the serious scientific territorial enclosure of the journals, since they do not aim at international level or even pursue such goals. Here, the authors put forward an idea of using a special phrase describing this phenomenon – ***scientometric implosion*** – and bring the term into circulation and use by scientometric researchers.

Definition of *implosion* is rooted in the natural sciences where it means compression of an object (Dvoryadkina and Kaibicheva, 2006). Meanwhile, when the object is subjected to endogenous and exogenous pressure, it can "explode." Projecting such a phenomenon on scientometrics and informetrics, it can be claimed the existence of a compressed space in the country, namely the journals being not ready to lose their own national auditorium break the country's boundaries while promoting scientific knowledge. However, as soon as they acknowledge the prospects of development, an explosion of citation takes place and journals and their articles become more popular.

Table 3 presents top 6 Armenian journals not indexed in any IDS according to their citations from the WOS. Although these journals are not indexed in any ISD, they are definitely worth being included.

Table 3. Top 6 Armenian journals according to citations received from the WOS (2013-2017)

№	Journal Name	Times cited (WOS)
1	Armenian Journal of Physics	49
2	Chemical Journal of Armenia	32
3	Biological Journal of Armenia	30
4	Proceedings of the YSU	28
5	Mathematical Problems of Computer Science	20
6	Issues in Theoretical and Clinical Medicine	19

References

Archambault, É. & Larivière,V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, 3 (79), 635-649.

Dvoryadkina, E., Kaibicheva, C. (2017). *Regional Peripheria: Place in Space.* Monograph. Ural State University of Economics. Russia.

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. *Journal of the American Medical Association*, 1 (295), 90-93.

Hoeffel, C. (1998). Journal impact factor. *Allergy*, 12 (53), 1225.

Detection of disruptive technologies by automated identification of weak signals in technology development

Geraldine Joanny[1], Sergio Perani[2] and Oliver Eulaerts[1]

[1] *Geraldine.JOANNY@ec.europa.eu; Olivier.EULAERTS@ec.europa.eu*
Joint Research Centre, European Commission, CDMA 00/P157, Brussels, BELGIUM

[2] *Sergio.PERANI@ext.ec.europa.eu*
Joint Research Centre, European Commission, IPR 44 00/038, Ispra, ITALY

Introduction

Text mining techniques applied to scientific literature and patents are commonly used to identify weak signals of new technological developments or new applications of existing technologies.

Our research focuses on developing a method and an IT tool to automatically detect these technologies or technological domains that exhibit certain characteristics of emergence. This is done through defining indicators that can be combined to calculate a weak signal score for an emerging issue or technology. For an organisation like the Joint Research Centre, acting at the interface between science and policy making, such a capacity to identify (pre-)emerging technologies is important to assess their possible future impact on the development or implementation of EU policies (see Moro, Boelman, Joanny, & Lopez-Garcia, 2018). Previous attempts by the JRC to characterise emerging technologies were based on network analysis (Joanny et al., 2015).

Methodology

We apply text mining techniques, developed in the context of the TIM project (www.timanalytics.eu), to big corpus of textual data such as scientific publications.

Step 1 - Generation of the dictionary

A dictionary of multi-word concepts is generated from the corpus of documents using text mining. A corpus composed of 5 years of scientific publications from the Scopus database was used for the current study. Single words, multi-word terms and acronyms are extracted from the title, abstract or author keywords of the publications. TF*IDF is then used for selecting the most relevant keywords. We also apply stemming (reduction of words to their stem) as an approximate method for grouping words with a similar basic meaning together. The resulting dictionary is a list of more than 4 million terms relevant for the corpus.

Step 2 - Generation of the initial datasets

Search queries are automatically generated for each of the concepts in the dictionary to retrieve documents from Scopus (1996-2018). As a result, datasets containing documents retrieved for each of the concept in the dictionary are obtained.

Steps 3 to 5

The methodology to be used for ranking the datasets (step 3), selecting datasets for analysis (step 4) and calculate a weak signal score (step 5) is investigated in this study.

After automated generation of the datasets, a methodology to obtain a relevant number of potential weak signals to be analysed has been devised:

- ranking: we first developed a method to prioritise the datasets to analyse using a simple and fast-to-calculate indicator;
- select: additional filters were applied to the datasets to discard weak signals of poor quality (noise, false positive, etc.);
- further ranking the cleaned datasets: composite indicators reflecting the characteristics of a weak signal for emerging technologies was devised and applied on the cleaned datasets to rank them.

Results

The most successful methods for ranking and selecting the datasets and the calculation of a weak signal indicator are presented hereafter.

Step 3 – Ranking the datasets

We investigated the use of an indicator "activeness[y]", with y = a number of years typically ranging from 1 to 7. Activeness[y] is computed as the ratio: [(number of documents for the last y years)/(total number of documents) x 100] for each of the datasets. A high activeness[y] score would indicate that a higher percentage of documents have been published over the last [y] years and could be used to rank the datasets. Various activeness[y] indicator were tested, and y=1, y=2, y=3 gave the best results for the analysis of weak signals.

Step 4 - Selecting the datasets for analysis

Filters are needed to reject datasets showing a high activeness score but not related to an actual

technology signal. Three filters proved to be useful to exclude datasets from the analysis.
- A simple filter on size is used to reject datasets that do not reach a certain minimum number of documents (for example 10).
- A filter relying on the "compactness" of the datasets is used to reject those datasets containing documents that are not semantically similar. Compactness refers to the percentage of documents that have a minimum threshold cosine similarity distance to other ten documents within a dataset and reflects the semantic coherence of documents within a dataset. The objective is to reject datasets with only one common concept from the dictionary but pertaining to completely different conceptual areas (e.g. documents related to a generalist conference where the only common term between the documents is the name of said conference). Datasets with low compactness are rejected.
- Finally, semi-automated manual filtering is used to reject datasets resulting from errors in the original corpus of data (e.g. spelling mistakes).

At the end of this step there is a manual check of the 250 datasets with the highest score for activeness[3] after filtering. The datasets definition (queries) might also be manually improved.

Step 5 – Calculating a composite indicator for weak signal of an emerging issue or technology

Our system allows the computation of many indicators or combination of indicators. In the present study, indicators were used on the selected datasets to detect characteristics of emergence in technology development. The following indicators were calculated and tested:

Activeness[2]
In addition to the ranking by activeness[3], we want to evaluate the usefulness of taking into account also the activeness in the last 2 years. Emerging technologies are expected to not only score high on the proportion of publications in the last 3 years but also for the past 2 years.

Activeness[1]
In a similar way, we calculate the proportion of articles in the last year to ensure that the weak signals we are considering are very recent. This indicator emphasizes the weak signal detection.

distinct_journaltitle
This indicator calculates the total number of different journals where the scientific publications in the dataset have been published. On one hand, if the number of different journals is very low it might be that the dataset pertains to a technological niche or that it is noise created by a particular issue. On the other hand, a very large number of journal titles could indicate that a technology is mature and has already diffused in many different S&T areas.

coverage_confproc
We assume that weak signals of technological development have been discussed first at international scientific conferences. They should therefore be a significant number of conference proceeding documents in the pre-emerging datasets. We designed an indicator that calculates the percentage of documents of the type "conference proceeding" for each of the datasets. This indicator is scientific-field specific and had to be used with caution.

After step 3 and step 4, the indicators described above were combined to prioritise even more the weak signals prior to a detailed analysis phase.
The "weak signal score" was calculated as the product of the indicators. Some fine-tuning is still necessary to assign different weights to each of the indicators but this score was useful to discern the technological weak signals in scientific literature.

A similar approach will be used for the detection of weak signals in patent applications.

Conclusions

An IT system was successfully developed and used for the automatic identification of weak signals of emerging technologies. The results will be fed into the horizon scanning mechanism of the JRC allowing experts and policy makers to evaluate the disruption potential of these signals and devise the adequate policy or regulatory responses.

References

Joanny, G., Agocs, A., Fragkiskos, S., Kasfikis, N., Le Goff, J.-M., & Eulaerts, O. (2015). Monitoring of technological development - Detection of events in technology landscapes through scientometric network analysis. In *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference*.

Moro, A., Boelman, E., Joanny, G., & Lopez-Garcia, J. (2018). A bibliometric-based technique to identify emerging photovoltaic technologies in a comparative assessment with expert review. *Renewable Energy*, *123*. doi:10.1016/j.renene.2018.02.016

Societal impact of scientific work in in the process of re-accreditation of higher education institutions and public scientific institutes in the Republic of Croatia.

Marina Grubišić[1]

marina.grubisic@azvo.hr

Agency for Science and Higher Education, Donje Svetice 38, Zagreb (Croatia)

Introduction

In this paper we will present the results of a study aimed in the direction of developing a theoretical framework for measuring societal impact of research. We will focus on a case study of the evaluation process for measuring performance of public higher education institutions (HEIs) at universities and public research institutes in Republic of Croatia. We expect to develop a hybrid framework, which will be a pondered combination of standard bibliometric metrics combined with a metric measuring the societal impact of research. For that we will follow (Spaapen and van Drooge, 2011) which proposes evaluation based on assessing a sample of societal interactions submitted by an institution under evaluation as representative for their practice. In this context, basing our case study on the reports submitted by Croatian HEIs and public research institutes in the process of reaccreditation is meaningful since there precisely such data is collected and reported by the institutions. Additionally, we will use reports issued by expert evaluators to conduct a content analysis as a basis of objective measurement and thus reduce a possible bias in the results. In the rest of the paper, we will review bibliometric indicators as a measurement tool, and then we will present a framework of the accreditation process in Croatia. Finally, we will show preliminary results and indicated initial conclusions as well as possible directions for further research.

Measuring the societal impact of research

Bibliometric evaluation is focused primarily on analysing the scientific impact. Bibliometric indicators typically do not measure the societal impact of research. Furthermore, the notion of the societal impact is not uniquely defined across national and international scientific systems and there is no accepted taxonomy of what the notion of societal impact should include.

We have developed a conceptual framework for measuring societal impact based on the Theory of productive interaction. Productive interaction can be categorized in three categories according to the taxonomy of (Spaapen and van Drooge, 2011):

- Direct interaction (DI)
- Indirect interaction (II)
- Financial interaction (FI)

We will validate the model by applying it on two scientific fields in Croatia.

Case study Croatia

In the Republic of Croatia, the social impact of scientific work has not been evaluated so far in the above-mentioned categories. We will concentrate on independent constituent parts of universities (faculties) and public scientific institutes in the fields of biomedicine and social sciences. These areas are a good example of the fields where scientific work has influence on the wider society.

After we formulate a framework of measurement and evaluation of social impact, qualitative research will be carried out on the reports of expert evaluators in the process of re-accreditation of higher education institutions and public scientific institutes in Croatia. The analysis includes only faculties of public universities since professional schools of higher education and polytechnics in the Republic of Croatia according to the Act on Scientific Activity and Higher Education(AZVO, 2010[1]) are not obliged to carry out scientific activity or do not have to have an accredited scientific license.

The evaluation was carried out in accordance with the Criteria for assessment of quality of higher education institutions within universities (AZVO, 2013[2]) which were used from 2010. to 2016. The analysis will focus on criteria evaluating the societal impact of scientific work according to the presented framework.

The public science system in Croatia formally includes both higher education institutions within the university as well as scientific institutes. The Agency for Science and Higher Education has carried out the process of re-accreditation of all public scientific institutes according to criteria for evaluating exclusively the scientific activity, i.e. evaluation process for issuing scientific license.

Evaluation of public scientific institutes was carried out in accordance with the Principles and criteria for evaluation of scientific organizations in the Republic of Croatia (AZVO, 2013[3]).

Methodology

The analysis was conducted using a qualitative data processing tool. All the above reports were analyzed with the software tool [4] according to the principle of productive interaction. A text is marked according to the number of instances when a coded interaction (grouped further in three categories) is detected in the body of text.

Table 1. Number and percentage of codes and cases

CATEGORIES		CODES		CASES	
DI	Professional conferences	3	0,02	3	0,094
DI	Professional bodies	21	0,14	21	0,656
DI	Management bodies	6	0,04	6	0,188
DI	Meeting stakeholders	24	0,16	24	0,75
DI	Collaboration with public services	21	0,14	21	0,656
II	Professional publications	6	0,04	6	0,188
II	Media presence	7	0,047	7	0,219
II	Social media presence	3	0,02	3	0,094
II	Reporting to local governance	12	0,08	11	0,344
FI	Commercial contracts	16	0,107	16	0,5
FI	Professional contracts	27	0,18	26	0,813
FI	Financing of students				
FI	Financing of teachers	4	0,027	4	0,125

Conclusion and further research

Table 1. presents the number and percentage of detected coded interactions and cases. The results are shown for the group of faculties of public universities and public scientific institutes in the scientific fields of biomedicine and social sciences.

The key terms detected for direct interaction are; participation in professional bodies and conferences, meetings with stakeholders, membership in management bodies and collaboration with public services. As we can see in Table 1. the majority of codes for direct interaction are; meeting with stakeholders in 75% of cases and participation in professional bodies in 65 % of cases.

The key terms for indirect interaction are; professional publications, presence in the media, presence on social networks and reporting to local governance. Table 1. shows that the majority of codes for indirect interaction are; reporting to local governance in 34% of cases and media presence of institution participation in 65 % of cases.

The key terms s for financial interaction are commercial and professional contracts, financing of students and financing of teachers. The dominant codes for financial interaction are; professional contracts in 81% of cases and commercial contracts in 50 % of cases.

Through this qualitative analysis, we obtain a first overview of the conceptual framework recognized by the expert committees as the social impact of scientific work at faculties of public universities and public scientific institutes in the social and biomedical scientific field. Based on these findings we will be able to better plan the scope of further more detailed research.

Acknowledgments

This research has been performed as a part of the Postgraduate doctoral programme Information and Communication Sciences at Faculty of Humanities and Social Sciences, University of Zagreb under the supervision of prof. dr. sc. S. Špiranec and as part of the work in Agency for Science and Higher Education under the supervision of prof. dr. sc. J. Havranek. The author gratefully acknowledges the support.

References

Bornmann, L. (2013) 'What is societal impact of research and how can it be assessed? a literature survey', *Journal of the American Society for Information Science and Technology*, 64(2), pp. 217–233. doi: 10.1002/asi.22803.

Bornmann, L. and Marx, W. (2014) 'How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparisons', *Scientometrics*, 98(1), pp. 211–219. doi: 10.1007/s11192-013-1020-x.

Derrick, G. E. and Samuel, G. N. (2016) 'The Evaluation Scale: Exploring Decisions About Societal Impact in Peer Review Panels', *Minerva*. Springer Netherlands, 54(1), pp. 75–97. doi: 10.1007/s11024-016-9290-0.

Spaapen, J. and van Drooge, L. (2011) 'Introducing "productive interactions" in social impact assessment', *Research Evaluation*, 20(3), pp. 211–218. doi: 10.3152/095820211X12941371876742.

[1].https://www.azvo.hr/images/stories/o_nama/Act_on_Scientific_Activity.pdf

[2]https://www.azvo.hr/en/evaluations/evaluations-in-higher-education/re-accreditation-of-higher-education-institutions-2010-2016

[3]https://www.azvo.hr/en/evaluations/evaluations-in-science/re-accreditation-of-scientific-organisations/re-accreditation-of-public-research-institutes

[4] https://provalisresearch.com/products/qualitative-data-analysis-software/

Can Anti-Cocitations Also Measure Author Relatedness?

Maria Cláudia Cabrini Grácio[1] and Dietmar Wolfram[2]

[1] *cabrini.gracio@unesp.br*
São Paulo State University (UNESP), Av. Hygino Muzzi Filho 737, 17525-900, Marília (Brazil)

[2] *dwolfram@uwm.edu*
University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201 (United States of America)

Introduction

Author co-citation analysis (ACA) provides a broader understanding of the intellectual structure of a scientific field and its theoretical and methodological currents as defined by the scientific community. With ACA one can visualize the intellectual connections between authors as established over time through citations given by members of the research community. As authors are co-cited in publications, links between authors are established and strengthened (White & McCain, 1998; Small, 2004; Cronin & Shaw, 2002). Despite the recognized contributions of ACA that allow researchers to visualize the scientific structure defined by active citers, these analyses are not able to identify the absences, incompatibilities and partitions also defined by the citers through the act of identifying and citing references.

Using a dual approach to the notion of author co-citation, which focuses on inclusion for evaluating the strength of relationships for grouping or partitioning authors, one may also be able to assess the strength of the relationship between two authors based on the mutual exclusion of scientific literature citations. By measuring the extent to which citers recognize authors in a mutually exclusive way by citing one author and not the other, we can conceive the notion of an anti-cocitation between two authors, defined as the count of citations to one author and not to the other per unit of published research. If there is a publication that cites author A and does not cite author B, this represents a unit of the author anti-cocitation relation. The higher the frequency of anti-cocitations between two authors, the more distant they are according to the understanding of the citing community. Thus, we presume that author anti-cocitation strength identifies the partition/mutual exclusion between two authors, as determined by the citing community.

In order to complement and refine the visualization of the theoretical-methodological structure of a scientific field, this research aims to evaluate the contribution of anti-cocitations to identifying not only the similarities between authors, but also the partitions or mutual exclusions between them. More specifically, this research presents a “proof of concept” by identifying and comparing the relatedness (co-citations) and mutual exclusions (anti-cocitations) between the 28 Derek de Solla Price Memorial Medal winners.

Methods

The corpus of articles used consisted of 2,966 papers indexed in the Scopus database between 2014 and 2018 published in the journals *Scientometrics*, *Journal of Informetrics*, and *Journal of the Association for Information Science and Technology*. We retrieved the citations for each medal winner in Scopus and each medalist’s cocitation and anti-cocitation frequencies with the other 27 authors. Next, we normalized the resulting cocitation frequency matrix using Salton’s Cosine measure. From the 28x28 asymmetric anti-cocitation frequency matrix we created a normalized symmetric anti-cocitation matrix by defining an anti-cocitation index (ACI) between author A and author B as:

$$\mathrm{ACI}_{(A,B)} = \frac{\sqrt{\mathrm{anticocit(A,B)} \times \mathrm{anticocit(B,A)}}}{\sqrt{\mathrm{cit(A)} \times \mathrm{cit(B)}}}$$

where cit(X) = total number of papers citing X; anticocit(X, Y) = total number of papers citing X and not citing Y. The resulting $\mathrm{ACI}_{(A,B)}$ value ranges from 0 to 1. An ACI value of 1 indicates the citations for each author are mutually exclusive. A value of 0 indicates complete overlap.

We performed a cluster analysis using Ward’s Method for each matrix (normalized cocitation and normalized anti-cocitation) to generate a dendrogram to compare the clustering of authors from cocitation (similarity) and anti-cocitation (dissimilarity) perspectives. The two resulting cluster outcomes were compared by means of the intra-group averages for cocitations and anti-cocitations to verify which matrix offered better results in terms of both high similarity and low dissimilarity.

Results

The cluster analysis performed on the normalized cocitation matrix produced four clusters of authors (C1, C2, C3, C4) (Table 1). We observe that C1 and C3 contain cocitation averages for member authors that are higher than those for authors who are outside of these clusters. Considering that cocitation averages among authors within C1 ranged from 0.26 and 0.39, all authors who are outside C1 had cocitation averages smaller than those of the authors inside this cluster, with the exception of Rousseau (cocitation average = 0.28). This characteristic is also observed in C3, with no exceptions. These findings show that these author clusters (C1 and C3) are consistent based on similarity. Conversely, C2 and C4 do not display similar characteristics to C1 and C3 since members of C2 and C4 do not consistently have larger cocitation averages than authors who are outside the group. This suggests that clusters C2 and C4, obtained through clustering based on cocitations, are not internally homogeneous in terms of cocitation average. Also in Table 1, we can observe that C1 and C3 contain the lowest intra-group anti-cocitation averages, i.e., a smaller proportion of mutually exclusive citations. On the other hand, also in relation to the anti-cocitation indices, C2 and C4 do not represent well-defined clusters given the lack of intra-group consistency when compared to values for authors in other clusters.

Table 1. Cocitation & anti-cocitation averages cluster analysis based on the cocitation matrix

	Avg.Cocitation Cosine Value				Avg. ACI Value			
Clusters	C1	C2	C3	C4	C1	C2	C3	C4
Schubert	0.39	0.19	0.12	0.05	0.58	0.78	0.86	0.84
Glänzel	0.42	0.24	0.19	0.06	0.54	0.70	0.75	0.75
Leydesdorff	0.37	0.24	0.21	0.06	0.59	0.69	0.70	0.74
Braun	0.29	0.14	0.07	0.03	0.66	0.85	0.93	0.96
Van Raan	0.37	0.22	0.17	0.05	0.61	0.74	0.80	0.82
Moed	0.33	0.20	0.13	0.03	0.65	0.77	0.84	0.91
Garfield	0.31	0.19	0.20	0.08	0.68	0.78	0.76	0.74
Zitt	0.26	0.11	0.15	0.02	0.67	0.89	0.85	0.97
Thelwall	0.23	0.17	0.09	0.02	0.76	0.81	0.90	0.95
Bar-Ilan	0.16	0.14	0.08	0.03	0.82	0.85	0.92	0.92
Ingwersen	0.09	0.09	0.10	0.03	0.87	0.89	0.89	0.96
Martin	0.23	0.18	0.11	0.05	0.76	0.77	0.87	0.85
Persson	0.22	0.12	0.16	0.04	0.75	0.88	0.84	0.92
Merton	0.18	0.11	0.13	0.03	0.79	0.88	0.86	0.94
Cronin	0.17	0.15	0.18	0.06	0.82	0.84	0.81	0.89
Narin	0.23	0.11	0.10	0.05	0.74	0.88	0.89	0.88
Irvine	0.13	0.12	0.03	0.02	0.75	0.76	0.96	0.98
Rousseau	0.28	0.18	0.17	0.04	0.71	0.80	0.81	0.90
Egghe	0.22	0.20	0.10	0.05	0.76	0.74	0.89	0.88
Vinkler	0.17	0.10	0.05	0.03	0.75	0.88	0.95	0.95
White	0.15	0.12	0.32	0.05	0.85	0.87	0.62	0.87
McCain	0.13	0.11	0.36	0.06	0.83	0.87	0.60	0.90
Small	0.24	0.13	0.30	0.09	0.74	0.87	0.66	0.79
Griffith	0.11	0.07	0.32	0.04	0.83	0.91	0.62	0.94
Brookes	0.05	0.05	0.08	0.05	0.87	0.90	0.84	0.95
Vlachy	0.03	0.02	0.00	0.02	0.83	0.93	0.99	0.97
Moravcsik	0.08	0.06	0.09	0.02	0.83	0.91	0.85	0.97
Nalimov	0.04	0.02	0.08	0.00	0.84	0.93	0.81	1.00

From the cluster analysis performed using the ACI matrix, five groups of authors were identified (Table 2). We observe that G1, G2, G3, and G4 display similar characteristics for intra-group consistency (highest cocitation values) to those observed for C1 and C3 in Table 1. Furthermore, only group G5, resulting from the cluster analysis based on the ACI matrix, shows less consistency with regard to cocitation average outcomes - a similar characteristic that is observed in clusters C2 and C4 in Table 1 for the cocitation outcomes.

Also in Table 2, we can observe that G1, G2, G3, and G4 for the anti-cocitation averages display similar characteristics for internal consistency (lowest anti-cocitation averages). Furthermore, only group G5, resulting from the cluster analysis based on the ACI matrix, displays less consistency for the anti-cocitation averages, which is a similar outcome to the characteristics of clusters C2 and C4 in Table 1 for the anti-cocitation outcomes.

Table 2. Cocitation & anti-cocitation averages cluster analysis based on the ACI matrix

	Avg. Cocitation Cosine Value					Avg. ACI Value				
Clusters	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5
Braun	0.26	0.15	0.07	0.08	0.08	0.70	0.83	0.93	0.91	0.90
Schubert	0.35	0.18	0.12	0.12	0.13	0.62	0.77	0.86	0.87	0.81
Rousseau	0.26	0.26	0.17	0.15	0.10	0.73	0.70	0.81	0.84	0.86
Vinkler	0.16	0.17	0.05	0.06	0.06	0.77	0.80	0.95	0.93	0.93
Zitt	0.22	0.08	0.15	0.11	0.06	0.73	0.91	0.85	0.88	0.94
Persson	0.20	0.08	0.16	0.10	0.09	0.77	0.91	0.84	0.90	0.88
Glänzel	0.39	0.20	0.19	0.20	0.15	0.56	0.70	0.75	0.76	0.74
Leydesdorff	0.34	0.18	0.21	0.20	0.16	0.60	0.69	0.70	0.75	0.71
Van Raan	0.33	0.20	0.17	0.20	0.15	0.65	0.69	0.80	0.78	0.78
Moed	0.29	0.18	0.13	0.22	0.12	0.69	0.74	0.84	0.76	0.85
Garfield	0.28	0.21	0.20	0.16	0.14	0.71	0.73	0.76	0.82	0.76
Irvine	0.11	0.41	0.03	0.05	0.08	0.80	0.00	0.96	0.92	0.84
Egghe	0.24	0.41	0.10	0.12	0.10	0.75	0.00	0.89	0.87	0.87
White	0.14	0.06	0.32	0.14	0.10	0.85	0.93	0.62	0.85	0.86
McCain	0.12	0.07	0.36	0.09	0.10	0.84	0.92	0.60	0.90	0.87
Small	0.22	0.10	0.30	0.08	0.12	0.76	0.89	0.66	0.92	0.82
Griffith	0.11	0.04	0.32	0.06	0.07	0.84	0.95	0.62	0.94	0.91
Thelwall	0.21	0.12	0.09	0.30	0.08	0.79	0.84	0.90	0.65	0.90
Bar-Ilan	0.14	0.08	0.08	0.29	0.08	0.84	0.91	0.92	0.69	0.90
Ingwersen	0.09	0.06	0.10	0.19	0.06	0.88	0.94	0.89	0.78	0.93
Brookes	0.05	0.06	0.08	0.05	0.04	0.88	0.87	0.84	0.90	0.93
Vlachy	0.03	0.03	0.00	0.01	0.02	0.85	0.94	0.99	0.96	0.95
Moravcsik	0.07	0.04	0.09	0.03	0.06	0.86	0.94	0.85	0.95	0.90
Cronin	0.16	0.10	0.18	0.18	0.10	0.82	0.88	0.81	0.81	0.87
Merton	0.16	0.10	0.13	0.11	0.07	0.82	0.89	0.86	0.88	0.91
Narin	0.19	0.16	0.10	0.05	0.10	0.79	0.80	0.89	0.94	0.86
Martin	0.22	0.22	0.11	0.13	0.11	0.77	0.55	0.87	0.86	0.83
Nalimov	0.03	0.01	0.08	0.02	0.02	0.86	0.98	0.81	0.95	0.93

Final considerations

The clustering outcomes based on the author anti-cocitation matrix produced groups that have better consistency than clustering based on the author cocitation matrix, considering that 4 of the 5 groups clustered using anti-cocitations are robust in their cocitation and anti-cocitation averages when compared to authors outside the groups, while only 2 of the 4 groups resulting from the cluster analysis based on the author cocitation matrix present this characteristic. Future research will expand our investigation and will examine how cocitations and anti-cocitations may be combined for more nuanced assessments of disciplinary structures and author relatedness.

References

Cronin, B. & Shaw, D. (2002). Identity-creators and image-makers: using citation analysis and thick description to put authors in their place. *Scientometrics,* 54, 31 – 49.

Small, H. (2004). On the shoulders of Robert Merton: towards a normative theory of citation. *Scientometrics*, 60, 71-79.

White, H.D. & McCain, K.W. (1998). Visualizing a discipline: an author co-citation analysis of Information Science, 1972-1995. *Journal of the American Society for Information Science,* 49, 327-355.

How open are journal articles with open access topic?

Carey Ming-Li Chen[1] and Wen-Yau Cathy Lin[2]

[1] *carey.mlchen@gmail.com*
Science & Technology Policy Research and Information Center, National Applied Research Laboratories, 14F, No.106, Sec.2, Heping E. Rd., Taipei city 10636 (Taiwan)
Graduate Institute of Business Administration, National Taiwan University, No.1, Sec. 4, Roosevelt Rd., Taipei City 106 (Taiwan)

[2] *wylin@mail.tku.edu.tw*
Department of Information and Library Science, Tamkang University, No.151, Yingzhuan Rd., Tamsui Dist., New Taipei City 25137 (Taiwan)

Introduction

Since the term "open access" (OA) was first formulated in *Budapest Open Access Initiative*, *Bethesda Statement on Open Access Publishing*, and *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* in the early 2000s, the scholarly community nowadays is moving OA2020 vigorously. We may experience the 3rd paradigm shift in scholarly publishing in a very near future but when we look down at the ground, we can't help wondering how many articles in the topic of OA are published in the mechanism of OA. This research aims to explore how open of journal articles with OA theme.

Several research objectives will be investigated including the percentage of OA and non-OA mechanism, publishing year trend, concentration of journals and countries of authors' affiliations. This research will help us understand how many OA issues related articles participate OA movement with practical actions.

Data collection and cleaning

The sample articles are searched by the following strategies from Web of Science (SCI-Expanded, SSCI) in March 2019.

> (TI=("open access" OR "OA journals"
> OR "OA articles" OR "OA publication"
> OR "full OA" OR "hybrid OA")) AND
> LANGUAGE: (English) AND
> DOCUMENT TYPES: (Article OR
> Review) Indexes=SCI-EXPANDED,
> SSCI Timespan=2008-2017

The preliminary search result is 823 articles. Because OA issues may appear in various fields of journals, OA abbreviations may also represent multiple meanings, for example, Open Access Endoscopy or Open Access endoscopy in medicine, and open access area in fishery. It is necessary to do further manually checking on title, abstract, and keywords in each article to ensure research validity. Ultimately, 377 articles are moved from the data set and the accurate sample number is 446 articles.

Research Results

What was the trend for OA topic publishing?

A total of 163 journal articles were published by OA mechanism, and the share is 36.54%. The Figure 1 demonstrates the trend of journal articles with OA topic and it breaks down the share of OA publishing. The number of article with OA topic has increased rapidly and the number of articles published in OA mechanism grown as well. In the beginning of the period, even these articles with OA topic, the authors seldom chose OA mechanism. However, the share of OA mechanism has grown eventually, and especially in 2016, almost half of articles with OA topic have chosen OA mechanism to let the articles be accessible more easily.

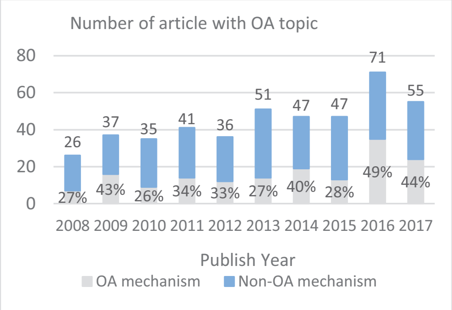


Figure 1. The number of articles with OA topic and its share of type of OA mechanism

Where were they published?

Table 1 presents the top 10 journal distribution of articles with OA topic and what percentage of those articles chose OA mechanism. The most popular journal for authors who worked on OA topic is *Learned Publishing*, the share of OA publishing is 63.89%. The second place is *Scientometrics*, however, for the authors who published OA topic articles in this journal, only 14.81% of them applied OA mechanism. Although *Scientometrics* is a hybrid OA journal, but there are other factors holding back the authors to adopt OA mechanism obviously. Nine ones in the top 10 journals are not so surprisingly belonging to the library and information science field, but there is a mega journal, *PLOS One*, did play an important role in introducing OA topic, it is the only one multidisciplinary journal.

Table 1. Top 10 journal of article with OA topic

rank	Journal Title	# of OA article published in OA mechanism (%)	# of OA article
1	*Learned Publishing*	23 (63.89)	36
2	*Scientometrics*	4 (14.81)	27
3	*Journal of the Association for Information Science & Technology*	8 (32.00)	25
4	*Journal of Academic Leadership*	0 (0.00)	20
5	*Online Information Review*	1 (6.25)	16
6	*Journal of Scholarly Publishing*	0 (0.00)	15
7	*Serials Review*	3 (21.43)	14
8	*College & Research Libraries*	13 (100.00)	13
9	*Information Research: An International Electronic Journal*	12 (100.00)	12
10	*Interlending & Document Supply*	2 (18.18)	11
10	*PLOS One*	11 (100.00)	11

Note: The journal title in grey column means they are full OA journal.

Who produced them and chose OA publishing?

To answer who involved in OA topic studies and published the articles in OA mechanism, this study examines the authors' country distribution from 163 articles. Considering the corresponding authors are the main one who is responsible for article submission and might also be the one who pays the article process charges (APC) from the financial sources, here we list the top 10 countries of corresponding authors based on the number of OA topic article in OA mechanism. The result shown in Table 2 and it indicates that the authors from USA, UK and Finland published the most articles with OA topic in OA mechanism. Although authors from USA produced the most articles in OA topic, the share is lower than UK and Finland. The most "open" authors might come from UK and Croatia, over 60% of articles which discussed OA topic and adopted OA mechanism. It demonstrates how they truly practiced open access sprit in talking about open access.

Who sponsored them?

To examine the funding source of these journal articles with OA topic and published in OA mechanism, this study analyses the funding acknowledgement contents and notices these 163 articles were funded by various type of funding organization. The top 3 funding organizations are Andrew W. Mellon Foundation, Deutsche Forschungsgemeinschaft (DFG) and UK Arts and Humanities Research Council, they sponsored 4, 3, 3 articles respectively. Compared to the result from Table 2, this result comes surprisingly. The top 1 funding organization is the USA private foundation instead of official funding bodies. DFG is another main funding organization in these 163 articles, and it is the main funding organization for German science system, but corresponding authors coming from Germany did not the main contributors in OA topics. Another surprising part is famous open access advocator and pioneer, Wellcome Trust, did not show up at the top list of funding source. However, it might be that Wellcome Trust published reports or policy documents more often than journal articles.

Table 2. Top 10 corresponding country

rank	Corresponding author country	# of OA article published in OA mechanism (%)	# of OA article
1	USA	46 (38.02)	121
2	UK	31 (60.78)	51
3	Finland	15 (57.69)	26
4	Canada	9 (37.50)	24
5	India	7 (41.18)	17
6	Italy	5 (45.45)	11
7	Germany	5 (33.33)	15
8	Netherlands	4 (44.44)	9
9	Croatia	4 (80.00)	5
10	Spain	3 (21.43)	14
10	Australia	3 (25.00)	12
10	China	3 (23.08)	13

Discussion & outlook

This preliminary study aims to examine how open of the journal articles with OA topic and the results show it has a lot of room to improve. However, this result might tell us another fact that even the journal articles which discussed about OA, they have other reasons to choose not publish in OA mechanism to let the concepts be accessible more easily and quickly. The reasons under the paywall might not just be relevant with unaffordable expense of APC or career promotion (Solomon & Björk, 2012), and these reasons are needed to be explored in the future. Otherwise, how should the concept of OA be introduced to those who are not in this field more easily?

References

Solomon, D., & Björk, B.-C. (2012). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the American Society for Information Science and Technology*, 63(1), 98-107. doi: 10.1002/asi.21660

Shepard’s Citations Revisited – Citation Metrics for Dutch Legal Information Retrieval

Gineke Wiggers[1] and Wout Lamers[2]

[1] *g.wiggers@law.leidenuniv.nl*
Leiden University, eLaw - Center for Law and Digital Technologies, Leiden (The Netherlands)

[2] *w.s.lamers@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS),
Leiden (The Netherlands)

Introduction

Legal information retrieval (IR) systems for smaller jurisdictions often do not have the features users have come to know from web search, such as PageRank-like impact measurements. To ensure good access to legal information, such a relevance factor should be added to ranking algorithms in legal IR. This paper presents the first complete citation index for Dutch legal publications.

Background

Frank Shepard is credited with creating one of the first citation networks, and the best-known legal one; the Legal Case Citator or Shepard’s Citations. It is immortalized in LexisNexis’ Shepardize function (which shows whether a case has been overturned in appeal). According to Garfield (1979), the Legal Case Citator influenced him in the creation of the Science Citation Index. The Science Citation Index introduced the notion of citation indexing in the context of journal articles, and has been extensively used for impact measurement of journal articles and research evaluation.

This paper combines the work of Shepard and Garfield to create a citation index for all types of legal publications, to be used in information retrieval.

Citations in Dutch legal publications

Because the (Dutch) legal domain has a different publication culture than other research domains, one of the questions that arises is what citations in legal publications represent. Research by Wiggers and Verberne (2019) shows that legal scholarly and professional publications cite each other frequently. This suggests that a citation in a legal publication does not measure pure scientific impact, but a more general form of impact on the entire legal field.

Characteristics of Dutch legal publications

Two characteristics on which the Dutch legal domain differs from other research domains are: (1) its strong national ties and (2) the often strong interconnection between research and practice.

Using citations in legal information retrieval systems has already been accomplished by systems like LexisNexis and WestLaw[1], but because these systems are created for a common law jurisdiction, which relies heavily on case law, it cannot simply be copied for civil law jurisdictions like the Netherlands, which rely mainly on legal codes with case law as an interpretative aid.

In the Dutch legal domain, the strong tie between research and practice is demonstrated by the fact that scholars and practitioners use the same legal information retrieval systems, and by the lack of distinction between legal scholarly and professional publications (Stolker, 2015). This lack of distinction could be one of the reasons why a complete citation-index (including journals) has not yet been established within the Dutch legal domain.[2]

Building the citation index

Data gathering

We built the citation index using data from Legal Intelligence, one of two large Dutch legal content integrators. For all publications in the system references to other publications are marked in the text by the Legal Intelligence system with an identifier, so that all possible notations of references are harmonized. This includes names of cases, based on the thesaurus of the system.

From this, we have created a dataset with all documents that cover the Dutch jurisdiction. Of these documents, we have gathered relevant metadata, such as title, publication date, law area (as specified by publisher) and identifier.

Based on the identifier, we searched the system to discover the number of other publications that reference this publication.

[1] http://lscontent.westlaw.com/images/content/L-355700_West-Search-brochure.pdf

[2] A proof of concept was presented by Wirt Soetenhorst in 2017, but we did not find any information that this citation index has been completed. Other citation indexes focus only on laws/case law.

Data cleaning

Based on preliminary results, we discovered that most documents contain a reference in the text to where the document itself can be found (e.g. journal and article number). We adjusted the citation index to remove these self-references.

Another problem we encountered are document references that are mapped in the thesaurus to a legal concept or term. If these documents (e.g. a legal case) are mapped as a synonym to a legal term, references to the term are not distinguished from references to the document. In some cases this behaviour is desired, such as when a case is referred to by party names. In other cases, where the case is mapped to a generic legal term or abbreviation, this causes incorrect citation counts. This occurred in a handful of instances and is solved in the thesaurus.

Normalization

The normalization of the raw citation count of the publications is based on the NCS score of the CWTS (e.g. Waltman et al. 2011) and the work of Rehn et al. (2014). It calculates a citation score normalized for time (based on year of publication), law area (as reported by publisher of the document, including government documents) and document type.

First results

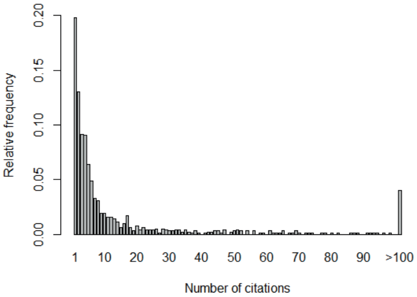


Figure 1. Distribution of documents on number of received citations

This first results show a highly skewed distribution of citations with a long-tail, which is in alignment with citations of journal articles in other research fields. The mean number of citations for documents that have received at least 1 citation is 46. The median is 4.

90% of the documents have a normalized citation score of less than 1, meaning they have less citations than the average for that year, law area and document type. 95% have a normalized citation score of less than 1.72, showing that a small number of publications receive a large number of citations.

Conclusions

We have created a legal citation index consisting of all Dutch legal publications. The normalized citation scores derived from this index can be used to improve ranking in legal information retrieval systems.

Because of the strong interconnection between research and practice, the impact on the legal field as a whole cannot be determined by citations only, as that would only show the impact on actors in the legal field that publish themselves. For that reason, to include a factor of impact in the ranking algorithm, additional information will be required, such as usage information. A factor of usage, or of recency, will also aid in dealing with recently added publications that have not had time to gather citations.

Acknowledgments

The authors would like to thank Legal Intelligence for providing the dataset and Brent Verpaalen for programming the citation index.

References

Bornmann, L., Bowman, B.F., Bauer, J., Marx, W., Schier, H., & Palzenberg, M. (2014). Bibliometric Standards for Evaluating Research Institutes in the Natural Sciences. In Cronin, B. & Sugimoto, C.R. (eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (p. 201-224). Cambridge, Massachusetts: MIT Press.

Garfield, G. (1979). *Citation Indexing: its theory and application in science, technology, and humanities.* New York: John Wiley & Sons, Inc.

Rehn, C., Kronman, U. & Wadskog, D. (2014). *Bibliometric Indicators – Definitions and Usage at Karolinska Institutet*. Stockholm, Sweden: Karolinska Institutet University Library

Soetenhorst, W.J. (2017). Een juridische citatie-index: het proof of concept is voorhanden. Nederlands Juristenblad 17(915), 1184-1186.

Stolker, C. (2015). *Rethinking the Law School: Education, research, outreach and governance.* Cambridge: Cambridge University Press.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. Journal of informetrics, 5(1), 37-47.

Wiggers, G., Verberne, S. (2019). Citation Metrics for Legal Information Retrieval Systems. *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval* (BIR 2019). Cologne, Germany.

Consistency Comparison of Four Typical Data Set Construction Methods for Domain Analysis in Bibliometrics

Yu Shao[1] and Guo Chen[2]

[1]*shaoyu@njust.edu.cn*
School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094(China)
[2] *delphi1987@qq.com*
School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094(China)

Introduction

Constructing domain datasets plays an important role in intelligence work. The reliability of data sets will directly affect the accuracy of analytical results (Feng & Leng, 2006). However, in practical, an ideal bibliometric dataset of one domain is usually unattainable according to Bradford's law.

Currently, one major task in bibliometric dataset construction is to create appropriate search queries for literature databases (Liu, 2019). Subject classifications, journals & conferences, keywords, and extended keywords are frequently used in search queries to obtain datasets. It is necessary to explore the consistency of these four types of datasets in domain bibliometric analyses, so that better understanding of the analytic results can be achieved, such as various kinds of ranking tasks, important citation identification, and hot spot detection.

In this paper, we choose "artificial intelligence" as our target domain to collect data. Six common indicators (subject classification, country, institution, author, keyword, and reference) are compared using overlapping ratio and Spearman correlation coefficient. Our result explicates to what extent these four datasets are consistent.

Data and methods

Data

The data sets used in our experiment are all collected from "Web of Science Core Collection" in Web of Science (WOS) database. The analysis period of this study is from 2007 to 2016. Descriptions of the four data sets are as follows:

- *WC dataset*: retrieved with "WC = Artificial Intelligence", containing 254,672 records.
- *Keyword dataset*: retrieved with "TS=artificial intelligence", containing 12,893 records.
- *Keyword list dataset*: retrieved with "TS=artificial intelligence" and 35 keywords most relevant to "artificial intelligence" given by experts, containing 203,201 records.
- *C-J dataset*: retrieved with 54 journals and 34 conferences on artificial intelligence recommended by Chinese Computer Federation, containing 115,531 records.

The authors' names are disambiguated and their institution and country information extracted from the C1 field of the bibliographic data. Keywords stemmed to avoid interference of grammatical form.

Methods

In bibliometric analysis, scholars tend to focus on high-frequency items, such as high-frequency keywords. Therefore, we set a number of different top values for the six elements (for example, for *keyword*, we compared the Top 50, 100, 200, 500 and 1000 keywords identified from our four datasets). Overlapping ratio and Spearman correlation coefficient are used to signify differences. The overlapping ratio indicates the consistency of identifying high-frequency items in a given domain, while the Spearman correlation indicates the consistency of rankings between them. The experiment is designed as shown in Figure 1.

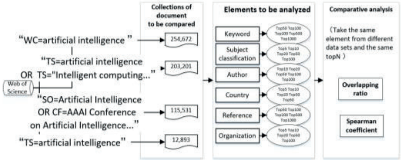

Collections of document to be compared
Elements to be analyzed
Comparative analysis
"WC=artificial intelligence "
"TS=artificial intelligence
OR TS="Intelligent computing..."
Web of Science
"SO=Artificial Intelligence
OR CF=AAAI Conference
on Artificial Intelligence..."
"TS=artificial intelligence"
254,672
203,201
115,531
12,893
Keyword
Top50 Top100 Top200 Top500 Top1000
Subject classification
Top5 Top10 Top20 Top50 Top100
Author
Top10 Top20 Top50 Top100 Top200
Country
Top5 Top10 Top20 Top30 Top50
Reference
Top50 Top100 Top200 Top500 Top1000
Organization
Top5 Top10 Top20 Top50 Top100
(Take the same element from different data sets and the same topN)
Overlapping ratio
Spearman coefficient

Figure 1. Flow chart of experiment

Results and discussion

General rules

Fig.2 shows that the overlapping ratios and ranking correlations of high-frequency items of six elements are relatively low between four datasets, which might suggest great disparities in results between studies with different dataset construction strategies adopted. Therefore, we should be cautious to employ dataset construction methods when carrying out bibliometric studies.

Moreover, the Spearman coefficient is lower than the overlapping ratio in all six elements. The Spearman coefficient can be even negative in some cases, which indicates that the rankings of high-frequency items, especially for authors and references, are sensitive to dataset construction methods.

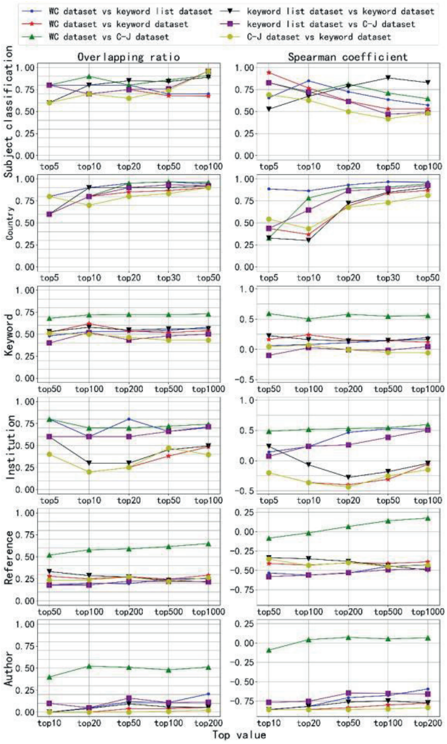


Figure 2. Comparing results between the four datasets in six elements using overlapping ratio and Spearman correlation coefficient

Different elements

According to Figure 2, the results are summarized as in Table 1. It shows that the six common analytical elements can be classified into two categories according to their content.

Table 1. Summary of results

Category	Field	Granularity	Overlapping ratio	Spearman coefficient	Differences between datasets
Producer-related	Author	Fine	Low	Low	Big
	Institution	Medium	Medium	Medium	Medium
	Country	Coarse	High	High	Small
Content-related	Reference	Fine	Low	Low	Big
	Keyword	Medium	Medium	Medium	Medium
	Subject classification	Coarse	High	High	Small

As shown in Table 1, with the increase of granularity of elements, the differences between the datasets gradually grow narrow. As coarse-grained elements, subject classification and country information are relatively stable in the four datasets. It shows that their requirements for the datasets are not very strict. However, WOS classifies all documents of the same journal or conference into one discipline. Shu (2019) found that 46.0% of journal articles, on average, come from other disciplines than that of their journals. Therefore, it is crucial to use an appropriate way to construct datasets when utilizing discipline information to perform our study. The datasets obtained from *author* and *reference* are quite different. It might suggest that if the *author* or *reference* element is needed, domain analytical results can differ greatly between datasets. The two comparison indicators suggest that the data set obtained from *keyword* element maintain medium level of differences from other datasets. In Figure 2, the six lines of *institution* are obviously divided into two parts, and the three lines at bottom imply that *keyword dataset* (about 13,000) differ greatly from other datasets. This means that if you use a small dataset to study institution information in domain bibliometric, you may get unstable results.

Different constructing methods

In Figure 2, lines marked with dots are always at the bottom of plots, suggesting that *C-J dataset* and *keyword dataset* are quite different. If the dataset constructed by both methods is used in domain analyses, the results may be relatively reliable. *WC dataset* is the most similar to *C-J dataset*, especially in terms of *keyword, author, reference and institution*. For instance, lines marked by triangles are always at the upper side of the plots of these six elements because WOS classifies all documents from the same journal/conference into one discipline, so *C-J dataset* is equivalent to a subset of *WC dataset*.

Conclusion

Through the quantitative comparison of four datasets constructed with different methods, we exhibit their inconsistency in most cases in bibliometric analyses. Therefore, we suggest bibliometric researchers to pay more attention to dataset construction to attain analytical results more reliable.

Ackonwledgements

This study is supported by the project of the National Social Science Fund for Youth of China (NO.16CTQ024).

References

Feng, L., & Leng, F. (2006). The survey on the data selecting method of information research. *Library intelligence*, 50(09), 94-96. (in China).

Liu, M. (2019). *Research on Data Set Construction Method for Discipline Domain Analysis.* Retrieved March 28th, 2019 from: https://mp.weixin.qq.com/s/ejBWwzj6LtjqeXQQ1HJcSA.(in China).

Shu, F., Julien, C. A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics,* 13(1), 202-225.

Exploring the teaching activities of the Italian universities through conditional efficiency analysis

Camilla Mastromarco[1], Pierluigi Toma[2] and Cinzia Daraio[3]

[1] *camilla.mastromarco@unisalento.it,* [2] *pierluigi.toma@unisalento.it*
University of Salento, Department of Economics, Management and Quantitative Methods - Ecotekne, via per Monteroni, 73100 Lecce (Italy)

[3] *cinzia.daraio@uniroma1.it*
Sapienza University of Rome, Department of Computer, Control and Management Engineering "A.Ruberti"- via Ariosto 25, 00185 Rome (Italy)

Introduction

Public entities behavior is generally modeled founding on methodologies typical of the economic sciences, however, empirical econometric or operational research are needed to obtain comparative evaluations and effective policy implications. In the case of universities, the quantity, quality and mix of services produced are largely determinable by autonomous decisions, influenced in particular by the preferences of the different categories of stakeholders. Italy, historically characterized by colleges and universities, has to invest in science and education in order to fully implement revitalizing strategy in terms of innovation and growth. Universities in Italy, as well as throughout the world, carry out different institutional activities, i.e. teaching, research and third mission. The aim of the research is to investigate the performance of university teaching by evaluating the efficiency of different faculty courses with an example carried out on the University of Salento that we plan to extend at the national level. For this purpose we use advanced and robust nonparametric tools recently developed in nonparametric efficiency frontier literature. The frontier in the performance assessment of university teaching activities consists of an extension in the use of traditional and limited series of indicators to other classes of indicators mineable from the information available through SUA-CdS data sheet and *ad hoc* surveys on graduates conducted by ANVUR (National Agency for Evaluation of the University and Research System) and/or MIUR (Ministry of Education, University and Research), in order to improve the quality of teaching monitoring and promote its dissemination through the different universities.

Brief literature review

Although the role of universities in the knowledge society is increasingly relevant, there are lack of systematic quantitative evidence at the micro-level. Bonaccorsi and Daraio (2007) examined original data from universities in six European countries, applying for the first time new generations of nonparametric efficiency measures on a large scale and providing micro-based evidence on the evolution of the strategic profile of universities in terms of scientific research, contract research, education and the third mission. De Witte and Lopez-Torres (2017) provides an extensive overview of the literature on efficiency in education.

The main critical points rising from the scientific literature attaining the Universities performance evaluation could therefore be summarized as follows (Daraio et al., 2015): mono-dimensionality; lack of statistical robustness; dependence on the university size and on the mix of subjects; lack of the input-output structure.

Data and estimation strategy

Data collection

The dataset used in this study is based on the official statistics produced by ANVUR and MIUR, including data on some indicators analyzed in Bonaccorsi and Daraio (2007) and Daraio et al. (2011) for all higher education institutions operating in the first and second level of university teaching. For the aim of this research, University of Salento data, referring to the period 2013/2016, were used.

iC27 is the Overall ratio between teachers/enrolled students used as input. The outputs are: iC16bis i.e. the Percentage of Students who continue at the 2nd year in the same course with at least 2/3 of the CFU for the first year; iC15bis i.e. the number of students who continue at the 2nd year in the same course with at least 1/3 of the ECTS for the first year; iC22, i.e. the percentage of enrolled students who graduate within the normal duration of the course, iC25 i.e. the percentage of final year students satisfied with the course study program; iC07bis is the percentage of graduates employed at three years from degree. The conditional factors are instead: iC10 i.e. the percentage of ECTS achieved abroad by regular students on the total amount of credits achieved by students within the normal duration of the course and iC09, i.e. the value of the Teacher Research Quality indicators for Master's Degrees (QRDLM).

Methodology

Let the production process of teaching activities in the Universities, with conditional variables, be characterized as:

$$\Psi^z = \{(x, y) | Z = z, x\ can\ produce\ y\}$$

As Daraio and Simar (2005, 2006, 2007a, 2007b) suggest, for each period *t*, the attainable set $\Psi_t^z \subset R_+^{p+q}$ can be characterized from the support of the following conditional probability:

$$H_{X,Y|Z}^t(x, y|z) = Prob(X \leq x, Y \geq y | Z = z, T = t)$$

Then, the time-dependent conditional output oriented technical efficiency measure of a University operating at $(x, y) \in \Psi_t^Z$ can be defined as:

$$\lambda_t(x, y|z) = \sup\{\lambda | (x, \lambda y) \in \Psi_t^Z\} \\ = \sup\{\lambda | S_{Y|X,Z}^t\ (\lambda y | x, z) > 0\}$$

where

$$S_{Y|X,Z}^t\ (y|x, z) = Prob(Y \geq y | X \leq x, Z = z, T = t)$$

For our analysis, we follow Cazals et al. (2002) to apply the order-*m* frontier, which is defined as the benchmark of the best practice among *m* universities drawn at random from a population of universities using fewer inputs than *X*. Then, by following Mastromarco and Simar's (2015) approach to environmental conditions and time, the time-dependent conditional output-oriented order-*m* efficiency scores can be computed as:

$$\lambda_{m,t}(x, y|z) = \int_0^\infty \left[1 - (1 - S_{Y|X,Z}^t(uy | X \leq x, Z = z))^m\right] du$$

We can disentangle the effects of "z" and "time" from the efficiency levels of universities by considering the ratios of conditional to unconditional order-*m* efficiency measures, which are the measures relative to the partial frontier of the conditional to the unconditional attainable sets:

$$R_m(x, y|z, t) = \frac{\lambda_{m,t}(x, y|z)}{\lambda_m(x, y)}$$

Moreover, we can investigate the *z* and *t* values associated with the universities by looking at the behavior of the above ratio as a function of *z* and *time*. For this purpose, we proceed by estimating the following nonparametric regression function:

$$R_{m_{i,t}} = g(Z_i, T_t) + u_{i,t}$$

where $u_{i,t}$ indicates the usual error term.

Preliminary results and conclusion

Results show that at the first academic level there are some courses that are efficient regardless of the output considered in the analysis and regardless of the efficiency assessment (unconditional, conditional, order-m, order-m conditional). For master's degrees instead there are more heterogeneous behaviors. When analyzing the percentage of enrolled students who graduate within the normal duration of the course and the number of students who continue at the 2nd year in the same course with at least 1/3 of the ECTS for the first year, course which result efficient are different from those that are efficient as regards the percentage of final year students satisfied with the course study program and the percentage of graduates employed in three years from degree.

These results, even if preliminary, represent a empirical contribution in the field of informetrics aimed at supporting policy making and strategic decisions in the Universities.

Further analysis will be carried out by increasing the number of analyzed universities, ensuring a stronger external validity of the research results.

Selected References

Bonaccorsi, A. & Daraio, C. (Eds.). (2007). *Universities and strategic knowledge creation: Specialization and performance in Europe*. Edward Elgar Publishing.

Daraio, C., Bonaccorsi, A., Geuna, A., Lepori, B., Bach, L., Bogetoft, P., ... & Fried, H. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy*, 40(1), 148-164.

Daraio, C., & Simar, L. (2005). Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of productivity analysis, 24(1), 93-121.*

Daraio, C., & Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications.* Springer Science & Business Media.

Mastromarco, C., & Simar, L. (2015). Effect of FDI and time on catching up: New insights from a conditional nonparametric frontier analysis. *Journal of Applied Econometrics*, 30(5), 826-847.

Witte, K. D., & López-Torres, L. (2017). Efficiency in education: a review of literature and a way forward. *Journal of the Operational Research Society*, 68(4), 339-363.

RISIS2: an innovative research infrastructure as a support for STI research community

Emanuela Reale* Grazia Battiato* Serena Fabrizio*

*emanuela.reale@ircres.cnr.it, grazia.battiato@ircres.cnr.it, serena.fabrizio@ircres.cnr.it
IRCRES - Research Institute on Sustainable Economic Growth of CNR, via dei Taurini 19, 00185 Roma, Italy

Background

In Europe, the 200 largest universities (out of more than 2.500) produce 80% of articles published by European universities. Similar studies dealing with nanotechnologies have shown that 80% of publications and an even greater percentage of patents were concentrated in just over 200 clusters worldwide. These asymmetric distributions call for a new type of dataset that keeps the identity and address actors, requiring thus to be built on open data. Following actors in different sources, in particular outputs and collaborations, geocoding addresses of organisations, and their authors and inventors have become critical resources for providing new evidence for policy-making.

RISIS2 objectives

The European Research Infrastructure for Science, Technology and Innovation Policy Studies (RISIS2) aims at building a data and services infrastructure supporting the development of a new and advanced generation of analyses and indicators through these achievements:

- The access to 13 science and innovation datasets.
- The production of 4 new datasets.
- The development of techniques to integrate different datasets.
- Hosting an integrated data platform for policy and research analysis.
- Providing training in using available data on research and innovation.

Vision for the STI field

The main pillars of RISIS are oriented to be identified as core support for the quantitative studies of the STI research community, the one dedicated to inform and build evidence for the policies that deal with preparing for the future, and this involves scholars interested in higher education, research and innovation. Add a new level of freely accessible aggregated data on very relevant political issues, challenges and company missions. Improve the use of our data sets and services, through the possibility of virtual transnational access with the opening of Risis Core Facility (RCF).

Risis Core Facility (RCF)

Through RCF, RISIS endows a unique entry point online, with which RISIS user can access a monitored and secured workspace.

RCF objective: to provide a ground-breaking infrastructure for Science, Technology and Innovation (STI) studies.

A workspace designed to distribute services to user interested in jointly exploiting different RISIS datasets and various Linked Open Data resources.

Access includes both ‘physical’ transnational access, (on site) and ‘virtual’ transnational access.

Main goals:

- to explore, retrieve and visualise results of data analysis for research purposes.
- to avoid a community connected to more and more open data enabling far wider exploitation.

A new opening to the growing number of open access datasets thanks to the D4Science capabilities articulated with the RCF.

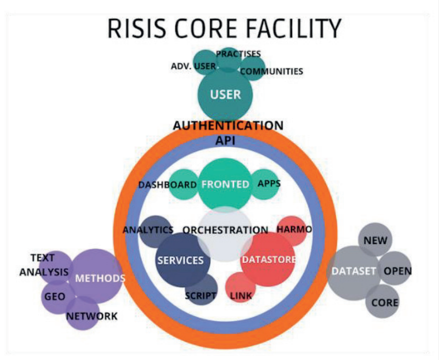


Figure 1. Architecture of the RISIS RCF

Datasets themes

RISIS datasets cover at least the EU and longer time periods (which is a critical dimension for indicator building), they represent core research, innovation and/or policy dimensions; and finally they have created a relevant interest from users in RISIS1 and beyond.

All datasets are organized around five main themes (familiae):

Firm Innovation dynamics
3 datasets concerning world largest firms (**CIB**), European venture capital backed firms (**VICO**) and fast growing mid-sized firms (**Cheetah**). They will be complemented by 2 datasets on social innovation and on trademarks.

European Integration
2 datasets dealing with European projects and trans-border programmes (**EUPRO** and **JOREP**) and European public research actors (with **ORGREG** serving a reference database incorporating the **ETER DB** on universities, PROs and university hospitals). Moreover, 3 new datasets are in development (**EFIL** on public funding instruments, **ESID** on social innovation, and **TM** Trademark on trademarks).

Knowledge dynamics
Output datasets enriched for analysis at actor and metropolitan area levels. Publications with the **WOS Leiden - ranking** and Patents and **PATSTAT - IFRIS** (get rid of accessible through CIB), with a demonstrator dataset for the study of emerging technologies (**Nano S&T DB**).

PhD and researcher careers and mobility
PROFILE and **MORE** will be complimented by a dataset on non-academic PhD careers.

Policy learning
A repository on science and innovation policy evaluations (**SIPER**) linked to the OECD-World Bank IPP platform. It will be complemented by a dataset on effective portfolios of public funding instruments on research and development.

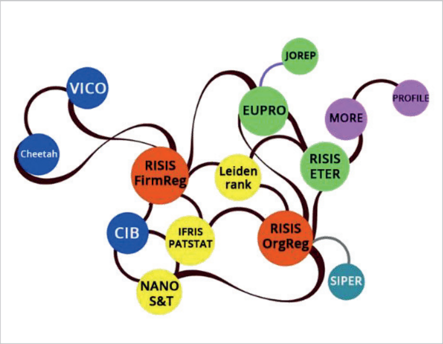


Figure 2. Datasets *familiae*

Training activities

Training provides basic and advanced knowledge on RISIS contents and infrastructure and how to use it, the methodologies to effort the datasets for research aimed to produce evidences relevant for policy making.

Training activities include the following:

Applied courses on datasets
On-line/on site about how to use the datasets for research aims.

Methodological courses
Dedicated to deepen specific methodologies and to enlarge the base of scholars using advanced quantitative methods.

Summer Schools
To train people using a mix of different datasets to address key problems/analyses that could be significant for policy purposes.

Who we are

RISIS2 Consortium is made of 18 members, 8 universities and 10 research institutes from 11 European countries (plus ISRAEL).
The Consortium is based on the developers and operators of both datasets and services.
www.risis2.eu

Acknowledgments

Funding from the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 824091 is acknowledged.

References

Gotsch, M., Hipp, C. (2012). Measurement of innovation activities in the knowledge-intensive services industry: a trademark approach, *Service Industries Journal*, 32(13), 2167-2184.

Joly, P.B., Gaunand, A., Colinet L., Laredo P., Lemarie, S., Matt, M. (2015). ASIRPA: A comprehensive theory-based approach to assessing the societal impacts of a research organization. *Research Evaluation*, 1-14.

Larédo, P. (2017). The next generation in STI studies, keynote speech for the 50th anniversary of science and innovation studies at the University of Manchester, Manchester, November 4, 2017, available on IFRIS website.

Lepori, B., Barré, R., Filliatreau, G. (2008). New perspectives and challenges for the design and production of S&T indicators. *Research Evaluation*, 17(1), 33–44.

Mazzucato, M. (2018). Mission-oriented research & innovation in the European Union, European Commission.

OECD (2015). Guidelines for collecting and reporting data on research and experimental development, Frascati Manual, Paris: OECD.

e-Lattes: A new framework in R language for analysis of the Lattes curriculum

Ricardo Barros Sampaio[1], Bruno Santos Ferreira[2], Antonio de Abreu Batista Junior[3] and Jesús P. Mena-Chalco[4]

[1]*rbsam@unb.br*
[2]*brunosfweb@gmail.com*
Colaboratório de Ciência Tecnologia Sociedade – Fundação Oswaldo Cruz (Brazil)
[3]*antonio.batista@ufma.br*
[4]*jesus.mena@ufabc.edu.br*
Universidade Federal do ABC – Centro de Matemática, Computação e Cognição (Brazil)

Introduction

Scientific publication and citation databases are spread amongst a great deal of research areas besides been a rich and continue source for scientific analysis. However, if one chooses to employ a research on the researchers themselves, and not only on their publication results, there a few data base examples that support such an endeavour. The Brazilian government have made it possible for its researchers to have such a database and for that all the analysis that might comes as a result of its examination.

For years the researchers curriculum database, called Lattes Platform, had its data not fully available for analysis, but that has changed in the past few years.

The Lattes Platform is a curricular information system, made available by CNPq (Brazilian National Council of Research), which allows curricular registration of researchers. Due to its extensive use by the Brazilian funding agencies, any researcher, Brazilian or not, aiming for funds from those agencies are required to have its own curriculum registered on the platform. Other agencies such as federal research institutions or higher education institutions have the updated curriculum as a requirement for career progression. Currently, the platform has more than five million enrolled resumes.

Due to its wealth of information and its increasing reliability and comprehensiveness, the Lattes curriculum (Lattes CV) has become an indispensable and compulsory element in the analysis of the merits and competence of funding projects in the area of science and technology in Brazil. The increasing availability of digital data on scholarly inputs and outputs in the Platform, from research funding, productivity, and collaboration, offers unprecedented opportunities to explore the structure and evolution of Brazilian science.

Often, compiling or summarizing bibliographic productions for a group of users of medium or large sized (eg, faculty group, postgraduate department), requires a great mechanical effort that is often susceptible to failure (Mena-Chalco & Cesar Junior, 2009).

R language Solution

E-Lattes consists of five modules. Figure 1 illustrates the relationships between them. Each module is responsible for a specific functionality in the R package and produces intermediate results that are inputs to other modules.

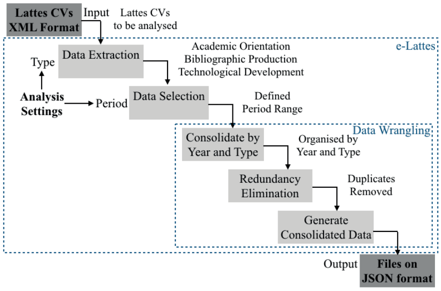


Figure 1. Implemented Modules on e-Lattes.

The Data Extraction module has a basic functionality of extract data from the CVs of the researchers, which comes in a XML format. The module extracts the general data (e.g., name, business address, abstract), in addition to all the bibliographic and academic production of the CVs. The data structure containing this information is used in the Data Selection module, but can be used by other modules (e.g., Data Wrangling). The specific data to be extracted is defined by the analysis settings, so depending on the analysis the result might contain only publications, academic orientation, technological developments (patents) or any combination of those.

The Data Selection module filters, by period, scientific publications and academic guidelines. The separation by year and type (e.g., periodicals, events, completed guidelines) is done by the Consolidated by Year and Type module. However, this module does not address any redundancy. This treatment is only possible by the Redundancy Elimination module. This module uses the string matching algorithm (Levenshtein) for deleting duplicates. For example, scientific articles with approximate titles that are found in the analysed CVs are considered only different instances of an article and so are counted only once. Because the

Lattes platform is a researchers data base, a great deal of duplication is encountered since all the collaboration papers are entered as many times are there are authors on the article.

Lastly, publications and unique guidelines, separated by year and type, serve as inputs to the Consolidated Data Generation module. Currently, with this data, it is possible to generate several bibliometric indicators, among them the academic profile of the researchers, the intersection of areas, academic seniority, bibliographic production and academic orientations of the group, for example.

E-Lattes is designed to handle JSON-formatted files for both execution parameters input and data output. This format is a lightweight, text-based and language-independent data exchange structure that allows extensive use of the data across different programming platforms nowadays (Crockford, 2017). This allows e-Lattes to be integrated with any interface or solution compatible with that format.

Web Solution

The e-Lattes Web project is a Web application that integrates with the R solution and makes its functionality available openly and easily. Through user-friendly interfaces, the application allows users to generate custom analyses, interact with data in real time, and generate maps and scientific indicators that can be used by students, researchers and managers in Science and Technology.

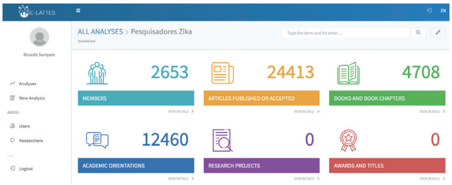


Figure 2. e-Lattes Web Application Interface

The application also provides a control panel in an easy-to-manipulate environment with didactic information on scientific, technological and academic production in an aggregate way for specific sets of professionals focusing on the institution's science and technology managers.

The e-Lattes Web design follows the architecture of a three-tier web application (Sanderson, 2008). This architecture model is known as Model View Controller (MVC) and is recommended as one of the best infrastructure models suggested for web-based applications (Leff & Rayfield, 2001). According to the MVC architecture: the Model layer must focus on the access and treatment of data, the View layer should focus on the treatment of the interaction with the users and finally, the Controller layer should intermediate the first two through services. In this perspective, e-Lattes Web is organized as follows: The Model layer contains all the access logic to the objects generated by the processing of the R solution; the view layer contains the user-visible interfaces; and the control layer, in addition to managing the integration between the View layer and the objects, handled by the Model layer.

The user-friendly web interface is responsible for interactions with users and was developed using HTML (in version 5) and JavaScript (in version 1.8.5). On the server side, its structure was developed using the PHP language (in version 7) under the Zend 2 framework standards.

All features in the R solution were modeled and made available as a web served in order to integrate the application with other external systems. Concerning to web services, one of the key features of a RESTful Web service is the explicit use of HTTP protocol methods as specified by RFC 2616. Thus, through the HTTP GET method, for example, an external application can retrieve data from an analysis without having access to the graphical interface. Similarly, the POST, PUT, and DELETE methods can be used to insert, update, or remove data from an analysis. Therefore, with the use of this protocol and the publication of Web Services, any application using the same standards may have access to the functionality of the R solution.

Conclusions

We have presented in this short paper a novel methodology to analyse scientific researchers registered at Lattes Platform. The proposed solution has been developed as a software package for R and is freely available for download through CRAN (The Comprehensive R Archive Network). Besides the R package we have also developed an open-access and easy-to-use WEB interface to help managers and users not familiar with R language to navigate and develop their own analysis.

References

Crockford, D. (2006). The application/json media type for javascript object notation (json) (No. RFC 4627).

Leff, A., & Rayfield, J. T. (2001). Web-application development using the model/view/controller design pattern. In Enterprise Distributed Object Computing Conference, 2001. EDOC'01. Proceedings. Fifth IEEE International (pp. 118-127). IEEE.

Mena-Chalco, J. P. ; Cesar Junior, R. M. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society, Porto Alegre, v. 15, n. 4, p. 31-39, 2009.

Sanderson, S. (2008). Architecture (pp. 51-99). Apress.

Discipline Impact Factor: its History and the Continuing Reasons for its Use

Vladimir S. Lazarev

vslazarev@bntu.by
Belarusian National Technical University, 16 Yakub Kolas St, Minsk, 220013 (Belarus)

Impact factor

In the article, which first introduced the word combination “impact factor” (Garfield, 1955), “impact factor” was still a full synonym of the word “impact”, and it did *not* relate to journal evaluation, as it happened much later (Bensman, 2007, p. 111): “Garfield <…> was to change this meaning when he created a measure he called the “impact factor” to determine which journals should be covered by the *SCI*. This term came to be defined as *the average number of citations to the papers of a given journal*”. According to E. Garfield’s definition, “this impact factor is the mean number of citations to a journal’s articles by papers subsequently published. It is determined by dividing the number of times a journal is cited (R) by the number of source articles (S) it has published” (Garfield, 1970, p. 5).

Discipline impact factor

The classical Garfield impact factor reflects the level of the use of an average paper of a certain journal *by all the journals representing technical and natural sciences* (being indexed by the Science Citation Index; later–by Web of Science). However, I believe that in order to organize a sufficient information service, it is much more important to know the level of use of an average paper of a certain journal (or of other serial) *not by all the journals* representing technical and natural sciences *in toto* (as reflected in the classical impact factor), but by the ones *specialized in that concrete discipline or a field of research* which is going to receive information services. After all, the provision of information services to specialists in a particular field of research is the task of a larger number of libraries than the information support of all natural and technical sciences “in general”.

In this regard, it is appropriate to recall that in the 70-ies years of the 20th Century Graeme Hirst introduced the so-called “discipline impact factor” (Hirst and Talent, 1977; Hirst, 1978). As Hirst (1978, p. 171) stated, “the discipline impact factor (DIF) is similar to the impact factor <…>, which measures the average number of times a paper in a given journal is cited, except that the DIF measures the number of times a paper in a journal is cited in the core literature of the given discipline”. I think this formulation is not a brilliant one because, in fact, the *specialized* journals (not the “core” journals) are implied to be the sources of citations in the study by Hirst and Talent (1977), and the term “*discipline* impact factor” itself implies the count of citations in *specialized* journals. For some disciplines among core journals there are journals of much wider specialization that the discipline itself. It is obvious that using them as a source of citations would result in involvement of citations that are *not* related to the discipline in question.

Since DIF was aimed at solving practical problems, relevant at that time for each research, university and college library, it should have been expected that it would become very much popular, would be used very frequently. However, it never happened. In fact, it was used surprisingly seldom. Apparently, this was due to the fact that the calculating of DIF required quite time-consuming computations, while the "classic" impact factor was presented in Science Citation Index in a ready form. Several papers, however, might be mentioned as specimens of the discipline impact factor use for determining appropriate lists of periodicals (e.g. Lazarev and Nikolaichik, 1979; Gould, 1981; Black, 1983; Lazarev, 1983; Kushkowski et al. 1998; Lazarev et al., 2017 etc.). There are also some papers in which just some minor elements of the Hirst’s methodology were used relating to the restricted number of “core journals” selection, but not to the application of DIF itself for determining extensive lists of necessary periodicals. The example is the paper by Jan and Zhu (2015).

Our experience of DIF application for serials evaluation of selections (Lazarev, 1983; Lazarev and Skalaban 2016; Lazarev et al., 2017; Lazarev et al., 2019 etc.) demonstrated that quite a substantial portion of journals that are being included in the list of serials to be determined in order to organize or amend information services of the specialists in a certain discipline or research field is being selected *exclusively* by means of DIF computation.

Do we still need it?

One of my papers was rejected by a reputed journal, whose editor wrote me that nowadays libraries buy access to huge databases (packages) and do not bother to determine the “best” journals, while it is much cheaper to buy the whole package than to buy separate journals.

Nowadays libraries really *mostly* buy access to huge databases (packages) and *do not bother* to determine the concrete necessary journals and other serials. And as bibliometric evaluation and selection of non-profile serials to be used by researchers *in a specific discipline* were usually performed exactly in order to select serials for the specialized library stock, there *seeme*d to be no more need in bibliometric evaluation of the non-profile serials value for researchers in a specific discipline (Lazarev 1998).

However, the following question still arises: "*Which* databases (packages) ought to be purchased? The answer might seem easy to a librarian who lives in a country where a regular *sufficient* financial support of university and research libraries is practiced. But in case of restricted, meager financing for database subscriptions, we are to spend our small money for sure. The point is we need to choose exactly the databases ("subscription packages") *with the best coverage of the relevant serials*, the databases (packages) that optimally meet *both* the requirements of containing more useful periodicals and of being cheapest to be purchased. As many as possible relevant periodicals ought to be accessed via these databases (packages) at the lowest financial cost. In order to arrange this, one is to check each "subscription package" for the presence of maximum number of necessary serials. In its turn, in order to fulfill the latter, one is to know concretely *which* periodicals are needed! And therefore, one is to start the procedure that is very much similar to the one that was practiced in the past for the selection periodicals immediately for acquisition to the library stock! (As for the Open Access journals, thought they *are* available, they ought to be identified as well!) So, we, librarians from the countries that cannot afford sufficient financial support of academic, university and research libraries, still *do* need in determining "best" journals and in good instruments for it. One of such efficient tools is the discipline impact factor.

References

Bensman, S. J. (2007). Garfield and the impact factor. *Annual Review of Information Science and Technology 41*(1), 93-155. https://doi.org/10.1002/aris.2007.1440410110

Black Jr., G. W. (1983). Core journal lists for behaviorally disordered children. *Behavioral & Social Sciences Librarian,* 3(1), 31–38. https://doi.org/10.1300/J103v03n01_04

Garfield, E. (1955). Citation indexes for science: A new dimension in Documentation through association of ideas. *Science*, 122, 108–111. https://doi.org/10.1126/science.122.3159.108

Garfield, E. (1970). What is a significant journal? *Current Contents,* (18), 5-6.

Gould, A. L. (1981). Verifying a Citation: Reference Use of OCLC and RLIN. *Reference Services Review,* 9(4), 51-60. https://doi.org/10.1108/eb048731

Hirst, G. & Talent N. (1977). Computer scienc journals–an iterated citation analysis. *IEEE Transactions on Professional Communication*, PC-20(4), 233-238.

Hirst, G. 1978. Discipline impact factor—a method for determining core journal list. *Journal of American Society for Information Science,* 29(4), 171–172.

Jan, E.J. & Y. Zhu. (2015). Identifying entities from scientific publications: A comparison of vocabulary- and model-based methods, *Journal of Informetrics*, 9(3), 455—465.

Kushkowski, J. D., Gerhard, K. H. & Dobson C. (1998). A method for building core journals lists in interdisciplinary subject areas. *Journal of Documentation*, 54(4), 477–488. https://doi.org/10.1108/eum0000000007179

Lazarev, V. S. & Nikolaichik V. V. (1979). Distribution of information on hematology in scientific journals. In *Sovremennye aspekty gematologii* [Modern aspects of hematology] (128-133). Minsk, Nauka i tekhnika Publ. (in Russian).

Lazarev, V. S. (1983). Comparison of the possibilities of various methods for selecting scientific journals that are most valuable for specialists (a brief review of the literature and the own data). *Nauchno-tekhnicheskaya informatsiya. Ser. 1* [Scientific and Technical Information Ser. 1], (6), 27–32 (in Russian).

Lazarev, V. S. (1998). On the role of bibliometrics in the knowledge society: bibliometric quicksand or biblometric challenge? *Newsletter to European Health Librarians*, (44), 17–18.

Lazarev, V. S. & Skalaban, A. V. (2016). The world major scientific periodicals to be used by researchers of renewable energy, local and secondary energy resources. *Energetika. Proceedings of CIS Higher Education Institutions and Power Engineering Associations*, 59(5), 488-502 (in Russian). https://doi.org/10.21122/1029-7448-2016-59-5-488-502.

Lazarev, V. S., Skalaban, A. V., Yurik, I. V., Lis, P. A. & Kachan, D. A. (2017). Selection of serial periodicals to support researchers (based on the example of scientific work on nuclear power). *Scientific and Technical Information Processing*, 44(3), 196–206. https://doi.org/10.3103/s0147688217030066.

Lazarev, V.S., Yurik, I.V., Lis, P.A., Kachan, D.A. & Dydik, N.S. (2019) Some methodological aspects of selection serials to be included in the information environment for researchers in a technical or natural science (by example of optoelectronics and optical systems), *Library Philosophy and Practice (e-journal)*, 2185. https://digitalcommons.unl.edu/libphilprac/2185

A Closer Look at Data Co-authorship:
Team Size Trends in 'Big Science'

Sarah Bratt[1], Jian Qin[1] and Jeff Hemsley[1]
[1]*sebratt@syr.edu*
Syracuse University School of Information Studies, 116 Hinds Hall, Syracuse, New York 13244 (United States)

Introduction

Scientific teams have grown over the past 30 years. The trend toward large teams solving complex problems has evidenced itself in hyper-authorship on papers, patents, and more recently, datasets (Cronin, 2001; Glänzel & Schubert, 2004). Research data have been called 'the backbone of scientific discovery,' the fuel of innovation, and the currency of science. However, little is known about dataset co-authorship trends. This study reports our findings of team size tends in data co-authorship in GenBank, an open research data repository. The analysis focuses on two primary dimensions of dataset co-authorship: (1) team size trends and mean co-authorship; and (2) proportion solo-authorship and co-authorship. Our results suggest a rapid shift from solo-authored to co-authored datasets and a steadily rising rate of mean and median team size over the course of our dataset, which fluctuates in response to 'outliers' – i.e. well-resourced, 'Big Science' teams (Borgman, 2015; Price, 1963). This study makes an important contribution to scientometrics by analyzing team size trends, providing a clearer understanding of an increasingly recognized form of scientific co-production: data co-authorship.

Methods

Data source

The study is part of a larger project on the impact of cyberinfrastructure on collaboration structures and dynamics (Qin, Hemsley, & Bratt, 2018). The data source used in this study is the metadata from *GenBank,* a large open research data repository that accepts only original contributions. GenBank was founded in 1982 and is maintained by the National Institute of Bioinformatics (NCBI) (Benson et al., 2013). The repository contains all publicly available nucleic acid sequences and protein translation *datasets* and their annotations, *publications* based on the datasets that are linked by the submitting author, and associated *patents*.

The GenBank metadata describes datasets, publications, and patents (1982-2018 at collection) was downloaded in flat files and parsed into a relational database. The author names were disambiguated using the Kaggle 2013 disambiguation solution and through triangulation of author names and other identifying information with the USPTO dataset, Semantic Scholar, and Web of Science (Chin et al., 2014).

Analysis

Co-authorship data for sequence data submissions from 1992 – 2018 was extracted from our database. The date range was selected to exclude earlier years, as they contained relatively low numbers during the initial period of GenBank's adoption. *Team size* is operationalized as the number of co-authors on a dataset. *Dataset* or *data submission* is defined as the submitted unit of data which is assigned a unique identification number. *Data co-authorship* is defined, intuitively, as the co-production of data represented by appearing as the name in the metadata of the data submission. Frequency counts were then computed for the number of datasets submissions per year and the authors per dataset – i.e., the "team size." Summary statistics of dataset submissions and team size over time (Table 1) were computed by generating the count of data submissions per year, the mean and median team size per year, and the ratio of solo-authorship to co-authorship on datasets submissions per year.

Table 1. Summary statistics for data submissions per year (non-cumulative) and mean team size in increments of 3 years in GenBank (1992 -2018). The sample mean ($\bar{x}$) represents 3-year increments.

Year range	# of data submissions	Mean team size	Datasets with > 1 author
1992-1994	21,067	$\bar{\mathbf{x}}$ = 1.12	$\bar{\mathbf{x}}$ = 4.2%
1995-1997	495,14	2.2	41.5%
1998-2000	83,867	4.3	66.7%
2001-2003	113,241	**7.9**	68.1%
2004-2006	129,587	**7.6**	71.5%
2007-2009	147,244	3.8	72.8%
2010-2012	155,186	3.9	78.2%
2013-2015	170,830	3.7	78.5%
2016-2018	139,893	3.7	76.8%

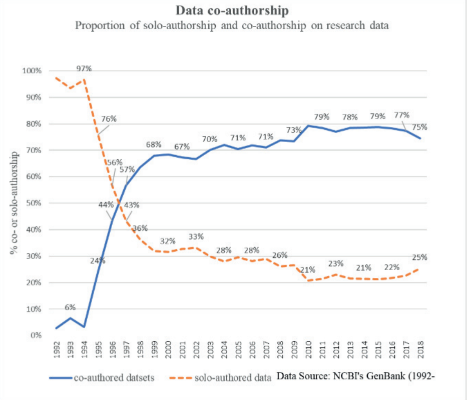
Results

Team size trends and mean co-authorship

Co-authorship on datasets, operationalized as *team size*, shows general upward trend especially from the initial years of the inception of GenBank. The average annual growth rate (AAGR) of mean team size was 7.5% (rounded) from 1992-2018. However, there were fluctuations in the mean team size, as shown in the highlighted portion of Table 1. From 2001-2003 and 2003-2006, the mean team size jumped from a relative stable ~4 members, to a mean size of nearly 8 members. Further investigation revealed that outliers in the maximum number of authors on a dataset spiked in these years, pulling the mean upward, with a maximum of 120 co-authors in the year 2000 and a max ranging from 202 – 214 in the year interval 2002-2005 (inclusive). The hyperauthorship on data submissions then subsides but does not dip below 103, compared to the early years (1992-1999) where max team size ranged from 11 to 88 authors.

Proportion co-authorship and solo-authorship

We found the proportion of co-authored datasets and solo-authored sharply "changed-over" in 1997 and stabilized in an inverse trend, with co-authored datasets comprising 75% of total submissions versus 25% of solo-authored datasets by 2018 (Figure 1).



Data Source: NCBI's GenBank (1992-

Figure 1. Proportion of data co-authorship vs. solo authorship on sequence submissions in GenBank (1992-2018).

Discussion

Cyberinfrastructure is a sign-post of data-driven science. Working with the data source of GenBank metadata reflects these trends. Our findings showed that the mean author trends rose, but with some fluctuations. We interpret this to reflect the sequencing of the Human Genome Project that was completed and published around 2002. The large genome projects are *development* projects that commonly require large teams, and attendant hyperauthorship on the intellectual products produced. We found that datasets, like publications, reflect hyperauthorship patterns, but that the patterns fluctuate according to the type of science conducted.

One potential limitation but interesting finding of the study is the decision to limit the period from 1992-2018. The removal of earlier years, while justified by the low submission numbers in the database's early years, indicates a remarkable aspect of cyberinfrastructure adoption. The sharp rise in submissions reflected in the early years of GenBank may reflect important trends in the adoption of cyberinfrastructure-enabled data repositories. Journals such as *Cell* started requiring manuscripts to include a data submission accession number at the time of submission to journals in late 1990's (Strasser, 2008), which also contributed to the upward trend. That is, the team size trends suggest that data co-authorship can be affected by project scale, publishing policies, and social practices, which is a research topic worth exploring in-depth.

Conclusion

Data co-authorship trends can have impacts on the design of information systems and science and innovation policy. In this study, we analyze 27 years of data-co-authorship trends using a novel data source: GenBank metadata. We provide a clearer understanding of a science community's history and trends of co-authorship and solo-authorship on research datasets, suggesting future directions for scientometric analysis of team size trends.

Acknowledgments

This project is supported by NSF award #1561348.

References

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic Acids Research, 41(Database issue), D36-42. https://doi.org/10.1093/nar/gks1195

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. MIT press.

Chin, W.-S. et. al (2014). Effective string processing and matching for author disambiguation. The Journal of Machine Learning Research, 15(1), 3037–3064.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? Journal of the American Society for Information Science and Technology, 52(7), 558–569.

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In Handbook of quantitative science and technology research (pp. 257–276). Springer.

Price, D. de S. (1963). Big science, little science. Columbia University, New York, 119–119.

Qin, J., Hemsley, J., & Bratt, S. (2018). Collaboration Capacity: Measuring the Impact of Cyberinfrastructure-Enabled Collaboration Networks. Science of Team Science (SCITS), 22, 636–655.

Strasser, B. J. (2008). GenBank–Natural History in the 21st Century? Science, 322(5901), 537–538.

A study of open access APC in Taiwan

Wen-Yau Cathy Lin

wylin@mail.tku.edu.tw
Department of Information and Library Science, Tamkang University, No.151, Yingzhuan Rd., Tamsui Dist.,
New Taipei City 25137 (Taiwan)

Introduction

There is no free lunch in academic publication and scholarly communication. Although open access (OA) movement advocates for free access of the results of academic research for readers, the burden of publication cost would have to be borne by the authors or subsidy of institutions, if the goals of maintaining the high quality of academic publication and free access to readers should both be achieved. Traditional academic journal publishers mainly adopt the toll access (TA) model where readers or their affiliations are charged for subscription fees, but now the publishers collect the article processing charge / article publication charge (APC) from authors or their affiliated institutions instead (Monson, Highby, & Rathe, 2014). Thus, a growing number of researches are discussing the topic of APC from the perspective of the country, because how many APCs each country pays to which publishers is the key decision making information for a country when formulating its OA mandate. (Borrego, 2016; Fukuzawa, 2017; Pavan & Barbosa, 2018; van Leeuwen, Tatum, & Wouters, 2018)

This study indicates that the ratio of papers published under the OA mechanism for the first time exceeded 15% during the past decade since 2009 in Taiwan. The journal articles indexed in Web of Science (SCI-Expanded & SSCI) which are with single or co-authors affiliated in Taiwan, are taken as the research objects. The study explores the ratio of OA paper to TA paper, and the number of OA papers that require APC charging. How many of these receive the research funds from Taiwan Ministry of Science and Technology (MOST)? And which journal publishes the highest number of papers, and what is the highest amount of APC? The results of the study can provide a reference for the government to formulate research subsidy policies and collection development policy of university and research libraries.

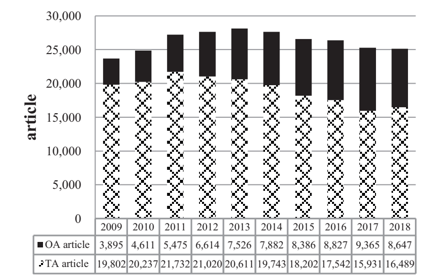
Methods

In this study, the SCIE and SSCI databases of Web of Science from 2009 to 2018 are used as data sources. Research dataset are collected during the season one of 2019. The research articles and review articles by authors with Taiwanese affiliation are identified. In addition, the articles which have authors affiliated in Taiwan as the main author (first and/or corresponding author) are further screened out. The funding acknowledgement contents of each article are also analysed. Since the APC frequently changes in various journals, and it is extremely difficult to figure out the exact price model of APC from years ago, therefore this study adopts the latest data collected in February 2019 when calculating APC.

Research Results

Share of OA and TA articles

Since 2009, the ratio of the number of journal articles involving authors affiliated in Taiwan which are published under the OA mechanism has increased from 16.44% to 37.02% in 2017, and slightly decreased to 34.4% in 2018. But whether this slight decrease is due to the incompletely updated data, or the true trend of decrease, still remains to be observed. More details are shown in Figure 1.



	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
■ OA article	3,895	4,611	5,475	6,614	7,526	7,882	8,386	8,827	9,365	8,647
▨ TA article	19,802	20,237	21,732	21,020	20,611	19,743	18,202	17,542	15,931	16,489

Figure 1. OT/TA articles by authors with Taiwanese affiliation between 2009 and 2018

Funding support of OA articles

Since the APC of the article is usually paid by the first author or corresponding author, the articles by the authors affiliated in Taiwan are further analysed. MOST is the most important official institution for academic research in Taiwan and is also the major research funding agency in the country level. As can be seen in Table 1, since 2013, more than 75% of OA articles have been sponsored with research funds, and the proportion of APC that is highly likely to be funded by MOST has even surpassed 50% since 2010.

Table 1. OA articles by main authors with Taiwanese affiliation and the funding support

Year	TW main authored article	With funding (%)	MOST funded (%)	NON-MOST funded (%)
2009	3,331	2,072 (62.20)	1,508 (45.27)	564 (16.93)
2010	3,869	2,641 (68.26)	1,961 (50.68)	680 (17.58)
2011	4,454	3,214 (72.16)	2,401 (53.91)	813 (18.25)
2012	5,413	4,012 (74.12)	3,084 (56.97)	928 (17.14)
2013	6,326	4,807 (75.99)	3,696 (58.43)	1,111 (17.56)
2014	6,542	5,038 (77.01)	3,837 (58.65)	1,201 (18.36)
2015	6,873	5,265 (76.60)	3,919 (57.02)	1,346 (19.58)
2016	7,259	5,612 (77.31)	4,115 (56.69)	1,497 (20.62)
2017	7,737	6,101 (78.85)	4,510 (58.29)	1,591 (20.56)
2018	7,187	5,696 (79.25)	4,197 (58.39)	1,499 (20.86)
total	58,991	44,458 (75.36)	33,228 (56.33)	11,230 (19.04)

Journals contain OA articles

Based on the OA articles involving with main authors affiliated in Taiwan between 2009 and 2018, Table 2 presents that *PLOS ONE* ranks first in terms of the total number of publications in ten years. But *Scientific Reports* published by Springer-Nature ranks first since 2016, and the number of articles published by *Medicine* and *Oncotarget* is also rapidly rising.

Table 2. Ranking of top 10 journals contain OA articles

Journal Title	‘09	‘10	‘11	‘12	‘13	‘14	‘15	‘16	‘17	‘18	total
PLOS ONE	19	8	1	1	1	1	1	2	2	2	1
Scientific Reports	--	--	477	151	31	11	3	1	1	1	2
Optics Express	1	1	2	2	3	3	6	6	14	21	3
Medicine	--	--	326	545	0	28	2	3	4	4	4
Journal of the Formosan Medical Association	2	3	3	7	7	6	7	6	7	7	5
Journal of the Chinese Medical Association	3	2	3	5	6	9	11	16	15	15	6
Taiwanese Journal of Obstetrics & Gynecology	5	5	6	4	5	7	10	9	8	16	7
International Journal of Molecular Sciences	46	34	18	11	8	10	9	5	5	3	8
Oncotarget	--	--	--	--	396	14	5	4	3	53	9
Sensors	8	5	8	8	10	12	13	12	10	9	10

Journals with highest amount of APC cost

Some OA journals are Diamond OA journals which do not levy APC, or only collect a relatively low amount of APC. Therefore, Table 3 determine the total amount of APC paid by main authors affiliated in Taiwan for specific journals between 2009 and 2018 by multiplying the number of articles by the APC. Only the top 5 journals are listed, and these journals are all Full OA journal, with a total amount of money reaching $17,923,364.

Table 3. Top 5 journals with the highest APC

Journal	APC ($)	Article #	Total amount ($)
PLOS ONE	1,595	4,546	7,250,870
Scientific Reports	1,790	2,172	3,887,880
Oncotarget	3,400	825	2,805,000
Medicine	1,950	1,266	2,468,700
International Journal of Molecular Sciences	1,803*	838	1,510,914

* CHF converted to US$

Discussion

According to the preliminary results of this study, the share of OA articles published by authors with Taiwanese affiliation has steadily increased and exceeded 30% in 2015, and the amount of APC paid in the past ten years for the top 5 journals is nearly $18 million. Since MOST of Taiwan is the most important official research funding agency, if the government is going to set out the national level of OA policy, the policy makers must face up to the serious and severe issue of APC. The university libraries have already required a high amount of journal subscription fees, and the research funding agency has also disbursed in the past and might require a higher amount of APC in the future.

Acknowledgments

This research is supported by the Ministry of Science and Technology of Taiwan (grant number: MOST 107-2410-H-032-068).

References

Monson, J., Highby, W., & Rathe, B. (2014). Library Involvement in Faculty Publication Funds. *College and Undergraduate Libraries*, 21(3-4), 308-329. doi: 10.1080/10691316.2014.933088

Borrego, Á. (2016). Measuring compliance with a Spanish Government open access mandate. Journal of the Association for Information Science and Technology, 67(4), 757-764. doi: 10.1002/asi.23422

Fukuzawa, N. (2017). Characteristics of papers published in journals: an analysis of open access journals, country of publication, and languages used. Scientometrics, 112(2), 1007-1023. doi:10.1007/s11192-017-2414-y

Pavan, C., & Barbosa, M. C. (2018). Article processing charge (APC) for publishing open access articles: the Brazilian scenario. Scientometrics, 117(2), 805-823. doi:10.1007/s11192-018-2896-2

van Leeuwen, T. N., Tatum, C., & Wouters, P. F. (2018). Exploring possibilities to use bibliometric data to monitor gold open access publishing at the national level. Journal of the Association for Information Science and Technology, 69(9), 1161-1173. doi:doi:10.1002/asi.24029

Readership of International Publications as Measured by Mendeley Altmetrics: A Comparison Between China and USA

Francis Houqiang Yu[1], Cathy Xueting Cao[2] and Biegzat Murat[3]

[1] *yuhouq@yeah.net* [2]*caoxueting0829@163.com* [3]*nulibiegzat@163.com*
School of Economics & Management, Nanjing University of Science & Technology, Nanjing (China)

Introduction

With the fast-changing landscape of online scholarly communication and quantitative scientific evaluation, altmetrics, more technically referred as social media metrics (Costas, 2018), are considered a promising toolkit for assessing the societal impact of research, as they offer novel ways to measure engagement with research output (Bornmann, 2014). Among them, Mendeley readership count is recognized as one of the most important data sources (Thelwall &Nevill, 2018). Different from downloads of PDF, the number of users that have saved a paper into their personal library is considered to be its Mendeley readership count (Mohammadi & Thelwall, 2014).

Compared with views data which reflect how many times the publication is visited or downloaded, Mendeley readership can provide more diverse types of context data, including reader's identity, discipline and country. Moreover, in addition to the immediacy advantage over citation indicators, it can reveal the usage by readers who read but never cite. Therefore, Mendeley readership data could be used to analyze the reading pattern in a more in-depth way and reflect broader impact.

This study makes use of Mendeley altmetrics to provide insights as regards how international publications of China and USA are used by various types of readers. As these two countries are strong scientific powers of which the evaluation attracts broad interest, Mendeley altmetrics will reveal the underlying usage pattern of their international publications, and provide a new perspective for understanding their impact.

Methods

Four disciplines were selected, two of them are from social science field, i.e. Information Science and Library Science (LIS) and Psychology (PSYC), the other two of them are from natural science field, i.e. Biochemistry & Molecular Biology (BIOC) and Mechanical Engineering (ENGI). For collection of bibliographic data, Web of Science database was used to retrieve all publications of China and USA in the year 2017. For collection of Mendeley altmetrics data, Webometric Analyst was used to extract Mendeley readership counts. Invalid records were removed. The cleaned dataset was used to do statistical analysis and comparative analysis mainly from four perspectives.

Result

Comparative analysis on the average number of Mendeley readership counts in four disciplines

As shown in Table 1, the proportion of publications with Mendeley readership of China is higher than that of USA. Among the four disciplines, LIS demonstrates the greatest difference between China and USA. In the other three disciplines, the number of Mendeley readers per publication of USA is higher than that of China, especially in BIOC.

Table 1. Comparison of Mendeley readership in four disciplines between China and USA.

	China			USA		
Discipline	N.P.R	%	M.	N.P.R	%	M.
LIS	534	87%	15	1976	36%	15
PSYC	2048	82%	17	20862	78%	17
BIOC	10703	87%	20	16017	64%	20
ENGI	7425	61%	8	3637	48%	8

*N.P.R represents the number of publications with Mendeley readership; % represents the percentage of publications with Mendeley readership over total number of WOS-indexed publications; M. (Mean) represents the number of Mendeley readership counts per publication.

Comparative analysis on the number of Mendeley readers per publication by identities.

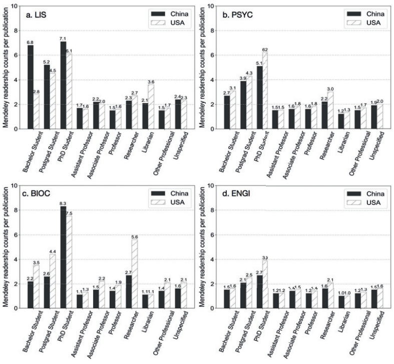


Figure 1. The distribution of Mendeley readership counts per publication by identities.

As shown in Figure 1, in LIS, on average each publication of China has 2.5 times as many bachelor student readers but only two-thirds as many librarian readers as that of USA. In BIOC, the number of PhD student readers per publication of China is higher than that of USA, while the number of researcher readers per publication of USA is twice as many as that of China. In PSYC and ENGI, the performance of publication of China as measured by Mendeley readership in all identifies is slightly poorer than that of USA.

Comparative analysis on Mendeley readers' source disciplines

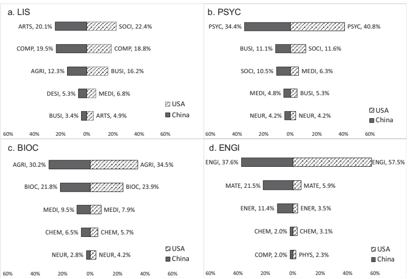


Figure 2. Source disciplines of Mendeley readers for publications of China and USA (Top five).

As shown in Figure 2, in PYSC and ENGI, the publications of both China and USA are read most by the readers from their own discipline. However, in BIOC, readers from Agriculture and Biological Sciences have the highest percentage, followed by the readers from its own discipline. While in LIS, the disciplines of the readers demonstrate the greatest difference between China and USA, suggesting that LIS publications of these two countries have impact on very different audiences.

Comparative analysis on Mendeley readers' source countries

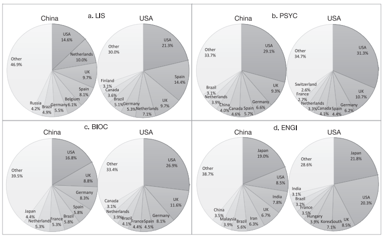


Figure 3. Source countries of Mendeley readers for publications of China and USA.

As shown in Figure 3, in LIS, PSYC and BIOC, readers from USA have the highest percentage for publications of both China and USA, suggesting that USA readers have much higher international visibility. Most importantly, for publications of China and USA in ENGI, readers from Japan, rather than from their own countries, have taken the highest proportion, suggesting that Japanese readers are very active to follow and use the research in ENGI area.

Conclusions

Main conclusions are drawn as follows. (1) In comparison, China has higher proportion of publications with Mendeley readership, but lower number of Mendeley readership counts per publication. (2) The number of Mendeley readership counts per publication by identities varies across disciplines and demonstrates discrepancy between China and USA, suggesting that publications of the two countries have yielded impact on different audiences. (3) In natural science field, source disciplines of Mendeley readers for publications of China shows the same pattern with that of USA, while in social science field, these two countries show quite different patterns. (4) USA Mendeley readers dominate in most disciplines of both China and USA, while Chinese Mendeley readers have low visibility. In ENGI, however, Japanese Mendeley readers have the highest percentage, publications of China have attracted more readers from developing countries like Brazil and India, while publications of USA have attracted more readers from developed countries like UK and South Korea.

These results are evidence that Mendeley altmetrics are useful in identifying underlying reading patterns of publications, revealing more nuanced and broad impact and comparing the performance of different countries.

Acknowledgments

The authors would like to sincerely thank Dujuan Xu for the help in data collection for the study. The research is supported by National Natural Science Foundation of China (NO.71804067), Humanity and Social Science Foundation of Ministry of Education of China (18YJC870023) and the Fundamental Research Funds for the Central Universities (No.30918013107).

References

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4), 895-903.

Costas, R. (2018). Towards the social media studies of science: social media metrics, present and future. Retrieved May 29, 2019 from: https://arxiv.org/abs/1801.04437.

Mohammadi, E. & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, 65(8), 1627-1638.

Thelwall, M. & Nevill, T. (2018). Could scientists use Altmetric. com scores to predict longer term citation counts? *Journal of Informetrics*, 12(1), 237-248.

Characterizing High-Quality Answers for Different Question Types on Academic Social Q&A Site

Lei Li [1,] Daqing He[2], Chengzhi Zhang [3, *]

[1] *leili@bnu.edu.cn*
Department of Information Management, Beijing Normal University, Beijing, China

[2] *dah44@pitt.edu*
Department of Informatics and Networked Systems, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[3] *zhangcz@njust.edu.cn*
Department of Information Management, Nanjing University of Science and Technology, Nanjing, China

Introduction

The sustainable development of social media has seen the emergence of many academic social network sites (ASNSs), which have attracted many scholars. The most popular ASNSs include ResearchGate.net (the largest), Academia.edu, and Mendeley.com (Van, 2014). ResearchGate provides a social Q&A platform for scholars to communicate with others by asking and answering questions, now recognized as academic social Q&A. Increasing numbers of scholars are raising questions on these sites, leading to a large increase in the number of answers. As on generic social Q&A sites, such as Yahoo! Answers, how to judge answer quality on academic social Q&A has become an urgent issue.

In our previous study, users of academic social Q&A and generic social Q&A sites determined answer quality according to different characteristics of the answers (Li et al., 2018). We also found some external factors that influence scholars to identify high-quality academic answers, such as the answer's discipline. This paper aims to explore whether are other external factors affect scholars' judgment of academic answer quality. Kim *et al.* (2007) point out that "users evaluate answers differently depending on the type of the question". For example, for opinion questions, questioners rate answers containing socio-emotional support most highly. Therefore, this paper will further explore whether question type affects the characteristics used by scholars to judge the quality of academic answers.

We take the dataset from our previous study (Li et al., 2018), comprising 973 answers to 101 questions in three representative disciplines on ResearchGate's Q&A. Adopting Harper *et al.*'s (2009) classification, the questions are divided into information-seeking and discussion-seeking. We then use logistic regression to explore the relationship between answers' characteristics and user-rated quality for both question types, aiming to identify the different characteristics of high-quality academic answers across these two question types.

To the best of our knowledge, this is the first study to explore the characteristics of high-quality academic answers across different question types. This study can help the academic social Q&A sites to recommend high-quality answers to users based on different question types, and then can contribute to providing scholars with an efficient academic information communication platform.

Method

Dataset

Besides the dataset of 973 answers to 101 questions on ResearchGate's Q&A, we also used our previous study's framework for characterizing high-quality academic answers. For all the answers in our dataset, values in respect of each characteristic have previously been determined. To measure answer quality, we use the number of "Recommendations" received on ResearchGate's Q&A. Answer quality is divided into three levels: low, medium, and high. Our previous study provides detailed information about our dataset and characteristic extraction (Li et al., 2018).

Based on Harper *et al.*'s (2009) question type classification, two coders used content analysis to unanimously divide the 101 questions into 47 information-seeking questions, 50 discussion-seeking questions, and 4 non-questions.

Regression analysis

After classifying the questions, the dataset was divided into two sub-datasets. Ordinal logistic regression (OLR) was then used for each sub-dataset to analyze the relationship between answer characteristics and user-rated quality across the two question types, because the answer quality value as the dependent variable is an ordered classified data. Before conducting OLR analysis, we reduced the right skewness of the data by log-transforming the continuous independent variables. We also conducted a multicollinearity diagnosis and parallel line test to confirm that our sub-datasets are suitable for OLR analysis. Finally, the results of OLR analysis for the two types of questions were compared to draw conclusions.

Results and Discussion

Table 1 shows the regression results between answer characteristics and answer quality for the two question types. The coefficient and significance values are reported for the purposes of comparison.

[*] Corresponding author

Table 1. Ordinal regression results for answer quality in responding to two question types

Characteristics (n=945, df=1)			**Information-seeking question（n=358）**		**Discussion-seeking question（n=587）**	
			Coefficient	**Sig.**	**Coefficient**	**Sig.**
Answerer's authority	**Answerer's history**	Answers provided	-0.858	0.196	**-1.276***	**0.035**
		Questions asked	-1.222	0.111	0.055	0.917
	Answerer's academic reputation	Answerer's publications	0.357	0.733	-0.315	0.693
		Publication reads	1.320	0.226	0.340	0.636
		Publication citations	0.533	0.635	0.382	0.648
		Impact points	-1.102	0.262	-0.245	0.704
		Institution's total impact points	-0.600	0.125	0.276	0.338
		Answerer's followers	2.036	0.072	**2.443****	**0.005**
Content characteristics	**Academic-related content characteristics**	Providing academic resources	**-0.531***	**0.034**	0.032	0.873
		Referring to basic theories	**0.926***	**0.021**	-0.336	0.181
		Providing research experience	-0.206	0.517	-0.131	0.639
	Non-academic content characteristics	Adding factual information	-0.188	0.461	0.024	0.906
		Providing opinions	-0.091	0.712	**-0.538****	**0.005**
		Consensus building	-0.045	0.877	0.263	0.182
		Social elements	-0.285	0.273	**0.384***	**0.033**
		Answer length	**2.988****	**0.001**	**2.006****	**0.001**

Notes: **: p<0.01; *: p<0.05

In the answerer's history category, there is only a significant negative relation between answer quality and the answerer's number of answers for discussion-seeking questions. Therefore, the more historical answers provided by a respondent to a discussion-seeking question, the less likely their answer to be recommended as high quality. Our previous research noted that ResearchGate's Q&A lists ongoing discussions between respondents and other respondents or questioners as new answers, so respondents to discussion-seeking questions usually have more historical answers. Therefore, this finding might be explained by most of the answers to discussion-seeking questions being complementary to previous answers, and so unlikely to be recommended as high quality.

In the answerer's academic reputation category, there is only a significant positive correlation between answer quality and answerer's followers for discussion-seeking questions. Thus, for this question type, as a respondent's number of followers increases, so does the likelihood of their answers being recommended as high quality. Conversely, for information-seeking questions, the answerer's authority does not affect users' judgment of answer quality. As reinforced by the further findings discussed below, in judging answer quality for this kind of question, answer content is the chief focus.

In the category of academic-related content characteristics, answers to information-seeking questions that refer to basic theories but include fewer academic resources are more likely to be judged as high quality. The finding that high-quality answers to information-seeking questions are those containing few resources is particularly interesting. Users seem to prefer theoretically based answers that provide detailed information directly, rather than pointing them towards other resources for answers. Reinforcing this phenomenon, we find that users also prefer longer answers to information-seeking questions.

Finally, in the category of non-academic content characteristics, answer quality for discussion-seeking question is also significantly positively correlated with answer length. In addition, high-quality answers to discussion-seeking questions include more social elements and do not contain too many personal opinions. Unlike for information-seeking questions, users prefer answers to discussion-seeking questions to include social elements, and rather surprisingly, few subjective opinions. With respect to the latter finding, users seem to prefer more objective answers in response to discussion-seeking questions.

Conclusion

Continuing our previous research, this paper further explored the different characteristics of high-quality academic answers for different question types. For discussion-seeking questions, users focus more on the answerer's authority and whether the answer contains social elements; for information-seeking questions, users focus more on whether the answer refers to the theoretical basis.

References

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature news*, *512*(7513), 126-129.

Li, L., He, D., Zhang, C., Geng, L., & Zhang, K. (2018). Characterizing peer-judged answer quality on academic Q&A sites: A cross-disciplinary case study on ResearchGate. *Aslib Journal of Information Management*, *70*(3), 269-287.

Kim, S., Oh, J. S., & Oh, S. (2007). Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective. *Proceedings of the American Society for Information Science and Technology*, *44*(1), 1-15.

Harper, F. M., Moy, D., & Konstan, J. A. (2009, April). Facts or friends?: distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 759-768). Boston: ACM.

Assessing citation network clustering as indicator normalization tool

Riku Hakulinen[1] and Eva Isaksson[2]

[1] Riku.Hakulinen@helsinki.fi, [2] Eva.Isaksson@helsinki.fi
University of Helsinki, Helsinki University Library, Research Services, PO Box 53,
00014 Helsinki (Finland)

Introduction

Attempts to use bibliometrics in assessing research performance requires a normalization procedure to cover publications from different fields. The classification behind this normalization has traditionally been broad categories based on scientific fields of journals. During recent years, due to the need for a more specific method, citation network clustering has been utilized to partition research output into (micro) research fields. Also, machine learning approaches with the aim of classifying publications based on their full text have emerged, but it seems that the time for these as a reliable basis for classification may not yet have come. In this work we study the effect of choosing a particular network clustering method on the bibliometric impact of University of Helsinki (UH).

The purpose of this effort is to test whether these clustering classifications that are network science wise perhaps coherent [Traag et al., 2019], produce also meaningful classifications, and to study how varying clustering methods and their parameters affect the UH results. We use publication data from several fields with different publication and citation practices and focus on testing Leiden and Louvain algorithms [Traag et al., 2019]. This results in statistics of classification robustness and, for example, a comparison of normalized citation scores (NCSs) for chosen sets of UH publications in varying clustering classifications. The NCSs here are produced by considering our dataset as the research publication output of the world.

Dataset

As a starting data we collected for this paper a set of 93.6 thousand publications from Web of Science, obtained through basic search ‘Topic = Social science’, limited to publication years 1990-2019. Citation counts from WoS were used in this work. A more comprehensive set of bibliographic data is collected for the poster from various sources including Scopus and Dimensions.

Method

Citation links are identified from bibliographic data and an undirected citation network is compiled from that as a list of publication index pairs. The network is used as input in the clustering tool *RunNetworkClustering*, provided by CWTS through GitHub: CWTSLeiden/networkanalysis. Further work on the data and clustering results is done using the statistical software R and the networks are visualized with VOSviewer.

Preliminary Tests

Algorithms

At least three clustering methods are tested on our data: Leiden algorithm with Modularity and Louvain and Leiden algorithms with Constant Potts Model (CPM) as quality functions. First observation was that with the still modest network of about 42k nodes and 100k links from the 93.6k publications, it was not trivial to find a suitable value for the resolution parameter. We used 0.00015 in CPM and 0.7 in Modularity.

Comparing NCSs

We calculated MNCSs for the UH set of 182 publications in 33 (Le/Mod), 51 (Le/CPM) and 58 (Lo/CPM) clusters. We also tabulated a comparison of three sets of NCS values for an exemplifying set of six publications from 31 UH publications sharing a cluster in each clustering result.

Themes within Clusters

Characterization of a network cluster content can have terms from a broader category or field, like Web of Science categories, which we used in the first test to label (in all three clustering results) the cluster containing the mentioned 31 publications.

Web of Science Categories as Label-sets

All three clusters produced by the three algorithms (between 5k and 7k publications in each), containing the 31 UH publications had the following five categories as the most numerous covering about half of all category designations:

[1] "Ecology"
[2] "Environmental Sciences"
[3] "Environmental Studies"
[4] "Geography"
[5] "Science & Technology"

When based on counts in the cluster, the order of these terms varied between types of clustering. More tests and possibly more specific terms are required to allow conclusions about the contents of these and other clusters.

Numerical Results

The results in Table 1 can be interpreted so that the normalizations based on classifications from all three clustering methods produce here similar, but different impact results.

Table 1. Example NCS values for arbitrary six of the 31 (see text) UH publications following normalization based on classification from three clustering methods and age of publication.

NCS	Publ_1	Publ_2	Publ_3
Leiden, Modularity	1.73	4.47	0.92
Leiden, CPM	1.69	4.35	0.90
Louvain, CPM	1.71	4.42	0.91
NCS	**Publ_4**	**Publ_5**	**Publ_6**
Leiden, Modularity	2.12	1.01	1.04
Leiden, CPM	2.06	0.98	1.01
Louvain, CPM	2.10	1.00	1.03

E.g., Publ_5 only gets above "world average" with Leiden/Mod –version of clustering, and MNCSs for the whole UH set were 1.28, 0.98 and 1.01 (Le/Mod to Lo/CPM). To have a better understanding from the perspective of an organization that purchases services based on these methods, this will be tested with more data and varying algorithm parameter values like the resolution.

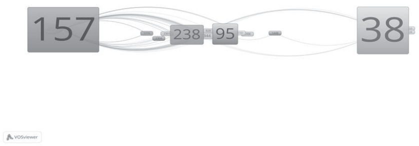


Figure 1. Visualization of Leiden/CPM cluster network. Includes about 42k publications in 426 linked clusters, compare to Figure 2. Cluster 38 contains 5181 publications of which 31 have an author with UH affiliation.

The large difference in amount of clusters in Figures 1 and 2 follows from that using CPM as the quality function produced about seven times more linked clusters with similar quality function values (~0.81) and with only 1.2 times larger total amount of clusters. The largest few clusters were of similar size in all three, but for CPMs the cluster sizes were more evenly distributed.

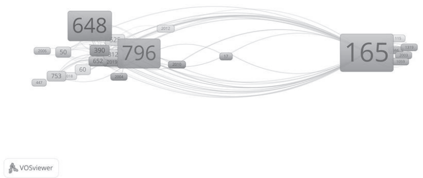


Figure 2. Visualization of Leiden/Modularity cluster network. Includes 64 linked clusters and is constructed from the same citation network as Figure 1. Cluster 165 contains 5715 publications of which 31 with UH affiliation.

Statistics and Conclusions

In addition to calculating average values and error bars using results from cluster analyses of the collected data, the aim is to connect some network properties like the number of vertices or even transitivity [Newman, 2002] with an indicator of quality of the clustering as classification. This will help to clarify results, e.g., the shown tentatively observed MNCS difference, in the context of UH publications.

Acknowledgments

The authors thank Ed Noyons from CWTS for discussing the application of Leiden algorithm for the University of Helsinki 2012-2016 publication dataset [Noyons & Mälkki, 2019].

References

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45, 167-256.

Noyons, E. & Mälkki, A. (2019). *Research performance analysis for the University of Helsinki 2012-2016/17.* Leiden: CWTS. Retrieved April 5, 2019 from: http://hdl.handle.net/10138/298733

Traag, V. A., Waltman, L. & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9.

Detecting Future Trends of Artificial Intelligence by Keyword Mapping in WoS and SCOPUS

Sejung Ahn[1] and Bitnari Yun[2]

[1]*sjahn@kisti.re.kr*
Korea Institute of Science and Technology Information, 66 Hoegi-ro, Dongdaemun-gu, Seoul 02456 (South Korea)
Data& HPC Science, University of Science and Technology, 217 Gajeong-ro, Yuseong-gu, Daejeon 34113 (South Korea)

[2]*kisti0746@kisti.re.kr*
Korea Institute of Science and Technology Information, 66 Hoegi-ro, Dongdaemun-gu, Seoul 02456 (South Korea)
Data& HPC Science, University of Science and Technology, 217 Gajeong-ro, Yuseong-gu, Daejeon 34113 (South Korea)

Introduction

Artificial intelligence (AI) technology is considered as an essential part in the era of the industry 4.0. Recently, the AI technology has been evolving very rapidly based upon the computing infra and sufficient big data for learning. Although there are many reports on the future prospects of artificial intelligence, it will be possible to grasp more precise status of future trends through scientometric analysis. In this study, we investigate the research area and detect the future trends of AI using bibliometric techniques. In addition, we discuss the difference between Web of Science and SCOPUS database through this study as well.

Data and Methods

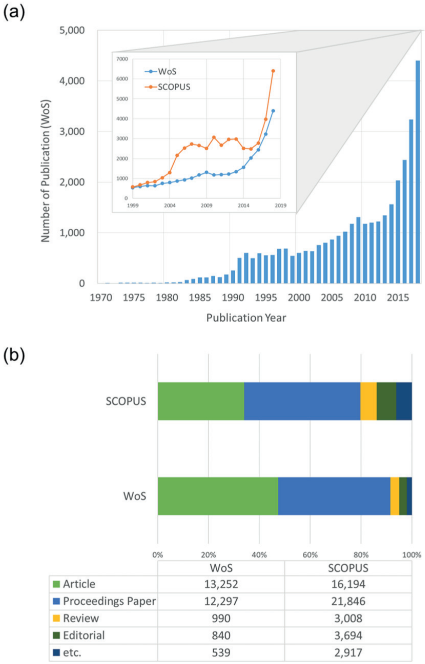
We used the Web of Science (WoS) database provided by Clarivate Analytics and SCOPUS database provided by Elsevier to collect the journal articles related on 'artificial intelligence'. Table 1 shows the search queries and results. The fundamental bibliometric analysis was carried out using VantagePoint of Search Technology, Inc. The keyword mapping and clustering analysis was performed using VOSviewer 1.6.10 of CWTS.

Table 1. Search queries and results

DB	Query	Time-span	Results
WoS	TS="artificial intelligence" (Indexes: SCIE, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI)	1970-2018	33,877
		1999-2018	27,918
SCOPUS	TITLE-ABS ("artificial intelligence") OR A UTHKEY ("artificial intelligence")	1999-2018	47,659

Result and Discussion

Figure 1(a) shows the number of AI related papers from 1970 to 2018 in WoS. It has steadily increased during 40 years and there are sharply increasing after 2010. The inset figure of Figure 1(a) shows the number of publications in WoS and SCOPUS from 1999 to 2018. The number of publications in SCOPUS are larger than in WoS, which is due to the differences of the included journals and search indexing. Figure 1(b) shows the document type composition of WoS and SCOPUS. It is investigated the ratio of proceedings paper in SCOPUS is much



	WoS	SCOPUS
🟩 Article	13,252	16,194
🟦 Proceedings Paper	12,297	21,846
🟨 Review	990	3,008
■ Editorial	840	3,694
■ etc.	539	2,917

Figure 1. (a) Number of publications of AI related papers in WoS by year from 1970 to 2018 (b) The document type composition of WoS and SCOPUS.

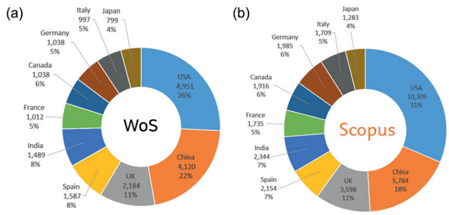


Figure 2. (a) Share of Top 10 countries of AI related papers in WoS and (b) SCOPUS.

higher than in WoS. Figure 2 represents the share of papers in the top 10 countries playing a leading role in AI. The composition of top 10 countries was the same in WoS and SCOPUS, and showed only a slight difference in ranking and ratio.

Keyword mapping and clustering of 33,877 WoS data were conducted to identify areas of research in the AI field. We used author keywords that appeared more than 10 times and identified 10 clusters by setting the resolution parameter of VOSviewer to 0.7. The major 5 research areas in AI technology field are ‘expert system’, ‘robotics’, ‘neural networks’, ‘machine learning’ and ‘deep learning’ as shown in figure 3. This mapping results reflects the current R&D status of AI technology sufficiently.

Meanwhile, we performed overlay mapping analysis using keyword-average-year in order to detect the future trend of AI. Overlay mapping techniques have often been used to detect the emerging technologies (Rotolo, Rafols & Leydesdorff, 2015). Figure 4 shows the result of overlay mapping for WoS and SCOPUS from 1999 to 2018. Both mapping and clustering results were similar, but there was a slight difference in the frequency and distance of the keywords. Also, we could confirm a little bit meaningful difference in research topography between WoS and SCOPUS through the keyword overlay map. As a result of the analysis, it was confirmed that 'big data' and 'deep learning' is the core AI research area in recent years. In particular, some biotechnology(BT) related keywords such as ‘health care’, ‘precision medicine’, ‘breast cancer’, and ‘drug discovery’ have recently become more prominent in the WoS overlay map.

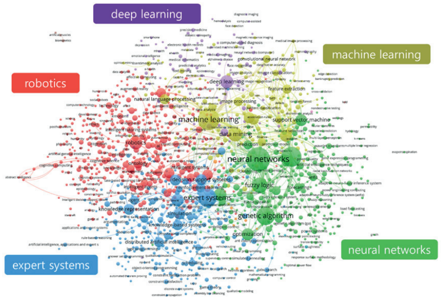


Figure 3. Network map of author keywords in WoS (1970-2018).

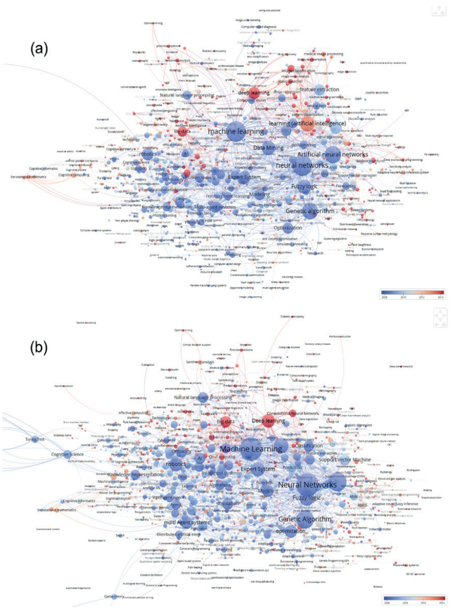


Figure 4. (a) Overlay map of author keywords in WoS and (b) SCOPUS in 1999-2018.

These BT related keywords are closely linked with the keywords such as ‘medical image processing’, and ‘image classification’ which are full-fledged applications of artificial intelligence technology. However, in the case of SCOPUS, the BT related keywords are shown weaker than the WoS. In particular, keywords related to AI algorithms such as 'convolutional neural network' are prominent.

Conclusion

In this study, we investigated the research trend of AI technology and identified a rapidly grow after 2010. The research areas in AI field were confirmed by the keyword mapping and clustering analysis. The overlay maps provided the emerging future trends of AI technology. It is expected that the rapid progress of AI algorithms and application to biotechnology will be active.

Acknowledgments

This study was funded by the Korea Research Foundation (NRF-2017R1E1A1A03070975) and the Korea Institute of Science and Technology Information (K-19-L03-C03-S01).

References

Rotolo, D., Rafols, I., Hopkins M. M. & Leydesdorff, L. (2017). Strategic intelligence on emerging technologies: Scientometric overlay mapping. *Journal of the Association for Information Science and Technology*, 68, 214-233.

On the Latent Shape of ICT research

Chiara Carusi[1] and Giuseppe Bianchi[1]

[1] *name.surname@uniroma2.it*
University of Rome Tor Vergata, Via del Politecnico 1, Electronic Engineering Dept., Rome (Italy)

Introduction

Applying a spectral clustering approach used in text analysis to the typical scientometrics problem of community detection, in a previous work by the authors (Carusi & Bianchi 2019) we exploited academic publications as a link between scholars and journals to detect and investigate research communities in a data-driven manner. As a follow-up of our recent work, the goal of this poster is to highlight what we believe is a very interesting and intuitive interpretation of the vector space in which our bibliometric analysis was cast, discussed hereafter in terms of the different targets and methodologies characterizing research activity.

The publication dataset

The results discussed in this poster cover 47,718 publications authored by at least one Italian faculty member in the Information and Communication Technology (ICT). Suitably aggregating the information retrieved, we built a (high-dimensional) 2,582x1,454 scholar-by-journal publication matrix, each entry set to the total number of papers that a given scholar has published in a given journal within the sixteen-year reference period 2000-2016.

The SVD latent space and the ICT communities

To project the scholar-journal publication dataset into a lower-dimensional, denser space, we resorted to its singular value decomposition (SVD) and considered the three main dimensions resulting from SVD. Computing the SVD of the publication matrix proved to be a very effective technique: (i) both scholars and journal were projected in the same metric space, and (ii) when addressing similar research topics, even two scholars that have never published a paper in the same journal were mapped as close (similar) points in the new space – and the same applies for journals. Thanks to this property of spatially reflecting similarities and differences between scholars and journals, the SVD space makes patterns in research activity clearly distinguishable (Figure 1). Via the spherical K-Means clustering algorithm, we finally detected 5 ICT communities, respectively addressing *computer science*, *electronics*, *telecom*, *controls* and *biomedics.*

The latent structure of ICT research

In addition to making the detection of communities easier, the latent metric space defined by the SVD of the publication matrix provides also a powerful machinery to investigate research activity in more detail. Indeed, the first dimensions of the SVD space appear to reflect the most essential aspects of research, and the coordinates of scholars and journals along the axes are the extent to which these aspects affect the activity of each scholar or are relevant for the scope of a journal. To guide the interpretation of the meaning behind the first three SVD dimensions, for each of them we will use, as a reference, the set of journals with the highest and lowest coordinates (Tables 1, 2 and 3).

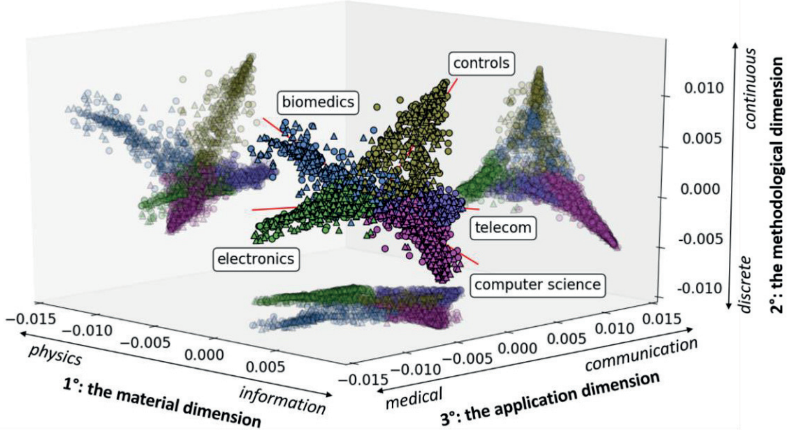


Figure 1. ICT scholars (circle markers) and journals (triangle markers) in the three-dimensional space obtained via SVD of the publication matrix, colored according to their community.

The material dimension

Starting from the first SVD dimension, the journals with the most negative coordinates appear to be related to physics - nuclear physics in particular - as opposed to information science journals (Table 1). Looking at Figure 1, this dimension clearly suggests a first partition of ICT research based on the tangible, physical nature of electronics and devices on the one side, and on the theoretical and abstract nature of logics, computation and combinatorics on the other. This dimension therefore clearly separates the *electronics* from the *computer science* community; conversely, the *telecom*, *controls* and *biomedics* communities all (reasonably) appear to have no definite physics- vs. information-based nature, lying halfway between these two opposite targets. However, to a certain extent, such "material" dimension affects also the *telecom*, *controls* and *biomedics* communities, even though only at a more detailed level. For instance, *telecom* journals range from remote sensing (*Rem. Sens. of Environment:* -0.0021, *IEEE Trans. on Geoscience and Rem. Sens.*: -0.0021, *Canadian J. of Rem. Sens.*: -0.0020) to computing and networking issues (*Performance Evaluation*: 0.0034, *Peer-to-Peer Netw. and Applic.:* 0.0032, *Pervasive and Mobile Comput*: 0.0031). Signal processing journals lie instead very close to the origin (*IEEE Trans. on Signal Proc.:* -6.9x10^{-5}, *IEEE J. on Selected Topics in Signal Proc.:* 1.3x10^{-4}, *J. of Signal Proc. Systems* 1.3x10^{-4}), as the relevant research activity equally draws from mathematics/ information theory and electronics engineering.

Table 1. ICT journals with the highest and lowest values on the 1st SVD dimension.

Journal	1stdim.
Int. J. of Algebra and Computation	0,0077
Higher-Order and Symbolic Computat.	0,0076
J. of Integer Sequences	0,0076
Electronic J. of Combinatorics	0,0076
Logical Methods in Computer Science	0,0075
Nuclear Physics A	-0,0052
Microscopy and Microanalysis	-0,0053
Physical Review C - Nuclear Physics	-0,0054
European Physical Journal C	-0,0054
Physics Letters B: Nuclear, Elementary Particle and High-Energy Physics	-0,0055

The methodological dimension

Moving to the second SVD dimension, ICT research seems to be "arranged" from a methodological point of view: the points with largest coordinates on the second axis account for control journals, which address dynamic systems and differential equations, as opposed to the negative coordinates associated to journals on programming methodologies (Table 2). Data seem to suggest that, independent of its physical vs information-theoretical nature, the second, crucial aspect of ICT research lies in the kind of methodologies employed, thus separating *controls* and *biomedics* from the other ICT communities (continuous-time vs. discrete mathematics).

Table 2. ICT journals with the highest and lowest values on the 2nd SVD dimension.

Journal	2nddim.
Math. of Control, Signals, and Systems	0,0111
J. of Mathematical Control and Inform.	0,0110
Systems and Control Letters	0,0103
Annual Reviews in Control	0,0101
Int. J. of Robust and Nonlinear Control	0,0101
Eur. J. of Combinatorics	-0,0052
Logical Methods in Computer Science	-0,0053
Electronic J. of Combinatorics	-0,0053
Higher-Order and Symbolic Computat.	-0,0053
J. of Integer Sequences	-0,0053

The application dimension

Finally, the third SVD dimension appears to focus on the practical problems targeted by research activity, distributing communication and medical applications as opposite extremes (Table 3). Such "application" dimension definitely separates the *telecom* community from *biomedics*, while leaving the other communities closer to the origin. Nonetheless, investigating such dimension at a higher resolution, *computer science* journals with negative and positive coordinates - even though rather small in terms of absolute values - consistently account for two very different specialties: bioinformatics (*BMC Evolutionary Biology*: -0.0049, *BMC Bioinform.:* -0.0035, *BioData Mining:* -0.0034) and distributed computing (*IEEE Trans. on Parallel and Distributed Systems*: 0.0021, *Int. J. of Pervasive Comput. and Commun.*: 0.0023, *J. of Internet Services and Applic.:* 0.0025).

Table 3. ICT journals with the highest and lowest values on the 3rd SVD dimension.

Journal	3rddim.
Foundations and Trends in Comm. and Information Theory	0,0074
IEEE Wireless Comm. Letters	0,0073
Physical Comm.	0,0072
IEEE Trans. on Wireless Comm.	0,0072
Int. J. of Satellite Comm. and Network.	0,0072
Diabetes	-0,0110
J. of Sports Sciences	-0,0111
J. of Applied Physiology	-0,0112
J. of Applied Clinical Medical Physics	-0,0117
European Journal of Applied Physiol.	-0,0120

References

Carusi C., Bianchi G. (2019). Scientific community detection via bipartite scholar/journal graph co-clustering. *J. of Informetrics*, 13(1), 354-386.

Debunking the Italian Scientific Sectors' classification system: preliminary insights

Giuseppe Bianchi[1] and Chiara Carusi[1]

[1] *{giuseppe.bianchi, chiara.carusi}@uniroma2.it*
University of Rome Tor Vergata, Via del Politecnico 1, Electronic Engineering Dept., Rome (Italy)

Introduction and motivation

Italy is one of the very few EU countries whose Ministry of Education, University and Research enforces a top-down scholar classification system. Such a system currently comprises 367 "scientific disciplinary sectors" (SDS), aggregated first into 88 macro-sectors and then into 14 high-level areas. Each Italian university faculty member belongs to one (and only one) specific SDS, which should best reflect her/his expertise and main field of research. The topmost importance of a truthful and consistent classification is amplified by *its pervasive usage in virtually all crucial university-related activities*: teaching accreditation (with strict conditions on ECTS credits delivered by specific SDSs), career-related assessments and habilitation/promotions (managed inside a given SDS, and involving as evaluators *only* professors from that SDS), per-SDS bibliometric thresholds, etc. And a faculty aiming at changing SDS label must undergo an extensive assessment involving several formal approval steps.

With such a crucial and pervasive role of the SDS classification in Italy, a number of questions naturally arise. To what extent is such an "a-priori" classification reliable and truthful, i.e. *is it really representative of the scientific communities involved?* And what is the level of overlap and intertwining among different SDSs, if any?

Contribution

To answer the above questions, we leverage the unsupervised data analysis methodology proposed in (Carusi & Bianchi, 2019). While our prior paper mainly focuses on the methodology itself, goal of this poster contribution is to explicitly compare its purely data-driven (hence aiming at being objective) findings with the Italian top-down SDS classification. Our preliminary results unveil potentially severe biases and structural overlaps in both broad areas investigated so far, namely SDSs revolving around ICT disciplines, and SDSs addressing Microbiology & Genetics.

Methodology and Datasets

Our spectral co-clustering methodology (Carusi & Bianchi, 2019) builds upon a bipartite graph that maps authors to the journals in which the authors have published their papers. In essence, each author is modelled as a vector in a (huge!) M-dimensional space where each coordinate accounts for the number of papers (zero or more) published on a specific journal. For the time range 2000-2016, we specifically studied two broad scientific areas: Information and Communication Technology (ICT dataset) and Microbiology & Genetics (MG dataset). Each dataset was constructed starting from the list of scholars (faculty members of Italian universities) belonging to the relevant SDSs. As summarized in Table 1, we used 8 SDSs for the ICT dataset (all 7 SDSs in the information engineering area plus an SDS dedicated, at least in principle, to theoretical computer science). In the MG dataset, we considered 11 SDSs belonging to 4 different areas (biology, chemistry, medicine, and even one SDS from agriculture) but which tackle, at least partially, microbiology and/or genetics.

Table 1. Scientific Disciplinary Sectors

ICT Dataset - SDSs	MG dataset - SDSs
INF/01 *Informatics*	BIO/11 *Molecular Bio.*
ING-INF/01 *Electronic Eng.*	BIO/13 *Experimental Bio.*
ING-INF/02 *Electromagnetics*	BIO/15 *Pharmaceut. Bio.*
ING-INF/03 *Telecommun.*	BIO/18 *Genetics*
ING-INF/04 *Control Sys. Eng.*	BIO/19 *General Microbio.*
ING-INF/05 *Information Processing Systems*	CHIM/10 *Food Chemistry*
ING-INF/06 *Electronic & Informatics Bioengineer*	CHIM/11 *Chemistry & Biotech. of Fermentation*
ING-INF/07 *Electrical & Electronic Measurement*	MED/03 *Medical Genetics*
	MED/04 *Experimental Medicine & Pathophys.*
	MED/07 *Microbiology & Clinical Microbiology*
	AGR/16 *Agriculture Microbiology*

Leveraging the observation that researchers in similar fields tend to mainly publish in the same set of journals, the problem of identifying scientific communities can be cast into a co-clustering problem on the bipartite scholar/journal graph. The number of clusters K can be varied to permit comparison with the "a-priori" Italian SDS classification at different resolution levels.

Analysis of the ICT sectors

As a starting point, Table 2 shows how the 8 considered SDSs map onto K=5 clusters resulting from our unsupervised (data-driven) classification algorithm. The matrix reports both number of scholars as well as % w.r.t. the Italian SDS. Clusters are descriptively labelled using the three most recurring keywords in the journal names belonging to the same cluster. Since the number of SDSs is greater than the coarser clustering resolution (here only 5 clusters), we expected aggregations, which are in fact confirmed by the table (INF/01 aggregated in the first cluster with ING-INF/05; ING-INF/02, 03 and 07 aggregated in the second cluster). The only minor bias so far is the non-aggregation of the Electromagnetics SDS (ING-INF/02) with the SDS on Telecommunications (ING-INF/03), despite for some calls for professorship they are aggregated in a same higher-level group by the Italian System.

Table 2. ICT confusion matrix

	knowledge theoretical logic	instrum. propagat. antennas	remote sensing commun.	control automatic robotics	medicine biology biomedical
INF/01	**637 (86%)**	7 (1%)	41 (6%)	7 (1%)	46 (6%)
ING-INF/05	**386 (77%)**	12 (2%)	38 (8%)	35 (7%)	30 (6%)
ING-INF/07	0	**117 (96%)**	0	4 (3%)	1 (1%)
ING-INF/01	10 (3%)	**298 (87%)**	10 (3%)	7 (2%)	16 (5%)
ING-INF/02	0 (0%)	**143 (86%)**	17 (10%)	1 (1%)	5 (3%)
ING-INF/03	1 (0%)	28 (9%)	**294 (90%)**	3 (1%)	2 (1%)
ING-INF/04	1 (0%)	0	1 (0%)	**257 (96%)**	8 (3%)
ING-INF/06	1 (1%)	3 (3%)	1 (1%)	3 (3%)	**111 (93%)**

More surprises emerge when clustering at a finer granularity. Indeed, we'd obviously expect the above aggregated SDSs to split apart as the chosen number of clusters grows. As shown in Figure 1, this does not appear to be the case, especially in relation to the two SDSs *INF/01 - Informatics* and *ING-INF/05 - Information Processing Systems*. While initially merged into a single **knowledge & theoretical logic** community, when the number of clusters increases from 5 to 9 both SDSs start to split into smaller communities related to research on **pattern recognition** and **pervasive multimedia**. However, as K increases, they surprisingly do not separate (as we would expect from different sectors), but appear to contribute equally to these more specific research subfields. In other words, these two SDSs seem to be deeply entangled, i.e. to behave as a single sector, each comprising the same 3 subsectors (case K=9).

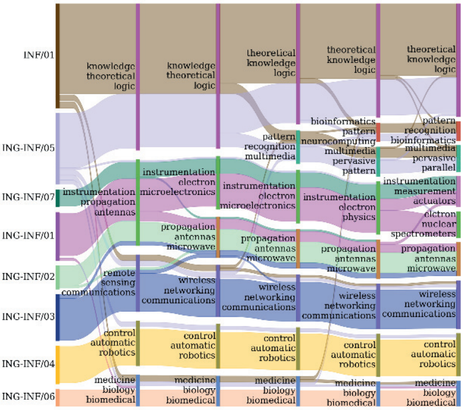


Figure 1. ICT alluvial diagram

Analysis of the Microbiology and Genetic sectors

Since an alluvial diagram would require extra space, we just report, in Table 3, the high-resolution case of 15 clusters, more than the 11 sectors accounted in the MG dataset. While some SDSs (e.g. CHIM/10, AGR/16, BIO/15, MED/03 and 07) have a quite sharp characterization, some overlap appears to emerge when comparing CHIM/11 and BIO/19. But an even more significant overlap appears to characterize the remaining BIO sectors (11, 13, 18), to the extent that it is hard to identify characterizing differences, despite the fine-grained projection over as much as 15 clusters. This is quantitatively confirmed by the correlation coefficients among the relevant rows, which range from a minimum of 0.68 to as much as 0.90 (BIO/13 vs BIO/18)! In front of such a high correlation, the question of which might be the most significant differentiating aspects among such SDSs appears thus a legitimate one. Especially for outsiders of such SDSs, such as the first author of this poster, which belongs to the Italian SDS *ING-INF/03 (Telecommunications)*, i.e., an SDS whose activity should not even tackle this poster's topics!

References

Carusi C. & Bianchi G. (2019). Scient. community detection via bipartite scholar/journal graph co-clustering. *J. of Informetrics*, 13(1), 354-386.

Table 3. Microbiology&Genetics confusion matrix

	food agricultural chromatogr.	natural product phytochem.	food microbiology technology	biotechnol. microbiology microbial	bacteriology microbiology antibiotics	antimicrobial infectious chemotherapy	virology vaccine hepatology	genetics medical human	genetics human prevention	plant aquatic marine	mechanisms bioinform. development	cell oncogene death	immunology leukocyte medicinal	blood ageing leukemia	artificial organs biomaterials
CHIM/10	59(82%)	12(17%)	0	1(1%)	0	0	0	0	0	0	0	0	0	0	0
BIO/15	6(9%)	62(90%)	0	0	0	0	0	0	0	0	0	1(1%)	0	0	0
AGR/16	0	0	113(77%)	32(22%)	0	0	0	1(1%)	0	0	0	1(1%)	0	0	0
CHIM/11	2(5%)	2(5%)	1(2%)	27(66%)	6(15%)	0	0	0	0	3(7%)	0	0	0	0	0
BIO/19	0	3(2%)	8(7%)	42(35%)	21(17%)	19(16%)	10(8%)	0	0	9(7%)	3(2%)	2(2%)	4(3%)	0	0
MED/07	2(1%)	4(1%)	1(0%)	0	16(6%)	208(72%)	32(11%)	1(0%)	0	0	0	1(0%)	11(4%)	7(2%)	7(2%)
BIO/18	0	1(1%)	3(2%)	8(4%)	2(1%)	3(2%)	1(1%)	33(17%)	40(21%)	18(9%)	41(22%)	32(17%)	3(2%)	3(2%)	2(1%)
MED/03	0	0	0	0	0	0	0	113(73%)	22(14%)	0	6(4%)	5(3%)	2(1%)	4(3%)	2(1%)
BIO/13	1(0%)	2(1%)	0	9(3%)	1(0%)	1(0%)	11(4%)	50(17%)	38(13%)	14(5%)	59(20%)	71(24%)	14(5%)	16(5%)	6(2%)
BIO/11	0	1(0%)	0	10(4%)	3(1%)	0	4(2%)	10(4%)	17(7%)	40(16%)	52(21%)	87(36%)	14(6%)	7(3%)	0
MED/04	2(0%)	8(2%)	0	1(0%)	2(0%)	9(2%)	35(7%)	9(2%)	26(5%)	3(1%)	32(7%)	188(38%)	112(23%)	53(11%)	9(2%)

Science, technology and innovation indicators to support research management: the case of Oswaldo Cruz Foundation (Fiocruz)

Marcus Vinícius Pereira-Silva[1], Fernanda Lopes Fonseca[1], Bruna de Paula Fonseca[1], Camila Guindalini[1], Rodrigo Ferrari[1], Paula Xavier dos Santos[1]

[1] *marcus.silva@fiocruz.br; ffonseca@cdts.fiocruz.br; bfonseca@cdts.fiocruz.br; cguindalini@cdts.fiocruz.br; rodrigo.ferrari@fiocruz.br; paula.xavier@fiocruz.br*
Fundação Oswaldo Cruz, *Av.* Brasil 4365, sala 7, 21040-900, Rio de Janeiro (Brazil)

Background

The mission of Oswaldo Cruz Foundation (Fiocruz) is to produce, disseminate and share knowledge and technologies to strength and consolidate the Brazilian Unified Health System (SUS), ultimately contributing to the health promotion and quality of life of the population. The Foundation is present in 10 Brazilian states and has an office in Mozambique. Nowadays, there are 16 scientific and technical units and 32 post-graduate programs in different areas of the health field, including Clinical Research, Development of Prophylactic and Therapeutic Vaccines, Molecular and Genetic Epidemiology in Health, Education and Health, History of Science, among others.

As a public and strategic institution, Fiocruz develops health research to generate benefits for society. However, there are few established mechanisms to assess the influence of the knowledge produced by the institution on the cultural, educational, economic, political and social fields. Most of the models that guide research evaluation and monitor processes are based on a productivity logic, in which quantitative data is used as qualitative indicators of research performance, neglecting the existent diversity of the various knowledge areas. To overcome these limitations, it is necessary to adopt new approaches to evaluate research impact.

This work aims to present the experience of developing the Fiocruz's Observatory in Science, Technology and Innovation (ST&I) in Health, as well as to highlight some of the institutional indicators produced in this context. The platform intends to contribute to Fiocruz's research management and ST&I policies formulation, through the production of indicators, studies, technical documents and news that support decision-making processes. It also aims to increase the social perception about the institution's potential, in terms of the achieved research and technological development advances.

Method

This paper is based on a case of participant observation of an ST&I indicators project to monitor and evaluate research and technological development of a public health institute. For the scientific production indicators, publications with at least one author affiliated with Fiocruz were extracted from the Web of Science (WoS), Scopus and SciELO databases. For patents, the Questel Orbit database was used. In both cases, the VantagePoint software was used for database harmonization, duplicate records removal and institutions standardization and Kibana e Elasticsearch for data visualization.

Results

During the pilot experiment in 2016, working groups from different units of Fiocruz produced diverse indicators, including bibliometric, demographic, scientific collaboration and technological development. Despite the progress achieved, the governance model was not effective. InCites and 'Plataforma Stela Experta' were contracted to support the development of these indicators, but due to financial constraints, the maintenance of the signatures became impracticable.

In 2018, the Observatory's governance was reformulated and it is currently coordinated by the Vice-Presidency of Education, Information and Communication and has an Executive Committee comprised of ST&I and bibliometric specialists, which is responsible for coordinating partnerships and the operational team.

The scientific production indicators of Fiocruz provided information about year, database, keyword, journal, funding and collaboration. It is also possible to filter and combine these indicators and access the articles through their DOI. Between 2010 and 2018, Fiocruz published 18,769 documents with a 48% growth rate over the years. A database analysis showed that Scopus (74%) and WoS (71%) had the highest number of records, while SciELO indexed just 17%. Among the 21 journals that published more than 100 Fiocruz's publications, 9 of them were foreign. The top five were: Ciência e Saúde Coletiva (886), Cadernos de Saúde Pública (780), PloS One (692), Memórias do Instituto Oswaldo Cruz (561) and PloS Neglected Tropical Diseases (407). United States, United Kingdom and France-based institutions were the most frequent international partners of Fiocruz, sharing authorship in 2,590 (14%); 1,182 (6%) and 622 (3%) publications, respectively. Among the 10 most frequent partner countries, just Argentina was a developing economy, with 449 co-authored publications. Overall, 5% of Fiocruz's publications were in collaboration with South American countries and 23% with seven major developed economies countries (G7) (UN, 2019).

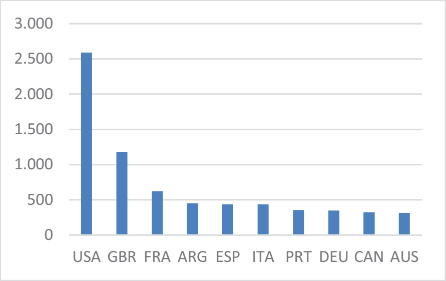


Figure 1. Top ten most frequent Fiocruz's partner countries.

The most frequent partners of Fiocruz were national public universities located in the Southeast Region of Brazil, including: Federal University of Rio de Janeiro (UFRJ), Federal University of Minas Gerais (USP), São Paulo University (USP), Rio de Janeiro State University (UERJ) and Fluminense Federal University (UFF), with 3,013 (16%), 1,583 (8%), 1,503 (8%), 1163 (6%) and 1,123 (6%) publications in collaboration, respectively. Our data also showed that the main sources of funding were public agencies, such as CNPq, Capes and Faperj.

Patents indicators provided information about number of patents, inventor name, partner institution, deposit region, classification and the link to access the document. Fiocruz has 197 patent families, 122 of them alive. Most of the patent families were filled without partnership (75%) and focused on drugs (52%), biotechnology (22%) and biological materials analysis (21%).

The Observatory's web portal (www.observatorio.fiocruz.br), relaunched in 2018, also integrates a document collection about institutional data and several contents that articulate and contextualize the indicators, such as reports, interviews, expert opinions, scientific articles and full texts of dissertations and theses.

Conclusions

Aligned with the open science movement and public transparency, the Observatory's is an important instrument for Fiocruz's research advancement and technological development. Despite the advances, some limitations must be acknowledged. Although most SciELO's articles derived from the health sciences area, in Portuguese and published by Brazilian authors[1], a more complete data source is necessary to consider other types of publications, such as books and book chapters. To accomplish this, an institutional tool to collect and standardize the data of the curricula registered in the 'Plataforma Lattes' is being developed[2].

The Observatory also intends to implement indicators of vaccines, biopharmaceuticals and diagnostic kits production, funding and education. In addition, considering that Fiocruz is currently elaborating an internal policy for managing and opening research data, the Observatory aims to develop indicators to measure the impact of such policy on research and technological development.

Finally, as the quantitative indicators should support qualitative analysis (Hicks et al. 2015), a group of specialists from different health areas was formed to analyse the data and propose new indicators considering the characteristics of each area. Journals published by Fiocruz also play an important role in the dissemination of the knowledge produced by the institution. For this reason, it is necessary to evaluate the rate of endogeneity with caution. Moreover, although Fiocruz has encouraged collaboration with South American and Portuguese speaking countries in more recent years (Ferreira et al., 2010), our data showed that the initiative has not yet reflected on higher collaboration rates. A large part of the collaborations to date has been with developed countries.

References

Ferreira J.R., Hoirisch C., Fonseca L.E & Buss P. M. (2016). International cooperation in health: the case of Fiocruz. *História, Ciências, Saúde-Manguinhos*, 23(2), 267-276.

Hicks D., Wouters P., Waltaman L., Rijcke S. & Rafols I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520, 429-431.

United Nations. (2019). *World Economic Situation and Prospects 2019*. New York: UN.

[1] Source: Scielo

[2] http://lattes.cnpq.br/

Global overview of patenting landscape in unmanned aerial vehicles

Gorry Philippe[1] and Maxim Kotsemir[2]

[1] *philippe.gorry@u-bordeaux.fr*
GREThA UMR CNRS 5113, Department of Humanities and Social Science, University of Bordeaux, Pessac (France)

[2] *mkotsemir@hse.ru*
Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow (Russian Federation)

Introduction

Unmanned Aerial Vehicle (UAV) or "drone" is defined as an aircraft without pilot on-board. While UAV originated in military applications after World War II, their use is recently expanding to commercial, scientific, recreational, agricultural and other applications. Market forecasts estimated UAVs to be a multi-billion dollars market within the next five to ten years (OECD International Forum on Transport, https://www.itf-oecd.org). Studying the UAVs development is interesting to understand combinatorial innovation and the economics interplay of defense and civil research.

Literature review

Although, UAVs is a high-tech field, few scholars have studied the dynamic of this today promising economic sector through patent indicators. Their research works are dedicated to specific aspects (see here as examples of such studies in e.g. Shiue and Chang, 2010; Liu et al., 2016; Kim et al., 2016). Our paper fills the gap in the analysis of the key technological domains and the key players in this field through patent landscaping. The current study is also the opportunity to use original data visualizations of the results of our analysis.

Methodology of the research

Our patent analysis was based on the worldwide collection of INPADOC (International Patent Documentation; EPO worldwide legal status database) using the Orbit (Questel®) SAS patent research platform. Patent bibliographic data were analyzed and visualized with the Orbit built-in analytic functions for the different information fields (priority or publication date and country, applicant name).

The data was derived at December 2017 and all metrics were based on patent family and priority date. The time span of our analysis covers 1995 – 2017 years. To detect the corpus of patent families for our study we run a complex query of keywords related with UAVs based on the set of keywords proposed in Kotsemir (2019) for the comprehensive analysis of UAVs publication trends in Scopus database. In the patent search, we run the combination of UAV-related adjectives like "unmanned" "unpiloted" and "unhabituated" plus terms like "aerial vehicle", aircraft, drone, "air vehicle", "helicopter" and also terms like "quadrotor", quadrocopter", "flying drone" etc. in the following bibliographic search fields: title, abstract and claims. In our query search we consider all word forms (i.e. singular and plural) of the searched keywords.

Results

Our analysis shows that until 2012 we can see quite stable dynamics of patenting in fields of UAVs but in 2014 – 2016 there was a burst of patent activity (Figure 1).

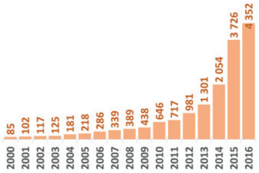


Figure 1. Number of UAV patent families by publication years in 2000 – 2016

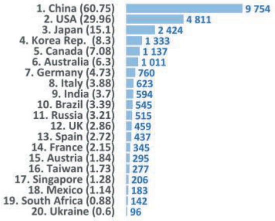


Figure 2. Top-20 countries by number of UAV patent families for publication years 2000 – 2016

The leading country in patent activity in UAVs is China, contributing to 60.7% of all patent families in UAVs for 2000-2016 publication years (Figure 2). Far behind China is the USA with almost 30% of contribution to global volume of patents in UAVs. Other quite important players in UAV patenting are Japan and South Korea. European countries lag far behind country-leaders with less than 5% of global number of patents for 1996 – 2015. We should note here that all BRICS countries are among top-20 countries by number of patents in

UAVs. Also, Asian countries show much stronger patent activity than European ones.

Figure 3 and Table 1 provide the sub-technological domains of global UAV R&D using the 35 domains defined by WIPO (World Intellectual Property Organisation) based IPC (International Patent Classification) codes. Key tech domains of global UAV patent landscape for 1996 – 2015 are "Transport" and "Control". It is the most "hot topic" in UAV patent activity. Other technical domain of importance is "Measurement".

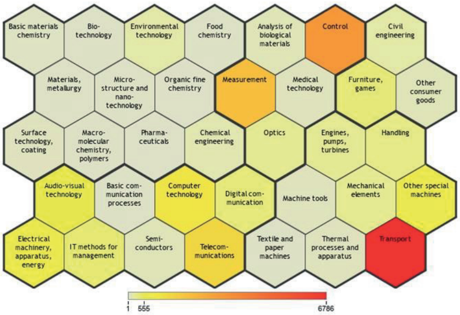
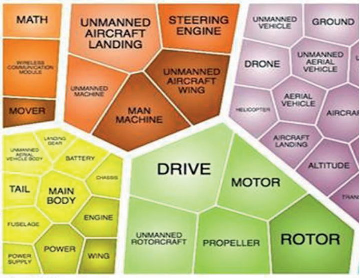


Figure 3. UAV patent families by WIPO Technology domains for 1996 – 2015

Note. Technological domains with the highest number of patent families are colored in red and orange.

Table 1. Top-10 Technology domains in UAV patents for 1996 – 2015

Technology domain (Number of patent families)
1. Transport (6 886); 2. Control (3 262); 3. Measurement (2 083); 4. Telecommunications (1 419); 5. Computer technology (1 050); 6. Other special Machines (728); 7. Electrical machinery, apparatus, energy (702); 8. Audio-visual technology (550); 9. IT methods for management (384); 10. Furniture, games (252)



© Questel 2017

Figure 4. Treemap clustering of technology segmentation concepts in UAV patent landscape (fragment)

Figure 4 shows the cluster of the underlying technical concepts in UAVs by measuring the shortest distance between the concepts and arranging them in hierarchical clusters. The topic segments of UAV patents are concentrated in topics related with parts of UAVs (like its engine, main body etc.) and also with tools (and methods) control of UAV (landing, flight etc.) (Table 2).

Table 2. Example of segment "Main body" (yellow segment in Fig. 4) in the technology segmentation treemap cluster map

Segment content (technology concepts)	N. of PFs
Battery	145
Chassis	131
Engine	148
Fuselage	1082
Landing Gear	291
Main Body	139
Power	139
Power Supply	144
Tail	180
Unmanned Aerial Vehicle Body	195
Wing	365

Note: "N. of PFs" means "number patent families".

Conclusions

Our research provided the overview of global patent landscape in field of UAVs for 1996 – 2016. China is the dominating country in patent activity. The "hot" technology domains of UAV patents are "Transport" and "Control". Topical segments of UAV patent landscape are concentrated by parts of UAVs, its control, and different aspects of application of UAVs. In the development of the study analysis of the leading firms and their collaboration through network analysis will be presented as well their competitive position using topographic map based on vector model of concepts extracted through semantic analysis. Further work should also include econometric modelling integrating other indicators such as research intensity measured by publications (Kotsemir, 2019) and macroeconomic indicators such GDP and defense budget.

References

Kim, D. H., Lee, B. K., & Sohn, S. Y. (2016). Quantifying technology–industry spillover effects based on patent citation network analysis of unmanned aerial vehicle (UAV). *Technological Forecasting and Social Change*, *105*, 140-157.

Kotsemir M. (2019). Unmanned aerial vehicles research in Scopus: an analysis and visualization of publication activity and research collaboration at the country level. *Quality and Quantity*, 1-31 (article in press).

Liu, Q., Ge, Z., & Song, W. (2016). Research Based on Patent Analysis about the Present Status and Development Trends of Unmanned Aerial Vehicle in China. *Open Journal of Social Sciences*, *4*(07), 172-181.

Shiue, Y. C., & Chang, C. C. (2010, May). Forecasting unmanned vehicle technologies: Use of patent map. In *2010 Second International Conference on Computer Research and Development* (pp. 752-755). IEEE.

National Research Council's Bibliometric Methodology and Subfields of a Scientific Discipline

Lawrence Smolinsky[1] and Aaron J Lercher [2]

[1] *smolinsk@math.lsu.edu*
Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (USA)

[2] *alerche1@lsu.edu*
30 Middleton Library, Louisiana State University, Baton Rouge, LA 70803 (USA)

Introduction

The US National Research Council (NRC) is the principle operating arm of the United States National Academies of Science and Engineering. The NRC has periodically evaluated doctorial programs and departments in the US issuing evaluations. This evaluation was last done in 2010 resulting in a report, Assessment of Research Doctorate Programs (National Research Council. 2011).

In the sciences, the NRC reported data based on 20 variables. The NRC writes, "Although there was some variation in the faculty responses, they were generally in agreement that publications and citations were the most important factors in program quality." (National Research Council, 2009, p. 12).

This poster examines the NRC methodology for these two important variables to find which subfields wield greater or lesser influence on a whole discipline than the raw numbers would suggest. We examine the question in terms of publications and citations. A straightforward approach would be to measure the percentage of discipline publications that fall in the subfield and the percentage of citations to the discipline that are credited to the subfield. A second method is follow the NRC methodology in weighting publications and citations by authorships.

Distribution of credit among multiple individual coauthors and articles has been discussed and studied many ways, which Osório (2018) helpfully categorizes by the different counting methods. But that is not what we are studying. We accept the scheme from the NRC as influential and important. We investigate its consequences on Physics.

Unfortunately, the NRC data was not available and does not classify subfields. Instead, we examine the publications of the American Physical Society's printed Physical Review journals for the year 2013. The American Physical Society's journals are important and give broad coverage, but they have no specific relationship to the NRC.

Subfield influence

What is the power or influence of one subfield on a discipline? One might scientometrically examine influence via bibliometrics, grant funding, scientists in the subfield, faculty hiring, faculty surveys, or how it is evaluated. The evaluation methods used by influential agents have broad consequences, e.g., faculty hiring.

Consider two highly cited articles. One has a single author and a second has three authors. These articles have quite different effects. The single authored article results in one scientist who has influence and prestige in his institution and discipline. The article with three coauthors may result in three influential advocates for their subfield and scientists in the subfield. The single authored article results in one advocate. A second effect is the impression of the value or influence of a subfield since these qualities are often associated to citation counts and productivity. Even though the NRC study does not directly address subfield influence it will have that effect.

Publication and citation spaces

We construct credit spaces, which represent the total credit that is to be assigned. We examine how the credit is split among some groups of subfields of physics as reflected in the Physical Review Journals' classification.

We use $\mathcal{P}$ for publication spaces and C for citation spaces. For a set of articles A or $\mathcal{P}_A$, let the number of coauthors of $a \in A$ be c_a and the number of citations to a be v_a. The number of articles in A is then $|\mathcal{P}_A| = \sum_{a\in A} 1$, the space of authorships for A is cw$\mathcal{P}_A$ with

(1) $$|\mathrm{cw}\mathcal{P}_A| = \sum_{a\in A} c_a.$$

The space of citations to the articles in A is C_A with $|C_A| = \sum_{a\in A} v_a$, and space of the coauthor weighted citations to the articles in A is cwC_A with

(2) $$|\mathrm{cw}C_A| = \sum_{a\in A} v_a\, c_a.$$

Each point in $\mathrm{cw}\mathcal{P}_A$ is a pair of an article and a coauthor to the article, and $\mathrm{cw}\mathcal{C}_A$ has one point for each triple of an article, an author of the article, and a citation to the article.

NRC methodology for publication credit assigns each point in $\mathrm{cw}\mathcal{P}$ to an individual scientist. Likewise, citation credit is awarded by assigning each point in $\mathrm{cw}\mathcal{C}$ to an individual scientist. The amount of total credit in the NRC weighting is given by formulas (1) and (2), rather than $|\mathcal{P}_A|$ and $|\mathcal{C}_A|$.

We examined subfields in groups as designated by the Physical Review Journals. These subfield groups are given in Table 1.

Table 1. Physical Review Journals (PR) and their coverage, which reflects the subfields measured.

PR	Coverage
A	Atomic, molecular, & optical physics and quantum information
B	Condensed matter & materials physics
C	Areas of experimental & theoretical nuclear physics
D	Elementary particle physics, field theory, gravitation, & cosmology
E	Statistical, nonlinear, biological, & soft matter physics

Table 2. PR denotes the Physical Review J. AP is the article proportion, $|\mathcal{P}_{\mathrm{PR}\text{-}}|/|\mathcal{P}_{\mathrm{PR}}|$. WAP is the Weighted article proportion is $|\mathrm{cw}\mathcal{P}_{\mathrm{PR}\text{-}}|/|\mathrm{cw}\mathcal{P}_{\mathrm{PR}}|$.The ratio is WAP/AP.

PR	AP	WAP	Ratio
A	0.195	0.054	0.277
B	0.331	0.122	0.369
C	0.077	0.103	1.338
D	0.224	0.680	3.036
E	0.173	0.041	0.237

Table 3. PR denotes the Physical Review J. CP is the citation proportion $|\mathcal{C}_{\mathrm{PR}\text{-}}|/|\mathcal{C}_{\mathrm{PR}}|$. WCP is the Weighted citation proportion is $|\mathrm{cw}\mathcal{C}_{\mathrm{PR}\text{-}}|/|\mathrm{cw}\mathcal{C}_{\mathrm{PR}}|$. The ratio is WCP/CP.

PR	CP	WCP	Ratio
A	0.158	0.038	0.241
B	0.384	0.120	0.313
C	0.079	0.262	3.316
D	0.262	0.555	2.118
E	0.117	0.025	0.214

Tables 2 and 3 show the credit given to each group of subfields according to each method. In the Article Proportion (AP) method, each article is valued equally. In the Weighted Article Proportion (WAP) method, an article is valued by the number of its coauthors. Analogous values for citations are in the Citation Proportion (CP) and Weighted Citation Proportion (WCP) methods.

The ratio shows the proportion of a subfield's actual publications or citations credited to the subfield in the weighted method. For example, condensed matter & materials physics (B) receives about 31% of the citation credit would get from the actual number of citations.

Conclusion

Collaboration has become valued for its own sake and there is evidence that collaboration also impacts traditional bibliometric variables. There is evidence that coauthored papers receive more citations (Onodera and Yoshikane, 2015). However, it is surprising to see how weighting by authorships skews the bibliometric data. The NRC methodology has surprising consequences for evaluation of the relative importance of subfields.

One possible objection is that the NRC did not gather its data for subfield analysis. But the same can be said for bibliometric databases. The SCCI was not begun for evaluating scientists or impact, but is routinely used for these purposes—including by the NRC. SCCI was a tool for the scientific community to find connected research and guide researchers. It was partly inspired by a legal index to court cases. Garfield wrote, "The legal 'citator' system provided a model of how citation index could be organized to function as an effective search tool" (1979, p. 7). One point of this research is the manner that influential bodies (e.g., NRC) awards credit has broad implications.

References

National Research Council. 2009. A Guide to the Methodology of the National Research Council Assessment of Doctorate Programs. Washington, DC: The National Academies Press.

National Research Council. 2011. A Data-based Assessment of Research Doctoral in the United States. Washington, DC: The National Academies Press.

Garfield, E. (1979). *Citation Indexing—Its Theory and Application in Science, Technology, and Humanities*. New York: Wiley & Sons.

Onodera, N., Yoshikane, F., 2015. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4), 739-764.

Osório, A. (2018). On the impossibility of a perfect counting method to allocate the credits of multi-authored publications, *Scientometrics*, 116(3), 2161–2173.

Sleeping Beauties in Mathematical Research

Samuel Hansen[1]

[1] *hansensm@umich.edu*
University of Michigan, Shapiro Science Library, Mathematics & Statistics Librarian, 919 S University Ave, Ann Arbor, MI 48109

Introduction

Mathematics is often called the language of science, but there is much which separate the two. For example there is the very nature of mathematical knowledge. Unlike scientific results, once a mathematician proves a theorem it is true for the rest of mathematics. Then there is the axiomatic non-empirical disposition of mathematics and its focus on describing a realm of pure abstraction versus science's focus on describing the real world.

Given these clear differences it is surprising that during scientometric study mathematics is often lumped together with science rather than being treated as its own entity. This poster paper represents a set of initial research results from a comprehensive bibliometrics study trying to fix this oversight. Specifically this poster paper will present an initial analysis of mathematical Sleeping Beauties, or research receives a spike of citations after years of relatively few.

Research Aging and Sleeping Beauties

Research aging, primarily through the study of references and citations ages, is an active area of bibliometrics research and has been for many decades (Anker, 1979; Glänzel & Schoepflin, 1995). Thanks to large-scale literature databases and citation indexes scientometric researchers have recently been able to do comprehensive aging studies of both references and citations (Zhang & Glänzel, 2017a; Zhang & Glänzel, 2017b). These large-scale databases have also allowed researchers to develop and formalize the idea of Sleeping Beauties (SBs).

The idea of SBs was first put forth by van Raan (2004) who identified a collection of papers which received fewer than two citations for years and then began to receive large numbers of citations. This work was expanded by Redner (2005) who conducted the first in-depth search of SBs, in the area of Physics. Both of these works were limited by arbitrary definitions of SBs, which drove Ke, Ferrara, Radicchi, & Flammini (2015a) to develop the "Beauty Coefficient" to provide a measure for how deeply research slept. They calculated the Beauty Coefficient for papers in Clarivate's Web of Science database and defined SBs as those papers with Beauty Coefficients in the top 0.1% of all research. This provided them with SB Beauty Coefficient thresholds in different disciplines, for mathematics it was 90.62(Ke et. al., 2015b, p. 19).

Methodology

The data used to conduct this study is from Clarivate's Web of Science citation database, 1900-2017. The Big Ten Academic Alliance has an agreement with Clarivate where they have provided the contents of their database in the form of XML documents with a license allowing for academic research by members of Big Ten academic institutions. These documents were parsed into a PostgreSQL database based off of a data model and python scripts created by the University of Indiana (Indiana University Network Science Institute, n.d.; Light, Halsey, & Herr, n.d.). The scripts used to extract, process, and calculate the Beauty Coefficient for the mathematical citation data are available via the University of Michigan Gitlabs instance (Hansen, n.d.).

Mathematical Sleeping Beauties

There are three mathematics subjects used by Web of Science to classify its contents, Mathematics, Mathematics, Applied, and Mathematics, Interdisciplinary Applications. These subjects are not applied in a mutually exclusive manner, and can overlap. These subjects are applied to 1,343,970 entries of all document types in the Web of Science database. Over half were considered Mathematics, with around 45% Applied and 15% Interdisciplinary Applications. There were 3847 cases of SBs, e.g. Beauty Coefficients higher than 90.62. Mathematics was most likely to generate a SB, with a rate around three times the other subjects.

Table 1. Counts of Mathematics Research and Sleeping Beauties in Web of Science by subject.

Table	Total	SBs
Mathematics	742541	3044
Applied	611160	743
Interdisciplinary Applications	199652	324
Total	1343970	3847

The ratio stays nearly the same when only mathematical research which has received more than 100 citations is considered. Interestingly this is true even though research classified as Mathematics is less likely than the other subjects to reach 100 citations, less than half as likely as Interdisciplinary

Applications. Since the number of citations, specifically the peak, plays a major role in the Beauty Coefficient this implies that the less applied or interdisciplinary a highly cited mathematics paper is the more likely it is go without citation for an extended period.

Table 1. Counts of Mathematics Research with citation counts of at least 100 and Sleeping Beauties in Web of Science by subject.

Table	Total	SBs
Mathematics	6485	938
Applied	6635	342
Interdisciplinary Applications	3995	174
Total	15745	1354

This analysis also identified a new SB (citation peak of 2017) with one of the highest known Beauty Coefficients, Clive Granger's "Investigating Causal Relations by Econometric Models and Cross-spectral Methods" (1969). With a beauty coefficient of just over 6737 it is only behind two of the SBs found by Ke et. al. (2015a, p. 7429). Related to their findings of a relationship between SBs and interdisciplinarity this work was assigned Web of Science subjects of Economics, Mathematics, Interdisciplinary Applications, Social Sciences, Mathematical Methods, and Statistics & Probability.

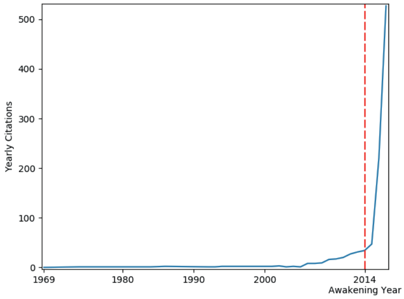


Figure 1. Citation History for Investigating Causal Relations by Econometric Models and Cross-spectral Methods with a dotted line indicating its awakening year.

Continuing Work

This analysis only represents the first step in ongoing research into the bibliometrics of mathematics research. This research will include thorough analyses of mathematical reference and citation aging, bibliographic coupling, and sub-discipline networks.

This work has also indicated there is still work to be done with regard to identifying different forms of SBs. The Beauty Coefficient is a useful measure but it cannot identify research which has an initial burst of citations, and then falls asleep, also known as all-elements-sleeping-beauties (Li, 2014), or papers which awake and fall asleep multiple times. A measure which could identify such SBs, perhaps related to peak analysis, could open up a new range of potential analyses, not only in mathematics, but for scientometrics as a whole.

References

Anker, A. L., Servi, P. N., Griffith, B. C., & Carl Drott, M. (1979). The aging of scientific literature: a citation analysis. Journal of Documentation, 35(3), 179–196.

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. Journal of Information Science, 21(1), 37–53.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica, 37(3), 424–438.

Hansen, S. (n.d.). Citation Aging Gitlabs Repository. Retrieved April 5, 2019, from GitLab website: https://gitlab.eecs.umich.edu/hansensm/citationaging

Indiana University Network Science Institute. (n.d.). Web of Science (WoS). Retrieved April 5, 2019, from Indiana University Network Science Institute website: http://iuni.iu.edu/resources/web-of-science

Light, R., Halsey, D., & Herr, B. (n.d.). Generic Parser Github Repository. Retrieved April 5, 2019, from https://github.com/cns-iu/generic_-parser

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015a). Defining and identifying Sleeping Beauties in science. Proceedings of the National Academy of Sciences, 112(24), 7426–7431.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015b). Defining and identifying Sleeping Beauties in science. Proceedings of the National Academy of Sciences, 112(24), 7426–7431.

Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition." Scientometrics, 100(2), 595–601.

Redner, S. (2005). Citation Statistics from 110 Years of Physical Review. Physics Today, 58(6), 49–54.

van Raan, A. F. J. (2004). Sleeping Beauties in science. Scientometrics, 59(3), 467–472.

Zhang, L., & Glänzel, W. (2017a). A citation-based cross-disciplinary study on literature ageing: part II—diachronous aspects. Scientometrics, 111(3), 1559–1572.

Zhang, L., & Glänzel, W. (2017b). A citation-based cross-disciplinary study on literature aging: part I—the synchronous approach. Scientometrics, 111(3), 1573–1589.

Research leadership flows and the role of proximity

Chaocheng He[1] and Jiang Wu[2]

2016201040025@whu.edu.cn
Wuhan University, School of Information Management, Wuhan, Hubei 430072 (China)
Jiangw@whu.edu.cn
Wuhan University, School of Information Management, Wuhan, Hubei 430072 (China)

Introduction

The increase of interdisciplinary communication, the acceleration of innovation and the growing complex of the research project, all work together to make it difficult for a single scientist or organization to complete a big science project. Therefore, steadily increasing research collaboration has become a trend, which motivate researchers to explore which measurable factors will promote research collaboration. Geographic and socioeconomic factors are the most common factors (Hoekman, Frenken et al. 2010).

The previous literature on proximity and research collaboration have certain limitations. Firstly, they consider all collaborations equally. However, the collaborative relationship with the first author and corresponding author can better reveal research collaboration because the first author and corresponding author often dominate and lead the research collaboration. Secondly, although Boschma (2005) identified five notions of proximity, prior studies have failed to systematically examine the relationships between these factors and research collaboration. Thirdly, few previous studies go deep into institutions level. It is microscopic institutions that have the primary mission of knowledge creation and diffusion.

Methods and Data

Research leadership

Research collaboration is a complex system where the first author and corresponding author often dominate and lead the research collaboration. For example, in biomedical field, the first author is assigned to carry out the research and write a paper. The other participants with more specialized roles sign as co-author. The corresponding author is responsible for both scientific and non-scientific contribution. Prior studies also understand the research group to which the corresponding author belongs as the research leader. Because the research group is the base unit. Furthermore, the corresponding authorship is highly valued in China. And in most cases for Chinese publications, first author and corresponding author belong to the same organization. Therefore, the institution to which the corresponding author belongs to is used in our study to be the research leader.

Measurement of research leadership

The prior studies mainly adopt the "full count" and the "fractional count" to measure research collaboration intensity (Berge 2017). Here, we make use of "fractional count", since it relates to the idea of contribution to knowledge production, rather than simply participating. We assume that a particular paper requires leadership mass 1, from the leading institution to all other participating institutions. The RL flow intensity $C_{ab,i}$ from institution a to institution b in paper i is expressed as

$$C_{ab,i} = \frac{1}{innum_i} \tag{1}$$

where $innum_i$ is the number of institutions in paper i. Thus, the RL mass $LM_{a,i}$ that leading institution a obtains from the paper i is

$$LM_{a,i} = \sum_{b=1}^{n_i-1} C_{ab,i} = \frac{n_i-1}{n_i} = 1 - \frac{1}{n_i} \tag{2}$$

where n_i represent the number of institutions in paper i. Here we don't take self-leading into consideration, so we sum up to n_i -1. And the RL flow intensity $C_{ab,i}$ from institution a to institution b in paper i is expressed as

$$C_{ab,i} = \frac{1}{corresnum_i} \times \frac{1}{instnum_i} \tag{3}$$

where $corresnum_i$ is the number of leading institutions in paper i. Therefore, the total RL flow intensity C_{ab} from institution a to institution b is calculated as

$$C_{ab} = \sum_{i=1}^{m_\mathrm{b}} C_{ab,i} \tag{4}$$

where m_b is the number of papers where a is the leading institution and b is a participating institution. And institution a's total RL mass is calculated as

$$LM_a = \sum_{b=1}^{B} C_{ab} \tag{5}$$

Data

We perform a data collection in Thomson Reuters's WoS Core Citation Database according to this search term "CU = A AND SU = B AND PY = C", where A is "PEOPLES R CHINA", B is research areas in "Life Sciences & Biomedicine" field, and C is 2013-2017. We focus on Chinese institutions because the corresponding authorship is highly valued in China. We focus on "Life Sciences & Biomedicine" because these fields require more on complex teamwork and the dominant role of the corresponding author is more pronounced. To avoid noise, we filter the institutions with positive RL mass in every year and finally obtain 244 institutions.

Model and variables

We adopt a gravity model to analyze the determinants of RL among different institutions. The basic idea of the gravity model stems from Newton's law of universal gravitation. Given the fractional count nature of our data and a large number of zeros (many institution pairs have no research collaboration), in line with previous studies (Plotnikova and Rake 2014), we adopt a Tobit regression model where we consider zero leadership as left censoring of the distribution. To explore the role and its dynamic evolution of proximity in shaping RL flows, we first conduct a cross-section estimate by the pooling data of 2013-2017, and then we perform cross-section estimates using two sub-period data. In addition, time lags are used to avoid endogeneity and reverse causality. The LM_i and LM_j refer to the period 2008-2012. The explanatory variables are lagged and capture information for the period 2008-2012 too. The equation to estimate is:

$$C_{ij} = \beta_0 + \beta_1 ln(LM_i) + \beta_2 ln(LM_j) + \beta_3 ln(gp_{ij}) + \beta_4 ln(cp_{ij}) + \beta_5 ip_{ij} + \beta_6 sp_{ij} + \beta_7 ln(ep_{ij}) \qquad (6)$$

Table1 description of variables

V	Description	Source
C_{ij}	RL flow intensity from institution *i* to *j* in the period 2013-2017	Web of Science
LM_i	RL mass of institution *i* in period 2008-2012 (Variable in logarithms)	Web of Science
LM_j	RL mass of institution *j* in period 2008-2012 (Variable in logarithms)	Web of Science
gp_{ij}	Geographical distance between institution *i* and *j*, in kilometers	Google Map
cp_{ij}	Cosine similarity between institution vector pairs in period 2008-2012	Web of Science
ip_{ij}	Dummy variable, which take value 1 when institution *i* and *j* are in the same province	Google Map
sp_{ij}	Dummy variable which takes value 1 if institution *i* and *j* have collaborated in period 2008-2012	Web of Science
ep_{ij}	Absolute difference in GDP per capital between the cities of institute *i* and *j* in period 2008-2012	NBS, China

As is shown in Table2, each variable's variance inflation factor (VIF) is lower than 4, indicating there is no significant multicollinearity in this study. Table 3 reports the estimation results of Tobit gravity model, Model (1) is a cross-section estimate by the pooling data of 2013-2017. Model (2) and (3) are estimation result of two sub-period data with two-year time-lag. Furthermore, we adopt the Chow test to determine the independent variables have significant differences in time series analysis (p=0.000). From Table 3, we can conclude that the RL mass of leading institution, the RL mass of participating institution, geographical, cognitive, institutional and social proximity are important factors that affect RL flows. These results remain robust to sensitive check of different sub-periods. In particular, leading institution's RL mass has a higher influence than participating institution's one. Leading institution's influence is decreasing, meanwhile participating institution's influence is growing. Geographical proximity, social proximity, and institutional proximity still have significant influence on the RL flows. However, their influence is decreasing. Notably, the influence of economic proximity in RL diffusion is getting smaller and even does not affect RL flows. On the other hand, the effect of cognitive proximity has an increasing trend and is more and more important.

Table 2 Descriptive statistics and correlations

	VIF	1	2	3	4	5	6	7	8
C_{ij}	-	1							
$ln(LM_i)$	3.13	.18	1						
$ln(LM_j)$	2.99	.15	-.00	1					
$ln(gp_{ij})$	2.58	-.17	-.04	-.04	1				
$ln(cp_{ij})$	1.63	.15	.43	.43	.03	1			
ip_{ij}	1.45	.18	.02	.02	-.74	-.04	1		
sp_{ij}	1.40	.27	.36	.30	-.18	.33	.18	1	
$ln(ep_{ij})$	1.34	-.17	-.06	-.06	.73	.01	-.78	-.16	1

Table 3 Estimation results

	2013-2017 Model (1)	**2013-2014 Model (2)**	**2016-2017 Model (3)**
$ln(LM_i)$	4.19^{***}	2.24^{***}	2.17^{***}
$ln(LM_j)$	3.12^{***}	1.67^{***}	1.83^{***}
$ln(gp_{ij})$	-1.75^{***}	-0.92^{***}	-0.83^{***}
$ln(cp_{ij})$	182.95^{***}	82.71^{***}	84.32^{***}
ip_{ij}	14.34^{***}	5.71^{***}	5.41^{***}
sp_{ij}	16.87^{***}	14.68^{***}	12.05^{***}
$ln(ep_{ij})$	0.38^{***}	0.14^{**}	0.48
_cons	-43.50^{**}	-25.89^{***}	-25.530^{***}
Chow Test	-	39.62^{***}	

Institutions should be encouraged to improve their own RL mass. The national government should pay more attention to transportation and information infrastructure establishment and encourage institutions to cooperate with partners of partners. Province governments should try to stay in keeping with others to make research policies.

References

Berge, L. R. (2017). Network proximity in the geography of research collaboration. *Papers in Regional Science* 96(4): 785-815.

Boschma, A. (2005). Proximity and innovation: A critical assessment. *Regional Studies* 39(1): 61-74.

Hoekman, J., K. Frenken and R. J. Tijssen (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy* 39(5): 662-673.

Plotnikova, T. & B. Rake (2014). Collaboration in pharmaceutical research: exploration of country-level determinants. *Scientometrics* 98(2): 1173-1202.

When gender doesn’t matter: the relationship between university’s presidencies and their research performance

Yuehua Zhao[1], Wen Lou [2*] and Ruofan Pi [2]

[1] *yuehua@nju.edu.cn*
School of information management, Nanjing University, Nanjing Jiangsu (China)

[2] *wlou@infor.ecnu.edu.cn,   alessia_pee@foxmail.com*
Department of information management, East China Normal University, Shanghai (China)

Introduction

Gender differences have always been of ongoing global concern. Even though the research performance gap between the two groups has tended to be smaller in younger generations (Arensbergen, Weijden, & Besselaar, 2012), the gender inequality remains stable world-wide (Paul-Hus, et al., 2015). In the meantime, research on university leaders have been conducted across many countries (Goodall, 2015). Our previous research (Lou, Zhao, Chen, & Zhang, 2018) proved that university leaders’ research performance could be influenced by administrative services in terms of the leadership experiences, leaders’ academic expertise, and their schools’ rankings (see Figure 1). In this paper, we examine the gender factor in relation to research performance and leadership role.

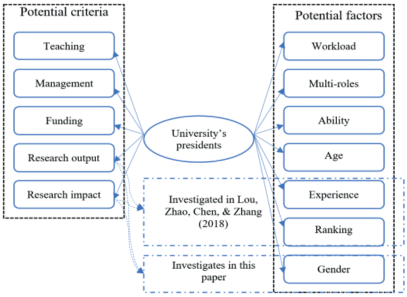


Figure 1. Research intention and implications.

Method

The profile information of the presidents of the top 500 universities from U.S. News Best Global Universities Rankings were collected in November 2018. We manually collected all publications and citations of 411 presidents from Web of Science. 58 presidents without publications in Web of Science and 31 presidents who were freshly established in 2018 were excluded.

The key to measure research performance differences is to compare a president’s research performance before and after taking the leadership role. We applied the four periods proposition in our previous study (Lou, Zhao, Chen, & Zhang, 2018) to distinguish the certain time before and after the service. We define the period “After” as the in-position period and “Before” as the reference period only if the latest period is longer than the in-position period. We define “After” as the reference period and “Before” as the latest period only if the latest period is shorter than the in-position period. Therefore, the following hypotheses are tested by statistical methods:

H1: There is no significant difference in the number of publications between the period “Before” and “After”.
H1a: There is no significant difference between female and male leaders in terms of the number of publications in the period “Before” and “After”.
H2: There is no significant difference in the number of citations between the period “Before” and “After”.
H2a: There is no significant difference between female and male leaders in terms of the number of citations in the period “Before” and “After”.

Findings

In total, 68 females and 343 males out of 411 university leaders with publications were included in this research. Total publications dropped 20% on average after taking the presidency and citations dropped massively as well (60% of Before times). As for different genders, the decrease percentage for each gender tended to be in line with the overall decrease percentage. Both male and female presidents faced minor loss to their research output.

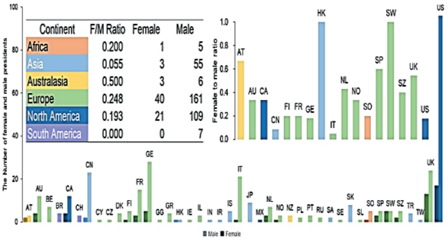


Figure 2. The number/ratio of female and male presidents on continent/country level.

In total, 411 presidents currently are from six continents and 42 countries and districts. In Figure 2, the global gender disparity is obvious. Male presidents outnumber female presidents in every continent and every country, except that Hong Kong and Sweden demonstrate less disparity with 1.0 ratios of female to male.

The relationship between research performance and the presidency

Figure 3 doesn't show clear differences in publication between before and after taking the job. However, the decline is more obvious on the citation side. We make use of Wilcoxon signed-rank test to examine the hypotheses H1 and H2. The results suggest that the presidency has a significant relationship with presidents' research works in relation to both research productivity and impact.

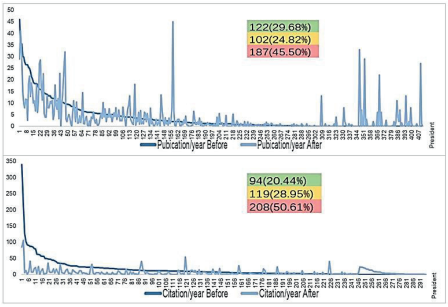


Figure 3. Distribution of annual publications and citations in the period Before and After.

Comparison between genders

Figure 4 shows the four trends of publications and citations for female and male presidents. It is not clear if there are differences in publication between before andafter taking the job. However, the decline is more obvious on the citation side.

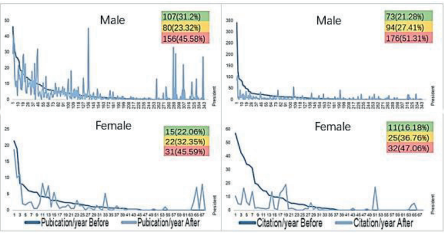


Figure 4. Distribution of annual publications and citations divided by genders in the period Before and After.

As for publication, the same proportions (approximately half) of female and male presidents experience a decrease of their research output. Surprisingly 31.2% of male presidents published more papers after taking the job than before. Female presidents were not as fortunate as male presidents. As for citations, more than half of the male presidents could not avoid the output decrease of presidency. The impact scale of female presidents is not as large as that of male presidents on the citation side. Mann–Whitney U tests were performed on H1a and H2a and results showed no significant differences in the gendered impact on research productivity nor on research impact.

Discussion and primary conclusion

Hypotheses H1 and H2 were rejected, and H1a and H2a failed to be rejected. The relationship between the presidency and research performance is critical. However, the gendered impact does not show much disparity.

There is no doubt that the decrease of research performance has a significant relationship with the presidency regardless of gender. The impact is more severe on research impact than on output. This could be explained by co-author teamwork bias. In our investigations of the publication details of the sampled presidents, we observed that, as time went by, presidents tended to co-author papers instead of first-author or single-author papers. Even though teamwork could keep up with the amount of research output, the quality of co-authored papers can be complex.

We confirmed gender disparity in science, even in academic leadership. Male presidents not only dominate research publications and citations but also outnumber female presidents in every continent and nearly every country. Gender disparity in terms of absolute and fractional numbers varies among regions. Yet gender disparity in the impact of presidency does not appear to be evident. In other words, such impact makes no difference between men and women. The findings drawn from this study are limited by the population sampled and may not be applicable to all academic settings. Further research will examine more factors to see whether the impact of gender differentiates within disciplines, nations, university levels, and individuals' experiences.

Acknowledgement

This research is supported by "Chenguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission [17CG19] and China Postdoctoral Science Foundation [2017M620142].

References

Arensbergen, P., Weijden, I., & Besselaar, P. (2012). Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, 93(3), 857-868.

Goodall, A. (2015). Universities and leaders: a causal link. *International higher education*, 45, 20–21.

Lou, W., Zhao, Y., Chen, Y., & Zhang, J. (2018). Research or management? An investigation of the impact of leadership roles on the research performance of academic administrators. *Scientometrics*, 117(1), 191–209.

Paul-Hus, A., Bouvier, R., Ni, C., Sugimoto, C. R., Pislyakov, V., & Larivière, V. (2015). Forty years of gender disparities in Russian science: a historical bibliometric analysis. *Scientometrics*, 102(2), 1541-1553.

Comparison of Social Science Papers and Books Based on Citation and Altmetric Indicators

Yang Siluo[1] and Yu Yonghao[2]

[1] *58605025@qq.com*
School of Information Management, Wuhan University, Wuhan 430072 (China)

[2] *1300016639@pku.edu.cn*
School of Information Management, Wuhan University, Wuhan 430072 (China)

Introduction

The diverse publication channels used by scholars significantly explain the difficulties of applying bibliometric methods to certain fields (Hammarfelt, 2014), especially in the social sciences where publication channels are diverse. Furthermore, single citation indicators are insufficient for a comparative evaluation of papers and books and for an illustration of their different bibliometric characteristics. The Book Citation Index (BKCI) introduced by Thomson-Reuters and altmetric.com which serves researchers with altmetrics data provide new indicators and perspectives to bibliometric researchers (Robinson-Garcia, Torres-Salinas & Zahedi, 2014; Zuccala, Verlevsen & Comacchia, 2015). Other previous studies (Bornmann, 2014; Hammarfelt, 2014) suggest that the altmetric indicators are valid for papers in specific disciplines. In the present study, we compared publications from SSCI and BKCI-SSH by using citation and altmetrics data to explore the overall situation in social science.

Method and Materials

Searching on the SSCI and BKCI-SSH, the search string "PY=(2013-2017) AND DT=(PROCEEDINGS PAPER OR ARTICLE OR REVIEW)" and "PY=(2013-2017)" provided us with a recall of more than 2,000,000 papers and 408,360 books(books and book chapters) on January 8, 2019. We subsequently downloaded these materials and extracted the digital object unique indicator (DOI) and citation indicator of all records for the following analyses.

Filtering of Publications for the Study

The altmetric indicators of these documents were acquired from the Altmetric.com platform by API-program with DOI, which recalled 897,302 paper records and 8,608 book records. The analyses and discussions of citation and altmetrics are based on both consistent datasets.

The Altmetric Attention Score is defined by altmetrics.com as "an automatically calculated, weighted count of all of the attention a research output has received". The attention score represents a weighted approximation of all the attention that altmetric.com picked up for a research output (not a raw total of the number of mentions).

Results

Time Series

Figure 1 shows the time series of the citation and altmetric score mean values of papers and books. Specifically, the figure indicates that papers are cited far more than books on average, whereas both curves of citation mean values present a significant decreasing trend. On the contrary, curves of altmetric score mean values ascend even though they are not as skewed as the citation mean curves. Altmetric data are the most frequently used in most recent publications, and they are valid for the most recent publication years, which are also suggested as "recent bias" in previous studies (Costas, Zahedi & Wouters, 2015). The altmetric scores of books share the same characteristic as those of papers.

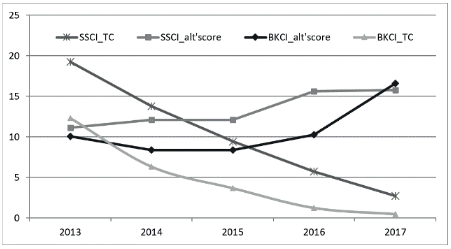


Figure 1. Citation mean and altmetric score mean values of papers and books in terms of publication years

Cumulative Distribution

The citation and altmetric score cumulative distribution of papers (from SSCI) and books (from BKCI) is illustrated in a log-log format in Figure 2, where Y-axis indicates the values of "citation"/"altmetric score," whereas X-axis indicates the cumulative number of documents whose "citation"/"altmetric score" value exceeds the corresponding Y-value. Figure 2 also displays that 5,169 papers and 16 books are cited 100 or more times. This high citation score may be considered an effect of the specific distribution over the disciplines, which future researchers can focus on (Leydesdorff & Felt, 2012). Notably, 108,645 papers and 4,615 books are never cited in these databases. The citation and altmetric score values of papers are higher than both of books, but the trend lines exhibit a similar distribution for both variables. Therefore, we suggest that papers and books hold similar bibliometric and altmetric characteristics.

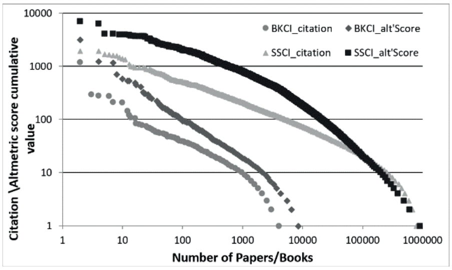
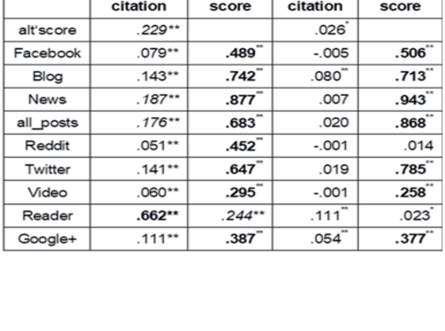


Figure 2. Citation and altmetric score cumulative distribution of 897,302 papers from SSCI and 8,608 books from BKCI on a log-log scale

Correlations

For comparative purpose, Spearman correlations among the citation and altmetric indicators of SSCI (papers) and BKCI (books) are exhibited in Table 1. The citation indicators of papers significantly correlate with certain altmetric indicators, especially altmetric score and Reader. However, the citation indicators of books show weak correlations with most altmetric indicators. Therefore, altmetric indicators can be applied to support the result of papers' citation analysis and forecast high-cited papers, but these indicators are not valid for books. Specific altmetric indicators of books are appropriate for a comprehensive analysis of their citation performance. Altmetric indicators determine altmetric score as the weighted approximation of other variables, thus it evidently correlates with most variables. This assumption is also the reason why altmetric score was used to represent other altmetric indicators in previous analyses.

Table 1. Spearman correlation analysis of citation and altmetric variables



	SSCI-citation	SSCI-alt' score	BKCI-citation	BKCI-alt' score
alt'score	.229**		.026*	
Facebook	.079**	**.489****	-.005	**.506****
Blog	.143**	**.742****	.080**	**.713****
News	.187**	**.877****	.007	**.943****
all_posts	.176**	**.683****	.020	**.868****
Reddit	.051**	**.452****	-.001	.014
Twitter	.141**	**.647****	.019	**.785****
Video	.060**	**.295****	-.001	**.258****
Reader	.662**	.244**	.111**	.023*
Google+	.111**	**.387****	.054**	**.377****

Conclusion

In terms of citation and altmetric indicators, papers perform better than books. Papers are cited far more than books on average, although papers exhibit a similar downtrend in terms of publication years. A similar result occurs in altmetric score indicators with an uptrend. Compared to the "hysteresis" of citation indicators, altmetric indicators display strong "recent bias," which means altmetric indicators are valid when assessing the most recent publications instead of old publications (Costas, Zahedi & Wouters, 2015). Nevertheless, papers or books have a similar cumulative distribution of citation and altmetric scores. Moreover, the citation score of papers has a significant correlation with altmetric score, different from the weak correlation between the citation and altmetric scores of books. Therefore, we suggest that altmetric indicators can represent a valid complement to citation for evaluating papers and books. Altmetric indicators may also be applied to predict high-cited papers. Further research must be conducted to determine the effect of the specific distribution of indicators on various disciplines.

Acknowledgments

This research is funded by the National Social Science Fund Key Project of People's Republic Of China (17ATQ009).

References

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. Journal of informetrics, 8(4), 895-903.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology, 66(10), 2003-2019.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. Scientometrics, 101(2), 1419-1430.

Leydesdorff, L., & Felt, U. (2012). "Books" and "book chapters" in the book citation index (BKCI) and science citation index (SCI, SSCI, A&HCI). Proceedings of the American Society for Information Science and Technology, 49(1), 1-7.

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: exploring the insides of Altmetric. com. arXiv preprint arXiv:1408.0135.

Zuccala, A. A., Verleysen, F. T., Cornacchia, R., & Engels, T. C. (2015). Altmetrics for the humanities: Comparing Goodreads reader ratings with citations to history books. Aslib Journal of Information Management, 67(3), 320-336.

Analysis of SSH Impact based on Citations and Altmetrics

Siluo Yang[1] and Mengxue Zheng[2]

[1] *58605025@qq.com* [2] *1984367517@qq.com*
School of Information Management, Wuhan University, Wuhan 430072, China

Introduction

Citation analysis remains a primary method of research evaluation, but such an analysis does not adequately reflect knowledge dissemination (Chen et al., 2015). Publication channels in the Social Sciences and Humanities (SSH) are diverse, thus using bibliometric analysis to evaluate academic performance in SSH is proven more problematic than in Natural Sciences (Nederhof et al., 1988). Bibliometric analysis must adapt to the peculiarities of SSH (Zhou et al., 2009), and the academic community has agreed that apart from the number of publications and citations, multi-source and multi-dimensional indicators should be involved in the evaluation (Moed & Halevi, 2015).

Altmetrics, which was first proposed in 2010 (Priem et al., 2010), has been regarded as the new possibility of impact measurement in the new social media environment. In this study, we evaluate and compare the impact of SSH publications by using citations and altmetrics. Compared with previous research, this study uses updated and large-scale data to reflect the current overall situation. On the basis of the particularity of SSH, we evaluate the performance from the perspective of the subject field. These topics cover the following:

(1) A comparative citation analysis of SSH publications based on selected fields in 2013–2017;

(2) A comparative altmetric analysis of SSH publications based on selected fields in 2013–2017;

(3) The correlations between citation and altmetric indexes.

Data and Methods

As illustrated in Figure 1, we downloaded a data set of Social Sciences Citation Index (SSCI) and the Arts & Humanities Citation Index (A&HCI) from Web of Science in January 2019 by using the search string "PY=2013-2017 AND DT=(ARTICLE OR REVIEW OR PROCEEDINGS PAPER)." We obtained a set of 1,327,924 records, which include three publication types in the period of five years. Subsequently, we obtained the altmetric data from Altmetric.com on the basis of the digital object unique identifier (DOI) search using API, which recalled 629,586 records.

We matched the data to four main fields comprising *Economics & Business Administration* (E&B), *Social, Political & Communication Science* (S, P&C), *Psychology,* and *Humanities & Art* (H&A) with 67 subject categories. This selection is based on the Global Institutional Profile Project subject map and the experience of bibliometric application of SSH (Zhou et al., 2008).

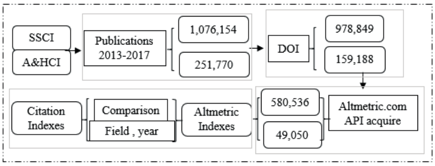


Figure 1. Data processing

Thirteen altmetric indicators obtained through Altmetric.com and three citation indicators are calculated for all selected publications. The three citation indicators are the following:

i) **Mean Observed Citation Rate** (MOCR). The ratio of citation count to publication count.

ii) **Field-Expected Citation Rate** (FECR). One expected citation rate of the corresponding fields is expressed and calculated in the same manner as mean expected citation rate (Glanzel et al., 2008).

iii) **Normalized Mean Citation Rate** (NMCR). NMCR = MOCR/FECR, and the result of which is a relative citation rate.

Results

Citation impact

Table 1 presents the NMCR values of the selected fields and FECR in the period of 2013–2017. First, the FECR value expectedly declines over time. Second, *Psychology* and E&B are the two fields with consistent relative citation impacts higher than the neutral value of 1.0. Moreover, *Psychology* ranks first according to the NMCR.

Table 1. NMCR of the selected fields and FECR in different years

Year	E&B	S, P&C	Psych-ology	H&A	FECR
2013	1.17	0.98	1.32	0.30	12.23
2014	1.14	0.97	1.34	0.29	9.18
2015	1.15	0.99	1.32	0.28	6.39
2016	1.15	0.98	1.33	0.27	3.87
2017	1.12	1.00	1.32	0.29	1.81
All	1.16	0.98	1.32	0.29	6.48

Figure 2 displays the distribution of the MOCR of the top 100 publications in different fields in the period 2013–2017, which does not change sharply over time. *Psychology* has the largest proportion, whereas H&A shares the least proportion.

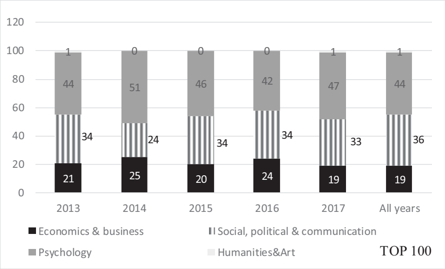


Figure 2. Distribution of the MOCR of the top 100 publications in different disciplines in SSH

Altmetric impact

Descriptive statistics including coverage rate, mean, maximum, and standard deviation are used in this study. The coverage rates are low, except for "readers_count" and "cited_by_tweeters_count".

Altmetric score, which means "a weighted count of all of the attention a research output has received," is compared with MOCR in each field. Figure 3 indicates that only publications with altmetric data are selected, suggesting that the presence of altmetrics is increasing over time, but citations are decreasing. With regard to the altmetric score, *Psychology* consistently acquires the most attention in SSH, whereas S, P&C comes second, and H&A places last.

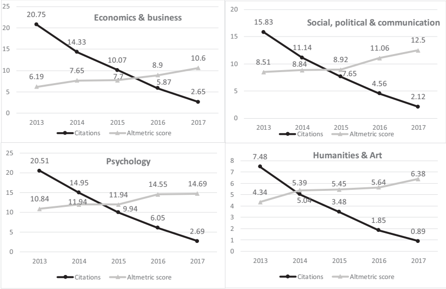


Figure 3. Mean altmetric score versus MOCR in different disciplines in SSH

Correlations between citation and altmetric indexes

Table 2. Spearman correlation analysis of citation and altmetric indexes

fbwalls	feeds	gplus	msm	posts
–.112**	.007**	.012**	.082**	.154**
rdts	**tweeters**	**videos**	**score**	**readers**
.002	.095**	.040**	.190**	.687**

**. Correlation is significant at the 0.01 level (two-tailed).

Table 2 exhibits the Spearman's rank correlation coefficient among citation and altmetric indexes in SSH. The coefficient indicates clear but weak correlations between citation and most altmetric indexes. However, the correlation between "readers_count" and citation appears especially strong, whereas that between "altmetric score" and citation appears relatively strong.

Conclusions

Disciplinary impact. Our results confirm that different disciplines in SSH perform differently in terms of citations and altmetrics. For example, *Psychology* consistently occupies the first place in each impact evaluation, whereas H&A constantly ranks last. Moreover, our findings are consistent with those of citation and altmetric analyses.

Citation analysis. The FECR value of each field, which equals to the mean citation of all fields, and the MOCR of each field sharply drop in recent published years, suggesting the need for a citation window.

Altmetric analysis. The presence of altmetrics increases over time, whereas recent publications exhibit better performance than old publications.

Correlations between citation and altmetric indexes. Clear but weak correlations exist between citation and most altmetric indexes, thus implying that altmetrics is not a possible alternative to traditional citation analysis, but a complement to citations.

Acknowledgments

We thank Stacy Konkiel from Altmetric.com for the data support. This research is funded by the National Social Science Fund Key Project of People's Republic of China (17ATQ009).

References

Chen, K., Tang, M., Wang, C. & Hsiang, J. (2015). Exploring alternative metrics of scholarly performance in the social sciences and humanities in Taiwan. *Scientometrics*, 102, 97-112.

Glanzel, W., Debackere, K., Meyer, M. (2008), 'Triad' or 'Tetrad'? On global changes in a dynamic world, Scientometrics, 74 (1), 59–76.

Moed, H. F., & Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology,* 66(10), 1988-2002.

Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., & Dekker, P. J. (1988). Accessing the useful of bibliometric indicators for the humanities and the social and behavioral sciences: A comparative study. *Scientometrics*, 15(5–6), 423–435.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010b). Altmetrics: A manifesto. Retrieved from: http://altmetrics.org/manifesto/.

Zhou, P., Thijs, B., & Glänzel, W. (2008). Is China also becoming a giant in social sciences? *Scientometrics,* 79(3), 593-621.

Exploring Linguistic Characteristics of Highly Browsed and Downloaded Academic Articles

Bikun Chen[1], Dannan Deng[2], Zhouyan Zhong[3], Chao Ye[4] and Chengzhi Zhang[5]

[1]*chenbikun@njust.edu.cn*
Nanjing University of Science and Technology, Nanjing (China)

[2] *zabel.deng@qq.com*
Nanjing University of Science and Technology, Nanjing (China)

[3]*594096489@qq.com*
Nanjing University of Science and Technology, Nanjing (China)

[4]*785771017@qq.com*
Nanjing University of Science and Technology, Nanjing (China)

[5]*zhangcz@njust.edu.cn*
Nanjing University of Science and Technology, Nanjing (China)

Introduction

Usage metrics of academic articles have become increasingly popular in scientometrics. But most researches focus on numerical analysis. Only a few researches analyse textual contents jointly with usage metrics (Chen 2018).

Usage metrics are required to be studied in a broader vision. The increasing availability of full text from scientific articles in machine readable electronic formats is an opportunity to greatly impact scientometrics. In-text citations (Boyack et al. 2018), entity metrics (Pan et al. 2018) and scientific writing styles (Lu et al. 2018) are the typical examples of full text analysis in scientometrics. Similarly, it is potential to introduce full text analysis to usage metrics. In the study, linguistic characteristics jointly with usage metrics are investigated: What's the linguistic characteristics (full text length, sentence length, lexical diversity, lexical density, et al.) of highly browsed and downloaded academic articles?

Data

Table 1. Data set.

Journal	# of Publications
PLoS Biology (BIO)	288
PLoS Medicine (MED)	171
PLoS Computational Biology (CBI)	1115
PLoS Neglected Tropical Diseases (NTD)	1372
PLoS Pathogens (PAT)	1181
PLoS One (ONE)	57361
PLoS Genetics (GEN)	1514

The data in this study consist of 63,002 full-text articles (only research articles pre-labeled by PLoS are kept) published from 2014 to 2015 in the PLoS journal family (detailed in Table 1). In PLoS, usage counts along with other metadata are collected between November 1st and November 7th, 2018. The PLoS journals are also indexed by PubMed Central (PMC) and Web of Science (WoS). In PMC and WoS, usage counts along with other metadata are also crawled between November 1st and November 7th, 2018.

Method

Highly browsed and downloaded academic articles in this study are defined by Top 20% papers ranked by HTML views and PDF downloads in PLoS and PMC platforms. In order to comparatively uncover linguistic characteristics of Top 20% papers, total papers and Bottom 20% papers are also incorporated. Indicators measuring linguistic characteristics are mainly selected from *Lu et al.(2018)* (detailed in Table 2). The indicator "Author Number" is inspired by *Chi and Glänzel (2017)*.

Table 2. Indicators measuring linguistic characteristics.

Indicator	Description	Formula
Author Number	Calculating total number of authors in each article	$\mathrm{TA} = \sum_{i=1}^{N} Author$
Full text Length	Calculating total number of words in each article	$\mathrm{TFL} = \sum_{i=1}^{N} Full\ text$
Sentence Length	Calculating average number of words in sentences of each article	$\mathrm{MSL} = \frac{\sum_{i=1}^{N} SL_i}{N}$

Lexical Diversity	Type-Token Ratio in each article	TTR $= \frac{\#\ of\ Distinct\ words}{\#\ of\ Tokens}$

Results

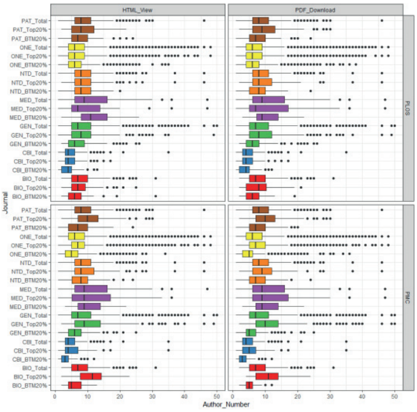


Figure 1. Author number (<= 50) distribution of PLoS and PMC platform

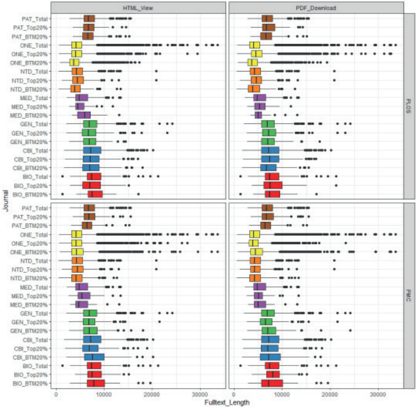


Figure 2. Full text length distribution of PLoS and PMC platform

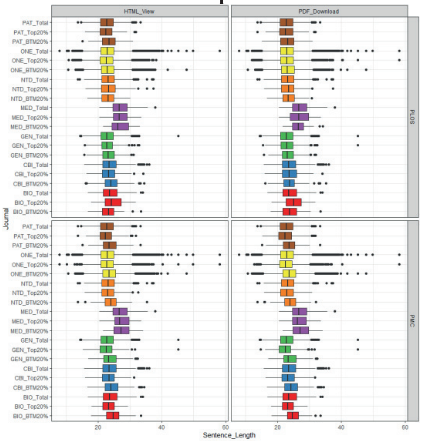


Figure 3. Sentence length distribution of PLoS and PMC platform

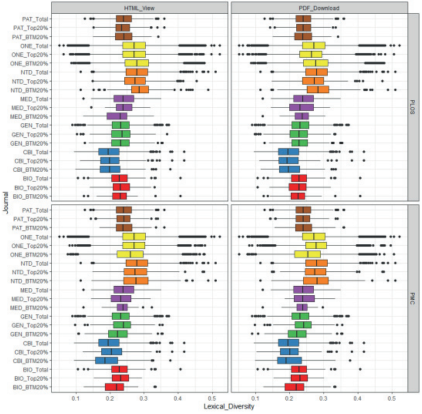


Figure 4. Lexical diversity distribution of PLoS and PMC platform

From Figure 1 to 4, they reveal different linguistic characteristics among Top 20%, total and Bottom 20% papers of PLoS and PMC platforms. Generally, Top 20% papers have more author number than total and Bottom 20% papers. Full text length, sentence length and lexical diversity are marginally different among Top 20%, total and Bottom 20% papers of PLoS and PMC platforms. Linguistic characteristics among specific journal are also different.

Acknowledgments

This paper is supported by Youth Program of National Social Science Fund in China (15CTQ035) and Research Training Program by NJUST (2018).

References

Boyack, K. W., Eck, N. J. V., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: a large-scale analysis. Journal of Informetrics, 12(1), 59-73.

Chen, B. (2018). Usage pattern comparison of the same scholarly articles between Web of Science (WoS) and Springer. Scientometrics, 115(1): 519-537.

Chi, P. S., & Glänzel, W. (2017). An empirical investigation of the associations among usage, scientific collaboration and citation impact. Scientometrics, 112(1), 403-412.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., & Schnaars, M., et al. (2018). Examining scientific writing styles from the perspective of linguistic complexity. Journal of the Association for Information Science and Technology, 70(5):462-475.

Pan, X., Yan, E., Ming, C., & Hua, W. (2018). Examining the usage, citation, and diffusion patterns of bibliometric mapping software: a comparative study of three tools. Journal of Informetrics, 12(2), 481-493.

From Macro to Micro: A Bibliometric Evaluation of Leading Scientific and Technological Achievements ——Taking the Novel Fermions in Solids as an Example

XIE Li[1] TAO Cheng[2] ZHANG Yuehong[2] CHEN Yunwei[1*] ZHANG Zhiqiang[1]

[1] *{xieli, chenyw, zhangzq}@clas.ac.cn* *Corresponding Author
Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu, 610041 (China)

[2] *{ctao, zhangyh}@cashq.ac.cn*
Bureau of Development and Planning, Chinese Academy of Sciences, Beijing, 100864 (China)

Introduction

In 2015, three teams—Institute of Physics of the Chinese Academy of Sciences(CAS-Inst Physics), Princeton University and MIT independently found evidence of the existence of Weyl fermions in solids simultaneously. This discovery was named Top Ten Breakthroughs of 2015 by Physics World magazine (IOP), and was selected by the American Physical Society as "The Eight Highlights of International Physics in 2015", the "Top Ten Breakthroughs in the Physical World 2015" by the British Physical Society. The three teams share this honor, but at the very beginning, who was the first discoverer caused controversy. In addition, CAS-Inst Physics has continued to make breakthroughs in the field of novel fermions in recent years such as the discovery of three-component fermions in the topological semimetal molybdenum phosphide in 2017 and find of the evidence for Majorana bound state in an iron-based superconductor in 2018. Thus, whether these achievements leading the development in its research field needs an objective answer.

The application of bibliometric methods in assessment is mainly take institutions, disciplines/ research fields, journals, papers, research groups, and individuals as objects separately, and the content ranges from academic competitiveness, influence, creativity to productivity (Ren, 2016; Zhao, 2014; Wu, 2018). But there is rarely assessment for the pioneering and leading achievements, and no evaluation takes different levels from macro to micro such as fields, institutions and related specific papers into consideration at the same time integratedly.

Data and methods

First of all, based on the analysis of the field research trends, we evaluate the research heat of the field in which the leading results are derived from the macro level. Secondly, through the comparative analysis of the competitiveness of the main research institutions in the field of leading achievements, it is possible to evaluate the research strength of the teams to which the results belong from the meso level. Finally, based on the citation analysis, we analyzes the specific research results from the micro level, and analyzes the international influence of the core results in a fine-grained manner.

The data were obtained from ISI-WOS core collection. The main retrieval date is February 10, 2018. Retrieval strategy based on subject and took full use of logical combination, wildcard, position limitation and noise elimination. The time span was not limited, and so is the document type.

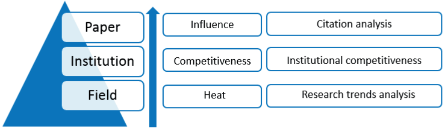


Figure 1. From Macro to Micro: A Bibliometric Evaluation of Leading of Scientific and Technological Achievements

Results

Field heat: global research of four types fermions growing rapidly

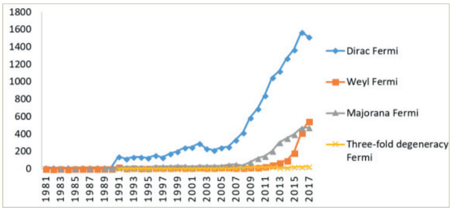


Figure 2. Global trends of fermions research

Institutional competitiveness: CAS has the superiority in fermions research papers

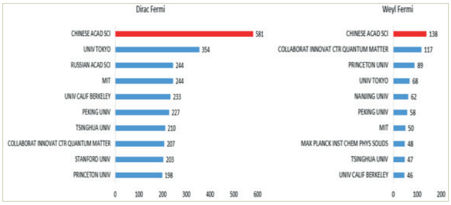


Figure 3. Global Top 10 institutions with the largest number of Dirac & Weyl fermi research

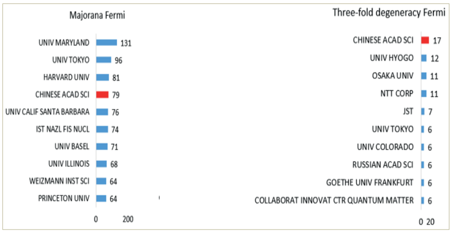


Figure 4. Global Top 10 institutions with the largest number of Majorana & Three-fold degeneracy fermi research

Paper influence: Citing papers are highly consistent across national/organization/journal distribution

Three core papers which independently claimed on founding evidence of the Weyl fermi's existance in solids are from CAS-Inst Physics, Princeton University, and MIT.

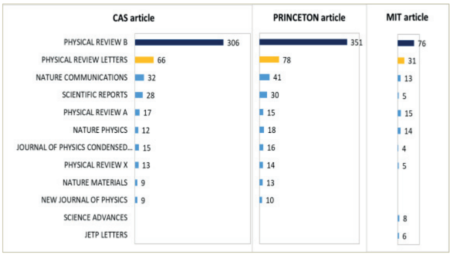


Figure 5. Top 10 journals that cite the three core papers

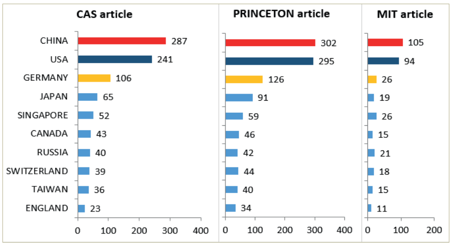


Figure 6. Top 10 countries/regions that cite the three core papers

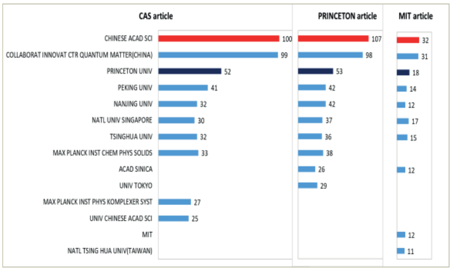


Figure 7. Top 10 institutions that cite the three core papers

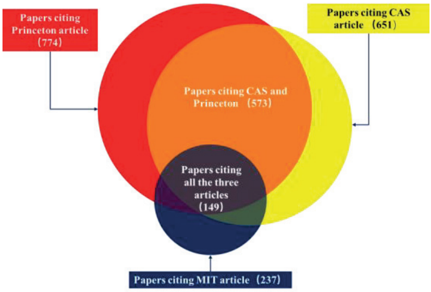

Papers citing Princeton article (774)
Papers citing CAS article (651)
Papers citing CAS and Princeton (573)
Papers citing all the three articles (149)
Papers citing MIT article (237)

Figure 8. Co-citations of the three articles

Noteworthily, since the global fermions research are quite concentrated and most institutions persist long-term cultivation, self-citation also make sense while studying the influence, thus were not excluded or studied separately.

Conclusions

The result shows that the CAS team has a deep and well-balanced foundation in the field of novel fermions in solids. CAS team is at the leading position in the field of Dirac fermions, Weyl fermions and three-component fermions based on the number of papers. The Weyl fermion article of CAS-Inst Physics, not only won the American physical society and the authoritative organizations such as the British Institute of Physics, but also shares more than three-quarters citations with Princeton university article, followed with high consistency of national distribution, organization distribution and journal distribution. The results indicate that the CAS team, together with the team of Princeton university and the team of MIT, are leading the pioneering work of discovering Weyl fermion.

Acknowledgements

This work is funded by the Special Project of Strategic Research and Decision Support System Construction of Chinese Academy of Sciences (GHJ-ZLZX-2019-31).

References

Ren Q.E. (2016). On Three Types of Bibliometrics Index Used for Academic Influence Evaluation of Publishing House. *Journal of Academic Libraries*, 34(5), 110-119.

Zhao, F., Ai, C.Y., You, Y., et al (2014). Bibliometric Evaluation of University Scientific Research ——A Case Study of Peking University. *Journal of Academic Libraries*, 32(1), 97-101.

Wu, A.Z., Xiao, L., Zhang, C.H., et al. (2018). Evaluation Methods and System of University Discipline Competitiveness Based on Bibliometrics. *Journal of Academic Libraries*, 36(1), 62-67,26.

Importance of research network analysis for early-career scientists

Akiko Ohata[1] and Kenichi Hagiwara[2]

[1] *ohhata.akiko@jaxa.jp*
Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency,
3-1-1, Yoshinodai, Chuo-ku, Sagamihara, Kanagawa, 252-5210 (JAPAN)

[2] *hagiwara.kenichi@jaxa.jp*
Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency,
3-1-1, Yoshinodai, Chuo-ku, Sagamihara, Kanagawa, 252-5210 (JAPAN)

Introduction

In Japan, the postdoctoral positions greatly increased in 1996. At present, however, the average age in postdoctoral researchers has become 36 years due to the difficulty of finding a permanent academic position. It causes a significant decrease in the number of students pursuing doctoral courses. This would be an obstacle for the development of science and technology. On the other hand, there has been insufficient discussions regarding an effective system or environment for early-career researchers to develop their abilities.

In this study, we compare the research publications of two groups of postdoctoral researchers. From the results on the difference between them, we indicate the importance of network analysis among researchers while developing their careers and abilities.

Comparison of the standard indicators in two doctoral fellow groups

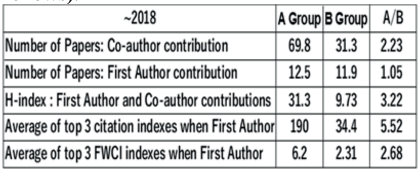
We compared the doctoral fellows in Group A, (Fellowship A: 3-year employment with rather high rewards), and in Group B (Fellowship B: 3-year employment with standard rewards). Fellowship A is the system outside Japan, while Fellowship B is the system in Japan. (Most of doctoral fellows in Group B are Japanese.) In both systems, they earned their postdoctoral positions in 2010 and their research fields are related to space and aeronautical science. We used SCOPUS® as the data base for this study. In this paper, we use the word "co-author" except for the first author of the published paper.

Table 1 presents a comparison of the standard indicators about the publications (article, conference paper and review) of the doctoral fellows in Group A and B. The following differences were identified:

(1) The average total number of papers in which doctoral fellows in Group A contributed as a co-author was much higher than that in Group B, whereas the average total number of papers in which doctoral fellows in Group A contributed as a first author was slightly higher than that in Group B.

(2) In terms of citation index and Field Weighted Citation Impact (FWCI) with respect to the first author contributions, and H-index including the co-author and the first author contributions, Group A exhibited a higher level than that exhibited by Group B.

Table 1: Average of the standard indicators for the papers by the doctoral fellows in Group A (16 doctoral fellows) and Group B (11 doctoral fellows).



~2018	A Group	B Group	A/B
Number of Papers: Co-author contribution	69.8	31.3	2.23
Number of Papers: First Author contribution	12.5	11.9	1.05
H-index : First Author and Co-author contributions	31.3	9.73	3.22
Average of top 3 citation indexes when First Author	190	34.4	5.52
Average of top 3 FWCI indexes when First Author	6.2	2.31	2.68

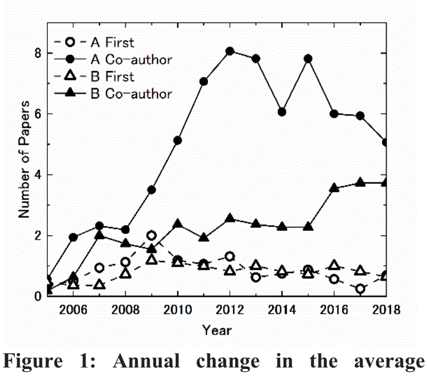
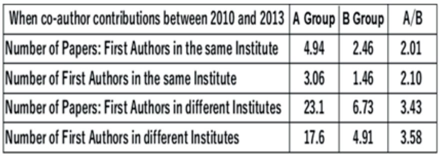


Figure 1: Annual change in the average number of papers by the doctoral fellows of Group A & B. Open symbols: Doctoral fellows are the first author. Filled symbols: Doctoral fellows are the co-author.

By considering the annual changes in the number of papers (with respect to the co-author and first author contributions by the doctoral fellows), the average total number of papers in which doctoral fellows in Group A contributed as a co-author, considerably increased during the doctoral fellowship period of interest (between 2010 and 2013) as shown in Fig.1.

Table 2: Average of the number of papers and the first authors' information when the postdoctoral fellows in Group A & B contributed as a co-author.



When co-author contributions between 2010 and 2013	A Group	B Group	A/B
Number of Papers: First Authors in the same Institute	4.94	2.46	2.01
Number of First Authors in the same Institute	3.06	1.46	2.10
Number of Papers: First Authors in different Institutes	23.1	6.73	3.43
Number of First Authors in different Institutes	17.6	4.91	3.58

With respect to the publications between 2010 and 2013 (this period being the focus was reasonable because the postdoctoral fellowship began in 2010.), the following differences can be clearly observed.

(1) In the case of Group A, the average number of papers in which the researchers belonging to other institutes (different from the institutions in Group A) were the first author, was rather high compared with the case of Group B.

(2) There are many contributors as a first author in the papers by Group A. Such a tendency became apparent when the researchers in other institutes became the first author.

From the above results, we noticed the importance of the collaborated researches both in the institute at which they studied and in other institutes. In the next section, we discuss the impact of collaboration on the citation index using the network analysis.

Network analysis in the collaboration works

As one of the key indices of research network analysis, the betweenness centrality could denote the key person. Figure 2 shows one example of the calculated betweenness centrality for the researchers (X1 ~ X9) collaborated with the doctoral fellow X of Group A. It was obtained by calculating the betweenness centrality for the published papers by the doctoral fellow X in each year.

In 2014, four researchers (X1, X2, X3, and X5) have a high betweenness centrality. Among the publications by the doctoral fellow X in 2014, X2, X3, and X5 became the authors to the papers with a higher citation index, while researchers (X7, X8, and X9) with a lower betweenness centrality did not contribute, as shown in Table 3. This fact shows that the collaboration with the researchers having a higher betweenness centrality may lead to the production of a paper with a higher citation index.

Thus, the network analysis for the papers can give us the information for establishing a better research collaboration to produce a paper with a higher impact.

Summary

We compared the research publications of two postdoctoral groups, Group A and B. Group A produced many papers with a higher impact. The average total number of papers in which doctoral fellows in Group A contributed as a co-author was much higher than that in Group B. In addition, many different researchers contributed as a first author in the publications by the doctoral fellows in Group A. Such a tendency became particularly apparent for the papers with other institutes.

Furthermore, we investigated the impact of collaboration on the citation index. From the betweenness centrality analysis, we found that in the papers with a higher citation index, the researchers with a higher betweenness centrality contributed. Therefore, network analysis to establish the better environment for early-career researchers is important as the collaboration works plays the important role for career development.

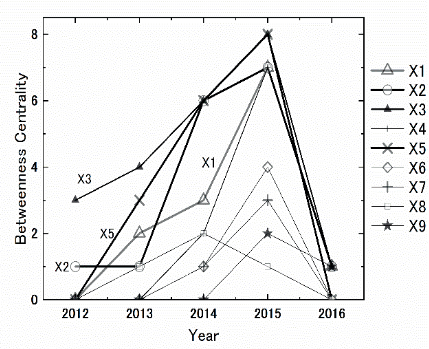
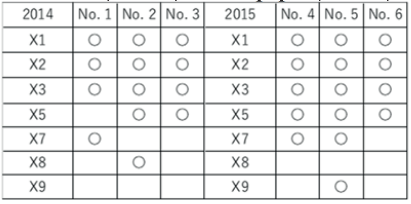


Figure 2: Calculated betweenness centrality with respect to 9 authors (X1~X9) for the papers by doctoral fellow X of Group A in each year.

Table 3: Other researchers' contributions to the papers by doctoral fellow X in 2014 and in 2015. No.1~3 and No.4~6 are three papers from the order of higher citation index among the publications in 2014 and in 2015, respectively. Circle shows the contribution of other researchers (X1~X9) to each paper (No.1~6).



2014	No. 1	No. 2	No. 3	2015	No. 4	No. 5	No. 6
X1	○	○	○	X1	○	○	○
X2	○	○	○	X2	○	○	○
X3	○	○	○	X3	○	○	○
X5		○	○	X5	○	○	○
X7	○			X7	○	○	
X8		○		X8			
X9				X9		○	

References

M. Okamoto, et. al (2018). The 2015 Survey on Postdoctoral Fellows Regarding Employment and Careers in Japan. http://www.nistep.go.jp/reportlist

Finding More Methodological Entities from Academic Articles via Iterative Strategy: A Preliminary Study

Yuzhuo Wang[1] and Chengzhi Zhang[2, *]

[1]*wangyz@njust.edu.cn*
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094(China)

[2] *zhangcz@njust.edu.cn*
Department of Information Management, Nanjing University of Science &Technology, Nanjing 210094(China)

Introduction

Research methods have always played a significant role in the history of science. As important tools of solving research problems, research methods promote development of a discipline, and enable researchers to solve problems efficiently.

Generally, research methods mean the ways or materials of solving problem and are constituted by methodological entities (MEs) which include models, algorithms, and so on. Currently, mention of algorithms (Wang & Zhang, 2018), datasets (Zhao, Yan, & Li, 2018), software (Pan et al., 2015) is investigated according to full-text of academic articles. However, existing work focuses on the specific type of methods, and scale of methods is small, which cannot provide scholars with a comprehensive review of methods. Therefore, an iterative strategy of finding more methodological entities in a special domain is proposed in this paper.

In this paper, we take natural language processing (NLP) domain as an example, and collect more MEs from academic articles in the domain. More specifically, we investigate two questions: How many frequently-used MEs are used in a specific domain? What are the most popular MEs in each type of MEs? If we answer the questions, we could construct a taxonomy of MEs and further evaluate or recommend related MEs for scholars, especially the beginners of a research domain. It should be noted that a methodological sentence means a sentence that contains MEs.

Methodology

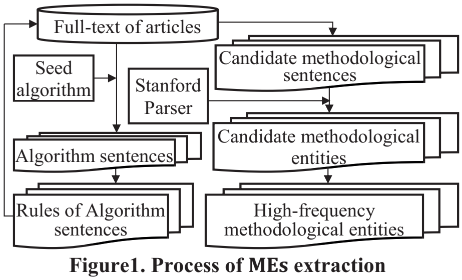

Full-text of articles
Seed algorithm
Stanford Parser
Candidate methodological sentences
Algorithm sentences
Candidate methodological entities
Rules of Algorithm sentences
High-frequency methodological entities

Figure1. Process of MEs extraction

As shown in figure 1, MEs are extracted from the full-text of academic articles by an iterative strategy.

(1) Data collection

Full-text of ACL (The Association Computational Linguistics) annual conference papers between 1979 and 2015 are downloaded from the ACL Anthology (http://www.aclweb.org/anthology), all the 4,568 papers are available in XML format.

(2) Collecting rules of algorithm sentences

We choose the top-10 data mining algorithms as seed words, since algorithms are usually published in scholarly articles, especially in NLP domain (Tuarob *et al.*, 2016). After that, we compile a dictionary about names of these ten algorithms, and extract sentences that contain algorithms from papers, the sentences are called "algorithm sentences". Four post-graduates who major in information science are invited to conclude rules of algorithm sentences. Rules refer to the word or phrase in sentences that indicates authors used algorithms, such as "based on". We try to use the rules to find more methods used by authors. A professor in the NLP domain reviews the rules and confirms the final 415 rules.

(3) Extracting candidate sentences and entities

Candidate methodological sentences are identified by rules of algorithm sentences. Stanford parser (https://nlp.stanford.edu/ software/lex-parser.shtml) is utilized to extract NPs (noun phrase) from candidate sentences, since the name of a ME is likely to be presented as a NP. For each NP, we count the number of sentences that contain the NP. NPs whose frequency beyond 10 are regarded as candidate MEs.

(4) Method entities filtering and classification

Four post-graduates manually identify MEs from candidate MEs. Various entities representing the same method are summarized into a group. For example, '*SVM*' and '*support vector machine*' are summarized into '*suppor*t vector machine' group. We do identify more kinds of methods through these rules. Then, MEs are manually classified into seven types according to their name, including Algorithm, Data &source, Index &measure method, Linguistic rule, Model, Tool, Other (Methods that failed to identify their categories are marked as '*Other*'). Finally, we get 237 frequently-used methods.

Results

(1) Rules of methodological sentences extraction

[*] Corresponding author.

Using the rules of algorithm sentences, we get 101,944 methodological sentences. We list the top-5 rules which identify the most sentences. According to table 1, "*Based on*" obtains the most methodological sentences. Three of the top-5 rules are related to "*use*", which means the function of MEs is "use".

Table1. Top-5 rules of methodological sentences extraction

No.	Rule	Number of sentences (ratio)
1	*based on*	2695(6.40%)
2	*using*	2098(4.99%)
3	*use*	1344(3.19%)
4	*used to*	2863(2.81%)
5	*training*	2787(2.73%)

（2）Distribution of different types of MEs

We get 237 MEs from 4,493 conference papers. As shown in table 2, among the seven types of MEs, algorithm and model are the two most popular types, which account for 22.75% and 21.46%, respectively. Type of most MEs can be identified, since proportion of '*Other*' is only 8.15%.

Table2. Proportion of different types of MEs

No.	Type	Count(ratio)
1	Algorithm	53(22.75%)
2	Model	50(21.46%)
3	Data & source	41(17.60%)
4	Index & measure	25 (10.73%)
5	Tool	24(10.30%)
6	Linguistic rule	21(9.01%)
7	Other	19(8.15%)
-	Total	233(100%)

（3）Top entities in different types of MEs

Table3. Top-3 MEs in different types

Type	No.	Name of ME	Count
Algorithm	*1*	*Support vector machine*	526
	2	*Expectation maximization*	309
	3	*Gibbs sampling*	300
Data & source	*1*	*Wikipedia*	348
	2	*Wall street journal*	309
	3	*Penn treebank*	291
Linguistic rule	*1*	*Context-free grammar*	308
	2	*Combinatory categorical grammar*	206
	3	*Probabilistic context-free grammar*	164
Index & measure method	*1*	*BLEU*	527
	2	*F-measure*	383
	3	*Net promoter score*	184
Model	*1*	*N-gram language Model*	672
	2	*Conditional random Field*	339
	3	*Maximum entropy*	329
Tool	*1*	*Giza++*	199
	2	*Srilm*	92
	3	*Word2vec*	65
Other	*1*	*Linear regression*	52
	2	*Linear programming*	51
	3	*Directed acyclic graph*	40

For each ME, we count the number of papers where the ME appears. Due to space limitation, we list top-3 MEs in table 3. In general, for each type, top-3 MEs are well-known methods. Taking the algorithm as an example, '*Support vector machine*', '*Expectationon maximization*' and '*Gibbs sampling*' have appeared in more articles, which means that the more classic the method is more popular in the field.

Conclusion and discussion

Using top-10 data mining algorithms as seed words, we get 237 frequently-used MEs from the full text of academic articles based on the Iterative method. Results indicate that 'Based on' is the most popular rule in the sentences referring to methods. Algorithm is the most popular type of high-frequency MEs in the NLP domain. Additionally, authors are willing to use well-known methods, like '*SVM*', since them appear in more academic articles.

As an elementary work, there are two limitations. First of all, we only use the rules of algorithm sentences, which adversely affect the identification of other type of methods, e.g. models and tools. Later we will supplement the rules of different types of methodological sentences. On the other hand, the MEs are filtered manually. Therefore, we will use machine learning method to automatically identify MEs from large-scale papers. In addition, we can continually apply the iterative strategy in the future. It means we will use various types of MEs acquired in this paper to get more methodological sentences, and conclude rules of various types of methodological sentences. Finally, more MEs will be extracted based on different kinds of rules.

Acknowledgement

This work is supported by Major Projects of National Social Science Fund (No. 17ZDA291), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX18_0351).

References

Pan, X., Yan, E., Wang, Q., & Hua, W. (2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. Journal of Informetrics, 9(4), 860-871

Tuarob, S., Bhatia, S., Mitra, P., & Giles, C. L. (2017). Algorithmseer: a system for extracting and searching for algorithms in scholarly big data. IEEE Transactions on Big Data, 2(1), 3-17.

Wang, Y., & Zhang, C. (2018). Using Full-Text of Research Articles to Analyze Academic Impact of Algorithms. *In: Proceedings of iConference2018*, Sheffield, UK, 25th-28th March, 2018.

Zhao, M., Yan, E., & Li, K. (2017). Data set mentions and citations: a content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, 69(1):32–46.

Exploring the Effects of Data Set Choice on Measuring International Research Collaboration: an Example Using the ACM Digital Library and Microsoft Academic Graph

Ba Xuan Nguyen[1], Markus Luczak-Roesch[1] and Jesse David Dinneen[1]

[1] *ba.nguyen @vuw.ac.nz, markus.luczak-roesch@vuw.ac.nz, jesse.dinneen@vuw.ac.nz*
School of Information Management, Victoria University of Wellington, Wellington (New Zealand)

Introduction

International research collaboration (IRC) is a construct that refers generally to scientific activities between individuals in different countries. IRC measurement is important because countries can and want to benefit from international collaboration (Guerrero Bote et al., 2013; Katz & Martin, 1997). Hence, ways to measure IRC are a focus of bibliometrics and informetrics research.

Many datasets are available to measure IRC and other facets of what has been framed as the "Science of Science" (Fortunato et al., 2018), but it has also been shown that performing the same measurement procedure on different datasets can lead to different results (De Stefano et al., 2013). The extent as well as the causes for such variances need to be adequately understood. We aim to contribute to this understanding by addressing the following research question: **what are the effects of data set choice on IRC measurement?**

Research data and operationalisation of IRC

In this preliminary investigation we consider bibliographic metadata from the ACM Digital Library (ACM) and the Microsoft Academic Graph (MAG) dataset. The ACM data is supplied by the Association for Computing Machinery[1] as resource for research purposes (coverage from 1951-2017), while the MAG data (Sinha et al., 2015) was downloaded from the Open Academic Society website[2] (coverage from 1965-2017). Since ACM is largely a domain specific bibliographic source in the computing sciences (CS), a subset of the MAG dataset was created to cover only papers related to this field (by filtering with the most frequent "field of study" CS terms extracted from the matched collection). We acknowledge that applying this heuristic implies a limitation because it might mean we are missing some papers. In addition, some papers in this collection that already exist in ACM are also excluded to ensure that the ACM and MAG data sets are distinct.

In this study we investigate co-authored publication records and use the information about authors' affiliations to derive distinct **bilateral relationships**. If there is more than one co-author from a country in one publication, only one bilateral relationship is counted between that country and any of the others. From the ACM set of 182,791 records we identified 121,672 that are co-authored, 15,686 of which international co-authors, whereas from the MAG set of 939,821 computer science publications we found 594,036 with co-authors, of which 32,909 had international co-authors. This resulted in 21,827 (ACM) and 45,068 (MAG) distinct bilateral relationships.

Analysis and results

First we observe a difference in the numbers of bilateral relationships between countries found in ACM and MAG. While the trend of both datasets in the last 10 years is similar, the MAG data shows a substantially different evolution compared to ACM before that time and has a significantly lower amplitude (see Figure 1).

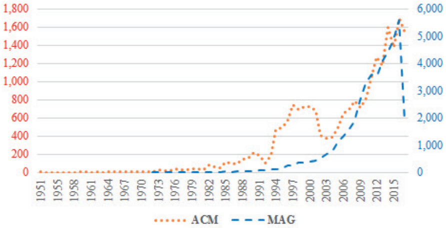


Figure 1. Total numbers of bilateral relationships over 1951-2017.

Comparing the top 10 countries ranked by total numbers of bilateral relationships over a period of 50 years (1966-2015), we find that the USA is consistently ranked first in both data sets. Other countries differ, however: for example China is ranked at the sixth position in the top ranks of ACM while being ranked second in MAG. Similarly, Canada is listed as the fourth in the former but sixth in the latter. To find out how consistent this ranking of countries is over time we perform an analysis of the rank order of countries based on international collaborations per year. To do this we first create a dataset of the annual IRCs per country (# of distinct countries: MAG - 164, ACM - 111), then derive an annual ranking of all countries by the amount of IRCs in the respective years (from highest to lowest), which we then inspect to find (a) the unique countries that are represented in both datasets (N=109) and (b) a reasonable cutoff point from which onwards we have a set of countries that are ranked in any of the following years. We set the cutoff to the year 2000, and derive a set of 30 countries that are fully covered from this year onwards until 2017, allowing us to rank these countries for all 18 years.

For each pair of rank vectors for these 30 countries we compute Spearman's and Kendall's rank order correlation coefficients, and the hamming distance. We also compute the mean and standard deviation (SD) for each of the rank vectors (so each country has one mean rank for its position in ACM and one for MAG). Again, the analysis of this data shows that the USA is consistently ranked first (and therefore has no correlation coefficients as the SD is zero). For all other countries we find that the mean of the hamming distance of the rank vectors is notably high (16.10, SD 2.37), which means that they are ranked differently in the two datasets for most years. Figure 2 displays the Spearman correlation coefficient (with confidence intervals) for these 30 countries. It highlights that there is basically no correlation present, which means that even the trend of the rank for countries (i.e. if a country rises over the years in one dataset it also rises in the other) is not consistent between the two datasets.

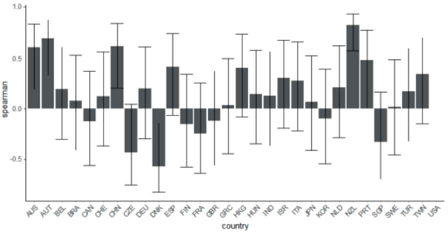


Figure 2. Spearman's rank order correlation coefficients (with confidence interval) for 30 countries for rank vectors from 2000-2017.

Conclusion

In this short paper we reported about our efforts to quantify and qualify the effects of data set choice on the outcomes of IRC measurement. We sought to provide empirical evidence that there are significant differences and to give some preliminary indicators for what cause and effect these may have. By performing an intuitive time series and rank order analysis **we found (1) inconsistent temporal coverage of the computer science domain in ACM and MAG data; (2) a similar upwards trend in bilateral scholarly relationships in recent years but with varying amplitude; and (3) significant movements in ranks for countries that are not consistent between the two datasets.**

We conclude that there exist differences that need to be better understood and require further investigation. The results presented here already have implications for our understanding of key activities in bibliographic analysis, such as temporal sampling when measuring IRC. Future work will need to dig deeper into the cause and effect relationships and we seek to undertake an analysis that clusters countries based on their rank variance to see if there are certain countries that are affected more or less than others by the differences in the data sets. Finally, the problem demonstrated here can also be looked at qualitatively to understand whether the data quality issues matter to actual data consumers such as policy makers.

References

De Stefano, D., Fuccella, V., Vitale, M. P., & Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks, 35*(3), 370-381.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Vespignani, A. (2018). Science of science. *Science, 359*(6379), eaao0185.

Guerrero Bote, V. P., Olmeda-Gómez, C., & de Moya-Anegón, F. (2013). Quantifying the benefits of international scientific collaboration. *Journal of the American Society for Information Science and Technology*, *64*(2), 392-404.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration?. *Research policy*, *26*(1), 1-18.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. P., & Wang, K. (2015, May). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243-246). ACM.

[1] ftp://pubftp.acm.org

[2] https://www.openacademic.ai/oag/

The role of the integrated impact indicator (I3) in evaluating the institutions within a university

Ivan Pilčević[1], Srđa Bjeladinović[2] and Veljko Jeremić[3]

[1] *ivan.pilcevic@gmail.com*
University of Belgrade, Faculty of Organizational Sciences, Jove Ilića 154, 11000, Belgrade (Serbia)

[2] *bjeladinovic.srdja@fon.bg.ac.rs*
University of Belgrade, Faculty of Organizational Sciences, Jove Ilića 154, 11000, Belgrade (Serbia)

[3] *veljko.jeremic@fon.bg.ac.rs*
University of Belgrade, Faculty of Organizational Sciences, Jove Ilića 154, 11000, Belgrade (Serbia)

Introduction

Since its introduction (Leydesdorff & Bornmann 2011; Leydesdorff & Bornmann 2012; Rousseau 2012), the integrated impact indicator (I3) has grabbed the attention of numerous researchers (Ye et al. 2017) and a significant number of papers have elaborated the benefits that I3 brings to impact measurement (Bornmann et al. 2019; Leydesdorff et al. 2019). Accordingly, we wanted to explore the potential of using the I3 class of indicators to evaluate departments/institutions belonging to a specific university. This is particularly important for developing countries that are facing continuous government cuts in higher education (Villarreal & Ruby 2018).

Methodology

As a case study, we analysed the dataset (Pilcevic et al. 2018) containing WoS indexed papers (SCIe and SSCI indexed journals) published by researchers with the University of Belgrade (UB) as the reprint authors. We evaluated the subset of papers published in 2014, assigning each paper with an I3N indicator obtained using https://www.leydesdorff.net/i3/ranking.htm. As elaborated in previous research (Pilcevic et al. 2018), each article was assigned to a particular faculty (31 faculties) and institute (11 institutes) within the UB. If the published work was the result of collaboration between institutions within the UB, it was affiliated with all the institutions that participated in its authorship. In addition, the results of the four leading (in terms of number of published articles) institutions in each administrative cluster (Faculties of Medical Sciences, Scientific Institutes, Faculties of Technology and Engineering Sciences, Faculties of Science) were merged so that the performance of each of the four administrative clusters could be closely observed.

Results

In total, 1829 papers were scrutinized. As we can see from Table 1, Scientific Institutes exhibit the best results (the median value of I3N is 1.900), which is even more obvious if we observe Figure 1.

Table 1. Descriptive statistics of I3N for each administrative cluster of institutions within the University of Belgrade

I3N	A	B	C	D
mean	1.786	2.448	2.021	2.166
median	1.400	1.900	1.500	1.700
st. dev	1.047	2.406	1.311	1.345
IQR	1.000	1.500	1.200	1.300

A - Faculties of Medical Sciences, B - Scientific Institutes, C - Faculties of Technology and Engineering Sciences, D - Faculties of Science

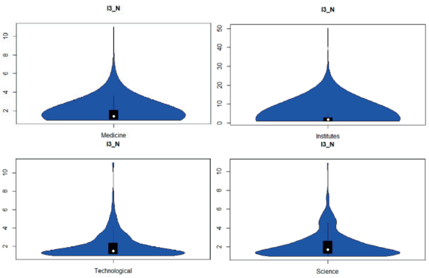


Figure 1. Violin plot of I3N for each administrative cluster of institutions within the University of Belgrade

Interestingly, each cluster of institutions shows considerable skewness of I3N, with Scientific Institutes having a particularly skewed distribution. Digging a bit deeper, within the administrative cluster of the Faculties of Medical Sciences, the Faculty of Medicine leads the way in terms of number of articles. The median value of I3N for the Faculty of Medicine is the same as for the entire cluster, 1.400. On the other hand, although the Faculty of Pharmacy published fewer articles than its peers in the Faculty of Medicine, it performed better in terms of I3N, with a median value of 1.700. Each of the leading institutes in the cluster of Scientific Institutes (Vinca Institute, Physics Institute, Sinisa Stankovic Institute, and ICTM Institute) showed a similar I3N performance. The performance of the Faculty of Biology and Faculty of Chemistry is particularly interesting: they published an almost equal number of papers, but the I3N median value for the Faculty of Chemistry is 1.800, while for the Faculty of Biology it is 1.300.

Conclusion

The I3N measure was able to shed new light on the impact of institutions within the University of Belgrade. Unfortunately, the UB's results are far from satisfactory. In the future this study could emphasize UB researchers who excel, which would contribute to the growing need to rank not only universities but also academic staff.

References

Bornmann, L., Tekles, A., & Leydesdorff, L. (2019). How well does I3 perform for impact measurement compared to other bibliometric indicators? The convergent validity of several (field-normalized) indicators. *Scientometrics*, Online First, doi: 10.1007/s11192-019-03071-6

Leydesdorff, L., & Bornmann, L. (2011). Integrated Impact Indicators (I3) compared with Impact Factors (IFs): An alternative research design with policy implications. *Journal of the American Society of Information Science and Technology, 62*(11), 2133–2146.

Leydesdorff, L., & Bornmann, L. (2012). Percentile ranks and the integrated impact indicator (I3). *Journal of the American Society for Information Science and Technology, 63*(9), 1901–1902.

Leydesdorff, L., Bornmann, L., & Adams, J. (2019). I3: A non-parametric alternative to the journal impact factor. *Scientometrics* (in press)

Pilčević, I., Jeremić, V., & Vujošević, D. (2018). Evaluating the scientific performance of institutions within the university: An example from the University of Belgrade leading institutions. *Journal of the Serbian Chemical Society, 83*(11), 1285–1295.

Rousseau, R. (2012). Basic properties of both percentile rank scores and the I3 indicator. *Journal of the American Society for Information Science and Technology, 63*(2), 416–420.

Villarreal III, P., & Ruby, A. (2018). Government Models for Financing Higher Education in a Global Context: Lessons from the US and UK. Available at: https://repository.upenn.edu/ahead_papers/3/

Ye, F. Y., Bornmann, L., & Leydesdorff, L. (2017). h-Based I3-type multivariate vectors: Multidimensional indicators of publication and citation scores. *COLLNET Journal of Scientometrics and Information Management, 11*(1), 153–171.

Research on Influence of Dataset Scale on Domain Analysis in Bibliometrics

Panting Wang[1] and Guo Chen[2]

[1] *118107022078@njust.edu.cn*
Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094(China)
[2] delphi1987@qq.com
Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094(China)

Introduction

In bibliometric research, it is essential to get appropriate dataset because the dataset directly determines the reliability of follow-up analysis and analysis results. An inevitable question about dataset in domain analysis is how many bibliography records we should collect. Obviously, it is inefficient for researches to collect too much bibliography records of some domain because unrelated ones are more likely to be brought in. Therefore, it is quite necessary for us to investigate the influence of dataset scale on bibliometric results.

Although the dataset collection methods of a given domain have been covered in some studies (Feng & Leng 2009), there has been no study yet on its scale. To reveal the influence of dataset scale on bibliometrics analysis, we conduct an experiment on the “artificial intelligence” domain, in which a standard dataset is sampled with five sizes, and then six typical elements are compared using overlapping ratio as well as spearman correlation. The overlapping ratio indicates the consistency of identifying important elements in a given domain, while the spearman correlation indicates the consistency of ranking them. According to the results, some suggestions are put forward.

Datasets and Method

Datasets

Firstly, we construct a standard dataset of “artificial intelligence” by retrieving "WC = Artificial Intelligence and PY = 1996-2017" in Web of Science (WOS), where we get 723,187 records.

Then, we used 5,000, 10,000, 20,000, 50,000 and 100,000 as sample sizes to sample the standard dataset. For each sample size, we sampled five times to ensure its reliability and then calculated the average index of them.

Methods

For the standard dataset and different sample datasets, we choose six elements commonly used in bibliometrics (subject classification, country, institution, author keyword, reference, and author) for comparison.

Given that in practice, people tend to focus on elements with high frequency (for example, prolific authors, hot-spot keywords). Therefore, we set different TOP N for each element in our experiment, as shown in Table 1.

Table1.TOP N Settings

Element	TOP N				
WC	5	10	20	50	100
Country	5	10	20	30	50
Institution	5	10	20	50	100
Keyword	50	100	200	500	1000
Reference	50	100	200	500	1000
Author	10	20	50	100	200

For evaluation, we focus on two common tasks in domain bibliometrics: (1) identifying important elements; (2) ranking important elements.

In identifying important elements, we check how many top N elements in the standard dataset are also identified in those sampled datasets. We utilized the overlapping ratio (OR) for evaluation (Shu et al.2018), which is calculated as follow:

$$OR = \frac{number\ of\ overlap\ items\ in\ Top\ N}{N}$$

In ranking important elements, we check the similarity of top N element ranks between the standard dataset with those sampled datasets. We utilized the Spearman coefficient (SC) for evaluation (Spearman 1904).

The overall flow chart is shown in Figure 1.

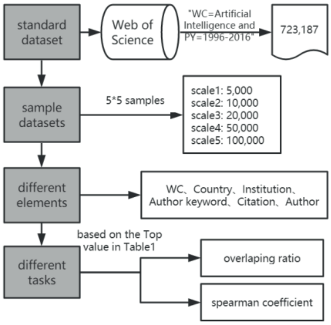


Figure 1. The flow chart of our experiment

Results

General rules

Figure 2 shows that larger sample datasets fit the standard dataset better in both analysis tasks, for each element and for each TOP N value. However, it also shows the law of diminishing marginal returns, indicating that expanding a large dataset is less efficient than a small dataset.

For WC, keyword, reference and author, the larger the number of elements to be identified or ranked (that is, the larger the TOP N is), the harder to fit them to standard datasets, which indicates that identifying or ranking more items are less reliable. However, country does the opposite; this might because the total country number is limited. For institution, there is no obvious rule.

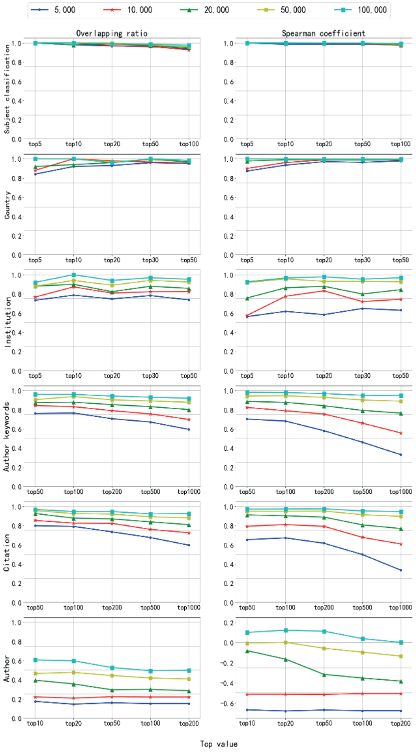


Figure 2. The fitting results of different elements in different tasks

Different elements

For WC and country, it is distinctly easy to fit them in both tasks. The OR and SC are both high even we only use 5000 records to fit a large domain.

For institution, keyword and reference, large datasets perform quite well and small datasets perform mediocre. However, top authors are hard to fit, especially in ranking them.

The result indicates that, when analyzing subject classifications and countries, the dataset is no need to be too large; when analyzing institutions, keywords and references, a large dataset is better and a small dataset may be acceptable; when analyzing authors, it is essential to expand the dataset as far as possible.

Different tasks

Overall, as shown in figure 2, the fitting result of ranking elements is consistent with that of identifying them. Because of the limited total number of WC and countries, the ranking effect is better than identifying. However, the ranking effect of important elements of institutions, keywords, authors and references is worse than identifying them. Ranking elements in bibliometric is more sensitivity to dataset scale, especially ranking authors.

Conclusion

Through the quantitative comparison between a standard dataset with five types of sampled datasets, we exhibit the influence of dataset scale on identifying and ranking important elements in domain bibliometrics. To fit a standard dataset with 700,000 records, author analysis requires a dataset as completely as possible, while other elements can be fit well with a small dataset with 50,000 to 100,000 records. Meanwhile, records collected from literature databases may be unrelated to the given domain (Shu et al.2019). Therefore, we suggest that the researchers should pay more attention to the construction methods of dataset rather than the size of it on domain bibliometrics research.

Acknowledgements

This study is supported by the project of the National Social Science Fund for Youth of China (NO. 16CTQ024).

References

Feng, L., Leng, F. (2018). The research about the boundary of data collections of domain analysis based on domain analysis needs and goals. Library intelligence, 53(24), 51-54. (in China)

Shu, F., Julien, C. A., & Larivière, V. (2018). Does the web of science accurately represent Chinese scientific performance?. Journal of the Association for Information Science and Technology.

Shu, F., Julien, C. A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics,* 13(1), 202-225.

Spearman, C. (1904). The proof and measurement of association between two things. American journal of Psychology, 15(1), 72-101.

Author's Name Recognition in Academic Full Text Based on BERT

Zihe Zhu[1],Chuan Jiang[2], Si Shen[3] and Dongbo Wang[4]

[1]*zihe.zhu@qq.com*
Nanjing Agricultural University(China)

[2] *jiangchuan_321@163.com*
Nanjing Agricultural University(China)

[3]shen*si@njust.edu.cn*
Nanjing University of Science and Technology(China)

[4] db.wang@njau.edu.cn
Nanjing Agricultural University(China); KU Leuven（Belgium）

Introduction

Named-entity recognition is one of the entity recognition research. Along with the development of deep learning many new solutions were provided for entity recognition. Habibi et al. (2017) applied LSTM-CRF model to text mining of medical corpus and achieved improvement of recall rate of name entity recognition in medical field. Devlin et al. (2018) from Google released the bidirectional encoder representations from transformers (BERT) that can understand the context. Besides, the author order of multi-authored papers can reveal subtle patterns of scientific collaboration (He et al. 2012). However, there's still a gap in the author extraction in the academic full text. In this study, three deep learning-based models were respectively used for author's name recognition from academic full texts in *Scientometrics* (2010-2018), the recognition results by using different models were compared as well.

Introduction to data source and methods and Analysis of the distribution

Data source

The data were sourced from 2,849 published articles full text on *Scientometrics* (2010-2018)。

Distribution of the author-entity

Based on the format of author's name entities in the texts, the citation of author's names was divided into 3 types, single-author, double-authors and multiple-authors. Then the corpus was labeled artificially, and the distributions of author's name entities are shown in Table 1.

As shown in Table 1, the total occurrence frequency of author's name entities was 138, 331. The occurrence frequency of single-authors was the highest, reaching up to 49,038 and accounting for 35.45%. Overall, different citation situations of author's name entities distributed evenly.

Tab.1 Author-entity types and distribution

Author's Name Type	Count	Percentage
Single-author	49,038	35.45%
Double-author	43,105	31.16%
Multiple-author	46,188	33.39%
Total	138,331	-

Methods

In this paper, three deep learning models - LSTM, LSTM-CRF and BERT - were used. Firstly, the LSTM model is a neural network consisting of three gates (input gate, forget gate and output gate) and one state cell, which makes up for the defect with recurrent neural network (RNN) where the contextual information stored is limited.

Secondly, LSTM-CRF model has the following working principle: On the basis of LSTM, the Softmax layer is replaced by CRF layer as the output layer, thus improving the prediction performance.

Thirdly, BERT is the pre-training language model proposed by Devlin et al. (2018). With BERT, pre-training is already performed on linguistic representation before the training begins. Our experiments were conducted based on the BERT pre-training model (multi-language version).

Extraction of the author named entity

In the training of LSTM and LSTM-CRF models, the corpus was pre-trained into an embedding of 200 dimensions. The number of hidden layers with deep learning training was 256; the number of Bi-LSTM layer was 2, batch size being 32, learning rate 0.001 and epoch 200. The number of epochs before early stopping was 10. Above were the values of hyperparameters.

In the training of BERT model, a large number of pre-experiments were conducted. Finally, the pre-training model with 11 layers, 748 hidden units and 12 self-attention heads was chosen. The values of the hyperparameters were as follows: max sequence length 128, batch size 32, case-sensitive, learning rate 0.001, epoch 200, and number of epochs before early stopping 5.

Then 10-fold cross-validation tests were performed using the three models respectively. The performance of each model was assessed by P, R and F value, and results are summarized as shown in Table 2.

Tab.2 Results of 10-Fold Cross-Validation

Num.	LSTM	LSTM-CRF	BERT
1	95.51%	96.00%	95.83%
2	95.95%	96.33%	96.32%
3	96.14%	96.62%	96.40%
4	96.54%	96.60%	96.74%
5	96.45%	96.67%	96.73%
6	96.07%	96.39%	96.59%
7	95.65%	96.44%	96.38%
8	95.65%	96.10%	96.23%
9	96.25%	96.44%	96.41%
10	96.12%	96.43%	96.45%
AVG_F	96.03%	96.40%	96.41%
Max_F	96.54%	96.67%	96.74%

Based on the experimental results, the maximum and average F-measures of ten tests were obtained, as shown in Fig. 1 and Fig. 2.

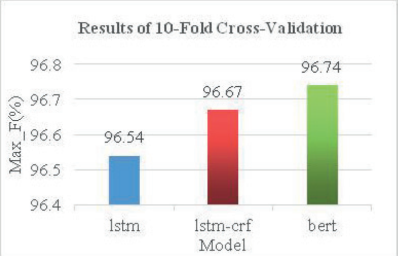


Fig.1 Histogram of maximum F-measure of three models (LSTM, LSTM-CRF and BERT) in 10-fold cross-validation tests

As shown in Fig. 1, the maximum F-measure of BERT reached up to 96.74%, which was higher than that with LSTM-CRF and also than that of LSTM. As shown in Fig. 2, the average F-measure of LSTM-CRF was only lower than that of BERT by 0.01%, and both were higher than that of LSTM. On the whole, the average F-measures of all three models were above 96%, indicating that deep learning had a very good effect in extracting author's name entities from academic full texts.

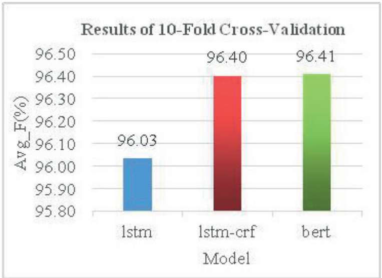


Fig.2 Histogram of average F-measure of three models (LSTM, LSTM-CRF and BERT) in 10-fold cross-validation tests

Conclusion

In this study, three deep learning-based models (LSTM, LSTM-CRF and BERT) were used for extracting author's name entities from academic full texts on *Scientometrics* (2010-2018). The average F-measures of all three models reached well over 96%. The optimal effect was achieved with BERT, for which the maximum F-measure was 96.74%. Moreover, the method in this paper can be helpful for the resarch of author's name in the full text, such as regularities of the names distribution and the author co-occurrence .

Acknowledgments

Researches are supported by the National Natural Science Foundation of China (Grant Number: 71673143).

Reference

He, B. , Ding, Y. , & Yan, E. . (2012). Mining patterns of author orders in scientific publications. Journal of Informetrics, 6(3), 359-367.

Habibi, M. , Weber, L. , Neves, M. , Wiegandt, D. L. , & Leser, U. . (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14), 37-48.

Hochreiter, S. , & Schmidhuber, Jürgen. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

Devlin, J. , Chang, M. W. , Lee, K. , & Toutanova, K. . (2018). Bert: pre-training of deep bidirectional transformers for language understanding.*arXiv preprint arXiv: 1810. 04805v1*

Research on Functional Structure Identification of Academic Text Based on Deep Learning

Youshu Ji[1], Qi Zhang[2], Si Shen[3], Dongbo Wang[4] and Shuiqing Huang[5]

[1] *2018114009@njau.edu.cn*
Nanjing Agricultural University(China)

[2]*grangergogo@gmail.com*
Nanjing Agricultural University (China)

[3]*shensi@njust.edu.cn*
Nanjing University of Science and Technology(China)

[4]*db.wang@njau.edu.cn*
Nanjing Agricultural University(China); KU Leuven(Belgium)

[5]*sqhuang@njau.edu.cn*
Nanjing Agricultural University(China)

Introduction

In conventional research of citation analysis, the position of the cited section and the citing section is not taken into consideration. However, citation behaviour varies according to its location, and citation distribution is an important aspect of citation analysis (Zhang, 2012; Hu, Chen& Liu, 2013). How to automatically identify the functional structure of documents has become an important problem which could be solved from the following two aspects. On one hand, the subheads can be used for preliminary judgment, but the subheads are not arranged in an orderly manner. In many cases, readers cannot judge the functional structure of the section through its header. On the other hand, we can judge the functional structure of one section more accurately by its content. Some researchers (Lu, Huang& Bu, 2018) do related research with conventional machine learning methods. However, the conventional machine learning method needs manual extraction of features, which requires a large workload. Our research chooses the deep learning model to automatically identify the functional structure of academic text based on section content, in an attempt to lay a foundation for the research on citation location distribution.

Introduction to Data Sets and Methods

In this study, papers published in *JASIST* on 2006-2016 are obtained. After removing incomplete and erroneous papers, 1192 valid papers are saved by section, and the title information of each section is also saved. In the end, there are 7134 sections obtained. On average, each paper consists of 6 sections, and each section averages 1193 words.

Then, according to section header and content, the functional structure of all the sections are classified into five categories, namely “Introduction(I)”, “Related Research(R)”, “Method(M)”, “Experiment(E)” and “Conclusion(C)”. After manual labelling, a relatively large-scale academic full-text corpus with precise functional structure markers has been constructed. The five categories of text structures appear 1275, 1116, 1173, 1856, and 1714 times, respectively.

In previous researches, the effect of conventional machine learning model on functional structure identification cannot meet expectations without extracting artificial features. In our experiment, conventional machine learning model and deep learning model are adopted to transform structure identification of academic text into text classification. The identification effect of functional structure of academic text is studied separately, under SVM (Support Vector Machine) model, Text-CNN (Convolutional Neural Networks) model, and BERT (Bidirectional Encoder Representations from Transformers) model. The SVM model is developed from researching the optimal classification hyperplane in the case of linearly separable of training data, and then extended to the data classification in high-dimensional space, which has better classification effect. Text-CNN (Kim, 2014) model uses multiple kernels of different sizes to extract key information in sentences, which can capture local correlations better. BERT (Devlin, Chang& Lee, 2018) is a pre-trained method for language representation, which adopts pre-trained unsupervised, deep bidirectional system to obtain a general language representation model.

Experiment

We select the BERT pre-trained model of 11 layers, 748 hidden units and 12 self-attention heads released by Google to carry out our formal experiment after a large number of pre-experiments.

We input our corpus into pre-trained model for fine tuning to fit the model to this task. The following results are obtained by 10-fold cross-validation:

Table 1. Identification performance for functional structure of academic text under different models

Model	AVG			MAX		
	P	R	F	P	R	F
SVM	62.40	62.77	62.59	63.79	64.73	64.26
CNN	69.77	70.06	69.91	74.38	74.24	74.31
BERT	**82.94**	**82.68**	**82.81**	**84.18**	**84.33**	**84.25**

It can be concluded from Table 1 that without adding artificial features, the SVM model has the worst effect, the Text-CNN model has a poor classification effect and the BERT model has the best classification effect. BERT is a short text classification model. The pre-trained model is used to truncate characters, and only the first 128 characters are intercepted in each section. In this case, the BERT model still achieves the best results. On one hand, it embodies the superiority of the BERT model. On the other hand, it also proves that the first part of the section plays a greater role in the identification of the functional structure.

Statistical Analysis

We present the identification result of BERT model which earns the best identification effect. The distribution of positive identification result and negative identification result are shown below.

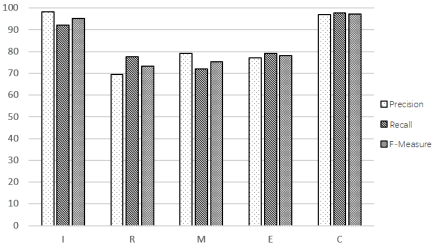


Figure 1. Positive identification result of each label (%)

Table 2. Negative identification result of each confusion

	Error type	Percentage
1	I——R/R——I	6.25%
2	I——M/M——I	1.79%
3	I——E/E——I	2.68%
4	I——C/C——I	0.00%
5	R——M/M——R	14.29%
6	R——E/E——R	**33.04%**
7	R——C/C——R	2.68%
8	M——E/E——M	**33.93%**
9	M——C/C——M	0.89%
10	E——C/C——E	4.46%

The precision, recall, and F1-score of each label are compared in Figure 1. This model has a better identification effect on the labels "I" and "C", namely, “Introduction” and “Conclusion”. Table 2 shows the negative identification result of BERT model, wherein the errors of the confused three types of text structure, namely “Related Research”, “Method” and “Experiment” account for two-thirds of all errors. The errors are caused by the different writing habits of the author and the different characteristics of research field thereof. A better distinction between these three categories is the focus of future research.

Conclusions and Future Research

In this research, the deep learning model BERT is selected. Without the need to manually extract features, the optimal F1 score of 84.25% is obtained in the functional structure identification, which has a good identification effect. However, there is still room for further improvement in three types of structure identification, namely, “Related Research”, “Method” and “Experiment”. The construction of this model provides a more efficient tool for the automatic identification of the functional structure of academic text, which helps to add text functional structure markers to unlabelled texts in future citation analysis, so as to solve the problem that the current citation location analysis is limited to manually labelled data sets.

Acknowledgments

I would like to thank the members of 748 Society of Nanjing Agricultural University for their support and help in acquiring and annotating JASIST Academic Full Text Corpus.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv*:1810.04805.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, 7(4), 887-896.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv*:1408.5882.

Lu, W., Huang, Y., Bu, Y., & Cheng, Q. (2018). Functional structure identification of scientific documents in computer science. *Scientometrics*, 115(1), 463-486.

Zhang, L. (2012). Grasping the structure of journal articles: utilizing the functions of information units. *Journal of the American Society for Information Science and Technology*, 63(3), 469-480.

A Longitudinal Study of Questionable Journals in Scopus

Jinseo Park[1], Jinhyuk Yun[2] and June Young Lee[3]

[1] *jayoujin@kisti.re.kr*
Future Technology Analysis Center, Korea Institute of Science and Technology Information, 66 Hoegiro,
Dongdaemun-gu, Seoul, 02456 (South Korea)

[2] *jinhyuk.yun@kisti.re.kr*
Future Technology Analysis Center, Korea Institute of Science and Technology Information, 66 Hoegiro,
Dongdaemun-gu, Seoul, 02456 (South Korea)

[3] *road2you@kisti.re.kr*
Future Technology Analysis Center, Korea Institute of Science and Technology Information, 66 Hoegiro,
Dongdaemun-gu, Seoul, 02456 (South Korea)

Introduction

The issue of predatory journals has been steadily debated with the rise of the open access journals as a matter related to the trust of the scholarly communication ecosystem. After German media reported "fake science" in 2018, the problem of fake conferences and journals has emerged as a social issue beyond academic fields in South Korea.

The predatory publishers and journals, predatory conference or predatory meetings, pseudo journals, hijacked journals and fake science, which emphasize the undesirable aspects of the scholarly communication ecosystem, imply different meanings in detail, but we will describe all of these terms as 'questionable'. 'Questionable' means a deviation from the normal practice of the scholarly ecosystem. At this point, some criteria should be provided to determine whether any academic journals or academic conference is normal or not. A typical blacklist is Beall's list, but the debate over whitelisting and blacklisting continues. Due to the lack of objective lists of which journals are questionable, it is very rare to find out the scale and scope of questionable journals. Nevertheless, previous studies have investigated the scale and trend of the questionable journals based on Beall's list. (Shen & Björk, 2015; Sterligov & Savina, 2016; Macháček & Srholec, 2017; Bagues, Sylos-Labini & Zinovyeva, 2019)

It is almost impossible to figure out the entire questionable journals. Shen & Björk (2015) estimated that the number of articles published by 8,000 questionable journals has increased from 53,000 to 420,000 in 2014. Our aim is not to estimate the 'real and accurate' volumes of questionable journals and the articles published by questionable journals. In this study, we will review the longitudinal trends of documents published by questionable journals based on comparable data, especially comparison by country and subject. We are also interested in South Korea's relative position in the global trend.

Data and Methods

We used Beall's list (original and update list until March 1, 2019) as the criteria for questionable journals. We used two methods for matching questionable journals with Scopus DB. First, we collected all ISSN in Beall's list (publishers and standalone journals, https://beallslist.weebly.com/) through crawling, and matched with Scopus DB. Second, we searched every publisher of Beall's list in Scopus DB. We identified 766 journals included in Scopus DB (1996~2018) matching the publishers and journals in Beall's list.

Results

Trends over year – Global vs South Korea

Figure 1 shows that the number of documents in Scopus and questionable journals (QJ). The Scopus documents increased linearly (blue), but QJ documents increased exponentially (red) and more rapidly than Scopus total documents.

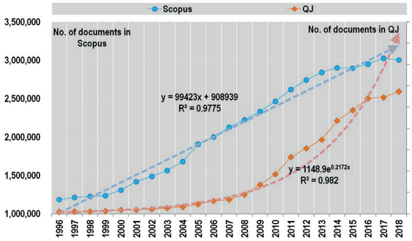


Figure 1. Number of documents in Scopus and Questionable journals (QJ)

Figure 2 reveals that the share of QJ documents in Scopus increase rapidly and the growth rate of South Korea (red) is higher than Scopus (blue).

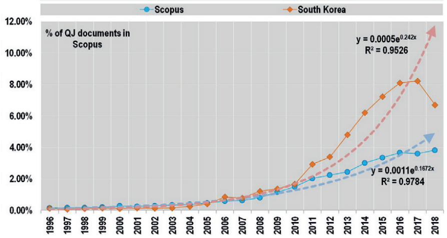


Figure 2. Share of QJ documents in Scopus – Global and South Korea

Cross-country comparisons

Figure 3 shows the share of QJ documents by country. Similar to Sterligov & Savina (2016) and Macháček & Srholec (2017), the situation in South Korea is the worst. Comparing the entire period (1996~2018) with the last five years (2014~2018), the matter of South Korea has become worse.

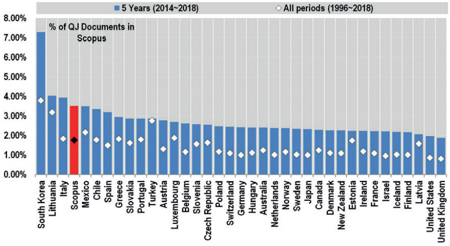


Figure 3. Share of QJ documents in Scopus by country (OECD) - 1996~2018 and 2014~2018

Cross-subject comparisons

There was a difference in share of QJ documents by subject area. In the whole period, the 'Pharmacology, Toxicology and Pharmaceutics' was 5.29%, 'Multidisciplinary' 3.93%, and 'Environmental Science' 3.59%. (Figure 4) Over the recent 5 years, the share of QJ documents in 'Pharmacology, Toxicology and Pharmaceutics' has also the highest at 10.49%.

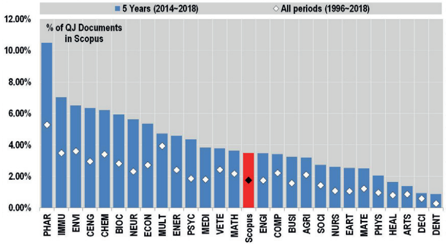


Figure 4. Share of QJ documents in Scopus by Subject Area - 1996~2018 and 2014~2018

Position of South Korea

We compared the share of QJ by subject between the all Scopus DB and the South Korea. Of 27 subject areas, only five subjects were lower than the average of Scopus. (Figure 5)

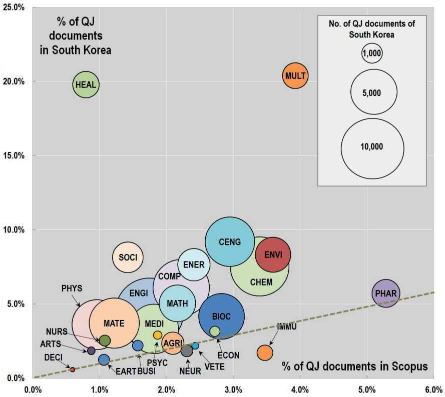


Figure 5. Share of QJ documents by subject - Scopus and South Korea

Conclusion

The term 'questionable journal' is a problem related to social trust in scholarly communication ecosystem and knowledge produced by the very same system. South Korea depends mainly on external authority for 'qualitative criteria' such as WoS or Scopus. We cautiously presume that the excessive QJ share of South Korea is a natural consequence of the individual researcher's rational choice steeped in this evaluation culture. In the future, it is necessary to study what institutional and cultural factors influence this increase and spread of questionable journals.

References

Bagues, M., Sylos-Labini, M. & Zinovyeva, N. (2019). A walk on the wild side: 'Predatory' journals and information asymmetries in scientific evaluations. *Research Policy*, 48:2, 462-477.

Macháček, V. & Srholec, M. (2017). *Predatory journals in Scopus*. A Project of the Economic Institute of the Czech Academy of Sciences. Institute for Democracy and Economic Analysis.

Shen, C. & Björk, B.-C. (2015). 'Predatory' open access: a longitudinal study of article volumes and market characteristics. *BMC Medicine*, 13:1, 230.

Sterligov, I. & Savina, T. (2016). Riding with the metric tide: 'predatory' journals in Scopus. *Higher Education in Russia and Beyond*. 1:7, 9–12.

Impact of National Research Assessment Exercises on Monographs and Scholarly Books authored by the Lithuanian Researchers

Eleonora Dagienė[1], Andrius Kriščiūnas[2], Gintarė Tautkevičienė[3] and Saulius Maskeliūnas[4]

[1] *e.dagiene@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

[2]*andrius.krisciunas@ktu.lt*, [3]*gintare.tautkeviciene@ktu.lt*
Kaunas University of Technology, K. Donelaičio g. 20, LT-44239 Kaunas (Lithuania)

[4] *saulius.maskeliunas@mii.vu.lt*
Institute of Data Science and Digital Technologies, Vilnius University, Vilnius (Lithuania)

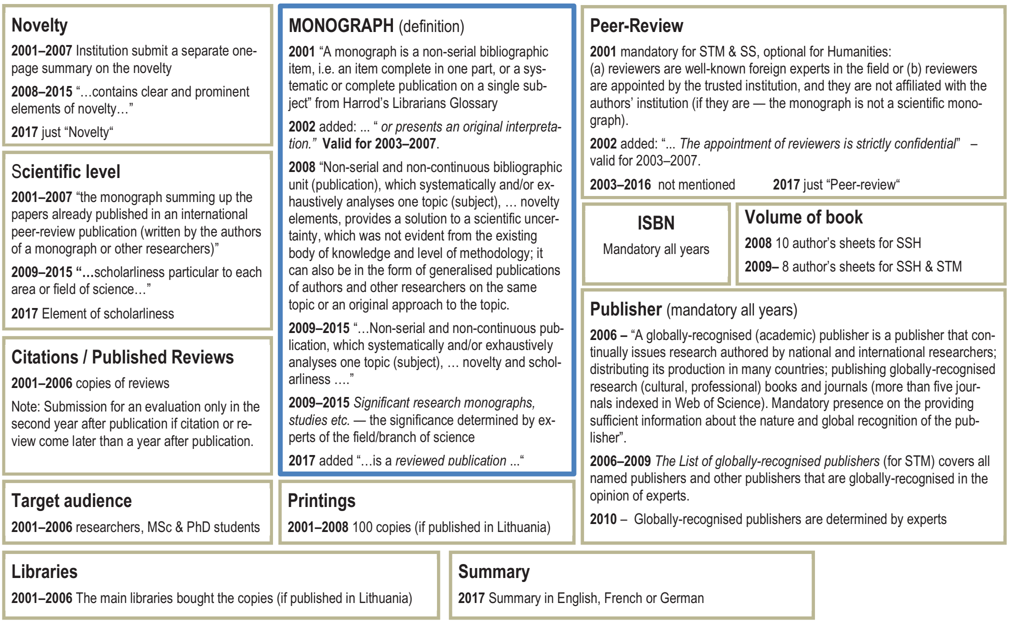
Introduction

For many years, scholarly books have been and continue to be an important channel of scholarly communication and a unit for research assessment. Books as a mean of scholarly publication are more frequently used in social sciences and humanities (SSH), yet remain important for the communication in science, technology and medicine (STM) (Bonaccorsi et al., 2017). However, in the context of research assessment, books and especially monographs are extensively discussed as an output in crisis due to the threats arising in relation to their quality and production (Basili and Lanzillo, 2018). In this work, we present a comprehensive and still on-going study on the assessment of books in Lithuania to fill in the gaps in the international knowledge on that topic. The goal of the study is to investigate the impact of National Research Assessment Exercises (RAE) on Monographs and Scholarly Books authored by the Lithuanian Researchers over the period of 2004 to 2016.

From 2005 to 2017, Lithuanian policy makers have designed and approved the evaluation methodologies after the books had been published. So, the research institutions were not able to prepare in advance for the evaluation. In 2018, the most current and valid Lithuanian legislative acts encourage participants of the research system to strive for the best results and choose the best publishers to disseminate their outcomes.

Methodology used in this research

Firstly, the method of document analysis was used to analyse how the evolution of requirements for research outputs such as monographs and other scholarly books occurs by the changes in the Lithuanian legal acts from 2001 to 2018.



Novelty

2001–2007 Institution submit a separate one-page summary on the novelty

2008–2015 "…contains clear and prominent elements of novelty…"

2017 just "Novelty"

Scientific level

2001–2007 "the monograph summing up the papers already published in an international peer-review publication (written by the authors of a monograph or other researchers)"

2009–2015 "…scholarliness particular to each area or field of science…"

2017 Element of scholarliness

Citations / Published Reviews

2001–2006 copies of reviews

Note: Submission for an evaluation only in the second year after publication if citation or review come later than a year after publication.

MONOGRAPH (definition)

2001 "A monograph is a non-serial bibliographic item, i.e. an item complete in one part, or a systematic or complete publication on a single subject" from Harrod's Librarians Glossary

2002 added: ... " *or presents an original interpretation."* **Valid for 2003–2007**.

2008 "Non-serial and non-continuous bibliographic unit (publication), which systematically and/or exhaustively analyses one topic (subject), … novelty elements, provides a solution to a scientific uncertainty, which was not evident from the existing body of knowledge and level of methodology; it can also be in the form of generalised publications of authors and other researchers on the same topic or an original approach to the topic.

2009–2015 "…Non-serial and non-continuous publication, which systematically and/or exhaustively analyses one topic (subject), … novelty and scholarliness …."

2009–2015 *Significant research monographs, studies etc.* — the significance determined by experts of the field/branch of science

2017 added "…is a *reviewed publication* ..."

Peer-Review

2001 mandatory for STM & SS, optional for Humanities:
(a) reviewers are well-known foreign experts in the field or (b) reviewers are appointed by the trusted institution, and they are not affiliated with the authors' institution (if they are — the monograph is not a scientific monograph).

2002 added: "... *The appointment of reviewers is strictly confidential*" – valid for 2003–2007.

2003–2016 not mentioned **2017** just "Peer-review"

ISBN

Mandatory all years

Volume of book

2008 10 author's sheets for SSH

2009– 8 author's sheets for SSH & STM

Publisher (mandatory all years)

2006 – "A globally-recognised (academic) publisher is a publisher that continually issues research authored by national and international researchers; distributing its production in many countries; publishing globally-recognised research (cultural, professional) books and journals (more than five journals indexed in Web of Science). Mandatory presence on the providing sufficient information about the nature and global recognition of the publisher".

2006–2009 *The List of globally-recognised publishers* (for STM) covers all named publishers and other publishers that are globally-recognised in the opinion of experts.

2010 – Globally-recognised publishers are determined by experts

Target audience

2001–2006 researchers, MSc & PhD students

Printings

2001–2008 100 copies (if published in Lithuania)

Libraries

2001–2006 The main libraries bought the copies (if published in Lithuania)

Summary

2017 Summary in English, French or German

Figure 1. Requirements for monographs as institutional outputs in 2001–2017

Secondly, the bibliometric analysis of the records of scholarly books submitted to RAEs was performed. The outputs as records were accumulated into the database managed by the Lithuanian Research Council (LRC) and used for RAE 2005–2017.

Finally, ISBN codes and information about publishers and countries of origin were specified using additional sources: (1) The Lithuanian Academic Electronic Library; (2) The National Bibliography Data Bank; (3) Worldcat, OECD; (4) Global Register of Publishers, International ISBN agency.

Following the specification of metadata, the further analysis focused on 4135 books with ISBN published in 2004–2016 that were submitted by institutions for the annual evaluation in 2005–2017.

Correlations between the types and numbers of books submitted for the assessment and the changes in methodologies

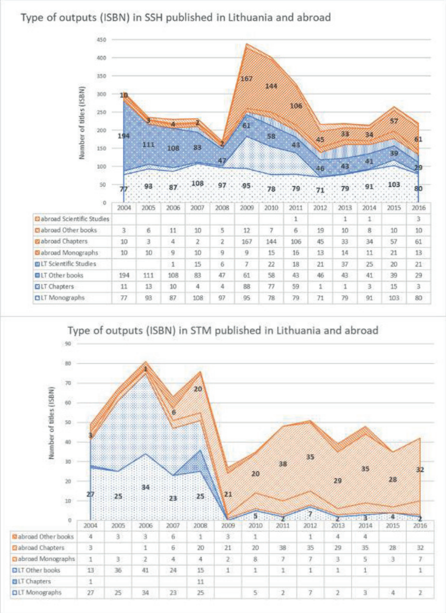
Over the analysed period (2001–2018), Lithuanian methodologies for research evaluation changed frequently and every new legal act had some changes in requirements for types of publications with ISBN.

The concept of a *monograph* was first mentioned in the Lithuania legislation in 2001, in *Regulations for requirements applicable to research monographs*. Because of page limitations, we cannot give a detailed explanation of the changes over the time. However, Figure 1 shows a high-level overview in the requirements for monographs over the period 2001-2017.

The methodologies for RAE were prepared by two separate groups of national experts in SSH and STM. Differences in the methodologies for SSH and STM made the impact on the types of publications (monographs, chapters in edited volumes, etc.), countries and publishers which issued the books authored by the Lithuanian researchers (Fig. 2).

The grouping of publications by type and area revealed that the greatest fluctuations occurred in SSH between 2009 and 2011 and in STM between 2008 and 2016. It can be linked to a new type that was introduced in the methodologies, namely, chapters/papers in edited books in 2009 (STM) and in 2010 (SSH). Since 2010 in STM, institutions received points and, respectively, more funding exceptionally for research monographs or book chapters issued by prestigious publishers. This explains why the level of monographs in STM is so low from year 2009. As can be seen in Fig. 2, while books in STM were mostly published abroad, most books in SSH issued in Lithuania. Chapters in edited volumes comprised more than 90% of the STM production since 2010 when book chapters started to be considered along with monographs. Very few STM monographs are published abroad.

Meanwhile, in the case of SSH disciplines, monographs, book chapters and other publications published in Lithuania or with non-prestigious foreign publishers earn less points only. The difference was not significant for books issued by global-recognised publishers. Such publishers are determined by experts since 2010. The definition used is quite simple and is presented in Figure 1.


Type of outputs (ISBN) in SSH published in Lithuania and abroad

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
abroad Scientific Studies								1		1	1		3
abroad Other books	3	6	11	10	5	12	7	6	19	10	8	10	10
abroad Chapters	10	3	4	2	2	167	144	106	45	33	34	57	61
abroad Monographs	10	10	9	10	9	9	15	16	13	14	11	21	13
LT Scientific Studies			1	15	6	7	22	18	21	37	25	20	21
LT Other books	194	111	108	83	47	61	58	43	46	43	41	39	29
LT Chapters	11	13	10	4	4	88	77	59	1	1	3	15	3
LT Monographs	77	93	87	108	97	95	78	79	71	79	91	103	80

Type of outputs (ISBN) in STM published in Lithuania and abroad

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
abroad Other books	4	3	3	6	1	3	1		1	4	4		
abroad Chapters	3		1	6	20	21	20	38	35	29	35	28	32
abroad Monographs	1	3	2	4	4	2	8	7	7	3	5	3	7
LT Other books	13	36	41	24	15	1	1	1	1	1	1		1
LT Chapters	1				11								
LT Monographs	27	25	34	23	25		5	2	7	2	3	4	2

Fig. 2. Publications in SSH and STM by type and publishing in Lithuanian and abroad

Conclusions

This research is on its starting point and still going on, so more detailed conclusions will be presented later. The results of this study point to the conclusion that changes in the research assessment system affect the researchers' choice between publishing a monograph or publication of smaller volumes, e.g., the article or chapter in an edited book.

The long-term requirement on publishers (e.g., publishing with the globally recognised publishers) led to an increase in the number of publications published abroad, especially in STM.

References

Basili, C., Lanzillo, L., 2018. Research Quality Criteria in the Evaluation of Books, in: The Evaluation of Research in Social Sciences and Humanities. Springer International Publishing, Cham, pp. 159–184.

Bonaccorsi, A., Daraio, C., Fantoni, S., Folli, V., Leonetti, M., Ruocco, G., 2017. Do social sciences and humanities behave like life and hard sciences? Scientometrics 112, 607–653.

Giménez-Toledo, E., Mañana-Rodríguez, J., Engels, T.C.E.E., Guns, R., Kulczycki, E., Ochsner, M., Pölönen, J., Sivertsen, G., Zuccala, A.A., 2019. Taking scholarly books into account, part II: a comparison of 19 European countries in evaluation and funding. Scientometrics 118, 233–251.

Interdisciplinary Research Based on Paper-level Classifications of Science- A Preliminary Case Study of Chinese Journals

Bikun Chen[1], Mengxia Cheng[2], Peiyao Li[3] and Yuefen Wang[4]

[1] *chenbikun@njust.edu.cn*
Nanjing University of Science and Technology, Nanjing (China)

[2] *1291361237@qq.com*
Nanjing University of Science and Technology, Nanjing (China)

[3] *980640538@qq.com*
Nanjing University of Science and Technology, Nanjing (China)

[4] *yuefen163@163.com*
Nanjing University of Science and Technology, Nanjing (China)

Introduction

Interdisciplinary research has become increasingly popular in scientometrics. Most researches up to now rely on journal-level classifications of science. By this approach, papers could be misclassified in journal classification systems (Shu et al. 2019), which may cause bias, especially if there is a significant proportion of multidisciplinary journals in the reference list (Zhang et al. 2010).

Different from international journal-level classification system (eg. WoS and Scopus Categories), Chinese bibliographic databases classify science at the paper-level using the Chinese Library Classification Scheme (CLC) (Chinese Library Classification 2010). CLC is also used by publishers in China to classify all publications including books, monographs, and journals. As a supplement to interdisciplinary research based on journal-level classification, it is meaningful to investigate interdisciplinary with CLC. In this study, interdisciplinary research is conducted with articles published in twenty Chinese journals at paper-level classifications of science.

Data

The data in this study consist of 55,894 articles (only research articles are kept) published from 2008 to 2018 in twenty core Chinese journals that belong to "Library, information and archival science" indexed by CSSCI (Chinese Social Sciences Citation Index) (http://cssrac.nju.edu.cn/index.html). The data are crawled between March 10th and March 24th, 2019 and manipulated between March 25th and April 5th, 2019.

Method

In Figure 1, relations of Chinese journal-level and article-level CLC codes are depicted. Every Chinese journal has a tier-2 CLC code and each article has at least one tier-2, 3, 4, 5 or 6 CLC code. Based on Figure 1 and the interdisciplinary theories constructed by Pierce (1999), two methods are designed to construct relations of CLC codes in different level.

Method 1 (undirected network): Co-occurrence of article-level CLC codes. When any article-level CLC code i and j co-occur in any article A, their co-occurring value is number one. The total relation intensity φ between CLC code i and j is the sum of CLC code i and j co-occur in any article A.

$$\emptyset_{i\,j} = \sum_{A} One(CLC_i, CLC_j)$$

Method 2 (directed network): Article-level CLC code points to journal-level CLC code. When any article-level CLC code k points to journal-level CLC code l in any article B, their value is number one. The total relation intensity φ between CLC code k and l is the sum of CLC code k points to l in any article B.

$$\emptyset_{k\,l} = \sum_{B} One(CLC_k, CLC_l)$$

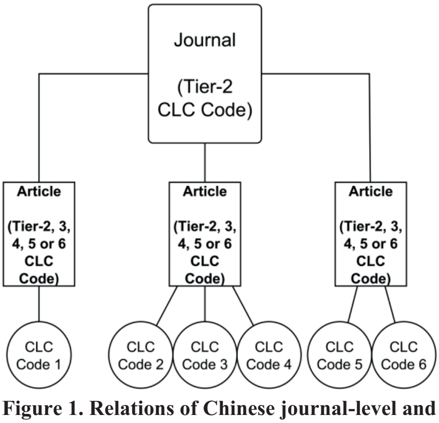


Figure 1. Relations of Chinese journal-level and article-level CLC Codes

Results

Due to the limited pages, only undirected and directed networks of tier-2 categories are shown below. Figure 2 is drawn by Force Atlas 2 layout algorithm and figure 3 is drawn by Circular layout algorithm (Nodes are ordered by decreasing output degree in counter clockwise direction). From Figure 2 to 3, each node indicates a category (labeled by CLC code), the node size indicates its weighted degree (undirected network) or weighted output degree (directed network), the links between nodes reveal relations of categories.

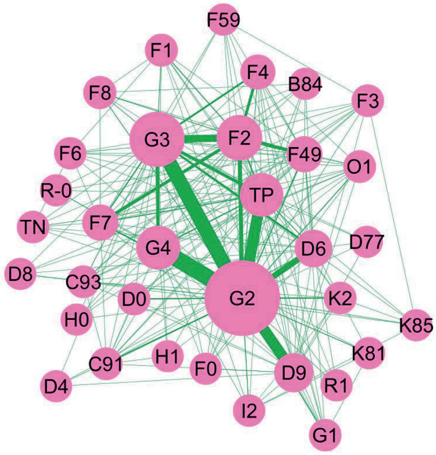


Figure 2. Tier-2 category undirected network (weighted degree >= 50)

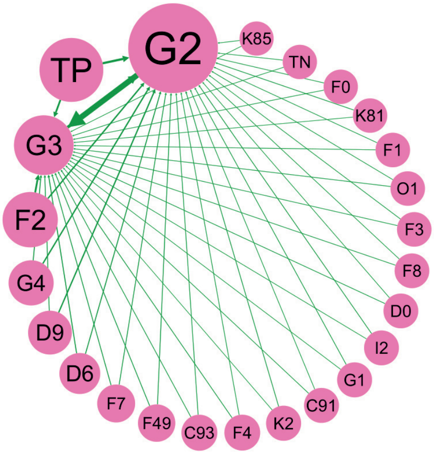


Figure 3. Tier-2 category directed network (weighted output degree >= 76)

In Figure 2, category "Information and Knowledge Dissemination" is most closely related to "Science & Science Studies", "Education", "Law", "Automation & Computer Technology", "Chinese Politics" and "Economy Planning & Management" (detailed in Table 1). In Figure 3, category "Automation & Computer Technology" contributed more knowledge to "Information and Knowledge Dissemination" and "Science & Science Studies" than other disciplines. Between "Information and Knowledge Dissemination" and "Science & Science Studies", the former contributed more knowledge than the other.

In this study, only twenty "Library, information and archival science" journals are investigated and the results are relatively limited. In further research, Chinese journals of all disciplines will be incorporated.

Table 1. Top 20 tier-2 categories of CLC in Figure 2 (ranked by weighted degree).

Code	Category
G2	Information and Knowledge Dissemination
G3	Science & Science Studies
F2	Economy Planning & Management
G4	Education
TP	Automation & Computer Technology
D9	Law
D6	Chinese Politics
F7	Trade Economy
F49	General Information Industry
F4	Industrial Economy
F8	Public Finance & Finance
F1	International Economy
C91	Sociology
O1	Mathematics
K2	Chinese History
C93	Management Science
D0	Political Theory
F3	Agricultural Economy
G1	International Culture
TN	Radio Electronics & Telecommunications

Acknowledgments

This paper is supported by Youth Program of National Social Science Fund in China (15CTQ035) and Research Training Program by NJUST (2018).

References

Pierce, S. J. (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer. Journal of the American Society for Information Science, 50(2), 271-279.

Shu, F., Julien, C., Zhang, L., Qiu, J., Zhang, J., & Lariviere, V. (2019). Comparing journal and paper level classifications of science. Journal of Informetrics, 13(1), 202-225.

Zhang, L., Janssens, F. A., Liang, L., & Glänzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. Scientometrics, 82(3), 687-706.

Zhongguo Tushuguan Fenleifa [Chinese Library Classification] (2010). (5 ed.). Beijing: National Library of China Publishing House.

Determining Citation Blocks using End-to-end Neural Coreference Resolution Model for Citation Context Analysis: a pilot study with seven PLOS journals

Marc Bertin[1] and Pierre Jonin[2] and Frederic Armetta[3] and Iana Atanassova[4]

[1] *marc.bertin@univ-lyon1.fr*
[2] *pierre.jonin@etu.univ-lyon1.fr*
Laboratoire ELICO, Université Claude Bernard Lyon 1
Bâtiment Nautibus 43 Boulevard du 11 novembre 1918 69622 Villeurbanne cedex (France)

[3] *frederic.armetta@univ-lyon1.fr*
Laboratoire LIRIS, Université Claude Bernard Lyon 1
Bâtiment Nautibus 43 Boulevard du 11 novembre 1918 69622 Villeurbanne cedex (France)

[4] *iana.atanassova@univ-fcomte.fr*
Centre Tesnière - CRIT, Université de Bourgogne Franche-Comté
30 rue Mégevand, 25030 Besançon Cedex (France)

Introduction and Research Problem

The study of citation contexts is an important element in understanding the function of citations and categorizing the relationships between works. We hypothesize that the space of citation contexts must be extended beyond the sentence and within a space delimited by criteria of a semantic or linguistic nature and not quantitative, i.e. according to a window delimited by numerical values. In this paper we propose the definition of citation blocks (CB) that are composed of one or more sentences that are linked by coreference clusters. The processing of semantic-pragmatic phenomena such as anaphora, cataphora and deixis is of central importance in the analysis and categorization of citation acts. Our aim is to define meaningful textual spaces for the analysis of citation contexts through the study of anaphoric relationships and more specifically coreferences.

A lot of research is based on the identification of textual spaces (TS) e.g. argumentative zones (Teufel, 1999) or the IMRaD structure with sentences which, from a linguistic point of view, represent a unit of meaning (Bertin et al., 2016). We can also choose the size of a window, variable or not, which determines the context around an in-text reference (Ritchie, Robertson and Teufel, 2008).

Research Problem

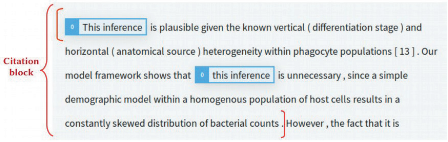
Coreferences and anaphoric relations use the notion of cohesion to define the nature of the anaphoric relationship. A referential object is called an anaphora when it refers to its antecedent. It may be a previously introduced expression but does not necessarily designate the same entity as that expression. The anaphora may be grammatical, lexical, nominal or pronominal in nature, but also adverbial, verbal, summarizing, associative, etc., underlining the complexity of this phenomenon. A coreference can be defined as a reference to the same entity whose context alone can establish the link between the two expressions. This can lead to the successive identification of corefential chains. Contextual and coreferential space from a linguistic point of view, refers to "the immediate environment" as the "linguistic context" for anaphors and "the immediate denunciation situation" for deictics.

Method

In order to determine the citation blocks (CB), we propose to study co-referential relationships in order to determine the size of this co-referential space. To identify all textual elements that belong to coreference clusters we annotated the dataset using AllenNLP libraries (Gardner, 2017), which implements end-to-end coreference resolution model (Lee et al. 2017). Coreference clusters are sets of text elements, that can be words or sequences of words. The elements of a coreference cluster can belong to the same sentence or to different sentences. In the later case, the coreference cluster establishes a link between these different sentences. As an example, figure 1 shows the textual space around an in-text reference. The expressions “this inference” belong to a coreference cluster and are in the first and second sentence. The citation block (CB) is thus delimited by these two sentences.

Dataset

For our experiment, we have processed a dataset that is composed of in-text references and their contexts chosen randomly from the 7 PLOS journals, wich 10,000 citation contexts from each journal.



This inference is plausible given the known vertical (differentiation stage) and horizontal (anatomical source) heterogeneity within phagocyte populations [13] . Our model framework shows that this inference is unnecessary , since a simple demographic model within a homogenous population of host cells results in a constantly skewed distribution of bacterial counts . However , the fact that it is

Figure 1. Coreference citation blocks.

Identifying the textual space (TS) around citations

For each in-text reference, we first identify the TS that can possibly be related to the reference through the use of coreference and anaphoric expressions in the following way: TS is composed of the sentence containing the in-text reference and all the following sentences until a new in-text reference is encountered within the same paragraph. In fact, we consider two types of boundaries that delimit the TS: paragraph breaks and the presence of other references. When a new in-text reference in encountered in a paragraph, we suppose that the sentences immediately following this reference could be related to it, provided that they do not contain other in-text references.

We consider that sentences that contain elements of the same coreference cluster should belong to the same citation block (CB). Given an in-text reference and its TS, we consider that the beginning of the CB is the sentence containing the in-text reference and the end of the CB is the last sentence in TS that is linked to this first sentence by the coreference clusters. In the case when the TS is composed of only one sentence, there is no need to identify the coreference clusters as the citation block is also composed of one sentence.

Results

We observe the trends in the different journals and section types of the IMRaD structure or articles. Table 1 presents the numbers and percentages of TS with 1 sentence (49.74%) and with two or more sentences. The latter are divided in two groups: TS without coreference clusters (9.16%) and TS with 1 or more coreference clusters (41.10%).

Table 1. Numbers of TS with 1 sentence and 2 or more sentences in IMRaD

	I	M	R	D	Total
Nb TS: 1 sentence	64.99%	39.45%	36.94%	47.53%	49.74%
Nb TS: 2 or more sentences:	35.01%	60.55%	63.06%	52.47%	50.26%
With 0 coreference clusters	9.90%	8.91%	7.12%	10.09%	9.16%
With 1 or more coreference clusters	25.11%	51.64%	55.94%	42.39%	41.10%
Total	100.00%	100.00%	100.00%	100.00%	100.00%

The further analysis will be done on the TS with 1 or more coreference clusters in order to delimit the citation blocks (CB) and evaluate the difference in the size of TS and CB that we obtain.

Discussion and Conclusion

The perspectives around this work focus on the problems of identifying coreference and anaphoric relationships with deep neural networks. The limits of this approach are the nature of the coreference resolution tools, which must be finer and offer more detailed analyses. It is necessary to evaluate and improve this identification by proposing learning dataset for the specific processing of scientific articles. This citation block model should eventually make it possible to better understand the nature of citation acts, to have a consensus on the spaces that carry information for the semantic categorization of citation contexts and to propose finer corpora dedicated to this task.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Bertin, M.; Atanassova, I.; Gingras, Y. & Larivière, V. (2016) The invariant distribution of references in scientific articles JASIST, 67(1), 164–177.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M. & Zettlemoyer, L. AllenNLP: A Deep Semantic Natural Language Processing Platform Proceedings of Workshop for NLP Open Source Software (NLP-OSS), ACL, 2018, 1–6

Lee, K., He, L., Lewis, M., & Zettlemoyer, L.S. (2017). End-to-end Neural Coreference Resolution. EMNLP.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In Joint Conference on EMNLP and CoNLL-Shared Task, 1–40. ACL.

Ritchie, A., Robertson, S., & Teufel, S. (2008). Comparing citation contexts for information retrieval. In 17th CIKM Proceedings, 213–222.

Teufel, S. (1999) "Argumentative zoning: Information extraction from scientific text" (1999) Citeseer, Citeseer, Phd thesis.

Evidence-based Nomenclature and Taxonomy of Research Impact Indicators

Mudassar Arsalan[1], Omar Mubin[1]
19316071@student.westernsydney.edu.au, O.Mubin@westernsydney.edu.au
School of Computing, Engineering and Mathematics, Western Sydney University, Sydney (Australia)

Abdullah Al Mahmud[2]
aalmahmud@swin.edu.au
Swinburne University of Technology Melbourne (Australia)

Introduction

Research Impact (RI) is a broad topic of scientometrics to support the progress of science and monitoring the influence of efforts made by the government, institutions, societies, programs and individual researchers. There are several documented and popular RI assessment methods developed by individuals and organizations for evaluating the research of a particular programme or general purpose. This intent has created the diversity in evaluation methods, frameworks and scope. Some methods focus only on the impacts related to academic recognition and use such as Bibliometric Measures. However, with the growing technology, academic networking, effective and targeted research strategies, and regular monitoring of RI are reducing the gap between the research producers and consumers. As a result, the horizon of RI is being broadened and covering other areas of impacts such as on economy, society, and environment.

Many individuals and organizations have introduced measures and indicators for assessing the RI. However, due to diversity in nature and scale of RI, not a single method is considered robust and complete (Vinkler, 2010). Therefore, new measures and indicators are being introduced on a time to time basis according to the interest and availability of resources of the method designers (CAHS, 2009). Additionally, higher availability of national and international funding for health sciences is critically influencing the science of RI assessment (Heller and de Melo-Martín, 2009). It means there are more indicators, measures, and frameworks for health-related research than any other areas of science. Resultantly, there is a huge gap available for generalizability and transformability of health related efforts to rest of the science.

This study aims to discover the evidence-based diversity of RI indicators and to develop a method, which can lead the generalizability and transformability of previous efforts. Nomenclature of RI indicators is developed based on divide and rule principal to achieve the generalizability. Additionally, taxonomical analysis is presented based on the nomenclature. This effort is a step forward to develop a robust and inclusive RI assessment method.

Method

We systematically searched the literature databases including Scopus, WebMD, ACM DL, IEEE Xplore, Web of Science and Google Scholar to collect research articles providing RI assessment indicators and methods. In many cases organizations published the methods and guidelines in form of technical reports therefore, grey literature was also considered.

A list of around 120 indicators was prepared for detail analysis. For deciphering the nomenclature, indicators were disintegrated based on their lexical and conceptual structures as discussed in the Results section. Nvivo 12 software was used to quantify the proportions of parts of nomenclature.

Results and Discussion

Nomenclature

The base of the cognitive structure of defined nomenclature in this study is the 'every indicator is a contextual-function to explain the impact'. The primary constructs of an indicator are function and context. Function refers to the 'correspondence', 'dependence relation', 'rule', 'operation', 'formula' or 'representation' as defined by Vinner and Dreyfus (1989). It explains the relationship between the two domains 'research' and 'impact'. In other words impact (y) is a function of research (x) (y=f(x)). At large, in scientometrics understanding, the functional operation can be 'improvement', 'recognition', 'reduction', 'replacement' etc. (see Table 1 for examples). The indicator is a subjective measure of a system dependent phenomenon which is always described in its contextual understanding by a system designer (Vinkler, 2010). Therefore, the indicator's function is always applied in a specific context. For instance, "improvement in patient care system", in this indicator, the patient care system represents the context of the healthcare system and it is critically important for researchers, funders, institutes and support organizations related to the health sciences (Trochim et al., 2011).

In many cases, an indicator is self-explanatory and well written in a proper construct-based structure such as 'development of mitigation methods for reducing environmental hazards and losses from natural disasters' (Grant et al., 2010). However, similar to an algebraic expression, sometimes constructs are obscured but well understood by the users. For instance, in 'Number of citations', function and contextual domain is missing but well understood as "increased number of bibliometric citations" (where Function is the addition, the contextual target is citations, and the domain is bibliometrics).

This contextual nomenclature of indicators allows focussing on context and function irrespective of the selection of the words and lexical structure of the indicator. Also, it strengthens the idea of contextual generalizability which is very helpful in extending the applications and scope of the indicators. For example, in ‘use of research in the development of medical technology’ (Function = development/creation, Contextual Target = Technology, and Contextual Domain = medical) can be generalized on variable domain such as ‘use of research in the development of technology’ (Function = development/creation, Contextual Target = Technology, and Contextual Domain = variable).

Table 1: Nomenclature of Indicator with Examples

Structure of Indicator = F + C (t + d)

Where I is Indicator, F is a function, C is context, t is target and d is a domain.

Functions (F)

Improvement / Addition / Reduction
This function of indicator explains the addition or enhancement of an existing phenomenon in quantitative or qualitative form. (**Example:** Improvement in economic gains such as increased employment, health cost cut (Weiss, 2007))

Creation
This function of indicator focuses on the creativity in form of the development of new knowledge, theory, technique, method, technology, approach, opportunity or any kind of workflow. (**Example:** Creation of prevention methods for clinical practice (Trochim et al., 2011))

Recognition
This function explains the recognition of effort in form of outstanding quality by the peers or experts such as in form of awards, promotions, meritorious selection and work showcasing etc. This recognition can be of the research, the researcher or the research institute. (**Example:** Receiving an award on research (Kuruvilla et al., 2006))

Obsoleting / Replacing
This function elaborates the policy, law, regulation to obsolete or disuse the existing phenomena to overcome the future negative impacts. (**Example:** Change in law to obsolete the existing method of drug approval (Maliha, 2018))

Context (C)

Target (t)
Contextual targets in research impact science include knowledge, service, policy, law, guideline, system, technology, procedure, method, framework, workflow, publication, patent, product, stakeholder, citation, literature gaps, intellectual challenges, scholarly issues, relationships, collaborations, networks etc. These are the key areas but usually partial in contextual understanding.

Domain (d)
The contextual domain is the main area of interest of the indicator system designer such as health, education, economy, environment, academia, medical science, chemistry, history, multidisciplinary etc. The main body of knowledge and elaboration of indicators are always from the domain language. The domain is the main component of the indicator which specialized the context and application of the indicator. However, the level of the domain is subject to the interest and perspective of impact evaluator.

Taxonomy

In analysed indicators, most of the indicators are functionally related to the improvements in the current state of affairs (63%), mainly focused on future research, services and methods (Figure 1). However, recognition of research (23%) in the form of bibliometric, rewards and other citations is also considerably highlighted in the literature-based list of indicators. Creativity and development (14%) are also the common influence of research, which is reflected in indicators mentioning the creation of new knowledge, technique, research teams, drugs etc. More than half (59%) of the indicators attempt to explore the impact in academic domain e.g. Where and how the research is recognized? What knowledge, methods and collaborations are formed? What challenges, issues and gaps are addressed? Knowledge domains related to the social systems and services are second in coverage (26%) that primarily focus on the healthcare, education and justice systems. Economic systems and services also have a good share (11%) in literature-based indicators. Although, during the last two decades the impact of research is improving the environment and sustainability has also emerged in various indicators but its representation is quite low.

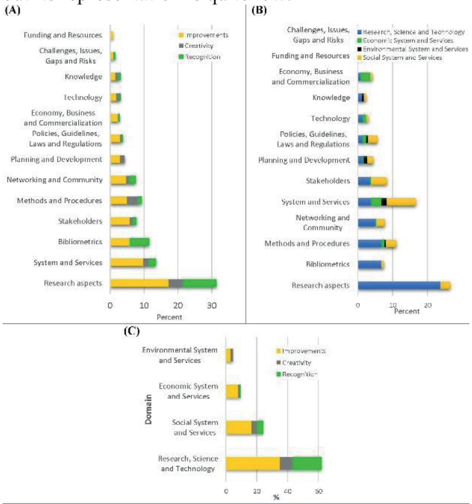


Figure 2: Cross-constructs distribution of Indicators Characteristics, (A) Functional Distribution of Target Areas in Indicators, (B) Domain Distribution of Target Areas in Indicators, and (C) Functional Distribution of Domains in Indicators

Conclusion and Future Direction

The general focus of the RI indicators is the use of research for improvement in the current state of affairs related to future research, services, technologies, policies and practices. This emphasis of research impact indicators can be broaden to all disciplines of science in future.

References

CAHS, 2009. Making an Impact: A Preferred Framework and Indicators to Measure Returns on Investment in Health Research. Canadian Academy of Health Sciences, Otawwa, Canada.

Grant, J., Brutscher, P.-B., Kirk, S.E., Butler, L., Wooding, S., 2010. Capturing RIs: A Review of International Practice. Documented Briefing. Rand Corporation.

Heller, C., de Melo-Martín, I., 2009. Clinical and Translational Science Awards: can they increase the efficiency and speed of clinical and translational research? Academic Medicine 84, 424-432.

Kuruvilla, S., Mays, N., Pleasant, A., Walt, G., 2006. Describing the impact of health research: a RI Framework. BMC Health Serv Res 6, 134.

Maliha, G., 2018. Obsolete to Useful to Obsolete Once Again: A History of Section 507 of the Food, Drug, and Cosmetic Act.

Trochim, W., Kane, C., Graham, M.J., Pincus, H.A., 2011. Evaluating translational research: a process marker model. Clin Transl Sci 4, 153-162.

Vinkler, P., 2010. The evaluation of research by scientometric indicators. Elsevier.

Vinner, S., Dreyfus, T., 1989. Images and definitions for the concept of function. Journal for research in mathematics education, 356-366.

Weiss, A.P., 2007. Measuring the impact of medical research: moving from outputs to outcomes. Am J Psychiatry 164, 206-214.

Does patentometrics represent valid patents?

Huei-Ru Dong [1,3] and Mu-Hsuan Huang[2,3]

[1] *141646@mail.fju.edu.tw*
Fu Jen Catholic University, Dept of Library and Information Science, New Taipei City, Taiwan

[2] *mhhuang@ntu.edu.tw*
National Taiwan University, Dept of Library and Information Science, Taipei, Taiwan

[3] National Taiwan University, Center for Research in Econometric Theory and Applications (CRETA), Taipei, Taiwan

Introduction

Patents are managed by the patent offices of various countries. Patent offices in different countries often have different patent validity periods because of differences in the degree of technological development or national economy at that time. Thomas (1999) analyzed the US patents issued between 1980 and 1985 and found that valid 4-, 8-, and 12-year patents accounted for 84.4%, 59.9%, and 39.4% respectively. Brown (1995) analyzed 605,000 US patents issued from 1982 to 1990, the valid 4-, 8-, and 12-year patents accounted for 82%, 69%, and 57%, respectively. Nikzad (2011) studied Canadian patents (CIPO), applied from 1990 to 2008, and found that 80% and 50% of the patents had validities of more than 4 and more than 10 years, respectively; in particular, nearly 90% of the 4-year patents were filed through the Patent Co-operation Treaty (PCT).

Furthermore, the proportion of valid patents in various patent offices, such as USPTO, UKPO, IGE, and DPMA, has been analyzed thus far. The relevant studies have found that the proportion of valid patents generally shows that the patents issued nearly to present, the higher the proportion of valid patents. Among the different patent offices, regardless of the proportion of valid patents in the short (4 years), medium (8 or 10 years), or long (12 or 14 years) terms, the USPTO has the highest proportion of valid patents, whereas the SIPO has the lowest proportion of valid patents (Brown, 1995; Dernburg & Gharrity, 1961; Hirabayashi & Myers, 1988; Nikzad, 2011; Thomas, 1999). Thomas (1999) reported that this is because USPTO's patent maintenance fees are relatively cheaper; moreover, the US market is larger than the European market and whence the patent owners are more willing to maintain patents.

The aforementioned studies have considered that valid patents consist a part of the patent database. However, in this study, we analyzed the total patents through patentometrics; here, we investigated the number and proportion of valid patents in the patent database. Furthermore, we discuss using valid patents rather than total patents through patentometrics.

Methodology

Here, we used patentometric methods, used to observe the perspectives of quantitative and qualitative analysis through patent data. Patentometric methods are widely used for competitor monitoring, technology assessment, R&D management, identification and assessment of potential sources for externally generating technological knowledge (particularly related to mergers and acquisitions), and human resource management (Ernst, 2003).

Data collection

The patent data used in this study are from the USPTO database because the United States has been one of the major markets of the world. For companies, filing USPTO patents represents a main strategic action and an influential symbol of global technological development. It can help them remain competitive in the market. In general, US utility patents are protected for a maximum of term of 20 years since the filing date; other terms include 4, 8, and 12 years. The patent data were downloaded from the USPTO website in March 2016. In total, 2,366,398 utility patents, filed between 1996 and 2010, are discussed herein.

Patent indicators

Citations per Patent

Citations per patent (CPP) is the average number of citations per patent.

Science Linkage

Science linkage (SL) is the average number of nonpatent prior-art citations per patent. The higher the SL value, the greater is the linkage to leadingedge or basic research.

Patent Breadth

Patent breadth (PB) is the average number of International Patent Classifications per patent. The greater the PB is, the wider is the influence of the patentee's industry and the more are the industries that are laid out (Lerner, 1994).

Result

This study collected 2,366,398 USPTO-issued patents, of which 1,479,164 (62.51%) were valid during sample collection in March 2016. Almost all patents issued after 2011 were valid during sample collection; therefore, only those issued before 2010 were analyzed.

Figure 1 illustrates that the number of USPTO patents gradually increased from 109,645 in 1996 to 219,613 in 2010, indicating that the number of patents in the 15-year-duration nearly doubled, except for patents from 2004 to 2007. Overall, the number of valid patents accounted for 62.51% of all patents; thus, the total number of patents and that of valid patents differed by approximately 37.49%. The proportions of valid patents issued during 2007–2010, 2003–2006, and 1996–2002 were approximately 85%–90%, approximately 70%, and approximately 50%, respectively. Moreover, the proportion of valid patents for the first patent maintenance in the fourth year, second patent maintenance in the eighth year, and the third patent maintenance in the 12th year after patent approval may be approximately 85%–90%, approximately 70%, and approximately 50%, respectively.

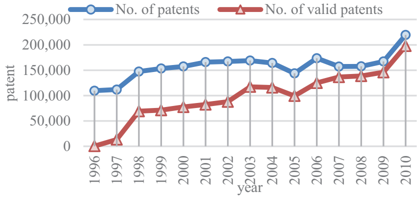


Figure 1 Number of valid and total patents in the USPTO

Compared with the proprtion of valid USPTO patent given by Thomson (1999) and Hirabayashi and Meyers (1988) for the period of the 1980s, the proportion of valid patents in this study demonstrated an increasing trend. However, compared with that reported by Brown (1995) for the 1980s, average proportion of valid USPTO short-term patents increased, but that of long-term patents decreased from 57% to <50%. This may be due to the long-term patent sample of Brown's study from only 1 year, whereas our long-term sample here covered 7 years.

Further, the t test was used to compare the differences between the valid and total patents by using the three indicators CPP, SL, and PB. Table 1 indicates that among the three indicators, the value of CPP is the highest. The CPPs of valid and all patents were 14.76 and 14.62, respectively. However, the SLs and PBs of were respectively 4.99 and 1.60 for valid patents and 4.34 and 1.59 for all patents.

Table 1 The t test results of valid and total patents

Indicator	Type	N	M	SD	Df	T	p
CPP	Valid	1,479,164	14.76	34.465	3051095.307	4.014	0.000
	Total	2,366,397	14.62	33.201			
SL	Valid	1,479,164	4.99	22.240	2842881.500	29.481	0.000
	Total	2,366,397	4.34	19.582			
PB	Valid	1,479,164	1.60	1.129	3088727.653	11.381	0.000
	Total	2,366,397	1.59	1.105			

*p< .05

This result is related to the characteristics of the three indicators: CPP represents the quality of the patent is an average of 14-15 patents. The SL represents the influence of the number of basic nonpatent literature on patents; it is influenced by an average of 4-5 basic documents. Finally, PB represents the number of technological fields that a patent can cover; most patents only covering one or two fields.

The t test results of the three indicators are significant; in other words, the analysis of valid patents differs from that of all patents. Furthermore, the value of the three indicators for valid patents is all higher than those for all patents. The patentometrics results are better for valid patents than for all patents.

Conclusions

This study mainly verified whether patentometrics of valid patents differ from those of all patents. The results demonstrated that the patentometrics of valid patents are better than those of all patents. Thus, the use of patentometrics can be considered for collecting valid patents, rather than all patents, while during patent analysis. Thus, collecting valid patents provides more representative analysis results. However, the differences occurring at different levels of analysis should be considered. These results may not be equally applicable to all technological fields or companies and thus further relevant research is warranted.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 108L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. MOST 108-3017-F-002-003- and MOST 107-2410-H-030-082-.

References

Brown, W. H. (1995). Trends in patent renewals at the United States patent and trademark office. *World Patent Information*, *17*(4), 225–234.

Dernburg, T., & Gharrity, N. (1961). A Statistical Analysis of Patent Renewal Data for Three Countries. *Patent, Trademark and Copyright Journal of Research and Education*, *5*(4), 340–368.

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information*, *25*(3), 233–242.

Hirabayashi, M. J., & Myers, J. S. (1988). U.S. patent expirations due to the nonpayment of the three and a half year maintenance fee. *World Patent Information*, *10*(3), 191–198.

Lerner, J. (1994). The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, 319–333.

Nikzad, R. (2011). Survival Analysis of Patents in Canada. *The Journal of World Intellectual Property*, *14*(5), 368–382.

Thomas, P. (1999). The Effect of Technological Impact upon Patent Renewal Decisions. *Technology Analysis & Strategic Management*, *11*(2), 181–197.

Investigating Citation of Algorithm in Full-text of Academic Articles in NLP domain: A Preliminary Study

Ruiyi Ding[1], Yuzhuo Wang[2] and Chengzhi Zhang[3,*]

[1] *dry@njust.edu.cn*
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094(China)

[2] *wangyz@njust.edu.cn*
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094(China)

[3] *zhangcz@njust.edu.cn*
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094(China)

Introduction

Nowadays, algorithms play an increasingly important role in scientific research. Algorithm are used and cited in academic papers, and studying these citations of algorithms enables people to get a comprehensive review about the use of algorithms in a specific domain. What's more, investigating the citation of algorithms also helps in evaluating and recommending related algorithms for scholars, especially the beginners of academic research.

Natural Language Processing (NLP) is a typical research field where algorithms are widely used. At the same time, studies have shown that in computer-related disciplines, the impact of conference papers is higher than journal papers (Lorcan, 2010). In addition, full-text content of papers provides us with more details about the citation of algorithms such as the location, motivation and emotions of citation. Therefore, this paper takes the field of NLP as a case to explore the citation of the algorithms based on the full-text papers of ACL annual conference, one of the most famous conferences in NLP domain.

Related works

Recently, with access of full-text databases, full-text of articles were used to study citation of knowledge entities such as data and software. Nicolas (2016) found that there were large differences in the citations of datasets in various subject areas. Existing work about algorithms concentrate on the mention rather than citation of algorithms and only study on top 10 data mining algorithms. Therefore, we want to investigate the citation of more algorithms, not just to the mention of specific algorithms.

Method

We use a dictionary-based approach to match sentences containing algorithms and a rule-based method to identify sentences containing reference symbols from the full text of the academic papers.

Data collection

The Annual Meeting of the Association for Computational Linguistics (ACL) is the highest level international academic conference in the field of NLP. We download papers of ACL annual conference between 1979 and 2015 from ACL Anthology (http://www.aclweb.org/anthology), all the 4，568 papers are available in XML format. 12 papers were excluded because of the formatting errors.

Algorithm dictionary construction

We apply a dictionary-based approach to identify algorithms from the full text of ACL articles. Through reading all the articles, we manually build a list of algorithms containing more than 1,800 terms, then use the full name of algorithms as queries to search on Google scholar and Wikipedia to acquire alias for each algorithm based on the descriptions of algorithms in the related papers and Wikipedia explanations. Finally, we obtain a dictionary containing 1,969 terms, including full name, abbreviation and alias of each algorithm (e.g. *NB* and Naïve-*Bayes* both refer to *Naïve Bayes*).

Citation sentence extraction

We define a sentence containing algorithms and reference symbols as a 'citation sentence'. Firstly, we use the algorithm dictionary to extract sentences containing algorithms, and we name these sentences '*algorithm sentence*'. After analyzing the form of reference symbols in the ACL articles, we conclude 2 composition rules of reference symbol: *Author (year)*, *(Author1, Author2, ..., year)*, and on this basis use a rule-based approach to extract the citation sentences from the algorithm sentences. Finally, we obtain a set of 15,495 citation sentences. However, some symbols of citation refer to other entities in the same sentence, such as a data set. Therefore, we limit the number of words between the algorithm and the

[*] Corresponding author.

reference symbol to 3, to filter out the non-correspondence sentences, and finally obtain 5,426 citation sentences after screening.

Result

Frequency of citation

In this article, the citation frequency refers to the times of every citing sentence of an algorithm in a paper. Most of algorithms were only cited once or twice, so we select the most cited top10 algorithms. Table 1 shows the result.

Table 1. Top 10 algorithm in citation frequency

Name	#Citation
Support Vector Machine (SVM)	230
Maximum Entropy (ME)	182
Conditional Random Field (CRF)	175
Hidden Markov Models (HMM)	174
IBM models (IBM)	161
Expectation Maximization (EM)	149
Latent Dirichlet Allocation (LDA)	136
Minimum Error Rate Training(MERT)	88
Neural Network (NN)	79
Decision Tree (DT)	73

It can be seen that in the ACL papers of 1979-2015, the most frequently cited algorithm is the '*Support Vector Machine (SVM)*', which has been cited for 230 times. It reflects the scholar's preference for '*SVM*' and the importance of it for NLP task. We speculate the reason is that '*SVM*' has a solid theoretical foundation and it is one of the most stable and accurate algorithms among all known famous algorithms. Compared to Wang (2018)'s work, our top10 algorithms in NLP field are different from the top10 algorithms of data mining. Only three are the same or similar: '*SVM*', '*EM*' and '*Decision Tree*' ('*C4.5*' is the most famous one). But in both top 10, '*SVM*' ranks first, showing that it has a wide range of applications and high academic influence in both NLP and general computer science field.

Time evolution of citation

We count the number of citations in each year to explore the temporal evolution of the algorithm citation. Figure 1 shows the trend of citation frequency of all algorithms in each year from 1979 to 2015.

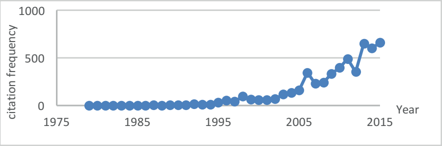


Figure 1. Citation frequency of all algorithms in each year

As shown in the figure, there is an overall upward trend. The citation frequency was very low before 1995, whereas it experienced a rapid increase after 2005. This reveals that during the past 10 years (2005~2015), the algorithms were gradually been widely used in the NLP field. In addition, considering that scholars' writing style may affects the citation frequency, we investigate the number of citing papers based on this.

Figure 2 shows the number of citing paper from 1979 to 2015. We define the number of citing paper as the number of articles that citing the algorithm, that is, no matter how many times the algorithm was cited in the paper is only recorded as one. Considering the difference in the number of papers accepted by the conference each year, we use the ratio of the number of citing papers to the annual total number of papers. It can be seen that the overall trend is also rising, and the increase after 2000 is relatively stable. The same as before, this also shows that the algorithm is becoming more and more important in NLP.

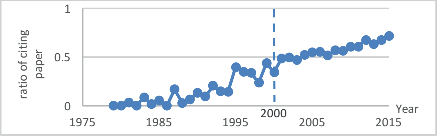


Figure 2. Ratio of citing paper of all algorithms in each year

Conclusion

Based on the ACL papers from 1979 to 2015, this paper investigates the frequency and the temporal evolution of the algorithms citations. Our results show that the algorithm with the highest citation frequency is '*SVM*'. In addition, during this period, there is a rapid growth in both the citation frequency and the number of citing papers, which indicates that the algorithms play an increasingly important role in NLP field. This study can provide a reference for the study on citations of knowledge entities.

Acknowledgments

This work is supported by Major Projects of National Social Science Fund (No. 17ZDA291), Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, ISTIC (No. ZD2018-07/01), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX18_0369).

References

Lorcan, C., Jill, F. & Barry, S. (2010). A Quantitative Evaluation of the Relative Status of Journal and Conference Publications in Computer Science. *Communications of the Acm Cacm Homepage*, 53(11), 124-132.

Nicolas, R.G., Evaristo, J.C. & Daniel, T.S. (2016). Analyzing data citation practices using the data citation index. *Journal of the American Society for Information Science,* 67(12), 2964-2975.

Mental health research in the countries of the Organisation of Islamic Cooperation (OIC), 2008-17

Grant Lewison and Richard Sullivan

grant.lewison@aol.co.uk, richard.sullivan@kcl.ac.uk
King's College London, Department of Cancer and Pharmaceutical Sciences, Guy's Hospital,
Great Maze Pond, London SE1 9RT, UK

Background and objectives

The 57 countries of the Organisation of Islamic Cooperation (OIC) suffer less than the world average from mental disorders (including Alzheimer's disease and self-harm), but their burden is growing more rapidly, and a few have burdens comparable with those of Europe and North America, see Table 1. We wished to see if their research outputs were commensurate with the disease burden, and if the distribution of the portfolio was appropriate for the challenge they faced.

Table 1. Percentages of total DALYs from mental disorders in some OIC countries in 2000 and 2015, the ratio between them, and corresponding percentages for other country groups.

Country	*ISO2*	*2000*	*2015*	*ratio*
Qatar	QA	12	17.5	*1.46*
UAE	AE	14	16.6	*1.18*
Iran	IR	9.7	12.7	*1.31*
Turkey	TR	9.0	12.0	*1.34*
Tunisia	TN	9.9	11.8	*1.19*
S. Arabia	SA	8.8	11.6	*1.31*
Lebanon	LB	8.2	10.6	*1.28*
Malaysia	MY	9.2	10.4	*1.13*
Morocco	MA	8.3	10.1	*1.22*
Jordan	JO	7.8	10.0	*1.29*
Algeria	DZ	7.4	9.0	*1.21*
OIC total	**OIC**	**3.5**	**5.0**	***1.44***
Canada,USA	CA,US	13.8	17.8	*1.29*
Europe 31	EUR	11.3	13.7	*1.21*
Rest of Wld	RoW	6.2	8.2	*1.32*

Methodology

Articles and reviews on mental disorders and from one or more OIC countries were identified in the Web of Science (WoS) by means of a complex filter based on title words and journal names, and downloaded to a spreadsheet. Five-year citation scores were also recorded. Their addresses were parsed to show fractional country contributions, and sub-filters were used to identify papers on particular disorders and in defined research domains. Outputs in 2015-17 were plotted against country GDP in 2014. International collaboration was also measured and preferred partners identified.

Results

There were 17,920 papers in the decade, and the OIC contribution was 15,170; the difference of 2750 (15%) represented foreign contributions. They came from Europe (7%), Canada + USA (5%) and the RoW (3%). In the decade, output quadrupled (see Fig. 1), and increased sharply in 2015.

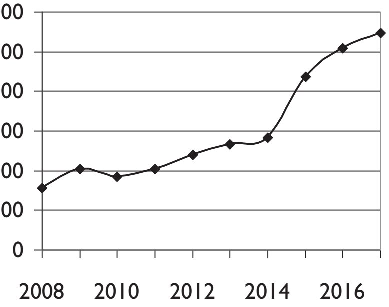


Figure 1. Increase of OIC mental disorders research papers in the WoS with time.

There was a rather weak correlation between country output and its wealth, see Fig. 2. Six OIC countries (notably Turkey, TR; Iran, IR; and also Tunisia, TN; Lebanon, LB; Jordan, JO; Uganda, UG) published more than five times what the trend-line predicted, but three (Indonesia, ID; Kazkhstan, KZ; Algeria, DZ) published < 20% of the predicted amount.

The amount of international collaboration varied greatly. It was almost 60% for Qatar and Uganda, but less than 10% for Iran and

Turkey. The USA was the preferred partner, followed at some distance by the UK, Australia, Germany and Canada.

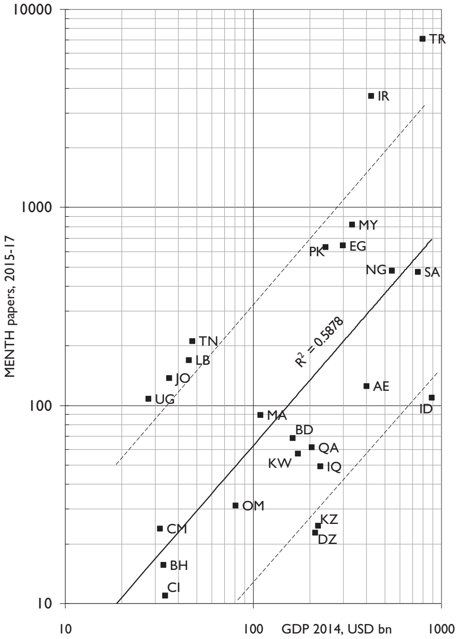


Figure 2. Correlation between OIC country outputs of mental disorders papers, 2015-17 and their wealth (GDP 2014, USD bn).

Fig. 3 shows the correlation between disease burden from the different mental disorders in 2010, and research outputs. Schizophrenia (SCH) and bipolar disorder (BIP) appear over-researched, but suicide and self-harm (SUI) is relatively seriously neglected (by a factor of about three).

In terms of citation performance, Uganda was the clear leader (12.3 cites per paper), followed by Lebanon (10.3). Iran performed fairly well (8.0), but Turkey's research was poorly cited (only 5.2), as was that of Nigeria (5.3).

Discussion

It is commendable that the overall output of papers has been expanding rapidly, but clearly several countries have been ignoring mental disorders in their biomedical research portfolio perhaps because of the stigma still attached to them. This is notable with respect to suicide and self-harm (Fig 3). There are surprisingly large differences between the OIC countries in terms of their willingness to collaborate internationally. Iran's relative isolation is understandable because of the sanctions regime imposed on it, but there is no such excuse for Turkey. One consequence is that its papers receive few citations and have little influence on researchers. Another is that, among the large OIC countries, it has almost the highest relative burden from mental disorders, and this is growing rapidly (by more than one third between 2000 and 2015, see Table 1.)

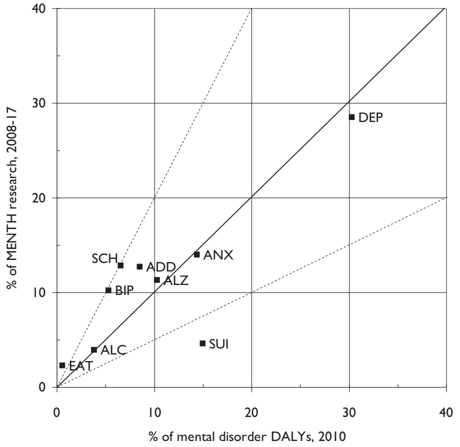


Figure 3. Comparison of mental disorders research outputs, 2008-17, with the overall burden of mental disorders in 2010 in OIC countries (WHO data).

There is a strong correlation between OIC countries' willingness to collaborate with third countries and their citation performance. This shows that scientific isolation is not a good strategy, and it is particularly disadvantageous for those countries who devote little effort to mental disorders research.

Since alcohol is proscribed or restricted in many Islamic countries, it is not surprising that little research is devoted to its misuse. Despite heavy consumption in Uganda, for example, the disease burden is still quite small so it appears that the ethos of avoidance of public drunkenness in Muslim-majority countries has proved effective in its control.

Identifying research areas for intensification of intraBRICS collaboration

Sergey Shashnov[1] and Maxim Kotsemir[2]

[1] *shashnov@hse.ru*

Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow (Russian Federation)

[2] *mkotsemir@hse.ru*

Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow (Russian Federation)

Introduction

In recent years research collaboration of BRICS countries in a wide range of subject areas has become a high priority for STI policymakers (see Sokolov et al., 2017). Meanwhile, recent studies in this field confirm that the intensity of intra-BRICS collaboration is quite low (See Khan, 2015; Finardi, 2015; Finardi and Buratti, 2016). Our study following the research of Shashnov and Kotsemir (2018) proposes an approach for detection of research areas with relatively low intensity of collaboration between BRICS countries. We also assess the potential for strengthening of intraBRICS collaboration in research areas with missed opportunities of cooperation between BRICS countries.

Methodology

The analysis is based on key indicators of intraBRICS research collaboration in Scopus in 2000 – 2017. As publications (taken as articles, review and conference papers) in intraBRICS collaboration we define publications whose authors are affiliated with at least two BRICS countries in Scopus. The focus of analysis on subject areas of intraBRICS collaboration is based Scopus classification. To compare intensity of IntraBRICS collaboration versus all international collaboration of BRICS countries we introduce an indicator "index of relative intensity of intraBRICS collaboration" (RIIC index further). This RIIC index is calculated for each BRICS country and for each of 27 Scopus subject areas as the ratio of "Share of subject area in total number of publications produced by country's authors in international collaboration with authors form other BRICS countries" to "Share of subject area in total number of publications produced in international collaboration (ICPs further) for individual country". Low (below 0.50) value level of RIIC index means that intensity of intraBRICS collaboration in specific subject area is much lower than the intensity of overall international collaboration of the country's authors.

Results

The results of our analysis show that BRICS countries are an important player in global science (Figure 1). China is closing the gap with the USA in terms of publication activity level. All BRICS countries (Russia to a somewhat lesser extent) show much higher growth rates of publications in Scopus than the USA, EU28 and entire world.

Country	2000	2017	Growth 2000-2017	2000	2017
	N. of publications			Share in a world	
BRA	14.7	72.0	4.90	1.2%	2.8%
RUS	34.0	85.4	2.52	2.8%	3.3%
IND	24.0	137.1	5.70	2.0%	5.3%
CHI	52.3	513.6	9.81	4.3%	19.9%
SAR	4.9	21.0	4.30	0.40%	0.81%
BRICS	129.2	821.0	6.35	10.7%	31.8%
USA	348.6	566.0	1.62	28.8%	21.9%
EU28	402.4	790.7	1.97	33.2%	30.6%
World	1 210.2	2 582.3	2.13	100%	100%

Figure 1. Basic indicators of publication activity of BRICS countries in Scopus in 2000 and 2017

BRICS countries are not important scientific partners for each other (Figure 2). The share of publications in intraBRICS collaboration in in total number ICPs is less than 5 per cent in China, less than 10% in Brazil and between 10 to 20 per cent in Russia, India and South Africa.

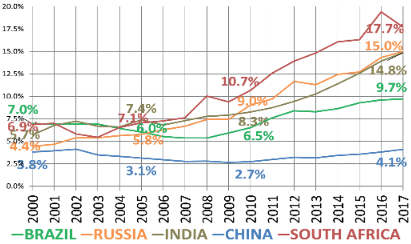
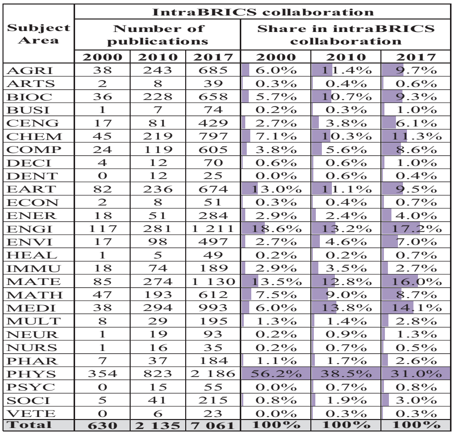


Figure 2. Share of publications in intraBRICS collaboration in the total number of ICPs of BRICS countries in Scopus

Nearly one third of intraBRICS collaboration is concentrated in "Physics and Astronomy" research area (Figure 3). The other important areas of

IntraBRICS research collaboration are “Materials science”, “Medicine” and “Engineering”. In social sciences and humanities, the level of intraBRICS collaboration is very low.

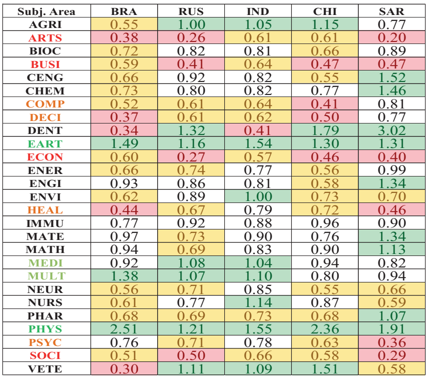


Subject Area	IntraBRICS collaboration					
	Number of publications			Share in intraBRICS collaboration		
	2000	2010	2017	2000	2010	2017
AGRI	38	243	685	6.0%	11.4%	9.7%
ARTS	2	8	39	0.3%	0.4%	0.6%
BIOC	36	228	658	5.7%	10.7%	9.3%
BUSI	1	7	74	0.2%	0.3%	1.0%
CENG	17	81	429	2.7%	3.8%	6.1%
CHEM	45	219	797	7.1%	10.3%	11.3%
COMP	24	119	605	3.8%	5.6%	8.6%
DECI	4	12	70	0.6%	0.6%	1.0%
DENT	0	12	25	0.0%	0.6%	0.4%
EART	82	236	674	13.0%	11.1%	9.5%
ECON	2	8	51	0.3%	0.4%	0.7%
ENER	18	51	284	2.9%	2.4%	4.0%
ENGI	117	281	1 211	18.6%	13.2%	17.2%
ENVI	17	98	497	2.7%	4.6%	7.0%
HEAL	1	5	49	0.2%	0.2%	0.7%
IMMU	18	74	189	2.9%	3.5%	2.7%
MATE	85	274	1 130	13.5%	12.8%	16.0%
MATH	47	193	612	7.5%	9.0%	8.7%
MEDI	38	294	993	6.0%	13.8%	14.1%
MULT	8	29	195	1.3%	1.4%	2.8%
NEUR	1	19	93	0.2%	0.9%	1.3%
NURS	1	16	35	0.2%	0.7%	0.5%
PHAR	7	37	184	1.1%	1.7%	2.6%
PHYS	354	823	2 186	56.2%	38.5%	31.0%
PSYC	0	15	55	0.0%	0.7%	0.8%
SOCI	5	41	215	0.8%	1.9%	3.0%
VETE	0	6	23	0.0%	0.3%	0.3%
Total	630	2 135	7 061	100%	100%	100%

Note: see full list and abbreviated titles of 27 Scopus subject areas at: https://dev.elsevier.com/tips/ScopusSearchTips.htm

Figure 3. Basic indicators of IntraBRICS scientific collaboration in Scopus

Figure 4 shows the values of RIIC index for BRICS countries. Areas with highest relative intensity of intraBRICS collaboration are “Physics and astronomy” and “Earth and planetary science”. Social sciences and humanities show the lowest value of the Index. In general, low level of Index is recorded for “Computer science”; “Decision sciences”, “Health professions” and “Psychology”.



Subj. Area	BRA	RUS	IND	CHI	SAR
AGRI	0.55	1.00	1.05	1.15	0.77
ARTS	0.38	0.26	0.61	0.61	0.20
BIOC	0.72	0.82	0.81	0.66	0.89
BUSI	0.59	0.41	0.64	0.47	0.47
CENG	0.66	0.92	0.82	0.55	1.52
CHEM	0.73	0.80	0.82	0.77	1.46
COMP	0.52	0.61	0.64	0.41	0.81
DECI	0.37	0.61	0.62	0.50	0.77
DENT	0.34	1.32	0.41	1.79	3.02
EART	1.49	1.16	1.54	1.30	1.31
ECON	0.60	0.27	0.57	0.46	0.40
ENER	0.66	0.74	0.77	0.56	0.99
ENGI	0.93	0.86	0.81	0.58	1.34
ENVI	0.62	0.89	1.00	0.73	0.70
HEAL	0.44	0.67	0.79	0.72	0.46
IMMU	0.77	0.92	0.88	0.96	0.90
MATE	0.97	0.73	0.90	0.76	1.34
MATH	0.94	0.69	0.83	0.90	1.13
MEDI	0.92	1.08	1.04	0.94	0.82
MULT	1.38	1.07	1.10	0.80	0.94
NEUR	0.56	0.71	0.85	0.55	0.66
NURS	0.61	0.77	1.14	0.87	0.59
PHAR	0.68	0.69	0.73	0.68	1.07
PHYS	2.51	1.21	1.55	2.36	1.91
PSYC	0.76	0.71	0.78	0.63	0.36
SOCI	0.51	0.50	0.66	0.58	0.29
VETE	0.30	1.11	1.09	1.51	0.58

Figure 4. Values of RIIC index for BRICS countries in Scopus subject areas for 2013 -2017

Considering dynamics, structure and concentration of country-partners of BRICS countries as well as positions of country-partners in global science for subject areas with low values of RIIC index we assess the potential for intensification of intraBRICS collaboration in these areas. Table 1 shows an example of such an assessment for research areas where values of RIIC index is below 0.50 for at least two BRICS countries. The results show that high potential for intensification of intraBRICS collaboration exists in ‘Business, management and accounting”, “Health professions” and “Social sciences” areas, while for “Arts and humanities” and “Dentistry” this potential is low.

Table 1. Example of schematic assessment of potential for intensification of intraBRICS collaboration

Subj. area	BRA	RUS	IND	CHI	SAR
ARTS	Medium	Weak	N/A	N/A	Medium
BUSI	N/A	Medium	N/A	Strong	Strong
DECI	Weak	N/A	N/A	Strong	N/A
DENT	Medium	N/A	N/A	N/A	Weak
ECON	N/A	Weak	N/A	Strong	Medium
HEAL	Strong	N/A	N/A	N/A	Strong
SOCI	Strong	Strong	N/A	N/A	Strong

Note: N/A means “not assessed”. We do not asses potential for intensification of intraBRICS collaboration for cells where RIIC index is higher that 0.50.

Conclusions

This study provided an overview of intraBRICS research collaboration and proposed an approach for detection of research areas with relatively low intensity of collaboration between BRICS countries. We also estimated the opportunities for strengthening of intraBRICS collaboration across research areas. Further analysis of potential for intensification of intraBRICS collaboration in areas with very low values of RIIC index collaboration is needed at the level of individual organisations. Here one should take into an account the level of concentration of leading organisations in selected research areas, the structure of their collaboration network and the place of partners from BRICS in these networks (see Moed et al., 2011 proposing an analysis with similar approach).

References

Finardi, U. (2015). Scientific collaboration between BRICS countries. *Scientometrics*, *102*(2), 1139-1166.

Finardi, U., & Buratti, A. (2016). Scientific collaboration framework of BRICS countries: an analysis of international coauthorship. *Scientometrics*, *109*(1), 433-446.

Kahn, M. (2015). Prospects for Cooperation in Science, Technology and Innovation among the BRICS Members. *Vestnik Mezhdunarodnykh Organizatsii-International Organisations Research Journal*, *10*(2), 105-119.

Moed, H. F., de Moya-Anegón, F., López-Illescas, C., & Visser, M. (2011). Is concentration of university research associated with better research performance?. *Journal of Informetrics*, *5*(4), 649-658.

Shashnov, S., and Kotsemir, M. (2018). Research landscape of the BRICS countries: current trends in research output, thematic structures of publications, and the relative influence of partners. *Scientometrics*, 117(2), 1115–1155.

Sokolov A., Shashnov S., Kotsemir M., Grebenyuk A. (2017). Identification of Priorities for S&T Cooperation of BRICS Countries. *International Organisations Research Journal*, 12(4). 32-67.

Model Entity Extraction in Academic Full Text Based on Deep Learning

Zhen Lei[1], *Dongbo Wang*[2]

[1] *19116129@njau.edu.cn*
Nanjing Agricultural University(China)

[2] *db.wang@njau.edu.cn*
Nanjing Agricultural University(China); KU Leuven（Belgium）

1. Introduction

1.1 Informetrics and natural language processing

Proposed in the end of the 20th century, informetrics have been divided into five main branches, including bibliometrics, scientometrics, informetrics, webmetrics and altmetrics. In the new era of big data and AI, improvements in statistics and evaluation of information by employing these techniques have been a hot topic in informetrics. In virtue of great advances in AI, the effectiveness of natural language processing has been significantly improved, which in turn facilitates analysis of information content in informetrics.

1.2 Named entity recognition

As a key part of information extraction and retrieval, the named entity recognition (NER) aims at identification and classification of components representing named entities in the text. Hence, NER is also known as named entity recognition and classification (NERC)(Nadeau, D., & Sekine, S. 2007.). For natural language processing, NER is an issue of sequence labeling that can be effectively solved using machine learning methods.

A named entity is a definite object of study. The MUC classifieds NER tasks into three major categories (named entity, time expression, quantity expression) and seven minor categorie(Chinchor, N. 1995;), while model entity is not included.

1.3 Model entity

In this article, labelled entities are denoted as model entities, which belongs to the Micro-Level Entities(Ding et al.,2013). Herein, the model refers to generalized model that contains abstract models (e.g., UTAUT), mathematical models (e.g., SVM), theories (e.g., ISP) and algorithms (e.g., PLS). The model entity is characterized by numerous classifications and variants. The numerous classifications can be attributed to continuously emerging models, theories and algorithms, while the numerous variants can be attributed to by improvement or integration of existing models, which leads to modified models with identical source, such as LSTM and BiLSTM.

2. Data set

2.1 Data source

The data set used in this study was from the *Journal of the Association for Information Science and Technology (JAIST)*. A total of 893 articles published in 2012 to 2016 were collected and their full texts were used as data for machine learning and deep learning model training and testing.

2.2 Data labelling

Model entities were labelled using the BEMS method. Model entities containing only a single word or abbreviation were labelled using the S-model; model entities consisting of two words were labelled by B-model and E-model at head and tail, respectively; model entities consisting of more than two words were labelled by B-model, M-model, and E-model at head, middle part and tail, respectively; words that are not model entity were labelled by O.

3. Model Entity Recognition

The data features for data learning may vary with the machine learning model, resulting in different effects. Herein, two machine learning models were introduced, followed by analysis and comparison of their effects.

3.1 CRF

The conditional random field is a classical model for sequence labeling in machine learning(Lafferty, J., McCallum, A., & Pereira, F. C. 2001). The results of model entity recognition in this study is as follows:

Table 1.ten-fold results with CRF

Number	Precision	Recall	Fb1
1	90.24%	68.12%	77.64%
2	91.01%	67.10%	77.25%
3	92.30%	67.59%	78.03%
4	91.59%	66.31%	76.92%
5	90.93%	66.41%	76.76%
6	92.25%	68.38%	78.54%
7	90.43%	67.98%	77.61%
8	91.35%	66.13%	76.72%
9	89.50%	66.33%	76.19%
10	91.32%	69.02%	78.62%

3.2 BiLSTM+CRF

The long short term memory (LSTM) network is a neural network model that has been widely applied for sequence issues. The LSTM network can mitigate gradient vanishing and explosion that are common in processing of relatively long texts by conventional RNN models(Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney., 2012). By superimposition of positive and reverse information using BiLSTM, model learning of context features. However, the BiLSTM model may lead to unexpected irregular sequences at the end, while the CRF model can effectively eliminate that issue. Therefore, combination of BiLSTM and CRF works perfectly for most data sets (Chen, Tao, et al.,2017).

Table 1. Ten-fold results with BiLSTM+CRF

Number	Precision	Recall	Fb1
1	82.84%	76.94%	79.78%
2	86.85%	74.02%	79.92%
3	81.16%	79.39%	80.26%
4	86.24%	72.55%	78.80%
5	85.17%	74.93%	79.72%
6	83.15%	77.34%	80.14%
7	85.57%	74.91%	79.89%
8	77.81%	80.27%	79.02%
9	81.34%	77.97%	79.62%
10	80.59%	76.70%	78.60%

3.4 Comparison evaluation

Based on experimental data of CRF model and BiLSTM+CRF model, the following conclusions can be drawn:

1) The accuracy of the CRF model remained at no less than 90%, while that of the BiLSTM+CRF model was 80%~90%.
2) The recall rate of CRF model was no larger than 70%, while that of the BiLSTM+CRF model was 70%~80%.

These can be explained by theories: As features were selected artificially in the CRF model, entity screening was highly harsh and the accuracy was guaranteed. In BiLSTM+CRF model, features were selected by continuous training of the word vector model and more model entity features can be obtained, resulting in increased entity recall rate but reduced accuracy. Overall, the BiLSTM+CRF model exhibits optimized performance and its F1 value can be up to 80.26 and doesn't require artificial feature selection (facile training). In practical applications, a high recall rate is of great significance for identification of entities in literature. The limits in accuracy of the BiLSTM+CRF model can be relieved by artificial selection. Moreover, improvements in data set quality would lead to further increases in model accuracy.

4. Application in bibliometrics

With natural language processing technologies, it is possible to go beyond conventional bibliometrics and deep into text contents. The model entity recognition allows us to investigate the correlations of different contents, methods, and objects. For instance, a conclusion that topics and methods of any two articles are similar can be made if it is identified that classifications and quantities of model entities involved in these two articles are within a certain range. Additionally, it is possible to predict that an article presents comparison of different models or improvement of a specific model based on classification and quantity of model entities.

5. Conclusions

In this article, we have proved that NER can be used in academic full texts for its fine effect(80.26% F-measue), which means we can extract model entities effectively from these texts for latter using. Despite the achievements of this study, limitations and future works should be mentioned.

5.1 Limitations

In this article, lemmatization of model entity was not involved. This may affect the accuracy of model entity recognition. Then, this study involved recognition of content feature, but not further analysis of the specific semantics. Additionally, model definition is not perfectly precise, which is reflected by inconsistence during labelling, resulting in negative effects on the accuracy of machine learning models.

5.2 Future works

In the future, we intend to focus more on using the entity retrieved to do more quantitative analysis and end up forming a mathematics model to describe the literatures about model entity. Also, we will make more effort on studying how to utilize NLP to contribute the study of informetrics.

Acknowledgments

Researches are supported by the National Natural Science Foundation of China (Grant Number: 71673143).

References

Chinchor, N. (1995, November). MUC-6 named entity task definition (version 2.1). In *6th Message Understanding Conference, Columbia, Maryland*. Chinchor, N., & Robinson, P. (1997, September).

Chinchor, N., & Robinson, P. (1997, September). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding* (Vol. 29, pp. 1-21).

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, *72*, 221-230.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Social media and library metrics and indicators: how can we measure impact on performance?

Francisco-Javier Calzada-Prado[1] and Carmen Jorge-García-Reyes[2]

[1] *fcalzada@bib.uc3m.es*
Universidad Carlos III de Madrid, C/ Madrid, 126, 28903, Getafe (Spain)

[2] *cjorge@bib.uc3m.es*
Universidad Carlos III de Madrid, C/ Madrid, 126, 28903, Getafe (Spain)

Introduction

Libraries and other cultural heritage institutions initially started to use social media as a cost-effective means to reach users and let them know about their services, activities and resources. More recently, they have realized that disseminating promotional messages should account for only a small part of their social media activity if they intend for these tools to work fully to their benefit in an environment of global competition for attention. They are also realizing the actual cost of planning, managing, deploying and evaluating a community-oriented and participatory social media presence may be proportional to the potential benefits to be achieved. In an era of accountability and quality management, libraries are wondering: what is the impact of our social media investments on our libraries' performance? The goal of this paper is to spark discussion on the measurement of conversion of social media efforts into library performance that will, ultimately, lead to the identification of a set of metrics and indicators that may guide libraries' use of social media and improve library performance and quality.

Social media metrics

No international consensus has yet been reached on the set of metrics that would best measure the efficiency of social media in any given context.

Among the most influential social media metrics are Kaushik's (*TrueSocialMetrics*), who recommends using four simple metrics: Conversation rate (comments, replies), Amplification rate (retweets, shares), Applause rate (favourites, likes), and Economic value (revenue+cost savings) (2011).

The Conclave (2013) suggests six metrics: Content & sourcing (data sources and research methods), Reach & impressions, Engagement & conversation (interaction, discussion), Opinion & advocacy (sentiment, action), Influence, Impact & value (outcome: effect, importance, ROI).

Measuring social media ROI (a type of return or impact metrics) has been found to be particularly challenging for most professional marketers (Sprout Social, 2018).

Agostino and Sidorova (2016) propose a social media performance measurement system framework based on a review of the literature that identified a selection of financial and non-financial indicators: social media ROI, network structure, interactions (likes, comments, shares), content/conversation, and users' sentiment/opinion. Although their framework provides a rationale for future research, its practical utility has not yet been tested.

Library use of social media metrics

Most of the social media metrics currently used by libraries are adapted from metrics favoured in corporate settings, and usually envisaged in the respective analytical tools. In a 2013 global survey, Liew, King and Oliver (2015) observed that most of the cultural heritage institutions responding had engaged in or were in the process of evaluating their social media activities. The obstacles to such assessment cited by respondents included "lack of resources", "shortage of skills", and "difficulties experienced with identifying metrics or measuring success". The challenge in social media evaluation is, indeed, as Showers suggests, to know "what we want to measure and why" (2015, p. 115). Matthews contends that any social media metrics selected by a library "should be able to measure four perspectives: exposure, engagement, influence, and results" (2018). In this vein, a significant contribution is that of González-Fernández-Villavicencio (2016), who compiled a set of social media metrics for library settings organised into six categories: Reach (popularity, size, visibility), Activity frequency (number of posts, uploads, etc.), Loyalty (website traffic from social media), Influence (users' brand perception: mentions, sentiment, reputation index), Engagement (comments, shares, views, downloads, etc.), and Conversion (return on investment: number of downloads of digital collections, downloads of tutorials, number of loans, etc.). Based on a selection of these metrics and indicators, the National Library of Spain found a strong link between their social media campaigns and a significant increase in digital collection usage and visits to their website (Carrillo Pozas, 2017).

Table 1. Selected social media and library metrics and indicators.

Social media metrics	ISO 2789:2013	ISO 11620:2014
▪ Audience: followers, subscribers. ▪ Activity: publications. ▪ Reach and impressions. ▪ User engagement: participation/interactions: likes, shares, comments, replies views, downloads. ▪ Loyalty. ▪ Influence and reputation: mentions, sentiment, advocacy.	▪ Services and use: General, Users, Loans, Renewals, Reservations, Interlibrary lending requests, Reference and informational questions received, Document delivery, Attendances at events and training, Physical visits, Number of searches, Number of accesses, Number of downloads, Use of the digitized collection, Number of virtual visits, Use of mobile services, Social network services, Content units on social networks, Usage of library-hosted interactive services, Usage of library social network services. ▪ Staff: Time spent on interactive services, Time spent on services for mobile devices, Time spent on library evaluation, Time spent on preparation of training lessons.	▪ Speed of reference transactions. ▪ Use of collection: Collection turnover, Loans per capita, Number of content units downloaded per capita, number of downloads per document digitized. ▪ Access: Library visits per capita, Percentage of external users, Percentage of the total library lending to external users, User attendances at library events and training lessons per capita. ▪ Collection cost: Cost per use, Acquisition cost per collection use, Cost per download. ▪ Staff: Percentage of user services staff, Percentage of library staff providing electronic services.

Library metrics and performance indicators

Two international standards assist libraries in the collection and interpretation of statistical data for describing library resources and their use, as much as institutional performance: ISO 2789:2013 (2013) and ISO 11620:2014 (2014). It is interesting to note that the former considers the number and usage of the library's social networks, while the latter addresses the issue less straightforwardly.

Linking social media metrics with library performance indicators

By way of background for the present discussion, Table 1 presents a selection of metrics and indicators from the three main sources considered in this paper: social media metrics, and ISO standards 2789 and 11620. The selection has been made according to their potential inter-relationships and impact on library performance.

On those grounds, a number of questions may be put forward to guide discussion and future research:

- What types of logical relationships might be established among social media and library metrics and indicators?
- How can metrics and indicators from different social media tools and analytics providers be reconciled and applied?
- Which social media metrics and indicators may be expected to impact library performance most prominently and might therefore be apt for inclusion in library assessment tools, and eventually even ISO 11620?

References

Agostino, D., & Sidorova, Y. (2016). A performance measurement system to quantify the contribution of social media: new requirements for metrics and methods. *Measuring Business Excellence*, *20*(2), 38–51.

Carrillo Pozas, A. (2017, January 24). Evaluando los medios sociales de la Biblioteca Nacional de España: métricas e indicadores. Retrieved April 5, 2019, from El Blog de la BNE website: https://perma.cc/96KZ-SPW6

González-Fernández-Villavicencio, N. (2016). *Métricas de la web social para bibliotecas*. Barcelona: UOC.

ISO. (2013). *Information and documentation - International library statistics (ISO 2789:2013)*.

ISO. (2014). *Information and documentation - Library performance indicators (ISO 11620:2014 (E))*. Switzerland: ISO.

Kaushik, A. (2011). Best Social Media Metrics: Conversation, Amplification, Applause, Economic Value [Blog]. Retrieved from Occam's Razor website: https://perma.cc/M8PQ-ABBZ

Liew, C. L., King, V., & Oliver, G. (2015). Social Media in Archives and Libraries: A Snapshot of Planning, Evaluation, and Preservation Decisions. *Preservation, Digital Technology & Culture*, *44*(1).

Showers, B. (Ed.). (2015). *Library analytics and metrics: using data to drive decisions and services*. London: Facet.

Sprout Social. (2018). The 2018 Sprout Social Index: Realign & Redefine. Retrieved April 2, 2019, from Sprout Social website: https://perma.cc/Z2G3-TRWU

The Conclave. (2013). *The Social Media Measurement Standards Guidebook*. Retrieved from https://perma.cc/T2UW-8VPJ

What kind of papers in the collection of highly cited papers in the Economic&Business field can obtain higher social influence?

Jiang Wu[1] and Xiao Huang[2]

[1] *jiangw@whu.edu.cn*
Wuhan University, School of Information Management, 430072, Wuhan, Hubei (China)

[2] *xiaoh@whu.edu.cn*
Wuhan University, School of Information Management, 430072, Wuhan, Hubei (China)

Introduction

Since the introduction of Altmetrics, some scholars have begun to pay attention to the similarities and differences between the academic influence indicators and social influence indicators and to find the relationship between the citations and the Altmetrics indicators(Gunther 2011, Li, Thelwall et al. 2012). The influence and evaluation of highly cited papers have always been the focus of scholars(Chen, Arsenault et al. 2015). By comparing the number of citations with the number of mentions in Twitter and Facebook, the research finds significant differences between the academic influence and social influence of highly cited papers.(Hassan, Imran et al. 2017). However, we don't know what kind of papers can obtain high social influence and what are the differences between high academic influence and high social influence papers. This study will improve the comprehensive influence analysis of highly cited papers , clarify the different perspectives of academics and public on highly cited papers and provide examples for researchers to evaluate scientific results.

Method and data

The influence of the paper on social media can reflect the attention of public to the paper. Such a focus may be different from that of academics. The research questions in this study are as follows:
Question 1: Is there a significant difference between the academic influence and social influence of highly cited papers?
Question 2: If there are significant differences, what factors are related to these differences?
Firstly, this study analyses the correlation between the number of citations and mentions in Twitter and Facebook to verify that the relationship of academic influence and social influence in the highly cited papers collection are significantly different. Secondly, this study obtains the top 10% cited papers and the top 10% mentioned papers in Twitter and then compares the two top-ranking papers by publication time, journal and topic distribution to identify the factors that contribute to the differences and explore how the factors affect the differences.

Data

In this study, highly cited papers in the Economic&Business field of the ESI(Essential Science Indicators) are selected as the research object. The Altmetrics indicators of the papers are obtained from Altmetrics.com with the unique identifier of the paper DOI in the ESI. There are 2,737 papers in the Economic&Business field, of which 2,688 papers include DOI. 2,349 papers with DOI have Altmetrics indicators. The data acquisition time is November 25, 2018.
The 2,349 papers are each mentioned on average 33 times in social media and online academic communities, much higher than the average number of mentions in the Altmetrics.com of 7.7. Therefore, the collection has both high academic influence and high social influence. The latest citations, abstracts, keywords and other information of these papers are crawled from the homepage. The data acquisition time is November 30, 2018.

Result

Correlation analysis

Four Altmetrics indicators including the number of mentions in Twitter are selected for correlation analysis with citations. The results are shown in Table 1.

Table 1. Correlation analysis results of citation and typical Altmetrics indicators

	1	2	3	4	5
Citation	1	-0.271** 0.000	-0.062** 0.002	0.612** 0.000	0.991** 0.000
Twitter		1	0.342** 0.000	-0.050* 0.016	-0.257** 0.000
Facebook			1	0.077** 0.000	-0.059** 0.004
Mendeley				1	0.621** 0.000
Dimensions					1

Table 1 shows that the citation is significantly negatively correlated with the number of mentions in Twitter and Facebook, which indicates that the relationship of social influence and academic influence of the highly cited paper collection is significantly different. And the coverage of the mentions in Twitter is much higher than Facebook, so it is reasonable to choose the mentions in Twitter to reflect the social influence of highly cited papers. Therefore, this study selects the top 300 cited papers and the top 300 mentioned papers in Twitter to compare the mechanism of academic influence and social influence by publication time, journal and topic distribution.

Publication time distribution

The publication year of each paper is extracted and the year distribution curves are plotted separately. The result is shown in Figure 1.

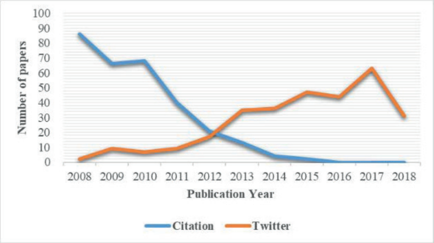


Figure 1. The distribution curve of publication time of the top 300 papers

Figure 1 shows that the trend of publication time curves for the two top-ranking papers is clearly opposite. Recent papers will gain more attention on social media and long-published papers will receive more citations.

Journal distribution

There are some differences of journals between high academic influence and high social influence. Select top ten journals with the highest number of papers in each category. The top ten journals with high academic influence are all economic journals, including the top journals in various sub-areas of economics. The top ten journals with high social influence include well-known multidisciplinary journals such as Nature, Science, and PNAS.

Topic distribution

This study compares the topic distributions of the abstracts of the two types of papers through the LDA topic model. The topics of abstracts are summarized based on the keyword probability and attribute. Each topic consists of several keywords. The results are shown in Table 2.

Table 2. Summary of typical keywords for abstracts of two types of papers

Citation	Keywords related theory and method	preferences、applications、evaluation、cocreation；methods、framework、propose、outputs、efficiency；organizational、review、theoretical；
	Both of them	meta、country、cultural、agricultural；experience、academics、market、topic；shareholders、governance、structures、commitments；China、tax、TAM；
	Keywords related empirical research	earnings、public、annual；marketing、service、customers；unemployment、indicators；distinct、balance、Entrepreneurial；
Twitter	Family	increases、household、food；financial、macroeconomic、family；
	Climate	Greenhouse、synthesis、reduction；environmental、air、damages；climate、change、emissions；
	Social Media	Twitter、Facebook、blogs、education；Twitter、diffusion、trust；Social media、institutions、prevalence；
	Regional development	distance、diversity、European、African、American；agricultural、inequality、economic、economic；
	Corporate research	workers、entrepreneurial；labor、manufacturing；Amazon、Mechanical；financial、employees；
	Health	obesity、weaknesses、care；
	Keywords related theory and method	product、choice； theory、model、empirical； factor、modeling、phenomena； marginal prices

Table 2 shows that there are significant differences in the topic distribution of the two top-ranking papers. The highly cited papers focus more on the research of theory and methods. Papers with high mentions in Twitter are more empirical studies, in which themes of family, climate, social media and health are closely related to the public life.

Reference

Chen, S., Arsenault, C., & Larivière, V. (2015). Are top-cited papers more interdisciplinary?. *Journal of Informetrics*, 9(4), 1034-1046.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research*, 13(4), e123.

Hassan, S. U., Imran, M., Gillani, U., Aljohani, N. R., Bowman, T. D., & Didegah, F. (2017). Measuring social media activity of scientific literature: an exhaustive comparison of scopus and novel altmetrics big data. *Scientometrics*, 113(2), 1037-1057.

Li, X., Thelwall, M., & Giustini, D. (2011). Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2), 461-471.

Assessing Promotion of Research Results in Media: Examples from Siberian Institutes

Denis Kosyakov[1], Inna Yudina[1], Zoya Vakhrameeva[1]

[1] *kosyakov@spsl.nsc.ru*
State Public Scientific-Technological Library of the SB RAS, Voskhod Str 15, Novosibirsk (Russia)

Introduction

Science communication for research organizations is more than part of public relations (Carver, 2014), institutions are driven by the need to justify the importance of their own activities (Bauer, Allum, & Miller, 2007), which ultimately affects funding. A significant driver is the need to promote the results of scientific research not only among the public but also in the professional community. Due to the rapid increase in the amount of scientific information (Bornmann & Mutz, 2015), scientists resort to all possible methods in order to draw peers attention to the results of their research (Wilkinson & Weitkamp, 2013). Back in the 1990s, Phillips, Kanter, Bednarczyk, & Tastad, (1991) drew attention to the relationship between the coverage of research results in the traditional lay press and subsequent citation indicators. In recent years, this trend has come to Russian science.

Research tasks and goals

The news report on the scientific publication has several goals: promotion of a research institution; promotion of personal brands of authors of scientific publications; increase of public importance and relevance of a research field; promotion of a scientific view of the world, public awareness, improving the quality and reliability of available information; promotion of the results of scientific research in the professional community.

A number of scientometricians study the degree of media coverage of research topics and areas since it characterizes well the public interest in the science field (Elmer, Badenschier, & Wormer, 2008; Holliman, 2004). We assume that the institutional level metrics should be related to the completeness of media coverage of scientific results of the organization, its authors, and research topics.

Methods and data

We selected mass media news reports based on research results of the institutions of the Siberian Branch of the Russian Academy of Sciences (SB RAS) as a subject for validation of the metrics proposed. The SPSTL SB RAS supports an aggregator of scientific news - Siberian Science News, which gathers media publications with references to Siberian research institutes and universities (Kosyakov et al., 2018). This project selects and gathers relevant news stories from a wide variety of sources. News reports related to the articles published in scientific journals for the period from the beginning of 2016 to September 2018 were selected from this newsfeed in semi-automatic mode by a number of keywords. The total number of news mentioning the institutes of SB RAS for this period was 5544, of which 301 messages were related to the results of scientific research. In total, 92 organizations got into consideration, however, the media activity of some of them was too small to be evaluated.

Named entities were compared with a list of Siberian authors of research articles obtained from the Russian Index of Science Citatitions (RISC) and additionally checked for affiliation with the abovementioned scientific organizations. For news reports and their versions in different media, the number of linked posts on social networks Facebook and VK were obtained. Data on the number of scientific articles indexed for 2016-2017, as well as the number of authors of these articles for every single institution, were also obtained from the RISC. Based on these data, three metrics were calculated for each institution:

- Media coverage of articles, equal to the ratio of the number of news reports to the number of publications indexed in the RISC for a particular year as a percentage.
- Media coverage of the authors is equal to the ratio of the number of unique employees of the organization mentioned in the news reports to the number of authors of publications indexed in the RISC for a particular year as a percentage.
- Media impact index equal to the sum of the number of news messages with a factor of 10, the number of reposts of news messages in the media with a factor of 4 and the number of posts in social media linked to the original news item or any of the reposts with a factor of 1.

Results

The analysis showed that the number of news mentioning Siberian research institutes is growing. This may be due both to an increase in the media activity of institutions, in particular, the establishment of PR departments and press services and to the general increase in the number of media and news. However, news on research results published in scientific journals occupy a modest place in this news feed. A total of 301 such news items and 3568 reprints were found. This averages over the entire period about 5.5% of the total number of mass media news reports with references to the institutes and a little more than 11% of reprints. The Krasnoyarsk Scientific Center (KSC) turned out to be the leader in terms of the number of news, for the entire study period. 71 original news reports and 1211 reposts related to the results of its research activities were published. The Institute of Cytology and Genetics (ICG), the Institute of Petroleum Geology and Geophysics (IPGG), the Institute of Geology and Mineralogy (IGM), and the Institute of Catalysis (IC) were also ranked in the top five with a noticeable gap from the leader.

The share of news based on scientific publications for the entire period under consideration reaches a little over 13% by the leader of this ranking, the KSC. For a few more organizations, this proportion is above 10%. It can also be noted that the media shows a noticeable interest in news related to scientific publications – the average number of reprints of such news is usually higher than the corresponding figure for all news reports.

The calculation of the coverage metrics described above is given in Table 1. We can observe visible progress in media coverage of scientific publications and authors. The higher output of the IMCB, which is small in the number of researchers, stands out. While large organizations held some of the high positions in the ranking (KSC, IPGG, ICG) small ones occupy the prominent place too.

Table 1. The degree of media coverage of scientific publications (PC), authors (AC) and media impact index (MI) of SB RAS Institutes (top 10 ranked by the publications coverage in 2017)

Institute	2016			2017		
	PC	AC	MI	PC	AC	MI
IMCB	4.00%	4.35%	45	5.66%	8.51%	483
ICKC	2.17%	1.00%	223	2.90%	3.08%	143
ICBFM	1.72%	1.58%	281	2.64%	2.90%	1 545
IAE	1.64%	5.06%	214	2.58%	4.79%	186
KSC	0.56%	0.58%	1 603	2.41%	3.91%	21 512
IPGG	1.11%	1.60%	8 749	2.35%	4.26%	1 465
ICG	0.68%	0.95%	239	2.26%	2.79%	8 208
IAET	0.99%	1.72%	1 055	2.24%	7.33%	3 692
ISEA	0.52%	0.00%	333	2.00%	2.27%	577
SIPPB	0.48%	0.76%	42	1.74%	2.65%	502

Conclusion

The study of media activity of research institutes of the Siberian Branch of the Russian Academy of Sciences shows an increasing interest in popularizing and promoting the brands of organizations, individual scientists and scientific results. The progress both in the level and in the completeness of the media coverage of research results published in scientific journals is clearly visible. The proposed metrics and the results of their calculations make it possible to identify the most successful practices, to identify weaknesses, to formulate recommendations on the most effective presentation and promotion of scientific results. The ongoing data collection on media publications mentioning institutions will expand the time range of analysis; more accurately identify trends due to the general environment and specific features of each individual organization.

Acknowledgements.

This work was supported by the Russian Foundation for Basic Research, Grant No. 18-011-00929 A.

References

Bauer, M. W., Allum, N., & Miller, S. (2007). What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda. *Public Understanding of Science*, *16*(1), 79–95. https://doi.org/10.1177/0963662506071287

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi.23329

Carver, R. B. (2014). Public communication from research institutes: Is it science communication or public relations? *Journal of Science Communication*, *13*(3), 1–4.

Elmer, C., Badenschier, F., & Wormer, H. (2008). Science for everybody? How the coverage of research issues in German newspapers has increased dramatically. *Journalism & Mass Communication Quarterly*, *85*(4), 878–893.

Holliman, R. (2004). Media coverage of cloning: A study of media content, production and reception. *Public Understanding of Science*, *13*(2), 107–130. https://doi.org/10.1177/0963662504043862

Kosyakov, D. V, Basyleva, E. A., Yudina, Y. A., Pavlova, I. A., Vasilieva, N. V, Dubovenko, V. A., & Guskov, A. E. (2018). Science news aggregation: media analysis and usage statistics. *Scientific and Technical Information, Series 1.*, (3), 11–17.

Phillips, D., Kanter, E., Bednarczyk, B., & Tastad, P. (1991). Importance of the lay press in the transmission of medical knowledge th the scientific community. *New England Journal of Medicine*, *325*(16), 1180–1183.

Wilkinson, C., & Weitkamp, E. (2013). A Case Study in Serendipity : Environmental Researchers Use of Traditional and Social Media for Dissemination, *8*(12), 1–9. https://doi.org/10.1371/journal.pone.0084339

Analyzing and Extracting Data Resource Entities in Full-text Papers

Qi Zhang [1], Youshu Ji[2], Si Shen[3] and Dongbo Wang[4]

[1] *grangergogo@gmail.com*
Nanjing Agricultural University (China)

[2] *2018114009@njau.edu.cn*
Nanjing Agricultural University(China)

[3] shensi@njust.edu.cn
Nanjing University of Science and Technology(China)

[4] *db.wang@njau.edu.cn*
Nanjing Agricultural University(China); KU Leuven（Belgium）

Introduction and Motivation

Due to the significant role of data resources in academic studies, automatic extraction of data resource entities in academic full-texts is not only helpful to the capture of target information, but also the fundamental research of datasets evaluation, datasets citations and datasets retrieval (Ding et al., 2013). However, most of the current entity extraction research of data resources focus on standard data sets, such as “TREC Robust 2004 collection”, while self-built data sets have not been thoroughly investigated as they don’t have specific name generally. In fact, this problem can be mitigated by tracing the sources of self-built data sets. For example, “Accordingly, we relied on Google scholar to create a corpus” pointed out that the data of the research were obtained from “Google scholar”, Li K (2018) has explored the impact of WOS as research tool and data resource, but they only focus on one of the data resources. On the other hand, present studies usually adopt ule-based extraction methods and CRF, but these methods are limited by low recall rates. Besides, they rely on the complicated rules raised by experts.

Hence, we established a data set consisting of 892 publications, which includes the label of the standard data collections and source of self-built data sets. Then, we analysed the distribution of the two kinds of data resources, and evaluated the performance of three entity extraction methods, namely CRF, Bi-GRU+CRF, BERT, for identifying data resource entities in academic full-texts.

Introduction to the data set

The data source used is the full texts of 892 academic articles published in 2012-2016 from *JASIST*, the total number of sentences is 285, 026. As mentioned previously, there are two kinds of data resources were labelled in our dataset. We have detailed labelling regulations, and the labelling was completed by graduate students who majored in Information Science. The basic information of our data set is presented in Table 1.

Table 1. The basic information of the data set

Sentences	Data resource entities	Articles containing data resources
285, 026	14, 973	741

Distribution of the data resource entities

The combination of data resource and its position could contribute to extracting of data re-use more accurately. Based on our previous studies of structure function of academic text, the 741 academic publications containing data resource entities were mapped into six structures, including introduction, method, related research, conclusion, references and others. We counted the locations of data resource entities, as showed in Fig. 1, 63% of all data resource entities were in the experimental and/or method chapters and these entities are likely to be directly used by the research, while entities in the related research part tend to be data sets used by other studies.

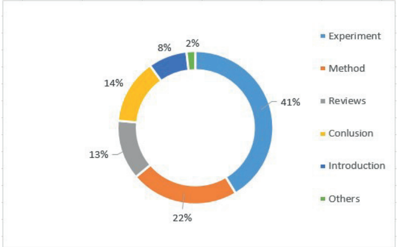


Figure 1. Distribution of data resource entities in academic publications.

Table 2. Top five standard data sets and data sources with highest document frequencies

	Data sources of self-built data sets	DFe	Standard data set	DFe

1	Web of Science	212	TREC dataset	18
2	Wikipedia	141	ClueWeb09	15
3	Scopus	136	OHSUME	9
4	Google Scholar	90	ISI data	8
5	PubMed	69	LETOR 3.0	6

Table 2 presents the top five data sources and top five standard data sets with highest document frequency in JASIST. As observed, establishment of self-built data sets is popular in academic publications, relatively few studies using standard data set. Meanwhile, representation of data source is relatively fixed, while standard data sets present various representations and may be exposed to inclusion and inclusive relations (e.g., ClueWeb09 is a part of TREC data set). Hence, data set classification will be further investigated in our future research.

Training And Testing CRF, Bi-GRU+CRF, BERT Models

Previous studies of entity extraction from academic full-texts tend to use small scale data sets and usually based on rule matching or machine learning model. Nevertheless, extractions based on rule matching or machine learning model are highly depending on rules or features. Deep learning models can obtain features by training of neural network parameters and have been widely applied in natural language processing.

CRF has been applied in related researches. We use CRF++ (Kudo, T.2010) for this experiment. In the Bi-GRU+CRF model, CRF is used as a sentence-level output optimization interface to obtain global optimum solution. BERT (Devlin, J., Chang, 2018) is a pre-training language model, which adopts the bidirectional self-attention mechanism and has refreshed the best achievements in various natural language processing tasks. In our research, the pre-training model with 11 layers, 748 hidden units and 12 self-attention heads provided by Google was employed after fine tuning. The main parameters of the two deep learning models are shown in Table 3.

Table 3. Main parameters of the two deep learning models in this study.

Model	Settings
Bi-GRU-CRF	Dimension of wordvec=200, hidden layers=256, learning rate=200
BERT	Learning rate=2.0E-5; max sequence length=256; batch size=64

The data set was split into train set and test set, and ten-fold cross-validation method was employed. For each model, the average of accuracy, recall rate and FB1 were regarded as the ultimate values.

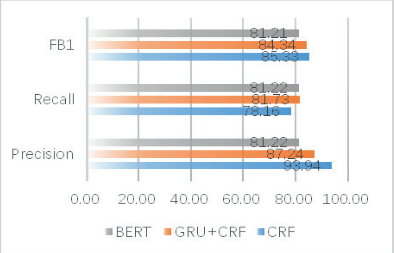


Figure 2. Extractions of academic data resources using the three models (%).

Among the three models, CRF has gotten high accuracy but low recall rate; Bi-GRU-CRF and BERT presented effective extractions acquisition with higher recall rates as they can automatically acquire context semantics information. Future studies should work on improving of high accuracy and high recall rate simultaneously.

Conclusion and Future Work

In this study, data resource entities in full texts of academic articles were classified as standard data sets and self-built data sets. Based on that, a data set containing 839 articles (285, 026 sentences) from JASIST was obtained and labelled. Then, we analysed the distribution of the two kinds of data resource in the publications. Finally, the Bi-GRU-CRF model and the BERT model were introduced into the extraction of data resource entities from full-text publications and the recall rate of both models increased. Future work will focus on overall improvements of extraction and more finely classification of data resource entities.

Acknowledgments

Researches are supported by the National Natural Science Foundation of China (Grant Number: 71673143).

References

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PloS one*, 8(8), e71416.

Li K, Rollins J, Yan E. Web of Science use in published research and review papers 1997–2017: a selective, dynamic, cross-domain, content-based analysis[J]. *Scientometrics, 2018*, 115(1): 1-20.

Kudo, T. (2010). CRF++: Yet another CRF toolkit. Available under LGPL from the following URL: *http://crfpp. sourceforge.net.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Research on Software Entities Extraction and Analysis Based on Deep Learning

Chuan Jiang[1], Qian Wang[2], Zihe Zhu[3], Si Shen[4] and Dongbo Wang[5]

[1] *jiangchuan_321@163.com*
Nanjing Agricultural University(China)

[2] *Jamie_wq@outlook.com*
Nanjing Agricultural University(China)

[3] *zihe.zhu@qq.com*
Nanjing Agricultural University(China)

[4] *shensi@njust.edu.cn*
Nanjing University of Science and Technology(China)

[5] *db.wang@njau.edu.cn*
Nanjing Agricultural University(China); KU Leuven(Belgium)

Introduction

As an important part of academic research, software entities can't comprehensively and accurately assess the impact from the perspective of metadata. Therefore, the key to assessing the impact of software entities is how to accurately extract software entities from academic full texts. Current research on entities measurement mainly includes from the perspective of entities network (Ding et al., 2013) and citation (Pan et al., 2015), and frequency (Yan & Zhu, 2015). However, there is still room for further improvement in the software entities identification method adopted. Moreover, the performance of deep learning in software entities extraction has not appeared in present researches yet. In this paper, software entities extraction experiment is performed on the full-text papers from JASIST by using machine learning, deep learning model and the latest NLP model BERT. Therefore, the purpose of this research is to improve the performance of software entities identification and provide precise identification methods for full-text entities measurement, and improve the measurement effect.

Introduction to data source and model and Analysis of the distribution

The data were sourced from 892 published articles on JASIST from 2012 to 2016. These articles were segmented into 285,026 sentences. Software entities were labelled in BEMS format, mainly including Web service, programming language, statistical analysis and literature management softwares.

After software entities labelled, the text structure was merged into an introduction, literature review, methods, experiments, conclusions and other six structures through artificial introspection. As shown in Table 1, statistical analysis was performed for text structure in terms of software entities in the academic full texts on JASIST. Since the current scientific research programs are largely driven by data, there are large amount of software entities in the experiment and method sections. These softwares are mainly intended for data acquisition and data mining. The conclusion section is mainly to use certain software to discover the potential law from data. Introduction and literature review sections generally review the history of using certain methodology or softwares for conducting scientific research activities.

Table 1 Distribution of software entities in text structure

Text structure	Freq.
Introduction	1441
Literature review	1703
Method	2880
Experiment	3351
Conclusion	1878
Other	270

Entities extraction is a crucial step in entities measurement of full texts. In order to select the optimal software entity extraction model, we compare the performance of machine learning and deep learning models. In this experiment, words were used as features and no other artificial features were incorporated. The data were divided into 10 labelled training sets and test sets. The 10-fold cross-validation experiments for software entities recognition were performed using CRF++ model, Bi-GRU-CRF model and BERT model respectively.

CRF++ (Kudo, 2010) model is a discriminant probabilistic undirected graph model. In the present research, basic feature template was used for software entities extraction.

BiGRU-CRF model consists of embedding layer, bi-directional GRU layer and CRF layer (Jiao et al., 2018). In order to avoid gradient explosion and disappearance, gradient cropping was adopted, and it was set to 5.0. The learning rate was initialized to 0.001, dimensionality of word embedding 200, and the number of hidden units 256. In order to avoid overfitting and to facilitate training speed, early stopping method was used. That is, the training would stop if the F-measure of the cross-validation set did not increase on 3 iterations.

BERT is a neural network model proposed by Devlin et al. (2018), and has achieved the optimal effect on 11 NLP tasks. Transformer is the core component of BERT. By using transfer learning, the output layer of Google's pre-training English BERT model was modified for software entities extraction. The number of hidden units was 768, self-attention heads 12, warmup_proportion 0.1, learning rate 2.0E-5, max_seq_lenth 256 and epoch 3.

Results

CRF++, BiGRU-CRF and BERT models were respectively used to recognize the software entities in the academic fill texts. The mean P, R and F-measures of 10-fold cross-validation were taken as the final results.

Table 2 Comparison of software entities recognition results

	CRF++	BiGRU-CRF	BERT
P	95.12%	89.50%	85.81%
R	78.68%	87.37%	85.10%
F	86.12%	88.40%	85.44%

The model recognition results are shown in Table 2. Of the three models, the CRF++ model had the highest precision but lower recall rate. This fact indicated that without adding artificial features, the CRF++ model could effectively recognize the frequently occurring entities, but lacked semantic understanding. May for this reason, the software entities with a lower occurrence frequency was not effectively recognized. While BiGRU-CRF and BERT models used the same embedding mechanism, incorporates the semantic information of words. Therefore, may for this reason, the two models effectively recognized the software entities based on semantics, resulting in increasing the recall rate. There was an additional CRF layer in the BiGRU-CRF model, which was capable of effective modelling of the transition probability of labels and avoiding label mistakes. May for this reason, the precision and recall of the BiGRU-CRF was higher than that of BERT. Table 3 shows the top 10 software entities with the highest occurrence frequency in the full texts of JASIST. It can be found that the software mainly involves statistical analysis, data mining, literature management and visualization and so on.

Table 3 Top 10 softwares with the highest occurrence frequency in the full texts of JASIST

Software	Freq.	Software	Freq.
Mendeley	724	Pajek	93
CiteULike	124	LIWC	89
SciMAT	116	SemRep	73
SPSS	102	CharaParser	59
VOSviewer	99	AnaCoTEx	59

Conclusion

In the present research, three models were used for software entities extraction experiments, respectively. It was found that the Bi-GRU-CRF model has the highest F-measure with 88.4%. The distribution and occurrence frequencies of software entities in the text structure were analyzed statistically. However, our research had certain limitations, one of which was the small sample size. For future research, the Bi-GRU-CRF model will be applied to other journals on information science. Large-scale software entities extraction and measuring impact of software entities will be carried out. Moreover, data mining and analysis will be performed on software entities diffusion and software entities network, so as to promote the development of informetrics of academic full texts.

Acknowledgments

Researches are supported by the National Natural Science Foundation of China (Grant Number: 71673143).

References

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PloS one*, 8(8), e71416.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805.

Jiao, Z., Sun, S., & Sun, K. (2018). Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *arXiv preprint arXiv:1807.01882.*

Kudo, T. (2010). CRF++: Yet another CRF toolkit. Available under LGPL from the following URL: http://crfpp. sourceforge. net.

Pan, X., Yan, E., Wang, Q., & Hua, W. (2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics*, 9(4), 860-871.

Yan, E., & Zhu, Y. (2015). Identifying entities from scientific publications: A comparison of vocabulary-and model-based methods. *Journal of informetrics*, 9(3), 455-465.

Identifying and evaluating strategic partners for collaborative innovation: A method based on a topic analysis of papers and patents

Yan Qi [1]* ZhengyinHu[2] Bin Xiang[3] ChunjiangLiu[4] HaiyunXu[5] YiWen[6]

[1] *qi.yan@imicams.ac.cn
Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences/Peking Union Medical College (CAMS&PUMC), No. 3, Yabao Road, Beijing, 100020 (China)

[2]huzy@clas.ac.cn; [3]xiangb@clas.ac.cn; [4]liucj@clas.ac.cn; [5]xuhy@clas.ac.cn; [6]weny@clas.ac.cn
Chengdu Documentation and Information Center, Chinese Academy of Sciences, No. 16, Nan‘erduan, Yihuan Road, Chengdu, 610041 (China)

Introduction

Economic globalisation has achieved many collaborative innovations such as strategic alliances, industrial clusters, and science parks, and owing to a rapid development, the problems incurred are increasing in number. For example, a homogeneous layout and similar development path make it difficult to achieve a differentiated development, which in turn exacerbates vicious competition; in addition, the correlation among innovation entities and industries is too weak to form an industrial symbiosis. These factors have led to high failure rates and the inability to achieve the expected synergy.

An effective selection of partners is a core factor affecting the collaboration performance, and guidelines or references have been developed (Hitt, et al., 2000). Most studies proposing quantitative indicators (Geum, et al., 2013) have been carried out using the bibliographic information of previous papers or patents without going deep into the literature content, whereas others considering the subject correlation (Xu, et al., 2016) are limited to only one type of innovation connotation (or one node in the innovation chain), namely theoretical research or technical development.

We believe that, for a specific problem, some institutions that have published numerous related papers are good at theoretical research, whereas other institutions that have numerous patents are good at technology development and product manufacturing. If a strategic co-operation can occur between these two types of institutions based on their complementary advantages, the transformation process of their theoretical research results into actual technological development will be promoted and the problem will be better solved, thereby avoiding vicious competition to a certain extent. Hence, in hopes of achieving the expected synergy and avoidance, as well as providing reference for further research, this study attempts to develop a systematic framework for collaborative innovation partner selection through a topical analysis of both papers and patents, which are a deep concern regarding the highly correlated connotation of science and technology innovation.

Methodology

Theoretical and methodological basis

In addition to the ideas mentioned above, there is another basis for this study. Scholars have pointed out that knowledge may flow from science to technology or from technology to science. We explored existing research on the science–technology relationship and proposed a new integrative index (Qi, et al., 2018) based on the topic analysis of theses and patents, based upon which we can obtain correlative science–technology themes that can be used as a foothold to carry out collaborative innovation, which can be viewed as the ‘specific problem’ mentioned above.

Method framework

We constructed the following framework:

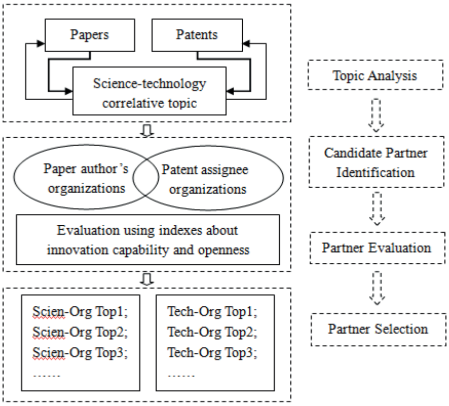

Papers
Patents
Science-technology correlative topic
Topic Analysis
Paper author's organizations
Patent assignee organizations
Evaluation using indexes about innovation capability and openness
Candidate Partner Identification
Partner Evaluation
Scien-Org Top1;
Scien-Org Top2;
Scien-Org Top3;
……
Tech-Org Top1;
Tech-Org Top2;
Tech-Org Top3;
……
Partner Selection

Figure 1. Framework and process flow.

Step1: Mining of correlative topics

The topics of two collections of papers and patents in a specific area are generated using topic modelling (e.g. LDA, PLDA, and PLSA), and the strongly associated topics are determined based on their high similarities. Patents and publications have different objectives and do not always share a similar wording, and thus term clumping is critical. Furthermore, experts need to be involved in a manual interpretation and evaluation to determine the appropriate theme for cooperation.

Step2: Evaluation index of alternative institutions

One correlative topic usually corresponds to multiple patents and papers, in turn corresponding to the institutions of multiple paper authors (Scien-Org) and patentees (Tech-Org). Therefore, further evaluation is needed regarding which is more suitable. We selected two criteria, the institution's innovation capacity (IA) and its attitude towards open innovation (IO), the dimensions and corresponding indicators of which are as follows:

Table 1. Criteria and indexes.

Institute	Criteria	Indexes
Scien-Org	IA	Pu, PuI, PuR, PuC, PuY
	IO	Puc,Pucr, PuCI, PuCY,PuCF
Tech-Org	IA	Pa, PaI, PaR, PaC, PaY
	IO	Pac, Pacr, PaCI, PaCY, PaCF

Pu is the number of publications of an institution on a topic, Pa is the number of patents, and PuCI is the number of institutional partners.

Step3: Institute ranking and matching

We propose a single integrated index, called CII, to conveniently sort the candidate institutions. As equations (1) and (2) show, different weight coefficients for different indexes (e.g. W_{11} and W_{12}), criteria (e.g. W_{21} and W_{22}), and their normalisation scores (e.g. Puc') are needed.

$$\text{Scien-Org: } \mathrm{CII} = \mathrm{W}_{11} * (\mathrm{W}_{21} * \mathrm{Puc}' + \mathrm{W}_{22} * \mathrm{Pucr}' \cdots) + \mathrm{W}_{12} * (\mathrm{W}_{31} * \mathrm{Pu}' + \mathrm{W}_{32} * \mathrm{PuI}' \cdots) \quad (1)$$

$$\text{Tech-Org: } \mathrm{CII} = \mathrm{W}_{11} * (\mathrm{W}_{21} * \mathrm{Pac}' + \mathrm{W}_{22} * \mathrm{Pacr}' \cdots) + \mathrm{W}_{12} * (\mathrm{W}_{31} * \mathrm{Pa}' + \mathrm{W}_{32} * \mathrm{PaI}' \cdots) \quad (2)$$

Each institution can be graded according to the CII value, and the matching can then be conducted based on certain principles. There may be other factors to be considered and suggestions proposed for the corresponding institutions to carry out research cooperation on a specific topic.

Case Study

For the years 2016 and 2017, 53 science topics were extracted from 6,985 papers, and 104 technology topics were extracted from 975 patents in the Hepatitis C virus (HCV) research field. We chose the topic of HCV detection, which corresponds to 430 papers and 347 author institutions, as well as 109 patents and 88 patentee institutions. The CII value was obtained using the following formula (3), and a partial ranking of the institutions is shown in Table 2:

$$\mathrm{CII} = 0.5 * \left(\frac{\mathrm{Pu}'}{\mathrm{Pa}'}\right) + 0.5 * \mathrm{PPuci}' \quad (3)$$

Table 2. Organisational ranking(partial).

Rank	Scien-Org	CII	Tech-Org	CII
1	Univ Roma Tor Vergata	0.56	Hoffmann-La Roche Co	1.00
2	Cairo Univ	0.54	JiNanUniv	0.67
…	…	…	…	…

Conclusions

Relatively speaking, the proposed method is targeted at specific scientific problems and innovation connotations located at different nodes along the innovation chain, which theoretically has a higher success rate and can avoid homogenised competition. The determination of related topics is most important, and domain experts are needed in the interpretation and selection of correlative topics worthy of cooperation. In addition, an extension of the document collection corresponding to the topic may be necessary for a better evaluation and ranking of alternative institutions. Evaluation indicators such as an innovation capacity or attitude are also critical. We believe that 'solving common problems and complementing each other in terms of capabilities' are guarantees for a successful cooperation. We are now focusing on the selection of partners at the institutional level for more resources and greater stability and may focus on the research team level in the future.

Acknowledgments

The work was supported by the Basic Scientific Research Project and 2017 'Peking Union Youth Research Fund' project of the Chinese Academy of Medical Sciences (Grant nos. 2017PT63008 and 2017330008).

References

Geum, Y., Lee, S., Yoon, B., et al. (2013). Identifying and evaluating strategic partners for collaborative R&D: index-based approach using patents and publications. *Technovation*, 33(6-7), 211-224.

Hitt, M. A., Dacin,M. T., Levitas, E., et al. (2000). Partner selection in emerging and developed market contexts: Resource-based and organizational learning perspectives. *Academy of Management Journal*, 43(3),449-467.

Qi, Y., Hu, Z. Y., Liu, Z. Q., et al. (2018). Exploration of a Science-technology relationship index and its measurement algorithm. *In Proceedings of 8th Global TechMining Conference*, Leiden:CWTS.

Xu, H. Y., Wei, L., Pang, H. S., et al. (2016).Methods to identify potential industry-university-research institutions cooperation partners. *Journal of The China Society for Scientific and Technical Information,* 35(5),521-529.

Online Attention of Scholarly Papers on Psychosocial Hazards - Job Stress, Bullying and Burnout

Witold Sygocki[1] and Małgorzata Rychlik[2]

[1]*wisyg@ciop.pl*
Central Institute for Labour Protection-National Research Institute, Czerniakowska 16, Warsaw 00-701 (Poland)
[2] *rychlik@amu.edu.pl*
Poznań University Library, Ratajczaka 38/40, 61-816 Poznań (Poland)

Introduction

This study focuses on the articles of Polish and Italian researchers related to psychosocial hazards. Workplace environment plays a major role in the performance and productivity of an employee and can exert enormous influence on physical well-being of man, or a lack of it thereof. In 2005 nearly one in four workers in the European Union (EU)-27 reported to be affected by work-related stress and it has become the second most reported health-related problem at work (Mellor, 2017 OSHWiki). It seems worthwhile then to examine whether those works that analyse the phenomena of job stress, bullying and burnout are met with enough appreciation and get proper attention on social media sites, within and beyond scientific communities. These similar phenomena are defined in different ways. Stress is usually defined as a perceived imbalance between the demands made on people and their resources or ability to cope with those demands; can be caused by multiple factors, one of that is work) (Mellor 2017, OSHWiki). Various terms are used to describe repeated and long-term negative treatment at work - bullying is the English term most commonly used by researchers all over the world, other ones "harassment" or sometimes "psychological harassment" is being used more widely. "Mobbing" is used in some countries and by some researchers (Vartia, 2016 OSHWiki). Burnout has been defined as a prolonged response to chronic emotional and interpersonal stresses on the job) (Ahola, 2013 OSHWiki).

Altmetrics is a method to assess the spread of scientific knowledge, e.g. sharing papers on Twitter, Facebook or blogs (Halevi & Schimming, 2018).

The main aim of this study was to examine if scholarly papers on psychosocial risks provide altmetric indicators and to compare papers written by Polish and Italian scholars. The publications in question are indexed in the Web of Science and Scopus and have a sizeable representation within the field of Occupational Safety and Health (OSH), including psychosocial hazards. There are institutes responsible for conducting research on OSH issues in either of the countries. In Italy, it is the National Institute for Prevention and Safety at Work, whereas in Poland the counterpart institution is the Central Institute for Labour Protection – National Research Institute.

On the basis of the obtained data, the authors attempt to provide answers to the following questions: RQ1. Do Polish and Italian scientific articles in the field of psychosocial risks have altmetric indicators? RQ2 Which altmetric indicators are most common, and which are the least frequent indicators in Polish and Italian works? RQ3 What are the average numbers of altmetrics per paper? RQ4 Do citation counts of articles correlate with Twitter mentions and Mendeley readers?

Methods

The study was divided into two stages. Stage 1 was to collect and filter the data obtained from the Scopus database. The metrics data collected from Scopus were related to the authors affiliated to Polish and Italian scientific institutions, included the keywords in the field of psychosocial hazards, and were limited to articles only (a simple search in Scopus "keywords": bullying, burnout, job stress and "*affiliation country*": Poland and Italy). The chronological scope of the study covered the years 2013-2018. Citation counts were collected for all papers. Only articles that were assigned a DOI were analysed (*N* Italy=594, *N* Poland=241).

The other stage of the study involved the use of the Altmetric Explorer (http://www.altmeric.com), which provided the present authors with altmetric indicators (Robinson-García, Torres-Salinas, Zahedi, & Costas, 2014). The data were collected on 27-29 of March 2019.

Results

The articles collected 23,069 altmetric indicators.

The highest number of altmertics was provided by Mendeley and Twitter, both for Italian and Polish papers.

Table 1. Altmetrics of Italian / Polish scholarly papers.

Keyword	**Number of papers with DOIs**		**Number of papers with altmetrics**		**Number of altmetric indicators**	
	N		*N*		*N*	
	IT	**PL**	IT	**PL**	IT	**PL**
Bullying	160	**29**	117	**24**	6226	**1515**
Burnout	196	**107**	84	**28**	5088	**1416**
Job stress	238	**105**	119	**37**	5537	**2044**
Total	594	**241**	320	**88**	16851	**4975**

The highest Altmetric Score (an indicator of the amount of attention that a research output has received) for Italian articles was 537 (bullying paper). This means that the article was in the 97th percentile of outputs of the same age and source. The highest Altmetric Score for Polish articles was 340 (job stress paper). The article was in the 90th percentile.

The results show a significant positive correlation between Mendeley readers and citation counts for both Italian and Polish articles (Thelwall, 2016), (Ortega, 2016). However, the correlation between Twitter and citation counts is statistically insignificant.

The average number of altmetric indicators per article was the highest for an article in Polish and was 151 (bullying paper). The highest average number of altmetrics for the Italian papers was 99 (bullying paper). The lowest average numbers of indicators were 17 for works in Polish (burnout paper) and 23 for the Italian papers (burnout paper), respectively.

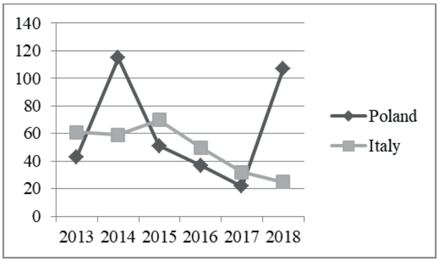


Figure 1. Average number of altmetrics per paper - job stress

Limitations

An important limitation of this paper is that in the study we chose arbitrarily only three keywords related to psychosocial hazards, even though a number of various terms are used to describe repeated and long-term negative treatment at work. By searching PubMed with the MeSH terms we also found other, far more expanded, entry terms related to the topic (e.g. job stress - 29 entry terms).

In addition, the key words: job stress, bullying and burnout had much higher representation in Scopus as compared to their derivatives or related terms (Scopus query: All countries, All years – query 20 May 2019; for the term "bullying" there are 14,400 records, whereas for the term "mobbing" the number is 1261; for "job stress" – 10,044 records and for "work stress" – 3,780.

It should be emphasised, that the purpose of this article is to show the tendency to non-traditional reach of articles, therefore, it was not that important to analyse all possible terms.

References

Halevi, G., & Schimming, L. (2018). An Initiative to Track Sentiments in Altmetrics. *Journal of Altmetrics*, *1*(1). http://doi.org/10.29024/joa.1

Ortega, J. L. (2016). To be or not to be on Twitter, and its relationship with the tweeting and citation of research papers. *Scientometrics*, *109*(2), 1353–1364. http://doi.org/10.1007/s11192-016-2113-0

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: exploring the insides of *Altmetric.com*. *El Profesional de La Informacion*, *23*(4), 359–366. http://doi.org/10.3145/epi.2014.jul.03

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. *Scientometrics*, *108*(1), 337–347. http://doi.org/10.1007/s11192-016-1973-7

Vartia, M. (2016). Harassment at work. *OSHWiki.* Retrieved 15:57, May 20, 2019 from https://oshwiki.eu/index.php?title=Harassment_at_work&oldid=245876.

Mellor, N. (2017). Psychosocial risks and work-related stress: risk assessment. *OSHWiki.* Retrieved 16:08, May 20, 2019 from https://oshwiki.eu/index.php?title=Psychosocial_risks_and_work-related_stress:_risk_assessment&oldid=247240.

Ahola, K. (2013). Understanding and Preventing Worker Burnout. (2013, April 26). *OSHWiki.* Retrieved 16:42, May 20, 2019 from https://oshwiki.eu/index.php?title=Understanding_and_Preventing_Worker_Burnout&oldid=237985

This information has been based on the results of a research task carried out within the scope of the fourth stage of the National Programme "Improvement of safety and working conditions" partly supported in 2017-2019 — within the scope of state services — by the Ministry of Labour and Social Policy. The Central Institute for Labour Protection – National Research Institute (CIOP-PIB) is the Programme's main co-ordinator.

Morphological Features and Citation Counts of Academic Books

Siluo Yang[1], Yiyi Yang[2] and Shaoyun Xiao[3]

[1]58605025@qq.com, [2]yangyiyi@whu.edu.cn, [3]412821663@qq.com
School of Information Management, Wuhan University, Wuhan 430072, China

Introduction

As an important carrier of inherited academic information, academic books are records that disseminate innovative achievements in a certain field. These books have always been regarded as important academic achievements. Moreover, citations can quickly and quantitatively measure researchers' performance and publications' academic value. The safest way to be frequently cited is to write a high-level and high-quality book. However, do other attributes surrounding the mere text also influence the citation counts of academic books?

Nair and Gibbert (2016) found that non-alphanumeric characters have small but important effects on the citations of journal papers. Gnewuch and Wohlrabe (2017) stated that a short title that contains non-alphanumeric characters easily achieves a relatively high citation count. Several studies on the external characteristics of journals exist, but research about the relationship between the external characteristics and influences of academic books is limited.

The Book Citation Index (BKCI) is designed to incorporate comprehensive book citation data. Many scholars have used data from BKCI to conduct various studies in combination with the influences and various measurement indicators of academic books.

The present study uses all the book records included in the SSH&S database of BKCI from 2013 to 2017 as sample. Moreover, the relationship among book citations, length attributes, character attributes, and types of books is analyzed.

Method and Data

Our paper downloaded the complete records for the period 2013–2017 from BKCI, including SSH&S. We obtained a total of 666,619 records after excluding erroneous ones.

Furthermore, we grouped the academic books according to their morphological features, such as "length attributes," "character attributes," and "types of books."

Subsequently, we employed the Statistical Product and Service Solutions to conduct stepwise regression analyses on the morphological features and citations. The dependent variable is book citations.

**Correlation is significant at the 0.01 level (two-tailed).

Result and Discussion

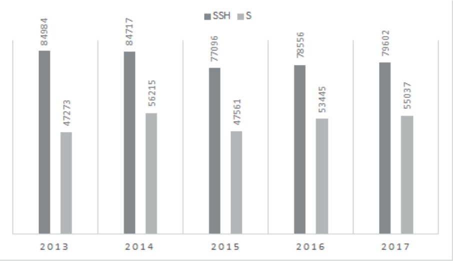


Figure 1. Record volume of SSH&S (2013-2017)

Table 1. Morphological attributes of academic books

Length attributes
Number of title words
Number of references
Number of pages
Character attributes
Number of non-alphanumeric characters
Type of books
Book or book chapter

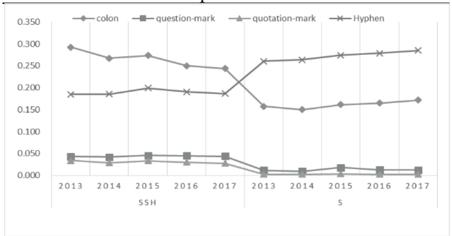


Figure 2. Relative frequency: colon, question mark, question mark and hyphen of SSH&S (2013-2017)

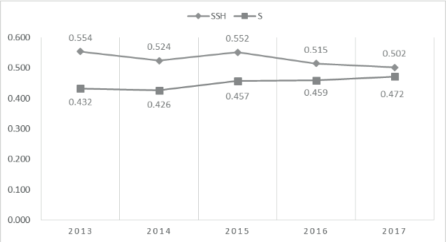


Figure 3. Relative frequency: Non-alphanumeric character of SSH&S (2013-2017)

The most common non-alphanumeric characters that appear in article titles are colon, question mark, quotation mark, and hyphen (Buter 2011). In Social Science & Humanities, colon is used most frequently, whereas hyphen is used most frequently in Science.

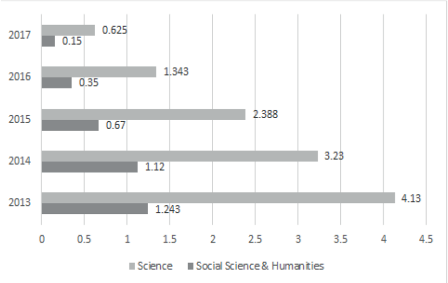


Figure.4 Mean: Citation of SSH&S (2013-2017)

Each year, the average academic books cited in Science is higher than that in Social Sciences.

Table 2. Regression results

	Coefficient	Beta	significance
Constant	1046.992		
Number of references	.022	.128	0.000
Published year	-.519	-.065	0.000
Number of pages	.007	.051	0.000
Number of non-alphanumeric characters	.129	.008	.000
Number of words		$.001^{e}$	0.665
Type of books		$.000^{e}$	0.959

Table 2 indicates that the number of references and pages have positive relationships with citation counts. To avoid the impact of publication age on the results, we conducted a stepwise regression analysis on a yearly basis by discipline.

Considering the massive data from each indicator's regression analysis, we only provided the partial results here.

Table 3. Regression results: types of books

SSH	Coefficient	Standard Error	Beta
2013	7.23^{**}	0.163	0.151
2014	5.659^{**}	0.211	0.092
2015	2.845^{**}	0.063	0.161
2016	1.018^{**}	0.030	0.122
2017	0.44^{**}	0.012	0.127

Table 4. Regression results: number of non-alphanumeric characters

S	Coefficient	Standard Error	Beta
2013	1.155^{**}	0.152	0.035
2014	1.288^{**}	0.092	0.059
2015	0.861^{**}	0.063	0.063
2016	0.431^{**}	0.037	0.051
2017	0.169^{**}	0.016	0.045

Table 5. Regression results: number of references

S	Coefficient	Standard Error	Beta
2013	0.07^{**}	0.002	0.190
2014	0.051^{**}	0.001	0.202
2015	0.043^{**}	0.001	0.256
2016	0.022^{**}	0.000	0.215
2017	0.01^{**}	0.000	0.221

Book citations are relatively short, thus many zeros are cited, resulting in R square < 0.1. This finding can only explain a small part of the records.

Conclusion

In Social Science & Humanities, "types of books" has a significant relationship with citation counts. The citation counts of books suggest more impact than those of book chapters.

In Science, the numbers of non-alphanumeric characters and references indicate a positive effect on citation counts.

Acknowledgments

This research is funded by the National Social Science Fund Key Project of People's Republic Of China (17ATQ009).

References

Buter, R., & van Raan, A. F. (2011). Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact. *Journal of Informetrics*, 5(4), 608–617.

Matthias Gnewuch, Klaus Wohlrabe. (2017). *Title characteristics and citations in economics Scientometrics*, 110(3), 1573-1576.

Nair, L. B., & Gibbert, M. (2016).What makes a good title and (how) does it matter for citations? A review and general model of article title attributes in management science. *Scientometrics,* 107(3), 1331–1359.

Linking individual-level to community-level thematic change: How do individual research trails match disjoint clusters of direct citation networks?

Jochen Gläser[1], Matthias Held[2] and Grit Laudel[3]

[1]*Jochen.Glaeser@tu-berlin.de*
TU Berlin, Hardenbergstr.16-18, 10623 Berlin (Germany)

[2]*held@ztg.tu-berlin.de*
TU Berlin, Hardenbergstr.16-18, 10623 Berlin (Germany)

[3]*Grit.Laudel@tu-berlin.de*
TU Berlin, Fraunhoferstr. 33-36, 10587 Berlin (Germany)

Introduction

Linking the micro-dynamics of individual research processes to the macro-dynamics of scientific fields is a key unresolved problem of science studies. Macro-level change is an emergent effect of micro-level decisions, which is why causation of macro-level changes goes through the micro-level even for macro-level causes (Gläser 2017). Establishing causal micro-macro-links would not only enhance the power of micro-level explanations but also contribute to our understanding of macro-level structures and dynamics as represented by the mapping of scientific fields.

The aim of the research presented in this poster is to contribute to the development and validation of bibliometric methods that establish micro-macro links. We link individual research trails of seven physicists working in atomic and molecular optics (AMO), five of which moved and two of which did not move to the topic of experimental Bose-Einstein condensation (BEC), to a macro-level clustering of AMO physics with the new Leiden algorithm. The individual research trails were verified in interviews. Our question is how the perceptions of topics by researchers working on them match clustering solutions obtained at different resolutions.

Data and Methods

Research trails

We constructed research trails for the five researchers by downloading their publications from the web of science, constructing bibliographic coupling networks (using Salton's cosine for bibliographic coupling strength) and choosing a threshold for the strength of bibliographic coupling at which the network disaggregates into components (Gläser and Laudel 2015). The 'manual' approach is preferable to clustering because the research trails serve as means of 'graphic solicitation' in interviews, for which instant visual recognition of different topics is essential. The components represent topics a researcher has worked on over time.

Interviews

The researchers were interviewed about their research topics beginning with their PhD topic, with an emphasis on thematic changes and the reasons for them. Developments in the interviewee's national and international communities were also discussed. The interviews lasted on average 90 minutes and were fully transcribed. Transcripts were analysed by qualitative content analysis.

Clustering

To construct the macro-level AMO dataset, we started delimiting the WoS by selecting all publications from journals in the subject category 'Physics, Atomic, Molecular & Chemical' published 1990-2005, excluding physical chemistry journals. We then expanded this dataset by (1) including publications from all other physics subject categories (in the same time frame) that cited at least two publications from our first delineation (extended to 1975-2005); and (2) by including publications from all other physics subject categories which have been co-cited with at least two papers from our first delineation. The direct citation network of this extended dataset has a giant component with 366,480 publications, which included all relevant 147 publications of the research trails. We applied the Leiden algorithm (Traag et al. 2018) for a coarse clustering and extracted the largest cluster with 96,137 publications including 146 of the research trails' publications. This served as our macro-level AMO dataset.

We applied the Leiden Algorithm to this dataset with two different resolution levels (6 e-5 and 2 e-4), minimum size 1000 and 500, discarding smaller clusters, which resulted in 10 and 31 clusters, respectively.

Labelling

Noun phrases have been extracted from titles and abstracts from all publications using part-of-speech tagging in python's 'nltk' package. To find characteristic terms for each of the clusters in both clustering solutions, these noun phrases were used in the mutual information-based labelling introduced by Koopman and Wang (2017).

Results

Matches

Most of the clusters in both solutions remained remarkably stable over time. None of the larger clusters began later than 1990 or ended earlier than 2005. Only very few clusters grew rapidly during the 15 years of study (e.g. clusters 5 and 0 in Figure 1), while the others showed little or no growth.

The BEC cluster (cluster 0 in Figure 1) was clearly identifiable through keywords and by the match with research trails. All researchers who switched to BEC had corresponding research clusters in their research trails. The case depicted in Figure 1 conducted research (PhD and postdoc) on different topics before becoming involved in BEC in the mid-1990s. He was then forced to work on different topics because he was still a dependent researcher.

Mismatches

While the switch to BEC and the decision not to switch to BEC could be clearly identified in the projection of research trails on the macroscopic cluster solutions, the non-BEC research trails were more widely distributed across topics (see e.g. the post-1997 publications in Figure 1). Interestingly, this does not occur in the 31-cluster solution, where all post-1997 publications are in just one cluster.

The publications of those researchers who did not switch to BEC are similarly distributed across clusters. In each case we find a bibliographically coupled cluster distributed across several of the macro-level clusters.

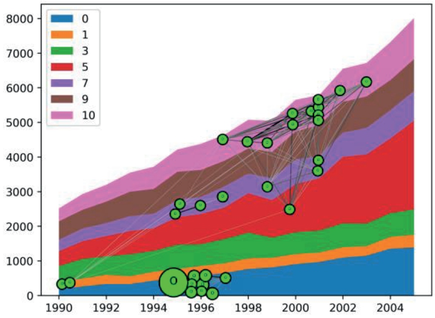


Figure 1. Projection of a physicist's research trail on the seven relevant clusters of the ten-cluster-solution

Discussion

The analysis of results is still preliminary. We see two explanations of the distribution of bibliographic-coupling clusters across macro-level clusters. First, a researcher's publications may be used (and thus cited) in other topics than those they work on, which is why the algorithm (which represents the collective perspective) puts them in a context that is different from that of the individual perspective. Second, the macro-level algorithm forces a separation of clusters that together would constitute a topic.

The possibility that a researcher's publications are located in fewer clusters of the higher-resolution clustering (with a total of 31 clusters) is a clear indication that more clusters obtained at a higher resolution do not necessarily correspond to a further differentiation of topics. Instead, the algorithm appears to recombine publications.

Conclusions

Topics in AMO physics (and most likely in all other fields) vary in the properties that makes them detectable and delimitable by the Leiden algorithm (and probably all other algorithms). Between 1995 and 2005, BEC was a fast-growing and self-referential topic that could easily be delineated. Other clusters at both resolution levels can less easily be thought of as topics.

If each topic reconstruction exercise simultaneously produces accurate and inaccurate representations of topics, discussing the validity of whole approaches is likely to be fruitless. Combining micro-level and content-based analyses with macro-level experiments of topic reconstruction might be a way forward towards identifying the variegated properties of topics and their connections to data models and algorithms.

Acknowledgments

This research was supported by the German Federal Ministry of Education and Research (Grant 01PU17003).

References

Gläser, J. (2017). A fight on epistemological quicksand: Comment on the dispute between van den Besselaar et al. and Butler. Journal of Informetrics 11(3), 927-932.

Gläser, J. and G. Laudel (2015). A Bibliometric Reconstruction of Research Trails for Qualitative Investigations of Scientific Innovations. Historical Social Research 40(3), 299-330.

Koopman, R., & Wang, S. (2017). Mutual information based labelling and comparing clusters. *Scientometrics*, *111*(2), 1157-1167.

Traag, V. A., L. Waltman and N. J. Van Eck (2018). From Louvain to Leiden: guaranteeing well-connected communities. arXiv:1810.08473 [cs.SI].

Comparing The Evolution of Research Subjects in Computer Science and Library & Information Science - A Case Study with NEViewer

Xiaoguang Wang[1], Wanli Chang[2] , Hongyu Wang[3,*] and Chen Zhang[4]

[1] *wxguang@whu.edu.cn* , [2] *wanlichang@whu.edu.cn* , [3,*] *wanghongyu@whu.edu.cn and* [4]*00009694@whu.edu.cn*
Center for Studies of Information Resources, Wuhan University, No.299 Bayi Street, 430072, Wuhan (China)

Introduction

Computer Science (CS) and Library & Information Science (LIS) are two main disciplines in iSchool. Therefore, the evolution of popular research subjects within two disciplines is usually related.

This study extracts from Microsoft Academic Graph (MAG), a large-scale Open Academic Graph (OAG) dataset, keywords in journals published from 2013-2017 of CS cited by Science Citation Index (SCI) and of LIS cited by Social Science Citation Index (SSCI). The distribution and evolution of research subjects in two disciplines are revealed and compared, using NEViewer (Xiaoguang et al. 2014), a visualization analysis tool for disciplinary topic networks.

OAG, as academic big data-sets, represents and organizes sci-tech papers as well as their relevant information regarding authorship, affiliations, and journals, etc. Semantic technologies such as ontology and resource description framework are used to fulfilled the representation and organization, and OAG allows open access on the Internet.

Data Preparation

At present, over 160 million papers and their metadata information are recorded in MAG, whose main data structure are illustrated in Table1.

Table 1. The main data structure of MAG.

Field	Note	Field	Note
id	MAG id	title	record’s title
authors	a list of authors’ name&affiliation	venue	journal name /Conf. name
year	publish year	keywords	a json list

This study imports the full data of MAG into Mysql by means of ETL process. Then, it uses the journal citation report (2017 edition) to filter the data with the field of “venue”, and retrieves (S)SCI papers in two disciplines ranging from 2013-2017. As a result, 133,673 papers in 495 core journals of CS and 9,248 papers in 88 core journals of LIS are input into the NEViewer for evolution analysis, and the experiment data is stored in GitHub [1]. To address the redundant information added by publishers in the field of “keywords”, this study uses a stop-word list to further filter the keyword list of data-sets, retaining data with 6 keywords or less.

Data Analysis and Result

NEViewer features functions of subject clustering and community partition and is applicable to analysis of the evolution of a variety of complex networks. With NEViewer, the study brings in the evolutionary manifold diagram of community networks in the discipline of CS and LIS, as in two figures below [1]. They illustrate both hotspot communities in 5 stages and the evolution graph of each community covering a period from 2013-2017. Each color block represents a hotspot community. A block with a larger area means more keywords in the community, and further more research has been done in the subject community. In a comparison of Figure 1 and 2, it can be readily seen that the research hotspots of CS concentrate intensively in 6 to 7 sizable communities, while those of LIS are scattered with only 2 to 3 sizable communities.

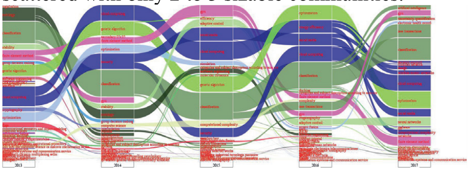


Figure 1. The subject community network evolution map in CS.

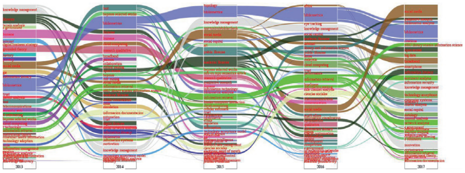


Figure 1. The subject community network evolution map in LIS.

Alongside Figure 1 and 2, hotspot communities at different stages (listed by size of the community) in NEViewer are revealed as in Table 2.

Table 2. Core Community Statistics.

CS	Year	LIS
Classification /cloud computing /ontology /genetic algorithm /optimization	2013	bibliometrics /libraries / knowledge management /humans /social media
Classification /security / cloud computing /genetic algorithm /ontology	2014	bibliometrics /research qualitative /humans / cloud computing /libraries
classification / cloud computing / genetic algorithm / social media /gpu	2015	bibliometrics /academic libraries /social media / knowledge management /public libraries
classification /cloud computing /optimization /social media /energy efficiency	2016	Bibliometrics /information retrieval /social media / knowledge management /cloud computing

[1] https://github.com/wanghongyu94/Experiment_ISSI2019_Poster

* Corresponding Author

classification /finite element method /cloud computing /optimization /computational modeling	2017	bibliometrics / social media / network analysis / open access /big data

The process of division and integration of some specific node at specific time is accessible in the hotspot evolution manifold diagram, enabling a clear understanding of evolution trajectory and trend of the research subject. “social media” is, as a research hotspot in both CS and LIS, taken as an example here as in Figure 3. The upper part of the figure represents the evolution of “social media” in CS. It indicates that research on social media begin with cloud computing and research on cloud computing is inextricably linked with social media. In 2017, the field of “social media” begins to see a division, emerging branches of classification, bibliometrics, knowledge representation, etc. And knowledge representation is an integration of classification and “social media”, which in turn indicates ever-increasing expansion of the research scope of “social media” and its growing intersection with classification, knowledge representation and bibliometrics.

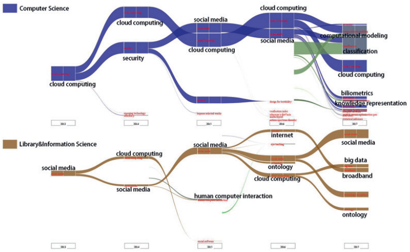


Figure 3. Evolution of social media hotspots in CS and LIS.

The lower part of Figure 3 represents the evolution of “social media” in the discipline of LIS. “Social media” began to be studied as a self-contained subject in 2013, and cloud computing cropped up as a branch in 2014, then human computer interaction in 2015, and internet, cloud computing, ontology among other branches in 2016. The research scope of “social media” has constantly expanded.

Table 3. Keywords of “cloud computing” and “social media” community in two fields.

CS		LIS	
Cloud computing	**Social media**	**Cloud computing**	**Social media**
Cloud computing	Social media	Cloud computing	Social media
Security	Information retrieval	Metadata	Social networks
Wireless sensor networks	Natural language processing	Web 2.0	Twitter
Privacy	Ontology	E-learning	Web2.0

According to the two indicators of Z-Value and P-Value in NEViewer, core keywords of cloud computing and social media in the field of CS and LIS are counted, as shown in Table 2. A synthetic analysis of Figure 3 and Table 3 indicates that cloud computing in CS involves research on social networks, privacy and security, which are also subjects of social media, and thereafter, social media emerges as a branch of cloud computing. By the same token, “social media” in LIS focuses on the research of data generated by social networks and social media. In the context of big data, processing and analyzing data through cloud computing has become hotspot, so cloud computing emerges as a branch following social media.

Conclusion

Three interesting things are identified in this study.
The first, through a comparison of research hotspots in CS and LIS, it can be found that the research hotspots in the former discipline are concentrated, indicating a clear connotation in its research; while the research hotspots in the latter are scattered, indicating an extensive denotation in its research.
Second, research hotspots in LIS see an earlier division than in CS. For example, in LIS, the research branch of “cloud computing” was produced by “social media” in 2014; in CS, however, the branch of “social media” was not generated from “cloud computing” until 2015.
Third, CS and LIS are inextricably linked. At the same time, the same hotspot in two disciplines may has different focuses. The discipline of CS focuses on technical research, while LIS is concerned with service research in the technical context. For instance, cloud computing in CS focuses on research on such technologies as network security, wireless sensor networks and privacy protection, while when it comes to LIS, it focuses on data organization and information services in the context of web2.0 and cloud computing environments.
The OAG has created new conditions for dynamic analysis across time and space, multi-dimensional analysis under multiple factors, multi-scale evolution analysis, automated intelligent analysis of research subjects. Through the fusion of multi-source heterogeneous OAG data, it is possible to further realize more pluralistic scientometrics analysis, such as the difference of research evolution between countries and the difference of research evolution between conferences’ articles and journals’ articles in the future.

Acknowledgments

This work is supported by National Natural Science Fund (No.71874129) and Major Project of National Social Science Fund (No.17ZDA292) of China.

References

XiaoGuang, W., Qikai, C. & Wei, L. (2014). Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. *Scientometrics*, 101(2), 1253-1271.

Drug Safety scientometrics overview highlights public health issues.

Philippe Gorry[1] and Enrique Seoane-Vazquez[2]

[1] *philippe.gorry@u-bordeaux.fr*
GREThA UMR CNRS 5113, University of Bordeaux, (France)

[2] *seoanevazquez@chapman.edu*
Dpt.of Biomedical & Pharmaceutical Sciences, School of Pharmacy, Chapman University, (USA)

Introduction

Drug safety

Although clinical trials are required for all new drugs before market approval, firms and drug regulatory agencies regularly face adverse drug reactions (ADR) in postmarketing phase which conduct to drug warning or withdrawal with deleterious impact on public health. Therefore, Drug Safety (DS) also known as Pharmacovigilance, is the science of monitoring ADR (Cobert, 2012). Despite that DS has a social welfare function, it has not been the study of any scientometrics analysis

Objectives

Our goal is to fill this gap by mapping the DS literature. This analysis might highlight some trends, the role of drug regulatory and firms in diffusing information, understand better some DS controversial events, help scholars in mastering pharmacovigilance literature, and it contribute to further development on literature-based discovery of ADR with the help of natural language processing.

Materials & Methods

Literature search on DS was run in Scopus® database until 31/12/2018, with a query across the title, abstract or keywords in all types of records using a string of 21 different keywords identified as MesH terms after mindmapping the MesH hierarchy. A corpus of 1,3350,138 documents was analyzed. In addition, a specific query on DS journals was run and 8,854 documents were isolated. Bibliometric analysis was done using Scopus embedded statistic functions and descriptive statistics analyses were run using Excel® and Xlstat® add-in package. Visualization of bibliometric network was performed with VOSviewer®, a network analysis software. "Reference Publication Year Spectroscopy" (RPYS) analysis (Marx, 2014). toward the identification of the historical roots of DS research field was done using CRexplorer®, a software for cited references analysis. delayed recognition (DR) publications were identified with the calculation of the Beauty coefficient (B), a parameter-free (Ke, 2015).

Results

Drug Safety research trends

A corpus of more than 1.3 M DS documents were gathered (72% articles, 14% reviews, 4% letters and 12% others). If publications on DS were published early during the 20th century with the founding of the FDA (1938), we focused our analysis after the Kefauver-Harris Amendment (1962) which revolutionize drug assessment after the thalidomide birth defect tragedy. Publications on DS started to rise with some delay in the early 70' and then, the number of publications stayed stable around 18.000 publications/year until 2001 (Figure 1).

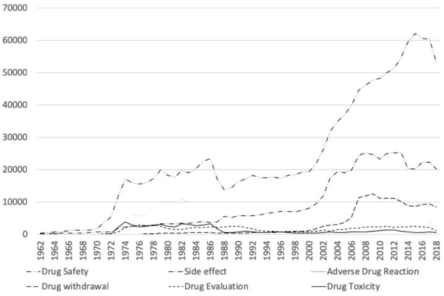


Figure 1. Trends of Drug Safety publications.

Then DS publications increased steady reaching 60.000 publications/year in 2015. Publications in DS journals represent only 1.57%. In the last 20 years, only 0,15% of the DS publications were sponsored by pharma firms. An analysis of press releases in Factiva® database identified a peak of media news covering DS between 2005-2009 at the time of anti-cox2 drugs withdrawal. The analysis of co-occurrence of words in the corpus of DS journals' papers visualize distinct keywords' groups centred around "Pharmacoepidemiology", "Safety", and "Pharmacovigilance", which changes along the last 20 years (data not shown). We undertake a detailed trends analysis of the main keywords. Figure 1 show the trends of the top 5 keywords by total publications numbers (side effects: 510,101; Adverse Drug

Reaction (ADR): 166,297; Drug withdrawn: 154,233; Drug evaluation: 82,913; Drug toxicity: 61,953). Some keywords in use at the beginning are no longer cited and *vice et versa.* Since the anti-cox2 controversy, the main keywords associated with DS publications is "Side Effect" and not anymore ADR.

Drug Safety publishing actors

The top 5 countries publishing on DS are in the order: USA (31%) United Kingdom (9%) Germany (7%), Japan (6%) and Italy (6%). However, the EU countries account as much publications than US underlying the importance of the EMA. Moreover, all BRICS countries ranked in the top 30 countries (data not shown). The DS research intensity was further explored at country level by looking at any correlation between DS publications, population size, health expenses by GDP and pharmaceutical expenses per capita (source: OECD) (Table 1).

Table 1. Correlation matrix (Pearson)

	DS pub.	Pop. size	Health % GDP	Pharma exp.
DS pub.	**1**	**0.911**	**0.762**	-0.194
Pop. size	**0.911**	**1**	**0.597**	-0.026
Health % GDP	**0.762**	**0.597**	**1**	**-0.490**
Pharma exp.	-0.194	-0.026	**-0.490**	**1**

NB: Values in bold are different from 0 with α=0,05

The top 3 publishing institutions are Harvard Medical School (US), INSERM (FR) and Veterans affairs (US) with more than 10.000 total publications each. NIH (US) is ranked only 10th and FDA (US) is lagging behind after 50th with 3763 publications. To be notice, the 1st pharma firm, Pfizer, ranked 12rd with 6891 publications (data not shown).

Drug safety historical roots

To gain historical insight on important past DS publications, we run a RPYS analysis. The deviation of the number of Cited References (CRs) from the median pinpointed years 1959, 1977, 2005 and 2012 as more significant than the others (data not shown). To identify the CRs responsible for these peaks, we sorted the list of unique CRs published by their citations' frequency. The top publication was a paper on "A method for estimating the probability of adverse drug reactions" (Naranjo, 1981).

Delayed Recognition of Drug Safety publications

To identify DR papers in DS, we calculate the "Beauty coefficient" for the top 400 most cited (between 10127 and 556 citations) papers on DS. We identified 8 candidate DR papers among which the publication of Naranjo (1981) (data not shown). We further explored an interesting case with B=393, an article describing "A rating scale for extrapyramidal side effects". This publication has been delayed for 23 years and its citations awakening with the rise of interest in extrapyramidal side effects (EPSE), a drug-induced movement disorder by antipsychotics.

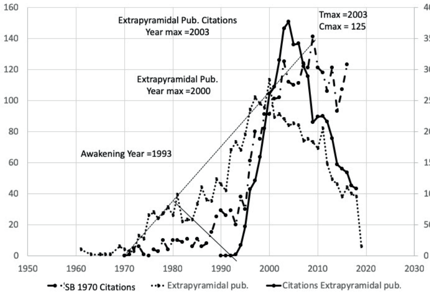


Figure 3. Delayed Recognition of EPSE.

Discussion

While DS is a major issue, we reported for the first time a landscape analysis of the literature. It reveals some characteristics: (1) research increase recently (2000') while major drug regulation backed to 1960' (2) it is a heterogenous research field which concepts evolved a lot in half century (3) most of publications are published outside core journals, probably by clinicians (2) the contribution of the industry is marginal, even if some firms are publishing a lot, raising questions about the industry responsibilities (4) US and EU countries are leading as expected because of the size of their drug market, but there is no correlation with pharmaceuticals expenses (5) US federal institutions are lagging behind the universities in reporting DS (6) anti-cox2 ADR controversy has awakened DS research field (7) identification of DR papers raise questions about resistance to the discovery of ADR and therefore its public health impact. Further studies are in need to measure if DR papers are linked to delayed drug withdrawn. Finally, our analysis supports the prospective for literature-based discovery in DS.

References

Cobert B, *Cobert's Manual of Drug Safety and Pharmacovigilance: Second Edition*, Jones & Bartlett Learning; 2012.

Ke, Q., Ferrara, E., Radicchi, F. & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *PNAS*, 112, 7426–743.

Marx, W., et al. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *JAIST*, *65*, 751–764.

Naranjo, CA et al. (1981). A method for estimating the probability of adverse drug reactions. *Clin Pharmacol Ther.* 30, 239-45.

Simpson GM, & Angus JW. (1970). A rating scale for extrapyramidal side effects. *Acta Psychiatr Scand* Suppl., 212, 11-9.

A Cleaning Method for various DOI Errors of Cited References in Web of Science

Shuo Xu[1], Liyuan Hao[1] and Xin An[2]

[1] *xushuo@bjut.edu.cn*, Leanne.H@qq.com
Research Base of Beijing Modern Manufacturing Development, College of Economics and Management, Beijing University of Technology, Beijing 100124, P.R. China

[2] *anxin@bjfu.edu.cn*
School of Economics and Management, Beijing Forestry University, Beijing 100083, P.R. China

Introduction

With the establishment of digital object identifier (DOI) system in 1997, managed by the International DOI Foundation (IDF), DOIs have been assigned uniquely to many digital objects. The DOI name is a case-insensitive alphanumeric string, and consists of two parts separated by a forward slash (Sidman & Davidson, 2001): a) a prefix beginning with the numeral 10 assigned by IDF or by DOI registration agencies, and b) a suffix assigned by the registrants.

It is well known that comprehensive bibliographic databases, such as Scopus and Web of Science (WoS), largely promote the development of scientometrics and informetrics. However, one should keep in mind that these databases are not free of errors (Franceschini et al., 2015), though data quality has improved significantly over the past decade. So far, errors have been found to happen to almost each field of publications. Of course, it is no exception for the DOI field. Franceschini et al. (2015) revealed that quite a few single DOI names were incorrectly assigned to multiple publications indexed in the Scopus database. The incorrect DOI names in the WoS database are also discovered by Zhu et al. (2019), but with different errors of duplicate DOI names (e.g., similar character are confused with each other, such as “O” vs. “0”).

By definition, each DOI name should be unique and must identify one and only one entity (Paskin, 1999). However, DOI errors present challenges for the accuracy of intelligence analysis on the basis of DOIs. In fact, apart from DOI errors described in Franceschini et al. (2015) and Zhu et al. (2019), it remains unknown that whether there are other types of DOI errors, how often each type of errors occur, and whether it is possible to automatically correct these errors. In this work, various DOI errors of cited reference in the WoS database are deeply analysed and a cleaning approach is put forward to alleviate the extent of DOI errors of cited references.

Dataset

The bibliographic data in the *gene editing* field was collected from the WoS core database on 25th January, 2018 from the library of Beijing University of Technology. The following search strategy is used in this study: “TS = (gene edit*) OR TS = (crispr) OR TS = (clustered regularly interspaced short palindromic repeats)”. The language is limited to English, and the document type includes *article*, *proceedings paper* and *review*. The publication year spans from 2000 to 2017. In total, the number of publications is 13,909. The number of the cited references with and without DOIs is 341,317 and 74,643, respectively. Due to the difficulty and workload of filling with the resulting DOI names for the latter, the cited references without DOIs are excluded from further analysis in this study.

Cleaning Method

Through careful analysis, this study finds that various DOI errors of the cited references exist in the WoS database. That is to say, DOI name of cited references in the WoS database are contaminated to some extent. As a matter of fact, due to the variety of DOI errors, it is not trivial to clean automatically DOI names. To the best of our knowledge, no software public available can competent for this cleaning task until now. Hence, a method for cleaning DOI names is proposed in this work, as shown in Algorithm 1. On the basis of manual curation rules, this approach is made up of one procedure (*Cleaning*) and three functions (*JoinDoi*, *TrimDoi* & *IsBracketMatch*). To facilitate the understanding, many data-types and built-in functions from Java programming language are explicitly utilized here.

The procedure *Cleaning* takes a cited reference (CR) field of an interested publication as input, splits it into multiple cited references (Line 2), and then try to separate DOI name(s) from other information one by one (Line 3-14). This study mainly focuses on various DOI errors, the cited references without the clue substring “, DOI” are discarded directly. The cited references with DOI name(s) are further grouped into two cases: those with multiple DOI names (Line 7-10) and those with single DOI name (Line 11). Note that it is very possible that for the former case (multiple *literal* DOI names), only one DOI name is actually output. The function *JoinDoi* devotes to removing the duplicate DOI names processed by the function *TrimDoi*.

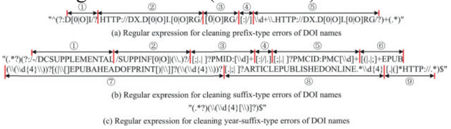
Algorithm 1 Cleaning the DOI names of the cited references in the WoS database

```
Precondition: RECORD_DELIMITER is the record delimiter, i.e., ";␣"
Precondition: DOI_DELIMITER denotes the DOI delimiter, i.e., ",␣"
Precondition: regexPrefix is the regular expression for prefix DOI errors, i.e., Figure 2(a)
Precondition: regexSuffix is the regular expression for suffix DOI errors, i.e., Figure 2(b)
Precondition: regexYear is the regular expression for suffix year DOI errors, i.e., Figure 2(c)

 1: procedure Cleaning(String refList)
 2:     String[] citedRefs ← refList.split(RECORD_DELIMITER)
 3:     for i ← 0 to citedRefs.length() do
 4:         int pos ← citedRefs[i].indexof(",␣DOI")
 5:         if pos ≠ −1 then
 6:             String doi ← citedRefs[i].substring(pos + ",␣DOI".length()).trim()
 7:             if doi.startsWith("[") and doi.endsWith("]") then
 8:                 doi ← doi.substring(1, doi.length() - 1)
 9:                 doi ← JoinDoi(doi.split(DOI_DELIMITER))
10:                 output i and doi
11:             else output i and TrimDoi(doi)
12:             end if
13:         end if
14:     end for
15: end procedure

16: function TrimDoi(String doi)
17:     String newDoi ← doi.toUpperCase().replaceAll("\\s+", "")
18:     Pattern patPrefix ← Pattern.compile(regexPrefix)
19:     Matcher matPrefix ← patPrefix.matcher(newDoi)
20:     if matPrefix.find() then newDdoi ← matPrefix.group(1)
21:     end if
22:     Pattern patSuffix ← Pattern.compile(regexSuffix)
23:     Matcher matSuffix ← patSuffix.matcher(newDoi)
24:     if matSuffix.find() then newDoi ← matSuffix.group(1)
25:     end if
26:     Pattern patYear ← Pattern.compile(regexYear)
27:     Matcher matYear ← patYear.matcher(newDoi)
28:     if matYear.find() then newDoi ← matYear.group(1)
29:     end if
30:     newDoi ← newDoi.replaceAll("\\\\", "").replaceAll("__", "_").replaceAll("\\.\\.", ".")
31:     newDoi ← newDoi.replaceAll("<.*?>.*?</.*?>", "").replaceAll("<.*?/>", "")
32:     if not newDoi.equals("") and newDoi.matches("10\\..*?/.*?") then
33:         if newDoi.matches(".*?[-|_]$") then return (newDoi, false)
34:         end if
35:         if IsBracketMatch(newDoi) then return (newDoi, true)
36:         else if IsBracketMatch(newDoi.substring(0, newDoi.length() - 1)) then
37:             return (newDoi.substring(0, newDoi.length() - 1), true)
38:         end if
39:     end if
40:     return (newDoi, false)
41: end function
```

The function *TrimDoi* tries to trim the DOI names by several regular expressions in Figure 1. Though most legal Unicode characters are allowed by ISO standard (ISO 26324:2012), it is very seldom that DOI names contain whitespace characters. Exceptions are still found in the WoS database. Hence, before cleaning further DOI names, all whitespace characters are removed (Line 17). Then, prefix-, suffix- and other-type errors of DOI names are cleaned sequentially. In addition, this function is also able to deal with several special cases (Line 30-31), such as forward slash, double underlines, double dots, XML tags, etc. In the end, if trimmed DOI names do not follow the specified characteristics (Simmonds, 1999) (Line 32), trimmed DOI name and false status are returned. Otherwise, if trimmed DOI names end with hyphen or underline symbol, these DOIs are also illegal (Line 33-34). Then the function *IsBracketMatch* is used to check whether the involved brackets match in trimmed DOI names or resulting substrings excluding the last letter (Line 35-38).



```
"^(?:D[0|O]I/?|HTTP://DX.D[0|O]I.[0|O]RG/|[0|O]RG/|[:|/]|\\d+\\.HTTP://DX.D[0|O]I.[0|O]RG/?)+(.*)"
```

(a) Regular expression for cleaning prefix-type errors of DOI names

```
"(.*?)(?:/-/DCSUPPLEMENTAL|/SUPPINF[0|O]|(\\.)?/|[;|.| ]?PMID:[\\d]+|[:|/|.]|[;|.| ]?PMCID:PMC[\\d]+|[(|.|;]+EPUB
|(\\(\\d{4}\\))?[(|\\[]EPUBAHEADOFPRINT[)|\\]]?(\\(\\d{4}\\))?|[.|;| ]?ARTICLEPUBLISHEDONLINE.*\\d{4}|[.|(]*HTTP://.*)$"
```

(b) Regular expression for cleaning suffix-type errors of DOI names

```
"(.*?)(\\(\\d{4}[\\)]?)$"
```

(c) Regular expression for cleaning year-suffix-type errors of DOI names

Figure 1. Regular expressions for cleaning various DOI errors.

Results & Discussions

Table 1 summarizes the distribution of various DOI errors in the *gene editing* dataset. From Table 1, one can see that the vast majority of DOI errors belong to the prefix-type error. In fact, the number of DOI errors with the prefix "DOI " is 4,968, which accounts for 92.39% DOI errors. Amongst the other errors, the number of illegal DOI errors is 154. To evaluate the performance of our cleaning method, the number of publications with multiple DOI names before and after cleaning is shown in Table 2. It is not difficult to see that the number of cited references with two and three DOI names is reduced drastically from 9,704 to 1,990 and from 45 to 33, respectively. This indicates that the quality of DOI names of cited references in the WoS database has been greatly improved.

Table 1. Distribution of various DOI errors in the *gene editing* dataset.

Prefix-type errors	Suffix-type errors	Other-type errors	∑
4,992 (92.84%)	221 (4.11%)	164 (3.05%)	5,377

Table 2. The number of cited references with multiple DOI names in the *gene editing* dataset.

No. of DOI names	2	3	4	5	8	15	∑
Before cleaning	9,704	45	1	3	1	1	9,755
After cleaning	1,990	33	1	3	1	1	2,029

Conclusions

As noted by Zhu et al. (2019), there is no simple way to recognize and thus to evaluate the extent of DOI errors in the Web of Science database. After careful analysis on the bibliographic data in the *gene editing* field, several classic DOI errors of cited references, such as prefix- suffix- and other-type errors, are identified. Then, a cleaning method of DOI names is put forward on the basis of regular expressions in this work.

Acknowledgments

This research received the financial support from Social Science Foundation of Beijing Municipality under grant number 17GLB074 and Natural Science Foundation of Guangdong Province under grant number 2018A030313695.

References

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015). Errors in DOI indexing by bibliometric databases. *Scientometrics*, 102, 2181-2186.

Paskin, N. (1999). The digital object identifier system: Digital technology meets content management. *Interlending & Document Supply*, 27, 13-16.

Sidman, D. & Davidson, T. (2001). A practical guide to automating the digital supply chain with the digital object identifier (DOI). *Publishing Research Quarterly*, 17, 9-23.

Simmonds, A. W. (1999). The digital object identifier (DOI). *Publishing Research Quarterly*, 15, 10-13.

Zhu, J., Hu, G., & Liu, W. (2019). DOI errors and possible solutions for Web of Science. *Scientometrics*, 118, 709-718.

A new approach to funding acknowledgment field in the Spanish case: can be used to identify gender gap in research funding?

Elba Mauleón[1] and Núria Bautista-Puig[2]
[1]mmauleon@bib.uc3m.es, [2]nbautist@bib.uc3m.es
Research Institute for Higher Education and Science (INAECU), University Carlos III of Madrid, 126 Madrid Str., 28903, Getafe (Spain)

Introduction

Acknowledgments in research publications is a way to express gratitude to the different entities who funded or contributed somehow to the research (Tang et al, 2017). These offer an overview of the funding landscape in which inputs and outputs form different researchers active in an area could be identified (Grassano et al., 2016). In addition, this information is considered one of the points of the reward triangle, with the authorship and citation (Costas and Leeuwen, 2012). Several studies have analyzed acknowledgments patterns from different fields: medical (Butler, 2001); nanotechnology (Shapira and Wang, 2010) or library and information science (Zhao, 2010). Some limitations has been highlighted from the literature: these acknowledgments are collected only when they include funding information (Costas and Leeuwen, 2012) or the lack of standardization (Grassano et al., 2016; Álvarez-Bornstein, 2017).

On the other hand, there are studies that analyze the differences between men and women for obtaining research funds. However, there are few studies that address the issue of gender gap funding by using the field 'Funding Acknowledgments' (FA) on the WoS.

Motivation and Methodology

Bearing in mind the importance of funding in research, on its visibility, impact and collaboration, we are going to analyse the acknowledgments patterns by gender in order to increase our knowledge about funding acknowledgments. The research study was based on publications collected from the WoS category 'Green & Sustainable Science & Technology', a sustainability area, in Web of Science (WoS) database. This category is related with environmental sustainability and has been selected as a sample of study. The period of analysis is from 2007 to 2018 and publications from Spain were considered (n=3,570). The gender of the researchers were assigned with a semi-automatic method by using an API Genderize[1], which determines the gender of a first name by indicating the country.

The objectives of this study are three-fold:

- Quantify the presence of men and women in scientific publications as a result of funded and non-funded research.
- Describe funding patterns in the area of ´Green Science & Technology Studies' by gender.
- Analyse funding effect in scientific publications by men and women through impact, visibility and collaboration.

Results

A total of 3,570 documents were identified during the period (2007-2018) in Green WoS category with, at least, one Spanish center. From this, 2,470 documents (69.19%) are with funding acknowledgments information included (FA) and 1,100 documents (30.81%) without. Of the 3,570 documents, sex was identified in 2,603 documents (a 75% del total). From those, 2.603, 68,73% are with funding acknowledgments (1,789), and 814 documents (31.27%) non- funded. This study analyses only the documents in which sex of all authors were identified. Some documents which were left outside can be caused because the name is not developed. This is caused because you only have the initial of the name and you could not identify their sex or because the name is ambiguous (e.g. with Chinese names). That is to say, 2,603 documents, a 75% of the total of the documents published for some Spanish centre in the Green category.

From the total of the documents in which sex was identified by all authors, a 29.43% was signed only by men, a 4.26% signed by only women and 66.3% by men and women. The lowest percentage of signed documents only by women is found in documents with funding (3.9%). However, the presence of women is higher in funded documents but non-funded, thanks to their presence in mixed papers (Table 1).

Table 1. Number of papers by gender.

	Only men	Only women	Men & Women	Total
Non funded	303 (37.22%)	41 (5.03%)	470 (57.74%)	814 (31.27%)
Funded	463 (25.88%)	70 (3.90%)	1,256 (70.21%)	1,789 (68,73%)
Total	766 (29.43%)	111 (4.26%)	1,726 (66.3%)	2,603

About the presence of men and women in scientific publications of funded and non-funded research on this area, the results show that the presence of

[1] Genderize (https://genderize.io/#overview)

women is higher in documents with funded than non funded (75.66% vs 24.34%). In this dataset, it predominates the documents signed as first author by men (in funded is 59% and in non-funded is 68%). Regarding size of the groups (number of authors/paper) is higher in documents with funded (4.64 vs 3.62).

Regarding collaboration, between non-funded is predominant the documents in without collaboration (40%). However, between the documents with funded, is higher the percentage of documents in international collaboration (45.5%). When there is funding, the percentage of papers signed by men and women together is higher in comparison with non-funded research in all types of collaboration (international, national and without). The percentage of documents signed only by men is always lower in documents non-funded in all types of collaboration.

In terms of visibility and impact, from the 1,145 documents published in the first-quartile, 835 (72.92%) are with funding vs 310 (27.07%) non-funded. Regarding impact, by measuring the number of citations per document, funded research has more impact (17.63 citation/paper) that non-funded research (13.3 citation/paper). If we analysed this information by gender, only men has a higher impact in funded (18.81 vs 15.23 in non-funded) and in documents with only men, the impact is higher in non-funded research (13.58 citation/paper).

Discussion and conclusions

This analysis follows the European Commission recommendations and the national and international data collection bodies on the presence of men and women in science in through Research Funding Organisations. The preliminary conclusions show that female presence is higher in funded research. The percentage of documents with some women (men and women plus documents signed only by men) is higher in documents funded that non-funded (74.22% vs 62.77%). These findings can show how gender equality is incorporated into policy funding. In this sense, the research teams incorporate women colleagues into their teams because they know this is a criteria considered as a positive value in the research process (H2020 programme).

Another fact that should be considered is the size of the teams (no. of authors). From the results, it can be seen that funded research was conducted in teams of greater size, denoting a higher collaboration. This higher presence for papers with funding being consistent with previous studies (Costas and van Leeuwen, 2012; Díaz-Faes and Bordons, 2014). Precisely the largest participation of women is in the largest groups.

Some limitations should be highlighted in this study. Regarding acknowledgments, it should be mentioned the limitation that funding sources are not always acknowledged by authors: nevertheless, this analysis could lead to a general overview of how funding affects scientific papers on this specific field. In addition, further research will be necessary to study more deeply the role of funding agencies (e.g. type of organization) in scientific publications. Despite the fact that now only Spain results are presented, more countries would be involved in the study with the aim to check if there is a particular trend in certain geographic areas. In this sense, it is intended to analyze more areas and in more countries. The main limitations of this study will be considered to obtain the indicators to extend the study.

Acknowledgments

This research was supported by predoctoral contract BES-2015-075461 from the Ministry of Economy and Competitiveness (MINECO).

References

Álvarez-Bornstein, B., Díaz-Faes, A.A., & Bordons, M. (2017). Relationship between research funding and scientific output in two different biomedical disciplines. STI indicators Conference 2017. Paris.

Butler, L. (2001). Revisiting bibliometric issues using new empirical data. Research Evaluation, 10(1), 59–65. doi:10.3152/147154401781777141

Costas, R., & van Leeuwen, T. N. (2012). Approaching the “reward triangle”: General analysis of the presence of funding acknowledgments and “peer interactive communication” in scientific publications. Journal of the American Society for Information Science and Technology, 63(8), 1647-1661.

Díaz-Faes, A. A., & Bordons, M. (2014). Acknowledgments in scientific publications: Presence in Spanish science and text patterns across disciplines. Journal of the Association for Information Science and Technology, 65(9), 1834-1849.

Grassano, N., Rotolo, D., Hutton, J., Lang, F., & Hopkins, M.M. (2017). Funding data from publication acknowledgments: Coverage, uses, and limitations. Journal of the Association for Information Science and Technology, 68(4), 999-1017.

Shapira, P., & Wang, J. (2010). Follow the money. Nature, 468(7324), 627–628. doi:10.1038/468627a

Tang, L., Hu, G., & Liu, W. (2017). Funding acknowledgment analysis: Queries and caveats. Journal of the Association for Information Science and Technology, 68(3), 790-794.

Zhao, D. (2010). Characteristics and impact of grant-funded research: A case study of the library and information science field. Scientometrics, 84(2), 293–306. doi:10.1007/s11192-010-0191-y

International references increase Chinese papers’ citation impact

Kaile Gong[1], Juan Xie[1], Ying Cheng[1], Yi Bu[2], Cassidy R. Sugimoto[2], and Vincent Larivière[3]

[1] gong@smail.nju.edu.cn; xiejuan9503@163.com; chengy@nju.edu.cn
School of Information Management, Nanjing University, Nanjing, Jiangsu 210023 (China)

[2] buyi@iu.edu; sugimoto@indiana.edu
School of Informatics, Computing, & Engineering, Indiana University Bloomington, Bloomington, IN 47408 (USA)

[3] vincent.lariviere@umontreal.ca
École de bibliothéconomie et des sciences de l’information, Université de Montréal, Montréal, QC H3C 3J7, (Canada)

Introduction

China has been the world’s largest producer of academic publications since 2016 (Tollefson 2018). Although English is the current scientific lingua franca (Gordin 2015), Chinese remains its predominance in China’s scholarly communication (Shu et al. 2018). The progress of science requires researchers to understand previous literature before doing their research, and the abundant Chinese academic resources can provide an easily accessible and understandable literature foundation for Chinese scholars. However, the language barriers, cultural identity and other factors make a part of Chinese scholars be lacking in international outlook and fail to incorporate international literature (Gong et al. 2019). Since contemporary scientific exchanges are international, and such internationalization has shown a positive influence on scientific impact (Sugimoto et al. 2017), we may hypothesize that a lack of use of international scientific literature has a negative effect on its scholarly impact. In this study, we take citation count as the indicator of academic impact, and the language of cited references as an indicator of using international scholarly literature.

Methodology

Data

The bibliographic and citation data of 37,801 papers published from 1998 to 2009 in 14 library & information science (LIS) journals were collected from the Chinese Social Science Citation Index (CSSCI), which is a China’s leading and authoritative database for scholarly citations in social sciences. All of the included papers are written in Chinese, and the first authors are all affiliated to China’s institutions.

Key variables

Key variables are shown in Table 1. Y is the indicator of papers’ academic impact; X1~X5 reflect different features of foreign language references.

Table 1 Key variables and definition

Code	Variable	Definition
Y	Five-year cumulative citation count	The number of citations, within CSSCI, received by the paper in five years after being published.
X1	Number of cited foreign language references	The number of non-Chinese references, including all document types, in the paper’s reference list.
X2	Number of cited foreign language journal articles	The number of non-Chinese journal articles in the paper’s reference list.
X3	Number of cited foreign language journal articles weighted by journal reputation	The weighted number of non-Chinese journal articles in the paper’s reference list. The weight is equal to zero if the cited journal was not indexed by Journal Citation Report (JCR) in the cited article’s publishing year; otherwise, it’s equal to the cited journal’s Journal Impact Factor Percentile (JIFP)[1)] in the cited article’s publishing year. The articles published before 1997 (earliest year covered by JCR) are calculated as 1997.
X4	Number of cited foreign language journal articles that belong to the same discipline as the citing paper	The number of non-Chinese journal articles, whose journal is classified as *Information Science & Library Science* in Web of Science or *Library and Information Science* in Scopus, in the paper’s reference list.
X5	Number of cited newer foreign language journal articles	The number of non-Chinese journal articles, whose age is smaller than the median age of foreign language journal articles cited in the same year, in the paper’s reference list.

1) http://help.incites.clarivate.com/incitesLiveJCR/glossaryAZgroup/g8/9586-TRS.html

Control variables

The control variables are shown in Table 2. All of them are verified to affect papers’ citation counts by previous studies (Tahamtan et al. 2016). X6~X8 are the paper-related factors, X9 is the journal-related factor, and X11~X14 are the author-related factors.

Table 2 Control variables and definition

Code	Variable	Definition
X6	Document type	Research article or review.
X7	Length	The number of pages in the paper.
X8	Early received citations	The number of citations, within CSSCI, received by the paper in the first two years after being published.

X9	Journal reputation	Whether the paper was published in a leading journal, which belongs to the *Catalogue of Leading Journals in Humanities and Social Sciences*[1], or a general-journal.
X10	Number of authors	The number of co-authors of the paper.
X11	Number of institutions	The number of institutions of co-authors of the paper.
X12	Type of first author's institution	Whether the first author's institution is a university, public library, or research institute.
X13	Level of first author's institution	985-[2], 211-[3] or general-university; national-, provincial-, or city-library; national-, provincial-, or city-research institute.
X14	Foundation	Whether the paper was supported by national-, provincial/ministerial-, city/school -, or non-foundation.

[1] http://skch.nju.edu.cn/regulation

[2] http://www.moe.gov.cn/srcsite/A22/s7065/200612/t20061206_128833.html

[3] http://www.moe.gov.cn/srcsite/A22/s7065/200512/t20051223_82762.html

Statistical analysis

The Mann-Whitney U test is used to analyse whether there is a significant difference in the five-year cumulative citation count (Y) between papers with foreign language references and those without. Multiple linear regression is used to analyse the relationships between citations and foreign language references. Papers with foreign language references are taken as samples in the regression, the dependent variable is $log_{10}(Y+1)$, and X1~X14 are the independent variables. All of the categorical variables are converted into dummy variables.

Results

Among all LIS papers included in this study, the papers with foreign language references are in the minority (37%), but the average five-year cumulative citation count of them ($\overline{\mathrm{Y}} = 2.71$) is higher than that of papers without such references ($\overline{\mathrm{Y}} = 1.55$). The result of Mann-Whitney U test (Z = -31.063, $p < 0.01$) demonstrates that the numbers of citations received by papers with foreign language references and those without differ significantly.

Table 3 shows that the regression is significant (F(10, 13923) = 1805.399, $p < 0.01$) and approximately 56.5% of variance in citations can be explained (R^2 = 0.565). Tests to see if the data meet the assumption of collinearity indicate that multicollinearity is not a concern (all VIFs < 5). After controlling widely-recognized factors, among papers with foreign language references, those citing more foreign language references (X1: $\beta = 0.026$, $p < 0.01$), more articles published in prestigious (X3: $\beta = 0.031$, $p < 0.05$) and own discipline's (X4: $\beta = 0.042$, $p < 0.01$) journals can receive more citations, but those citing more articles published in low-level journals belong to other disciplines cannot (X2: $\beta = -.054$, $p < 0.01$). In addition, the timeliness of cited foreign language journal articles has no significant effect on the citing papers' citation counts (X5: $p > 0.1$).

Table 3 Results of the regression analysis

Independent Variable	Standardized Coefficients (β)
Features of cited foreign language references	
X1	$.026^{***}$
X2	$-.054^{***}$
X3	$.031^{**}$
X4	$.042^{***}$
Paper-related factors	
X6 (Review)	$.026^{***}$
X7	$.049^{***}$
X8	$.725^{***}$
Journal-related factor	
X9 (Leading-journal)	$.041^{***}$
Author-related factors	
X13a (985-university)	$.029^{***}$
X14 (National-foundation)	$.015^{***}$
$R^2 = .565$ F(10, 13923) = 1805.399	Sig. = .000

** $p < 0.05$, ***$p < 0.01$, N = 13934, all VIFs < 5

Conclusion

We find that the inclusion of international literature helps to enhance Chinese papers' academic impact. It may because of the instrumental function of references -- signalling readers works they may be unaware of (Merton 1988). In this case, papers with international references become a window for domestic scholars to understand the work of international peers and thus play a key media role in the diffusion of knowledge from the international to the local. In short, diversity and openness can facilitate scientific research.

References

Gong, K., Xie, J., & Cheng, Y. (2019). Multi-dimensional analysis of the internationalization of references in Chinese journal papers: A study on library & information science. *Information Science, 37*(3), 127-135. (in Chinese)

Gordin, M. D. (2015). *Scientific Babel: How science was done before and after global English*. Chicago: University of Chicago Press.

Merton, R. K. (1988). The Matthew effect in science, ii: cumulative advantage and the symbolism of intellectual property. *ISIS*, 79(4), 606-623.

Shu, F., Julien, C. A., & Lariviere V. (2019). Does the web of science accurately represent Chinese scientific performance?. Forthcoming in *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24184

Sugimoto, C. R., Robinson-Garcia, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature,550*(7674), 29-31.

Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics, 107*(3), 1195-1225.

Tollefson, J. (2018). China declared world's largest producer of scientific articles. *Nature*, 553(7689), 390.

Multi-affiliations in scientific collaboration between G7 and BRICS countries

Sichao Tong, Ting Yue

tongsichao@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences.33 Beisihuan Xilu, Zhongguancun, Beijing (China)

Introduction

The researchers affiliated with multiple institutions are increasingly seen in the current scientific activities, e.g., one researcher obtaining the specific fellowships jointly funded by several institutions. Huang and Chang (2018) shows high percentages of publications written by multi-institutional authors in genetics and high-energy physics. For 'frontline' researchers, time-flexible academic positions is growing as a more attractive options (ESF, 2013). With direct links with several institutions, a researcher can be consequently recognized as a bridge between institutions, facilitating cooperation and exchange (ESF, 2013; Hottenrott & Lawson, 2017). Furthermore, these researchers are more often found in highly cited publications, reflecting their positive influence on scientific impact (Hottenrott & Lawson, 2017; Huang & Chang, 2018; Sanfilippo, Hewitt, & Mackey, 2018).

As many studies demonstrated, collaboration play a positive role in bringing high impact (Jones, Wuchty, & Uzzi, 2008; Narin, Stevens, & Whitlow, 1991; Persson, Glanzel, & Danell, 2004). Compared with general collaboration, serving as a natural bridge for knowledge transfer between affiliations, what's the general overview of publications with multi-affiliated authorship? Is there any heterogeneities by disciplines? Will multi-affiliated researcher bring higher impact for publications?

Methodology

We considered 802,164 institutional collaborated publications of 2015 retrieved from Web of Science, with all author address records, as our full set of data. We divided it into two datasets, the set of domestic institutional collaborated publications, and another set of international institutional collaborated publications, the following is abbreviated as DIC and IIC. 466,897 DIC publications and 335,267 IIC publications were published in 2015. First we give two definitions for multi-affiliated author.

- A multi-institutional author refers to an author affiliated with two or more institutions, the following is abbreviated as MI.
- A multi-national author refers to an author affiliated with two or more countries, the following is abbreviated as MN.

Then for DIC publications, we considered two groups: DIC publications with MI authorship, and the compared group of DIC publications which do not contain MI authorship. The former one accounts for about 44% (204,829) in DIC publications. Yet for IIC publications, we also considered two groups: IIC publications with MN authorship, and the compared group of IIC publications which do not contain MN authorship. The former group shows a share of 36.5% (122,506) in IIC publications.

For citations, we used 3-year windows. Mean Normalized Citation Score (MNCS) was applied to measure the citation impact.

ESI category has been used in this study.

Results

As the example set by discipline in Fig. 1, publications with MI authorship and publications with MN authorship show higher percentages in SPACE SCIENCE and most life science disciplines, also for MATHEMATICS, ENGINEERING and COMPUTER SCIENCE, they generally show lower. In SPACE SCIENCE, these two percentages both approach to 50%. While MATHEMATICS relatively shows the lowest percentages, 28.1% and 23.6%.

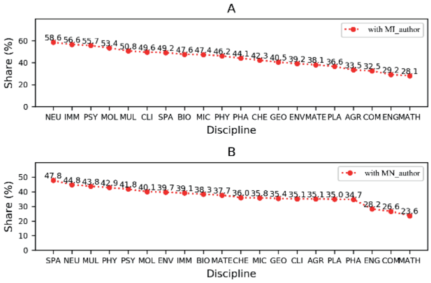


Figure 1. Share of publications by discipline.

Fig. 2 offers a general overview of share of publications by publication types in two datasets, for G7 and BRICS. The share of publications with MI authorship in DIC dataset for France scores far above other G7 countries. This result may be related to the intense collaboration between different research sectors in France (European Commission, 2017). In IIC section, G7 have little difference in the share of publications with MN authorship, vary between 34.2% and 40.2%. For BRICS, South Africa, Russia and

China show higher share of publications with MI, MN authorship, respectively, while India and Brazil show lower values.

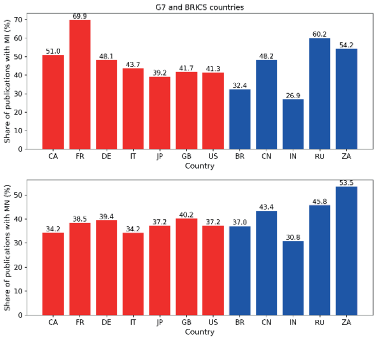


Figure 2. Share of publications for G7 and BRICS countries.

As observed in Fig. 3, disciplines show heterogeneities. In DIC dataset, MNCS of the two groups gain a largest gap (0.2) in MATERIALS SCIENCE. In SPACE SCIENCE, MNCS for publications of MI authorship scores below others. Regarding to IIC dataset, in MATHEMATICS, MATERIALS SCIENCE and SPACE SCIENCE, the former MNCS scores much higher, with a wide gap. Most life science disciplines show slight differences.

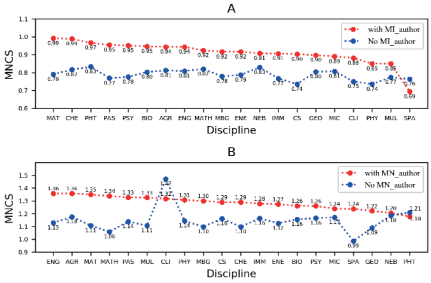


Figure 3. MNCS values by discipline.

Table 1 gives MNCS for G7 and BRICS. The gap between IIC publications with MN authorship and IIC publications with no MN authorship is relatively wider in BRICS countries. It also exists a phenomenon that researchers from BRICS countries often co-affiliated with high S&T level countries. They may have some correlation.

Table 1. MNCS for G7 and BRICS countries.

Countries	DIC		IIC	
	MI	NoMI	MN	NoMN
CA	0.96	0.86	1.48	1.45
FR	0.90	0.77	1.41	1.39
DE	1.03	0.81	1.55	1.39
IT	0.92	0.87	1.58	1.44
JP	0.81	0.60	1.30	1.18
GB	1.11	1.00	1.55	1.46
US	1.22	1.08	1.53	1.39
BR	0.55	0.46	1.25	1.03
CN	0.91	0.84	1.51	1.27
IN	0.62	0.60	1.32	1.09
RU	0.41	0.27	1.09	0.89
ZA	0.65	0.54	1.25	1.07

Conclusions

This study explores publications with multi-affiliated authors from the view of scientific collaboration. Results by disciplines show that the share of publications with multi-affiliated authorship has the correlation to characters of disciplines, collaboration in scientific activities may facilitate the appearance of multi-affiliated authors. Multi-affiliated authors can bring higher citation impact, compared with general collaboration. Multi-national authors may play stronger role serving as bridge for knowledge transfer in BRICS countries, which have relatively lower citation impact .

References

ESF. (2013). New concepts of researcher mobility—a comprehensive approach including combined/part-time positions. In Science Policy Briefing 49: Strasbourg: European Science Foundation.

Hottenrott, H., & Lawson, C. (2017). A first look at multiple institutional affiliations: a study of authors in Germany, Japan and the UK. Scientometrics, 111(1), 285-295.

Huang, M. H., & Chang, Y. W. (2018). Multi-institutional authorship in genetics and high-energy physics. Physica a-Statistical Mechanics and Its Applications, 505, 549-558.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. Science, 322(5905), 1259-1262.

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific Cooperation in Europe and the Citation of Multinationally Authored Papers. Scientometrics, 21(3), 313-323.

Persson, O., Glanzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics, 60(3), 421-432.

Sanfilippo, P., Hewitt, A. W., & Mackey, D. A. (2018). Plurality in multi-disciplinary research: multiple institutional affiliations are associated with increased citations. Peerj, 6, e5664.

European Commission. (2017). European Human Resources Strategy for Researchers (HRS4R).

Semi-automatic taxonomy development for research data collections : the case of wind energy

Haakon Lund[1] & Anna Maria Sempreviva[2]

[1]*hl@hum.ku.dk*
University of Copenhagen, Department of Information Science, Njalsgade 76, DK-2300 Copenhagen (Denmark)

[2]*anse@dtu.dk*
Technical University of Denmark, Department of Wind Energy, Frederiksborgvej 399, DK-4000 Roskilde (Denmark)

Introduction

A metadata scheme and related controlled vocabularies, taxonomies, for the wind energy sector, have been proposed by the FP7 Project Integrated Research Programme in Wind Energy, IRPWind (Sempreviva et al., 2017). The goal was twofold: on one hand to comply with the principles of Findable, Accessible, Interoperable and Re-usable data (FAIR) (Wilkinson et al., 2016) introduced by the European Commission to support the open data policy for EU funded research projects. On the other hand, to answer to the growing concern within the research communities on how to identify and locate the vast amount of already available and future data from the ongoing digital transformation for research data management purposes. Research data management is increasingly adopted by funding agencies at national level as well. The faceted IRPWind taxonomies were developed by expert elicitation where a group of domain experts collaborated to establish e.g. a hierarchy of terms describing the WE topics. This process does demand extensive use of human resources and in a future perspective is not sustainable. Here, we propose using alternative methodology to create a semi-dynamic taxonomy, updated in time with new research trends, that relies on the analysis of keywords provided by authors of articles in domain journals. To test the method, we sat the goal of reproducing the IRPWind taxonomy of the topics in the wind energy sectors. For this purpose, we use keywords provided by authors to tag papers in the Wind Energy journal (ISSN 1099-1824) ISI Journal Citation Reports © Ranking: 2017:43/97 (Energy & Fuels) 2017:22/128 (Engineering, Mechanical). Impact factor 2.938. The Wind Energy journal does in its scope align with the topics covered by the IRPWind taxonomy.

Methodology

The co-occurrence analysis of author keywords from research papers has long been established as a viable way of identifying new trends in research and the development of a scientific domain (Woon and Madnick, 2009) and within the community of bibliometrics (Romo-Fernandez et al., 2013). This based on the assumptions that author provided keywords do express recent trends in research and therefore can provide valuable input to necessary taxonomy updates. The identification of research trends in a specific research domain is closely related to the identification of new terms to include in a domain specific taxonomy. Woon and Madnick (2009) suggest the use of keyword co-occurrence of author generated keywords for automated taxonomy construction.

We extracted 5717 keywords from 1159 papers published in Wind Energy journal covering the period from 1998 to 2018. Due to different forms of terms, equivalence and incoherencies in the choice of keywords by authors e.g. Speed /velocity, blade/turbine blade; wind turbine/Energy conversion system etc., only 2917 unique keywords were retained of which 356 occurred 3 times or more.

First, we clustered the filtered keywords based on the analysis of their co-occurrence and visualized the clusters using the integrated bibliometric tool, VOS – viewer (Van Eck & Waltman, 2010). Then, we used the resulting clustering maps to identify the core themes in the temporal development of wind energy (see figure 1). Last, a growth indicator (Woon, Henschel & Madnick, 2009), as a proxy for trends, was calculated based on term frequency expressed as weighted average publication year. The actual growth indicator was calculated as:

$$\theta_I = \frac{\sum_{t\in[firstyear,endyear]} t.TF_i\,[t]}{\sum_{t\in[firstyear,endyear]} TF_i\,[t]}$$

Where θ_I is the growth potential for keyword *I* and *TF*$_i$*[t]* is the term frequency for term *I* and year *t.* A recent year suggests more prevalence of the topic.

Results

To evaluate the viability of using author provided keywords as candidate terms for taxonomy updates

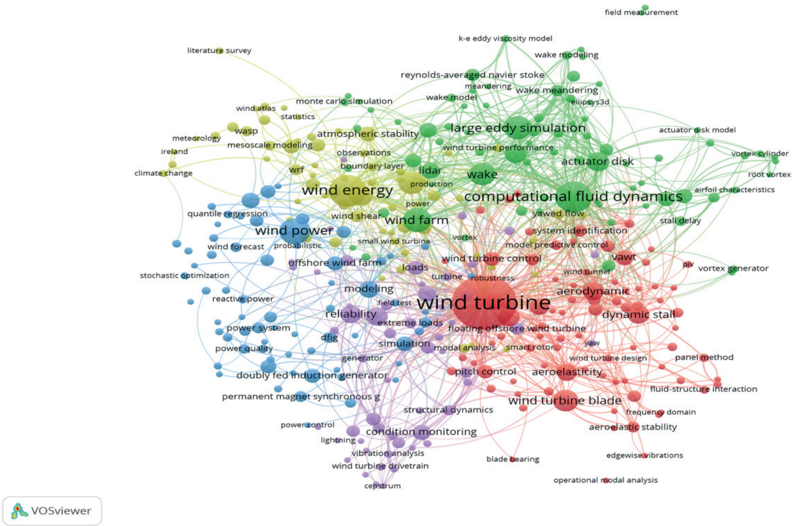


Figure 1. Keyword co-occurrence 1998 – 2018. 356 keywords occurring 3 or more times where analysed. A cluster size of 50 and a resolution of 0.5 were applied in VOS - viewer.

we calculated the overlap of existing IRPWind taxonomy terms with terms found by co-occurrence analysis.

Table 1. Overlap of IRPWind taxonomy terms and author keywords. L1 to L4 indicates the hierarchical levels in the IRPWind taxonomy for topics.

IRPWind level	L1	L2	L3	L4	Sum
#IRPWind terms (topics)	5	28	36	6	75
#author keywords (%)	3 (60%)	10 (36%)	13 (36%)	1 (17%)	27 (36%)

Conclusions

The resulting clusters were comparable to the IRPWind taxonomy of the WE topics. In research fields lacking metadata schemes and taxonomies, the use of author keywords instead of expert elicitation to arrange suitable taxonomies has pros and cons. An advantage is that using uncontrolled vocabularies allows detecting trends in scientific disciplines. Also, the procedure does not demand extensive use of human resources, as experts only will supervise the automatic procedures.

A shortcoming is that authors might use different words to identify the same activity, topic, instrument or variable depending on their field of activity e.g. electrical engineers use mostly wind power plant instead of wind farm. Also, the cumulated amount of author keywords will be a mix of terms identifying different categories e.g. activities, variable, topics, instruments etc., that must be semantically filtered in a number of acknowledged categories and meaningfully clustered.

References

Romo-Fernandez et al. (2013). Co-word based thematic analysis of renewable energy (1990–2010) *Scientometrics*. *97*(3), 743-765

Sempreviva, A.M. et al. (2017). *Taxonomy and meta data for wind energy R&D. Work Package 2-Deliverable D2.3.* IRPWind http://doi.org/10.5281/zenodo.1199489

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538

Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data,* 3:160018 doi: 10.1038/sdata.2016.18

Woon, W.L., Henschel, A. and Madnick, S. (2009) A framework for technology forecasting and visualization. *2009 International Conference on Innovations in Information Technology*. IEEE, 155-159. http://dx.doi.org/10.1109/IIT.2009.5413768

Woon, W.L. and Madnick, S. (2009). Asymmetric information distances for automated taxonomy construction. *Knowledge and Information Systems*, *21*(1), 91-111.

European Tertiary Education Register (ETER): Evolution of the Data Quality Approach

Cinzia Daraio, Renato Bruni, Giuseppe Catalano, Alessandro Daraio, Giorgio Matteucci[1], Monica Scannapieco[2]
Daniel Wagner-Schuster[3], Benedetto Lepori[4]

[1]daraio@dis.uniroma1.it, bruni@diag.uniroma1.it, giuseppe.catalano@uniroma1.it, alessandro.daraio@gmail.com, matteucci@diag.uniroma1.it *DIAG, Sapienza University of Rome (Italy)*, [2]monica.scannapieco@istat.it, *Italian National Institute of Statistics (Italy)*, [3]daniel.wagner-schuster@joanneum.at, *JOANNEUM RESEARCH, Institute for Economic and Innovation Research (Austria)*, [4]benedetto.lepori@usi.ch, *Università della Svizzera italiana, Faculty of Communication sciences (Switzerland)*

Introduction

The ETER project created a database on Higher Education Institutions (HEIs) in Europe, concerning their basic characteristics and geographical location, staff, finances, education and research activities. This database is publicly available at: https://www.eter-project.com.

The ETER database is targeted to include 37 countries composed by the 28 EU Member States, plus EFTA countries (CH, IS, LI, NO) and other five EU candidate countries (AL, ME, MK, RS, TR). The current ETER coverage includes the perimeter (i.e. the list of HEIs) and descriptors for all countries, while quantitative data are available for 35 countries, with the exclusion of Montenegro and Romania. The Belgium data collection is limited to the Flemish region. ETER data have been collected for six years (2011-2016).

The ETER database includes **3,198** unique HEIs over all years. For the academic year 2015/2016, 22.1 million undergraduate and graduate students and around 688 thousand PhD students are accounted in ETER.

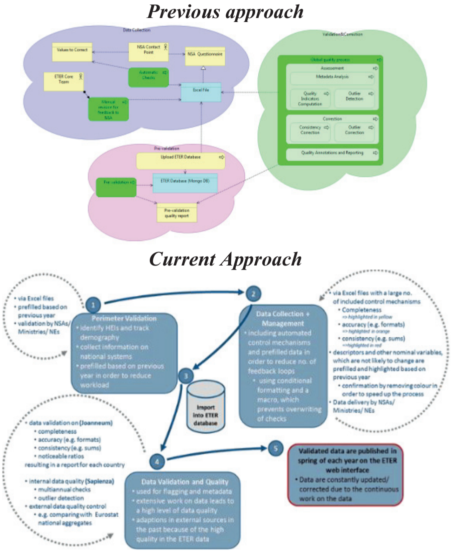
ETER data have been provided by National Statistical Authorities (NSAs), Higher Education Ministries or Higher Education Agencies, based on national statistical databases or higher education information systems. They have been complemented by descriptors and geographical information mostly collected by the ETER consortium. Data for DK, IS, LU, MK, TR have been collected directly by the consortium based on official data published online: their coverage is partial and they have not been validated by the respective NSAs.

The degree of completeness in ETER varies among countries and variables. Descriptors are available almost for all HEIs (completeness level between 0.96 and 1). Among quantitative variables, the completeness level ranges between very high for educational activities (e.g. students enrolled ISCED 5-7: 0.81), to very low for some financial variables breakdowns (e.g. other core budget that is available in 18% of the cases). Variables on staff are in an intermediate position (e.g. total academic staff FTE 0.56) with the exception of the breakdown by ISCED-F.

Relevance of Data Quality

Data quality is a relevant interdisciplinary issue, studied in statistics, management and computer science. Poor data quality greatly reduces their value. Validation and data quality controls are central tasks in ETER, facing challenges rose by the specific nature of ETER data: i) micro-data at institutional level with high level of heterogeneity instead of aggregate data, ii) second level data collection based on data collected nationally largely without a common reference framework. The latter implies also a limited control on the overall data collection process.

The goal of this work is to present the ETER data quality system. This approach (Daraio et al. 2018) combines different methods, including a systematic analysis of internal quality of data (format accuracy, completeness, consistency, timeliness), and advanced statistical methods for outlier detection and analysis of comparability, based on metadata for checking external validity by comparing ETER data with other data sources. The data quality methodology has evolved over time (see Figure 1 for an overview).


Previous approach

Current Approach

Data Collection
Values to Correct
NSA Contact Point
NSA Questionnaire
ETER Core Team
Automatic Checks
Excel File
Manual revision for feedback to NSA
Validation&Correction
Global quality process
Assessement
Metadata Analysis
Quality Indicators Computation
Outlier Detection
Correction
Consistency Correction
Outlier Correction
Quality Annotations and Reporting
Pre-validation
Upload ETER Database
Pre-validation
ETER Database (Mongo DB)
Pre-validation quality report
1
Perimeter Validation
• identify HEIs and track demography
• collect information on national systems
• prefilled based on previous year in order to reduce workload
• via Excel files
• prefilled based on previous year
• validation by NSAs/ Ministries/ NEs
2
Data Collection + Management
• including automated control mechanisms and prefilled data in order to reduce no. of feedback loops
• using conditional formatting and a macro, which prevents overwriting of checks
• via Excel files with a large no. of included control mechanisms
• Completeness => highlighted in yellow
• accuracy (e.g. formats) => highlighted in orange
• consistency (e.g. sums) =>highlighted in red
• descriptors and other nominal variables, which are not likely to change are prefilled and highlighted based on previous year
• confirmation by removing colour in order to speed up the process
• Data delivery by NSAs/ Ministries/ NEs
3
Import into ETER database
4
Data Validation and Quality
• used for flagging and metadata
• extensive work on data leads to a high level of data quality
• adaptions in external sources in the past because of the high quality in the ETER data
• data validation on (Joanneum)
• completeness
• accuracy (e.g. formats)
• consistency (e.g. sums)
• noticeable ratios
resulting in a report for each country
• internal data quality (Sapienza)
• multiannual checks
• outlier detection
• external data quality control
• e.g. comparing with Eurostat national aggregates
5
Validated data are published in spring of each year on the ETER web interface
• Data are constantly updated/ corrected due to the continuous work on the data

Figure 1. Overall ETER Data Quality Approach

Changes and developments

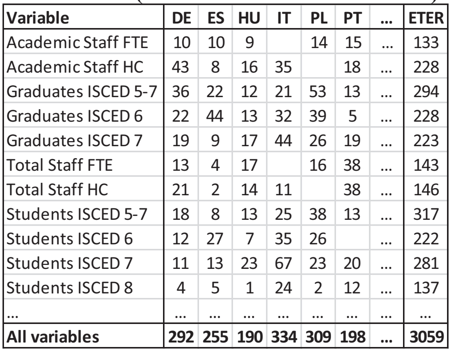
New types and more flexible checks have been introduced. For the multiannual analyses, the following checks are performed:

- Check of the *discontinuity*: useful to identify large changes and therefore to capture the volatility of variations over time. It is based on the computation of the variance of deltas (i.e. the variations) and their normalization that uses the power of the geometric mean (PGM). The power is used to implement the level of scale invariance chosen, that is the level of variation of the check with respect to the size of the variable. To this purpose, a scale invariance parameter (SI) is used to account for intermediary cases between fully scale invariant and fully scale variant cases. The SI parameter has been set at 0.5 (intermediate case), but can be changed manually from 1 (fully scale invariant) to 0 (fully scale variant).
- Check of the *variance* of deltas: identifies fluctuating trends and therefore captures the changes of direction in the time trend. It is based on the calculation of the sum of positive deltas and the sum of negative deltas. After that, the proposed algorithm calculates the product of the two sums of deltas and normalizes it using the power of the geometric mean in order to implement the desired level of scale invariance.

Overall, applying this methodology to 19 ETER variables over all reference years (2011-2016) 3,059 cases have been highlighted and checked in detail, spread in 33 countries (see Table 1). The distribution by country in general follows the size of the country in terms of number of institutions in ETER, with the six larger countries (DE, ES, IT, PL, TR, UK) accounting for one half of the cases.

In terms of variables, more than two third of cases concern student population (students and graduates) but volatility emerges in all variables considered.

Table 1. Outcome of the multiannual checks: cases detected (selected countries and variables)



Variable	DE	ES	HU	IT	PL	PT	…	ETER
Academic Staff FTE	10	10	9		14	15	…	133
Academic Staff HC	43	8	16	35		18	…	228
Graduates ISCED 5-7	36	22	12	21	53	13	…	294
Graduates ISCED 6	22	44	13	32	39	5	…	228
Graduates ISCED 7	19	9	17	44	26	19	…	223
Total Staff FTE	13	4	17		16	38	…	143
Total Staff HC	21	2	14	11		38	…	146
Students ISCED 5-7	18	8	13	25	38	13	…	317
Students ISCED 6	12	27	7	35	26		…	222
Students ISCED 7	11	13	23	67	23	20	…	281
Students ISCED 8	4	5	1	24	2	12	…	137
…	…	…	…	…	…	…	…	…
All variables	**292**	**255**	**190**	**334**	**309**	**198**	**…**	**3059**

The detected cases can be grouped into three categories:

1. Breaks in time series already known and flagged, both as consequence of demographic events or methodological discontinuities (e.g. change in the classification of curricula);
2. Country systemic issues, involving a large number of HEIs in one country and therefore pointing to breaks in time series which have not been notified or flagged before (e.g. new accounting method for contract staff);
3. Individual cases, which may be the consequence of mistakes in data reporting or special characteristics of the institution (e.g. recently founded HEIs).

The Consortium, together with NSAs, controlled individually and corrected or flagged all cases in categories 2 and 3.

Ratios to detect comparability problems that are mostly country-specific (see Table 2) have been also proposed and implemented in the last wave of data collection. Over 2,000 cases have been detected in the first test.

Table 2. Ratios for consistency analysis

Code	Name
R1	Enrolled Students / Academic Staff
R2	Academic staff / Total staff
R3	Personnel expenditure / Total staff
R4	Personnel expend. / Total expenditure
R5	Total expenditure / Total revenue
R6	Basic Governm. funds / Total revenue
R7	Graduates 5-7 / Enrolled students 5-7
R8	Graduates 8 / Enrolled students 8

Acknowledgments

Work done with the support of the European Commission ETER Project (No. 934533-2017-AO8-CH).

Selected References

Batini, C., & Scannapieco, M. (2016). *Data and information quality*. Springer.

Daraio C., Scannapieco M., Catarci T. and Simar L. (2018), ETER Data Quality Report, October 2018.

Lepori, B., Bonaccorsi, A., Daraio, A., Daraio, C., Gunnes, H., Hovdhaugen, E., Ploder, M., Scannapieco, M., Wagner-Schuster, D. (2018), Implementing and Disseminating the European Tertiary Education Register – Handbook for data collection. Brussels.

Scannapieco, M., Missier, P., & Batini, C. (2005). Data quality at a glance. *Datenbank-Spektrum*, 14, 6-14.

The State of Open Access in Germany:
An Analysis of the Publication Output of German Universities

Neda Abediyarandi and Philipp Mayr

neda67a@gmail.com; philipp.mayr@gesis.org
GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

Introduction

Starting with the Berlin declaration in 2003, Open Access (OA) publishing has established a new era of scholarly communication due to the unrestricted electronic access to peer reviewed publications. OA offers a number of benefits like e.g. increased citation counts (Gargouri et al., 2010) and enhanced visibility and accessibility of research output (Tennant et al., 2016). The OA movement with its powerful mandating and policymaking has been very successful in recent years. Relatively little is known about the real effects of these activities in terms of OA publication output of institutions on a larger scale (Piwowar et al., 2018). The aim of this article is to investigate to what extent the OA fraction of the publication output of German universities has increased in the last years. To answer this question, we analysed and compared total number of publications which have been published by researchers of the largest German universities. We compared the numbers of OA versus closed publications for 66 large German universities in the time span of 2000-2017.

Methodology

We follow the classic definitions and classify publications into three categories: Green OA, Gold OA and Closed. Closed access journals allow papers to be read by users with a subscription to the journal (Prosser, 2003). There are two major ways for peer reviewed journal articles to OA, publishing in pure OA journals (gold OA) or archiving of article copies or manuscripts at other web locations (green OA) (Björk et al., 2014).

For the analysis we used Web of Science (WoS) and UNPAYWALL[1] (Piwowar et al., 2018) to extract and analyse our data. To identify German university affiliations in WoS, we used data from the Competence Centre for Bibliometrics, in particular the result of the project “Institutional address disambiguation” (Rimmert et al., 2017).

We first selected 66 German universities which have more than 1,900 publications in WoS in a period of 17 years (2000-2017). In the following, we matched all WoS publications of these 66 universities with UNPAYWALL publications. We considered matching based on DOI and title to get precise results. We got round 34% matched publications because a larger number of DOIs for publications in WoS was missing (especially between 2000 and 2002). In the WoS dataset each publication can be affiliated with some authors. To remove redundancy, we randomly allocated each specific publication to one of its authors, in other words, if a publication is written by several authors from different universities; we counted just for one of them. In Table 1 we list the 10 German universities with the most matched WoS publications from 2000 to 2017.

Table 1. Total number of matched WoS publications by top 10 German universities (2000-2017).

University	Matched WoS articles
Heidelberg Univ.	72,556
LMU Univ.	67,525
Charité Berlin Univ.	63,949
Technical Munich Univ.	63,641
Bonn Univ.	54,671
Nuremberg Univ.	53,289
Karlsruhe Univ.	51,266
Hamburg Univ.	48,880
Freiburg Univ.	47,574
Technical Dresden Univ.	47,137

Approach

In the following, we are investigating the percentage of publications of German universities published in gold, green and closed access. In order to answer this question, we analysed our extracted data in two different aggregations.

1. Comparing the number of publications for the top 10 German universities: We analysed and compared the total number of gold, green and closed access publications for the top 10 German universities (see Table 1 and Figure 1) in terms of matched WoS articles.

[1] The UNPAYWALL dataset includes millions of articles in which publications were separated based on their access type (Green, Gold and Closed). https://unpaywall.org/

2. Comparing groups of German universities: We grouped 66 German universities into three different groups based on total number of their published WoS publications from 2000 to 2017.

The different groups of universities are the following:

- *Group 1*: 22 German universities which have published more than 31,000 publications (this includes the top 10 universities from Table 1).
- *Group 2*: 22 German universities which have published more than 12,000 and less than 31,000 publications.
- *Group 3*: 22 German universities which have published more than 1,900 and less than 12,000 publications.

We compared the total number of gold, green and closed access publications which were published by each mentioned group in year 2000, 2010 and 2017 separately (see Figure 2). To verify our analysis, we compared our data with the recent CWTS Leiden Ranking from May 2019[2]. We found a good match between our and the Leiden numbers for the German universities.

Results

The total numbers of gold, green and closed access publications for top 10 German universities from 2000 to 2017 are shown in Figure 1. Our findings show that all top 10 German universities still tend to publish most publications within the closed access model. If we compare with Figure 2, we see that the ratio of closed access publications is decreasing, but in the year 2017 still 50% and more of the WoS articles are published in closed access.

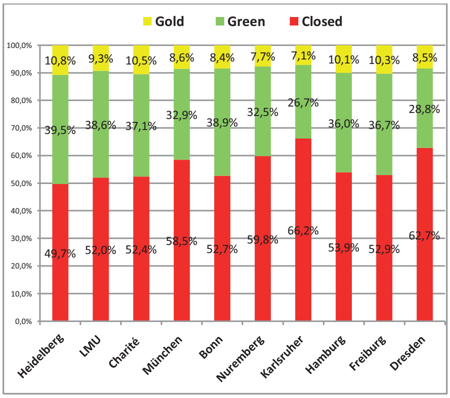


Figure 1. Comparison of total number of gold, green and closed access publications in the top 10 German universities (2000-2017).

We found that the top 10 German universities published more gold/green access publications rather than the others. Figure 2 shows the percentage of gold and green access publications for each group are significantly increasing in the last 7 years.

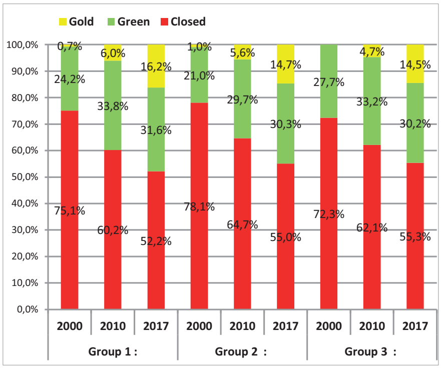


Figure 2. Three groups of German univ. based on their total number of matched WoS publications (2000, 2010, 2017).

Future Work

As a next step, we plan to analyse the effects of concrete OA mandating in Germany and abroad on the number of green and gold OA publications, their citation advantages and possible enhanced research visibility. In the future, we plan to compare the OA situation in Germany with other European countries and institutions all around the world.

Acknowledgement

This work was supported by BMBF project OASE, grant number 01PU17005A.

References

Björk, B.-C., et al. (2014). Anatomy of green open access. JASIST, 65(2), 237–250.

Gargouri, Y., et al. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLoS ONE, 5(10), e13636.

Piwowar, H., et al. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375.

Prosser, D. C. (2003). From here to there: a proposed mechanism for transforming journals from closed to open access. Learned Publishing, 16(3), 163–166.

Rimmert, C., Schwechheimer, H., & Winterhager, M. (2017). Disambiguation of author addresses in bibliometric databases - technical report.

Tennant, J. P., et al. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review. F1000Research, 5, 632.

[2] http://www.leidenranking.com/

Unveiling the path towards sustainability: is there a research interest on sustainable goals?

Núria Bautista-Puig[1] and Elba Mauleón[2]
[1]*nbautist@bib.uc3m.es*, [2]*mmauleon@bib.uc3m.es*
Research Institute for Higher Education and Science (INAECU), University Carlos III of Madrid, 126 Madrid Str., 28903, Getafe (Spain)

Introduction

Society in the 20th century has faced the effects of an unsustainable production model that led to global awareness of this topic. Different summits and conferences were held in which sustainability and sustainable development (SD) were the core discussion: this fact can be interpreted as a growing compromise from countries to work together on that issue and a topic that has become as policy priority setting. One of the most important was the Millennium Summit celebrated in 2000 which led to the establishment of the 8 Millennium Development Goals (MDGs). These goals were criticised for not be adequately aligned with human rights standards and principles and were relevant for poor countries only (Fukuda-Parr, 2016), and rich countries were supposed to add solidarity and assistance with finances and technology (Sachs, 2010). As a result, developing countries were made substantial progress towards the achievements of the goals. A framework with 17 Sustainable Development Goals (SDGs) indicators was established later with the Agenda 2030 (Rio, 2012). These indicators are an improved adaptation of the MDGs, and as Griggs stated (2013), they gave more attention to issues of environmental sustainability than the previous one. As well, these goals have a more ambitious vision because it assumes every country should assume responsibility and goes beyond international cooperation, but also focusing on eliminating discrimination and inequalities within the countries (Leal Filho et al., 2019).

Sources and methodology

The main objective of this paper is to analyse the scientific research on this subject and propose a methodology for classifying SDGs. Figure 1 illustrates the research methodology of this study.

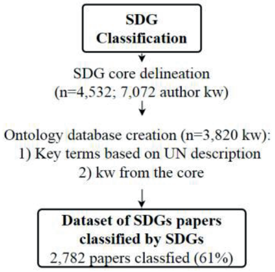

SDG Classification
SDG core delineation (n=4,532; 7,072 author kw)
Ontology database creation (n=3,820 kw):
1) Key terms based on UN description
2) kw from the core
Dataset of SDGs papers classified by SDGs
2,782 papers classfied (61%)

Figure 1. Research methodology

Results

4,532 documents were detected from 2000 to 2017 and this can be defined as the core of SDGs research. From this, 3,328 documents were from Millennium Development Goals (MDG) and 1,426 from Sustainable Development Goals (SDG) with an overlap of 218 documents between both. The cumulative average growth rate (CAGR) of scientific production during the period is 42.36% and, this growth, can be explained by the significant penetration of this topic in the Academia and research. During the period of study, the major increase of the scientific production can be related with the last years (2014-2017) with an increase of 110.39%, which coincides with the official launch of SDG in 2015.

By checking the co-occurrences map based on keywords, on the first years (2000-2005), it is observed that topics were related with health and well-being with terms such as malaria or mortality, but also with poverty or education. Besides, only keywords related to 'developing countries' or 'Africa' appeared, denoting the scope of these goals were focused on developing countries. In a more recent co-occurrence map (2012-2017), a more variety of topics was identified. Special emphasis should be taken to a node that includes terms such as policy or politics, which denote the interest to involve policymakers' decisions.

An ontology was created based on central keywords in the SDG United Nations description (e.g. 'sanitation' was classified into 'SDG6'). From the core research output, 7,072 author keywords were identified in the period and also were targeted into the goals. Only keywords with a clear focus on the SDGs were preliminary classified (e.g. 'Malaria' was classified into 'SDG3 Good Health and Well-being'). 3,820 keywords constituted the ontology and 2,782 papers (61%) from the core were classified. One paper could be multi-assigned. Then a co-occurrence network based on document relations on the SDGs was created (Figure 2). This shows the relations between SDGs papers and is divided into four clusters: Cluster no. 1 is related with urban settlement, its components and dynamics: sanitation (SDG6), responsible consumption and production (SDG12) and also with water from oceans, sear or marine resources (SDG14). Cluster no. 2 is related to governance and partnership for achieving SDGs and has the strongest edges between the nodes, denoting a strong relation of these SDGs to have the topic of this goals published together. Besides, it has a strong connection with health node (SDG3). Cluster no. 3 is more environmental-related (climate-change aspects, land or energies); as well, it is linked with 'Industry, innovation and infrastructures' (SDG9) showing their link with the effects of the industry on the environment, with impact effects. Cluster no. 4 is related with socio-economic aspects (poverty, education) and labour aspects (work) and to reduce inequalities (SDG10).

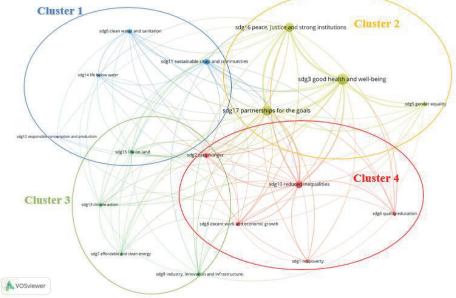

Cluster 1
sdg6 clean water and sanitation
sdg11 sustainable cities and communities
sdg14 life below water
sdg12 responsible consumption and production
sdg16 peace, justice and strong institutions
Cluster 2
sdg3 good health and well-being
sdg5 gender equality
sdg17 partnerships for the goals
sdg15 life on land
sdg2 zero hunger
Cluster 3
sdg13 climate action
sdg10 reduced inequalities
Cluster 4
sdg4 quality education
sdg8 decent work and economic growth
sdg7 affordable and clean energy
sdg1 no poverty
sdg9 industry, innovation and infrastructure;
VOSviewer

Figure 2. Co-occurrence map of SDG classification in scientific output (2000-2017).

Discussion and conclusions

Achieving long-term sustainability has become a challenge for all countries. As Salvia (2019) pointed out, the success on SDGs will rely on the strengthened collaboration of its actors. The involvement of the countries in the sustainable goals and their inclusion in global policy debates and national policy planning denotes the interest of global leaders in to achieve a more sustainable growth.

Is there a SDGs research?

This growth continues the findings of the research by Hassan et al. (2014) regarding SD scientific output, demonstrating the recent interest on this topic by the research community. It should be remarked the growth on the last 4-year period (increase of 100.39%) which coincides with the launched of the SDGs.

What topics are addressed in SDG scientific core production?

From the keyword co-occurrence map, it is stated that MDGs were more developing countries focused and did not consider environmental issues in comparison with SDGs. Topics were more related to health (SDG3), poverty (SDG1) and education (SDG4). With SDGs, the main difference is these ones implies the involvement of all countries with 'no one left behind' lemma. A new node include policy, denoting the involvement of policymakers. As well, with the growth and expansion of urbanization worldwide (Ramírez et al., 2016), cities and communities (SDG11) has appeared as a node on scientific literature.

What are the SDGs relations between papers in the core scientific literature?

An ontology classification approach by SDGs is presented in order to determine which ones are the most focused in the scientific literature and relations between them. Partnership (SDG17) and the strong institutions (SDG16) are the most predominant, showing a strong connections between these goals. This goes in line with Waage et al. (2015), which stated the SDGs will mainly act in governance and partnership among the subscribing states. As well, a cluster of environmental-related can be identified, with SDG related to environment and its interactions (SDG13- Climate action, SDG15- Life on land).

Acknowledgments

This research was supported by predoctoral contract BES-2015-075461 from the Ministry of Economy and Competitiveness (MINECO).

References

Hassan, S. U., Haddawy, P., & Zhu, J. (2014). A bibliometric study of the world's research activity in sustainable development and its sub-areas using scientific literature. Scientometrics, 99(2), 549-579.

Leal Filho, W.; Tripathi, S. K.; Andrade Guerra, J. B. S. O. D.; Giné-Garriga, R.; Orlovic Lovren, V.; Willats, J. (2019). Using the sustainable development goals towards a better understanding of sustainability challenges. International Journal of Sustainable Development & World Ecology, 26(2), 179-190.

Ramírez, J. F. R., Ibañez, A. M. A., & Montenegro, D. F. (2016). Los discursos de la sostenibilidad: análisis de tendencias conceptuales a partir de mediciones bibliométricas. Questionar: Investigación Específica, 4(1), 82-96.

Sachs, J. D. (2012). From millennium development goals to sustainable development goals. The Lancet, 379(9832), 2206-2211.

Salvia, A. L., Leal Filho, W., Brandli, L. L., & Griebeler, J. S. (2019). Assessing research trends related to Sustainable Development Goals: local and global issues. Journal of Cleaner Production, 208, 841-849.

Fukuda-Parr, S. (2016). From the Millennium Development Goals to the Sustainable Development Goals: shifts in purpose, concept, and politics of global goal setting for development. Gender & Development, 24(1), 43-52.

Waage, J., Yap, C., Bell, S., Levy, C., Mace, G., Pegram, T., ... & Mayhew, S. (2015). Governing the UN Sustainable Development Goals: interactions, infrastructures, and institutions. The Lancet Global Health, 3(5), e251-e252.

A Study on Grasp of Research Trend based on Abstract Analysis: Using the Theses of X-ray Exploration Satellite "SUZAKU"

Yuji Mizukami[1], Kyosuke Nakamura[1], Akiko Ohata[2], Kesuke Honda[3], Junji Nakano[4]

[1]*mizukami.yuuji@nihon-u.ac.jp*
[1]Nihon University, College of Industrial Technology, 1-2-1 Izumi, Narashino, Chiba 275-8575 (Japan)

[2]Japan Aerospace Exploration Agency, Institute of Space and Astronautical Science, Kanagawa (Japan)

[3]The Institute of Statistical Mathematics, Tokyo (Japan)

[4]Chuo University, Department of Global Management, Tokyo (Japan)

Introduction

Various observation data obtained from artificial satellites and from the explorers are expected to be widely used in the government, academia and other industry. Meanwhile, since the space development project is a large-scale, its contribution is required to be explained in all sessions including the National Assembly. It is necessary to provide objective explanation materials on the contribution. Therefore, this paper aims to derive an analytical method in preparing objective information which is an auxiliary explanation to the contribution. In the analysis, text mining analysis is applied to the abstract of the thesis to extract research trends by year to verify the effect of the method. The theses to be analysed which is the related theses of the X-ray exploration satellite "SUZAKU" are collected from the bibliographic databases.

Analysis data

The analyzed satellite is the X-ray exploration satellite "SUZAKU". The satellite was launched from the Uchinoura Space Observatory with the MV rocket 6 in 2005, and has been observed in 2015 after 10 years of operation.

There are three reasons why this satellite was selected. As the first reason, it is a relatively recent satellite, so it is considered to be suitable for collectiong the current situation. As the second reason, there is a 10-year operation period, there is a lot of data that needs to be collected. The last reason is that the operation has already been completed, and there is no further increase in measurement data, and permanent analysis is possible.

The papers to be analyzed were collected using Clarivate Analytics's bibliographic database Web of Science-Core collection (WOS).

Table 1 shows the search criteria for papers. The paper to be analyzed is an English-language paper whose topic includes "SUZAKU", and is for 18 years from 2001 to 20019. As a result of the search, 1419 papers were analyzed.

Table 1 Search condition

WoS Search condition : (TS=(SUZAKU) OR TS=("ASTRO-E II") OR TS=("ASTRO-E 2") OR TS=("ASTROE II") OR TS=("ASTROE 2") OR TS=("ASTROEII") OR TS=("ASTROE2") OR TS=("ASTRO-EII") OR TS=("ASTRO-E2")) AND **Language:** (English) AND **Document type:** (Article OR Review)
Index=SCI-EXPANDED, SSCI, A&HCI, ESCI Time span=All

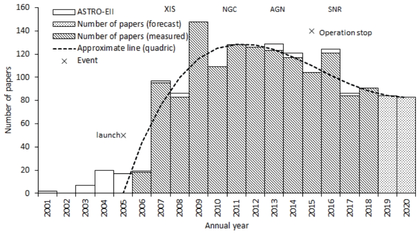


Figure 1 Annual change in the # of papers

Figure 1 shows the annual change of "SUZAKU" satellite related papers. "ASTRO-EII" was the name of the project before launching the "SUZAKU" satellite. Also, after JAXA(Japan Aerospace Exploration Agency)'s launch, the project name has been changed to "SUZAKU" by the convention of JAXA. In this paper, the papers retrieved by "ASTRO-EII" are excluded from analysis because they are considered not to be papers using Suzaku satellite observation data. Among the actual papers, 18 papers have been published in 2006, the year after launch, and 149 papers in 2009 show the largest number and then the numbers decline.

Results

Word extraction analysis results

Table 2 shows the results of word extraction analysis for the abstract information of the dissertation. The left side is the top 20 most frequently occurring nouns, and the right side is the top 20 most frequently occurring proper nouns.

Table 2 Word extraction analysis

Noun						Proper noun					
Rank	Word	Quantity	Rank	Word	Quantity	Rank	Word	Quantity	Rank	Word	Quantity
1	x-ray	3350	11	Region	760	1	SUZAKU	1608	11	Compton	172
2	emission	2023	12	Galaxy	711	2	XMM-Newton	322	12	NuSTAR	162
3	observation	1535	13	Flux	660	3	AGN*	313	13	ICM*	155
4	spectrum	1505	14	Temperature	655	4	Galactic	285	14	Telescope	149
5	source	1197	15	Accretion	556	5	Chandra	276	15	Gamma	144
6	component	916	16	Gas	551	6	NGC*	271	16	kT	129
7	datum	894	17	Disk	530	7	Seyfert	243	17	Si	121
8	energy	888	18	Density	504	8	XIS*	240	18	Spectrometer	110
9	result	887	19	Absorption	496	9	SWIFT	211	19	MG	101
10	cluster	807	20	Time	489	10	SNR*	197	20	Imaging	100

* Those are shortened forms of the analysis method using X-rays.

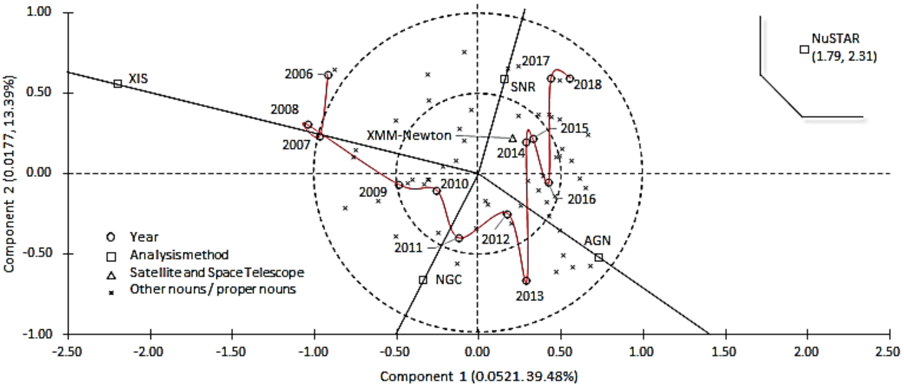


Figure 2 Correspondence Analysis of Dissertation Information

In noun section on the left side, many words related to X-ray analysis are shown. As for the proper nouns on the left side, AGN which is an analysis method of X-ray analysis is 3rd, NGC is 6th, XIS is 8th, SNR is 10th, and ICM is 13th.

Result of correspondence analysis

Figure 2 shows a scatterplot of the correspondence analysis between the abstract information and the dissertation year information. This figure is a scatter diagram of the first component and the second component, and explains 52.87% (← 39.48% + 13.38%) of the whole.

Annual information is from 2006 to 2018, and analysis methods such as XIS (X-ray Imaging Spectrometer), NGC **(**New General Catalogue**)**, AGN **(**Active Galactic Nuclei**)**, and SNR (Active Galactic Nuclei) are indicated. Also, based on the distance from the center, strong characteristics are shown for three years from 2006 to 2008, 2009, 2013, 2017, 2008 are intermedidate characteristics, and others are weak characteristics.

Conclusion

In order to express the degree of utilization of measurement data of satellites and spacecrafts by the number of papers and their features, we tried to apply text mining analysis using the information of X-ray exploration satellite "Suzaku" project. As a result of analysis, it was found that the trend of research on Suzaku-related papers has shifted its features to XIS, NGC, AGN, and SNR in 13 years. From this result, it is thought that Suzaku satellite was able to operate flexibly according to the needs of research.

Acknowledgments

This study was funded by the JSPS KAKENHI Grant-in-Aid for Scientific Research (C) (Grant Number 17K04710).

This study was funded by the Institute of Statistical Mathematics joint research program (2019-ISMCRP-1026).

References

Yuji Mizukami et al., An International Research Comparative Study of the Degree of Cooperation between disciplines within mathematics and mathematical sciences, Springer, Behaviormetrika, Vol.1, 19 pages, On-line, 2017

Ministry of Education, Culture, Sports, Science and Technology (2017), Benchmarking for Scientific Research 2017 <http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-RM262-FullJ.pdf> (Last Access: 2019 April 3)

Can Twitter hashtags be used for field delineation? The case of Sustainable Development Goals (SDGs)

Núria Bautista-Puig[1] and Jonathan Dudek[2]

[1] *nbautist@bib.uc3m.es*
Research Institute for Higher Education and Science (INAECU), University Carlos III of Madrid, 126 Madrid Str., 28903, Getafe (Spain)

[2] *j.dudek@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333 AL (The Netherlands)

Introduction

Society in the 20th century has faced the effects of an unsustainable production model of global impact. Against this background, the Millennium Development Goals (MDGs) were established in 2000, followed by the definition of 17 Sustainable Development Goals (SDGs) in order to advance the path towards sustainability. This topic has consequently risen to priority-setting status in the policy context, creating a collective awareness with an inclusive perspective at all levels of society. Achieving the SDGs has become a challenge for all countries, involving a multitude of stakeholders (Brown, 2006).

Research can be a response to topics that are considered a societal challenge and multi-faceted such as the SDGs. As well, the delineation of a field is crucial for decision-support studies: it allows to understand actors involved and to analyse the dynamics of a field.

This study proposes a delineation procedure to retrieve scientific publications centred around the SDGs. Our goal is to complement an ontology-based approach with an approach that is based on Twitter data. Twitter is seen as a relevant resource in the dissemination of scientific literature (Robinson et al., 2014) and one interesting point in this is the use of hashtags. Previous studies have analysed hashtag usage (Romero et al., 2011), but not applied this element to field delineation. We seek to understand how hashtags might be used for field delineation, asking: Can Twitter be used for identifying and delineating publications related to the 17 SDGs?

Sources and methodology

This study is based on records of scientific production in the Web of Science (WoS); as a second source, we referred to the database of Altmetric.com for Twitter records of scientific publications. The test of delineation of publications according to the 17 SDGs followed two different steps:

First, we identified publications in the WoS of SDGs as a seed of publications on this topic: these were determined by searching SDGs and MDGs on title, abstract and keywords. Then, an ad-hoc ontology was created for each SDG with a total of 3,825 terms. This was based on the selection of key terms from the description of the SDGs by the United Nations (United Nations, 2019) as well as the keywords taken from the initial seed of publications.

In the second part, we searched for tweets containing the hashtags '#MDG' and '#SDG' as well as hashtags referring to the different goals (e.g., '#SDG1') in the Twitter data by Altmetric.com. Consequently, any publication referred to in those tweets was collected.

Results

Hashtags as a retrieval element

By using the search strategy in the WoS, 4,725 documents were retrieved from 2000 to 2017. In the Twitter hashtag approach 1,300 unique documents could be collected. Considering the different SDGs, the hashtags that retrieved the most publications were SDG1 (No poverty) and SDG3 (Well-being and Health), denoting the importance of these topics in social networks (Table 1). The overlap between the results of both retrieval methods is 333 distinct documents, meaning that 75% of publications identified based on hashtags were not included in the set of publications identified with the seed.

Table 1. Publications per hashtag (publications can be assigned to multiple hashtags).

	Hashtag retrieval	SDGs seed
#SDG	994	224
#MDG	381	137
#SDG1	114	9
#SDG2	22	1
#SDG3	68	10
#SDG4	26	4
#SDG5	14	1
#SDG6	52	8
#SDG7	10	3
#SDG8	2	0
#SDG9	2	0

#SDG10	9	0
#SDG11	10	1
#SDG12	10	0
#SDG13	17	0
#SDG14	35	5
#SDG15	3	0
#SDG16	15	1
#SDG17	18	3

Topical comparison of retrieval methods

Regarding the classification of scientific publications according to the first approach, the most occurrences were found with 2,472 documents (52.32%) classified as SDG17 (Partnership for the goals). This is followed by 2,075 documents (43.92) into SDG3 (Health and well-being), 1,831 (38.75%) into SDG16 (Peace, justice and strong institutions) and 1,208 (25.57%) into SDG11 (Sustainable cities and communities). It should be considered that each paper could be classified in more than one SDG.

Regarding documents collected by the hashtag strategy, of the 1,300 documents, 933 documents (71%) were assigned to one or more specific SDG (i.e., not only to a general hashtag like "#SDG"). Considering the different goals, 557 were classified into SDG3 (Health) (42.85%), 424 documents (32.62%) into SDG16 (Peace, Justice and Strong Institutions), 392 (30.15%) into SDG17 and 290 (22.31%) into SDG10 (Reducing inequalities). Health seems a prominent topic in both scientific production and the interest apparent on Twitter; however, peace and justice seem to have more relevance on Twitter than in the research community. Partnership for the goals (SDG17) seems to be the most prominent topic among scientific publications included.

Comparing labelling accuracy

In order to determine whether hashtags were properly assigned, the SDGs from hashtags were compared with the SDG-labels assigned based on the ontology. For this validation, only specific hashtags (e.g., "#SDG1") were considered (*n* = 280 unique documents). From this, 36% of the documents were positively classified to the same goal according to the ontology. By checking differences by goals, SDG3 (Health and well-being) is the one that has been classified positively (41 documents), SDG6 (Clean water and sanitation) comes second (31 documents) and SDG4 (Quality education) third (19 documents). See Table 2 for all results.

Table 2. Documents positively assigned

	No Match	Match
#SDG1	111 (39.64%)	3 (1.07%)
#SDG2	12 (4.29%)	10 (3.57%)
#SDG3	27 (9.64%)	41 (14.64%)
#SDG4	7 (2.50%)	19 (6.79%)
#SDG5	9 (3.21%)	5 (1.79%)
#SDG6	21 (7.50%)	31 (11.07%)
#SDG7	5 (1.79%)	5 (1.79%)
#SDG8	2 (0.71%)	0
#SDG9	4 (1.43%)	0
#SDG10	8 (2.86%)	1 (0.36%)
#SDG11	9 (3.21%)	1 (0.36%)
#SDG12	10 (3.57%)	0
#SDG13	8 (2.86%)	9 (3.21%)
#SDG14	19 (6.79%)	16 (5.71%)
#SDG15	1 (0.36%)	2 (0.71%)
#SDG16	7 (2.50%)	8 (2.86%)
#SDG17	13 (4.64%)	5 (1.79%)

Discussion

This paper explores the potential of Twitter hashtags for retrieving scientific publications related to the SDGs. Results show a considerable interest in SDGs on Twitter: 1,300 papers were retrieved by using respective hashtag information. This dataset overlaps in 333 documents with the seed delineated by the search strategy in the WoS. This suggests that Twitter can be used as a retrieval tool for topically classified scientific publications.

Only 36% of the publications mentioned in tweets coincided with the ontology-based classification. This indicates that hashtags are not reliably assigned to publications by Twitter users; hence, we conclude that this method cannot be used for classifying publications along the lines of a specific SDG. Nonetheless, hashtags may give evidence of a publications' topical relatedness to SDGs in general. The value of such "Twitter-crowd-sourcing" of publications should be investigated further.

Acknowledgments

This research was supported by predoctoral contract BES-2015-075461.

References

Brown, L.R. (2006): *Plan B 2.0: Rescuing a Planet Under Stress and a Civilization in Trouble.* W.W. Norton & Company, New York

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: exploring the insides of Altmetric.com. *El Profesional de la Informacion, 23*(4), 359–366. doi:10.3145/epi.2014.jul.03

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 695-704). ACM.

United Nations (2019). SDGs description. Available at: https://www.un.org/sustainabledevelopment/

Using Full-text of Academic Articles to Find Software Clusters

Heng Zhang[1] , Shutian Ma[2] and Chengzhi Zhang[3, *]

[1]zh_heng@njust.edu.cn
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094(China)

[2]mashutian0608@hotmail.com
Department of Information Management, Nanjing University of Science &Technology, Nanjing 210094(China)

[3]zhangcz@njust.edu.cn
Department of Information Management, Nanjing University of Science &Technology, Nanjing 210094(China)

Abstract

Scientific software is making contributions to modern science. To meet huge academic demands such as data analysis, modelling, visualization and so on, various software has been developed to help different steps in scientific work. In order to reveal the connections between scientific software, we conduct cluster analysis among scientific software based on the full-text data of 23,120 articles published in PLOS ONE. Firstly, we select some popular software whose mention times are over 50 to be our candidate software list for clustering analysis. Secondly, Word2Vec is applied to learn distributed representation for each software. Then, we apply Affinity Propagation to cluster software and tune different parameters to obtain better results. Silhouette coefficient is computed here to evaluate clustering performance under each parameter setting. According to our optimal results, software clusters with specific functions can be found. And software which have strong linkage between each other are mainly have functions in common.

Keywords: Scientific Software, Software Clustering, Distributed Representation

Introduction

Scientific software is a critical component in academic researches. It can analyze data, simulate the physical world and visualize the results, one single scientific work will need the support of several software with specific functions. As the unsung heroes (Chawla, 2016), researchers are paying attention to investigate what kind of important role it plays in the advancement of science. For example, Callahan et al. (2018) developed u-Index metric to measure the impact of informatics tools and databases. Smith et al. (2016) discussed on software citation principles that may encourage relevant policies for software citation across disciplines and venues. Since more and more full-text of literature has becoming accessible, some studies are focusing on utilization of text mining for this topic. Duck et al. (2016) identified mentions of databases or software in the PubMed Central full-text corpus through text mining. However, few of the researches are relevant about mining the connections between different software. So, in this paper, we try to reveal the connections between software by cluster analysis.

Methodology

Framework of our study

The main purpose of this paper is to reveal the connections between scientific software. As shown in Figure 1, firstly, from PLOS ONE, we collected 23,120 articles published in 2017. Our original software list comes from the previous work (Pan et al. 2015) and we further filtered out those that was mentioned in less than 50 articles. So, we get 260 software to be the candidate software list for cluster analysis. Secondly, Word2Vec (Mikolov et al. 2013), which can learn high-quality word vectors from huge corpora, is applied to learn distributed representation for these 260 software using full-text. Then, we apply Affinity Propagation (AP) to cluster software using the vector data. Finally, we analyze the characteristics of top-5 clusters which contain the largest number of software. Since the clustering is conducted using software vectors learned by Word2Vec, we want to investigate that if there really exist any relations between those software pair with strong linkage. Here, strong linkage refers to high cosine similarity between software pairs based on their distributed representations.

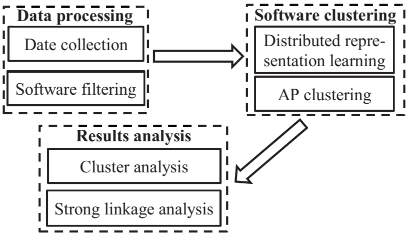


Figure 1. Framework of our work

Clustering evaluation

After clustering, we apply silhouette coefficient to evaluate the performance. Bigger silhouette coefficient value means better cluster result. Affinity Propagation is a clustering algorithm based on similarity matrix of data points (Frey et al. 2007). In

[*] Corresponding author.

this algorithm, *preference* is an important parameter, which controls how many exemplars are used. We set *preference* values from 0 to 0.9 and 0.1 as interval. and the clustering result with the maximum silhouette coefficient was used for the final analysis.

Results analysis

Cluster analysis

Silhouette coefficient of all different clustering results are shown in Table 1. We then further analyze the clusters obtained when the *preference* value is 0.4. The top-5 clusters with the largest number of software are shown in Table 2.

Table 1. Silhouette coefficient of clustering results in different *preference* values

preference	Silhouette coefficient	preference	Silhouette coefficient
0	0.1139	0.5	0.1445
0.1	0.1234	0.6	0.1256
0.2	0.1311	0.7	0.0992
0.3	0.1492	0.8	0.0417
0.4	**0.1576**	0.9	0.0058

Table 2. Top-5 clusters with the most software in the optimal clustering result

Clusters	Software
1	*SPSS, Prizm, Stata, SigmaPlot, Systat, MedCalc, StatSoft, G*Power, PASW, OriginLab, Minitab*
2	*ImageJ, Image ProPlus, NIS Element, AxioVision, Imaris, Aperio, MetaMorph, Feature Extraction, LAS AF, Leica Application Suite, Volocity*
3	*BLAST, SMART, Pfam, STRING, SignalP, Blast2GO, MEME, PANTHER, TMHMM, InterProScan*
4	*MEGA, MrBayes, RAxML, BEAST, STRUCTURE, FigTree, PAUP, TreeAnnotator, Modeltest*
5	*Clustal W, Geneious, MUSCLE, BioEdit, MAFFT, FASTA, Clustal X, Clustal Omega, Sequencher*

Firstly, we find that the software in each cluster is similar in function. In Cluster 1, software are mainly used for statistical analysis. Image processing software are gathered in Cluster 2. Software in Cluster 3 are relevant to protein research while software in Cluster 4 are used more in the study of heredity and evolution. Function of most software in cluster 5 is about multiple sequence alignment for DNA.

Strong linkage analysis

In order to understand more detailed relationships between software, we take software Clustal W as an example and analyze the other software which shows strong linkage with it. There are three software have high cosine similarity (>0.8) with Clustal W, they are Clustal Omega, MUSCLE and Clustal X. In gene sequencing domain, all of them are used for multiple sequence alignment. Besides, Clustal X, Clustal Omega and Clustal W are different versions of Clustal. MUSCLE is an alternative software of Clustal W. But in terms of the accuracy and speed of multiple sequence alignment, MUSCLE is better than Clustal W.

Conclusions

In this paper, we conduct AP clustering for 260 popular software using full-text from PLOS ONE. According to our experimental results, our method can find software clusters with specific functions and the software tend to have functional similarities between each other within each cluster. Since we used software mentioned times when selecting software for analysis, hidden research topics over current publication collection can also be inferred from these popular software based on their functions, such as protein analysis and DNA alignment.

Acknowledgments

This work is supported by Major Projects of National Social Science Fund (No. 17ZDA291), Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. SJCX18_0136), and Qing Lan Project.

References

Singh Chawla, D. (2016). The unsung heroes of scientific software. *Nature, 529*(7584), 115-116.

Callahan, A, Winnenburg, R. , & Shah, N. H. . (2018). U-index, a dataset and an impact metric for informatics tools and databases. *Scientific Data*, 5, 180043.

Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. 2016. Software citation principles. *PeerJ Computer Science 2*: e86. https://doi.org/10.7717/peerj-cs.86

Duck G, Nenadic G, Filannino M, Brass A, Robertson DL, et al. (2016). A Survey of Bioinformatics Database and Software Usage through Mining the Literature. *PLOS ONE 11*(6): e0157989. https://doi.org/10.1371/journal.pone.0157989

Pan, X, Yan, E, Wang, Q & Hua, W. (2015). Assessing the impact of software on science: a bootstrapped learning of software entities in full-text papers. *Journal of Informetrics, 9*(4), 860-871.

Mikolov, T, Chen, K, Corrado, G, & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Frey, B. J, & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972-976.

Specialized User Attention on Twitter: Identifying Scientific Topics of Interest among Social Users of Science

Jonathan Dudek[1] and Rodrigo Costas[2]

[1] j.dudek@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333 AL (The Netherlands)

[2] rcostas@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333 AL (The Netherlands)
DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University (South Africa)

Introduction

Over the recent years, scientific publications have received considerable attention on the social media platform Twitter. Accordingly, tweet counts became part of a range of alternative indicators of scientific impact, also known as 'altmetrics'. Like that, measuring the reception of scientific publications beyond classical metrics (like citation counts) and potentially originating from non-academic members of society becomes feasible. However, with tweet counts alone it remains unknown who exactly has shared a publication, i.e. whom attention for a scientific publication originates from, and how such attention may differ with different groups of users of scientific production.

Having in mind user groups which are not from academia, the question raised in this study is: What do non-academic users tweet about when they share scientific publications? And which scientific topics stand out in the attention on Twitter when comparing specifically non-academic user groups with the tweeting behaviour of the general tweeting population? Pursuing these questions, this study tries to understand how the attention of social users of science concentrates and reveals topics of special societal interest.

Methodology

Identification of Social Users

As a first step, we identified a set of what we term *social users.* Those we define as non-academic and institutional, i.e. non-individual Twitter accounts. For this purpose, we considered Twitter accounts which have shared a traceable link to a scientific publication in the form of a tweet. Such activity is captured by Altmetric.com; we used a dataset comprising a set of tweets from June 2011 to October 2017, stored in the local database of CWTS. Herein, a total of 4,117,887 different Twitter users can be identified.

Firstly, we removed accounts by universities and other academic institutions (e.g., university hospitals). Also, we removed Twitter accounts which might potentially belong to academic individuals (for this step, we refer to the identification procedure described by Costas, Van Honk, and Franssen, 2017.) Furthermore, in order to be included, Twitter accounts needed to show an URL in their Twitter profile, have provided their country of origin, and had set English as account language.

Following these conditional criteria, we applied two different approaches: The first one based on a query by keywords for Twitter accounts. Assuming different types of institutional social users, keywords were generated for searches among the Twitter descriptions, i.e. the Twitter-bio displayed on Twitter accounts. For the development of keywords, we considered different types of institutional social users, namely: Governmental/public institutions & projects; IGOs; councils and think-tanks; NGOs; associations; media/news networks; foundations; publishers (non-science); libraries; museums; companies; agencies and consultancies. Any Twitter account that returned one of the keywords searched for in its Twitter description was included for further examination.

In the second approach, any Twitter account bearing the "verified"-tag was included. This labelling is provided by Twitter for "authentic" accounts "of public interest" and refers to accounts from, among others, "music, acting, fashion, government, politics, religion, journalism, media, sports, business" ("About verified accounts", n.d).

Remaining accounts by individuals were eliminated by searching user bios for terms typical of individuals (e.g. father, mother, lover, etc.); also, the set finally was cleared manually from any individual users. In total, this resulted in 6,958 unique Twitter accounts.

Identification of Fields of Interest

As a second step, we investigated the share of publications that had been tweeted by a social user per scientific field. For the delineation of scientific fields, we refer to the publication-level classification procedure described by Waltman and Van Eck (2012). Accordingly, we used the classification of publications in the Web of Science database into micro-fields, as stored in their local version at CWTS. This covers a total of 19,162,082 publications (with the additional restriction of having a DOI) that are linked to one of 4,047 micro-fields.

We counted the numbers of publications per field shared by any Twitter account, and by a social user. Then, we calculated the proportions of these numbers to the total number of publications tweeted by either any Twitter account, or a social user. Hereafter, we investigated micro-fields prominent among social users. Results are reported in the next section.

Results

Of the total set of publications, 3,139,052 publications were tweeted at least once by any Twitter account. Contrastingly, 132,848 publications were tweeted by at least one social user (with the set of social users being a subset of the total group of Twitter accounts).

A Spearman correlation test of the numbers of tweets from either all Twitter users or all social users per micro-field shows a significant, positive association (*rs* = .90). However, a few outliers could be detected in the plot of fields – leaning either towards the tweeter group of social users, or to the group of all Tweeters. This means that for certain micro-fields, the relation of tweeted publications per field to the total amount of publications tweeted deviates across user groups. Comparing the first 20 micro-fields showing the highest relations among social users with the equivalent set of the first 20 micro-fields from all Twitter users reveals ten micro-fields exclusively shared by the former. These are listed in Table 1.

Table 1. Micro-fields prominent among social users, compared to shares by all Twitter users.

Micro-field Id	Publications tweeted per field/All publications tweeted	Publications tweeted by social users/All publications tweeted by social users
299	0.18%	0.79%
1220	0.09%	0.78%
611	0.12%	0.68%
388	0.13%	0.53%
2040	0.03%	0.50%
1297	0.11%	0.41%
384	0.12%	0.40%
153	0.16%	0.38%
33	0.16%	0.38%
383	0.05%	0.33%

The micro-fields in Table 1 include labels as for example, *"relative age effect; home advantage; tennis; blood flow restriction; small sided game"* (Id 299); *"fibromyalgia; chronic fatigue syndrome; fatigue; chronic widespread pain; exercise"* (Id 1220); *"maternal mortality; pregnancy; stillbirth; mother; antenatal care"* (Id 611); *"dinosauria; reptilia; squamata; theropoda; afe"* (Id 388); *"htlv; human t cell leukemia virus type; patient; adult t cell leukemia lymphoma; tax"* (Id 2040); *"crispr; genome editing; crispr cas9 system; zinc finger nuclease; plant"* (Id 1297); or *"hospital cardiac arrest; cardiopulmonary resuscitation; therapeutic hypothermia; resuscitation; survival"* (Id 384).

Discussion

This study set out to identify topics occurring in scientific publications that are of special interest to users from society. Observing outliers in the sharing behaviour of such users on Twitter, compared to that of undelimited user attention has revealed certain research topics of potential, specialised attention. Those may indicate areas of research that are considerably relevant for certain societal actors.

The selection of social users as such is limited; hence, results may reflect a bias towards respective areas of activity and interest. Further research needs to expand this set of social users; also, more fine-grained insights into the differences between sub-groups are needed (e.g., differences in the sharing behaviours of NGOs and foundations).

Acknowledgments

This study is part of the KNOWMAK project which is funded by the European Union's Horizon 2020 programme under grant number 726992.

References

About verified accounts. (n.d.). Retrieved from: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Costas, R., Van Honk, J., & Franssen, T. (2017, December 15). Scholars on Twitter: who and how many are they? Conference Paper, *International Conference on Scientometrics and Informetrics,* 2017, Wuhan, China. Retrieved February 01, 2019, from the arXiv database.

Waltman, L., & Van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.

Assessing algorithmic paper level classifications of research areas: exploring existing human labeled datasets.

Alexis-Michel Mugabushaka[1]

[2] *Alexis-Michel.MUGABUSHAKA@ec.europa.eu*
European Research Council Executive Agency (ERCEA) COV 24/161, B-1049 Brussels,
Belgium

Introduction

Arguably most, if not all bibliometric analyses rely implicitly or explicitly on classifications of research fields. They are used as units if analysis in their own right (e.g. to track progress in science) or for normalization of bibliometric performance indicators.

The most commonly used classifications rely on classification of research journals (and conferences series).

Their limitations have led to efforts to create algorithmically classifications at the level of papers. We can distinguish between two types of algorithmic paper-level classifications.

On one hand, there are those which seek to build a relatively small numbers of "research areas" than can be used in bibliometric normalization as surrogates for journal level-classifications. They include:

- The Leiden Paper-Level Classification, created using direct citation graph with a community detection algorithms. This classification has about 4000 "micro-fields) (Waltman, L., et al. (2012) and Traag et.al (2018).
- Dimensions classifications which are created using machine learning approach to classify research papers to one or several fields of the Australian and New Zealand Standard Research Classification (ANZSRC). (Hook et. al 2018).

One the other hand, we have classifications aiming to provide a more granular level of research topics that papers deal with. Recent examples are:

- The Scopus Research Topics created using citation graphs. It has about 100.000 topics. Boyack, K. W., & Klavans, R. (2010).
- Microsoft Academic Graph which offer a hierarchical classifications of research fields generated by hierarchical topic modelling. It has about 200.000 fields. Shen et al (2018).

The boundaries between the two categories are blurred (especially in case of hierarchical classifications line the MAG one) but we argue that they serve different purposes.

All those classifications are gaining in popularity in bibliometric studies but relatively little efforts goes into their assessment. Bornmann (2018) undertook an exploratory assessment of the Dimensions classification using his own papers and uncovered serious reliability and validity issues.

More recently Waltman et. al (2019) discussed principled approaches to assess the accuracy of clustering of research papers. They proposed to use different (and independent) "relatedness measures" and compare them among others.

In this poster, we focus on the first categories of classifications and explore the use of existing, human labeled datasets to assess them.

Data, Methods and Assumptions

For this exploratory analysis, we use an openly accessible datasets of the Leiden research classifications[1] with about 15 Million papers assigned to one of 4047 "micro" research fields. Publications are uniquely identified by their Web of Science identifier (UT).

As a human curated classification of research papers, we use the publications submitted by their research organisations in the context of the UK Research Assessment Framework (REF)[2]. This exercise is organized in 36 so called "unit of assessment" which are effectively groupings of related research disciplines.

Because REF is an important mechanism in resources allocation great care is taken to curate data that institutions submit and we assume that papers submitted to one of those panels are representative of those fields and that the classification to the "right" unit of assessment can be trusted.

The focus here is to assess how algorithmic classifications (in this case the Leiden classification) perform, in comparisons with standard journal based classifications in "recovering" the research fields (operationalized here as units of the assessment in REF). We use the Web of Science Research subject categories fields and the Scopus All Science Classifications fields. After matching the datasets, we got 117.775 research papers assigned to all four classifications: (REF unit of assessment (36 categories), Leiden

[1] http://www.leidenranking.com/information/fields
[2] https://results.ref.ac.uk/ (Results & submissions > Download submission data)

Paper Level micro-field (3.546 in the dataset), Web of Science (151 categories in the dataset) and AJSC (319 fields in the dataset). The smallest class is Classics with 118 papers and the largest Clinical Medicine with over 11.000 papers.

In assessing the classifications (against the REF categories), an important assumption is that research classification, at any level, should be able to group publications in a more general but distinctive categorisations which are hand-curated and widely recognized. The REF classification qualifies as such.

To assess this performance we measure the extent to which papers in a given cluster of the Leiden classification are concentrated or spread over the REF categories (using Gini Coefficient and Herfindahl–Hirschman Index). The intuition behind this measure is that well-constructed cluster fall in fewer REF categories. This approach has been used in similar context by Boyack et al. (2011).

Results and discussions

The figures 1 and 2 show that clusters of publications in the Leiden classification fall in fewer categories of REF categories than those from the journal based classifications.

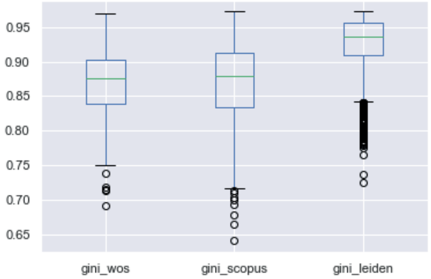


Figure 1. Gini Coefficient

It is fair to assume that this higher performance is due – among others – to their ability to disentangle not only papers from multidisciplinary journals (which are more likely also to be submitted to REF) but also from other discipline-specific but still general journals.

Outlook

The poster presents first results of exploring the use of hand curated classifications to assess the quality of algorithmic classifications. In future work we plan to explore other datasets for benchmarking, test other performance measures and compare different algorithmic classifications.

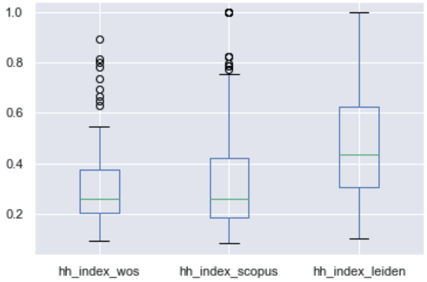


Figure 2. HH-index .

Acknowledgments

I thank Dr Joao Martins and Dr Diego Chialva (both at ERCEA) for fruitful discussions.

References

Bornmann, L. (2018). Field classification of publications in Dimensions: a first case study testing its reliability and validity. Scientometrics, 117(1), 637-640.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?. JASIST, 61(12), 2389-2404.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., ... & Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. PloS one, 6(3), e18029.

Hook, D., Porter, S., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. Frontiers in Research Metrics and Analytics, 3, 23.

Shen, Z., Ma, H., & Wang, K. (2018). A Web-scale system for scientific knowledge exploration. arXiv:1805.12216.

Traag, V., Waltman, L., & van Eck, N. J. (2018). From Louvain to Leiden: guaranteeing well-connected communities. arXiv:1810.08473.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science., 63(12), 2378-2392.

Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2019). A principled methodology for comparing relatedness measures for clustering publications. arXiv:1901.06815.

Financial Market Forecasting using Online Information: Research Stream Analysis based on Citation Network

Chaoqun Wang[1], Zhongyi Hu[1], Raymond Chiong[2], Ke Dong[1]

[1] *chaoqunwang@whu.edu.cn; zhongyi.hu@whu.edu.cn; dk8047@163.com*
School of Information Mangement, Wuhan University,
Wuhan 430072, China

[2] *raymond.chiong@newcastle.edu.au*
School of Electrical Engineering and Computing, The University of Newcastle,
Callaghan, NSW 2308, Australia

Introduction

Accurate financial market prediction is one of the most challenging tasks in financial market research. Recently, terabytes of digital information related to financial markets and individuals have been produced via online news platforms, forums, blogs, social media, and so on. Instead of focusing solely on developing sophisticated forecasting models based on technical analysis or/and fundamental analysis, researchers have moved their attention to making full use of the large amount of online information.

Although the literature related to financial prediction using online information is burgeoning, there is limited understanding on the research stream of financial prediction using online information. In this paper, we mainly present a citation analysis to provide new insights from a meta-perspective to help better understand the most influential publications and research streams of financial market forecasting using online information. To the best of our knowledge, this is the first bibliometric study that examines existing studies on financial market forecasting using online information.

Methodology

In this study, we carried out bibliometric analysis to discuss the literature related to financial market forecasting using online information. Specifically, two leading databases including Web of Science^{TM} (WoS) Core Collection and Scopus were used to collect as many publications in the focused area as possible and initially 238 and 429 articles were collected from WoS and Scopus, respectively. By eliminating duplications and irrelevant papers, we got a total of 441 papers eventually.

Results and discussions

Top ten influential papers

The total number of citations of a paper in the collected references, known as the local citation score (LCS), is used to evaluate the paper's impact. Based on LCS, the top 10 frequently cited articles are summarized in Table 1.

The most influential article is that by Bollen, Mao et al. (2011), having 120 local citations from 441 collected articles. This paper has successfully applied the public mood extracted from Twitter to improve the prediction of DJIA. Both the second and fourth articles focused on investigating whether Internet message boards can be applied for stock market prediction. The third influential article by Da, Engelberg et al. (2011) proposed a measure of investor attention using search frequency in Google and showed that this measure is correlated with prices of Russell 3000 stocks. The other papers mainly focused on investigating online news for stock price forecasting, indicating that financial news and stock prices have both attracted massive attention in this area.

Research stream clustering based on co-citation network

To find the relationships between these articles and identify key research streams, we clustered the articles' co-citation network using CiteSpace. Based on clustering, the co-citation network is divided into 58 clusters, and 17 of them have more than 10 members. Among them, the top 10 clusters with their size, silhouette, and label are shown in Table 2. The size denotes the number of articles in a cluster. Silhouette is a measure of how similar an object is to its cluster (cohesion) compared to other clusters (separation). Silhouette values range from -1 to 1, where a high value means that the object is well matched to its cluster but poorly matched to neighboring clusters. Label (LLR) in Table 2 represents the name of a cluster and is automatically labelled through the Log-Likelihood Ratio (LLR) in CiteSpace.

As shown in Table 2, the biggest cluster has 44 members and a silhouette value of 0.99, with the theme labelled as "*internet stock message board*". The most active citer of the cluster is by Antweiler and Frank (2004). The second largest cluster has been labelled as "*social network; financial analysis*". The most active citer is the article by Da, Engelberg et al. (2011). They used investor attention for financial market prediction. Except for news and social media, online search volumes (e.g., Google Trends) have also attracted much attention (Smith 2012, Heiberger 2015). The 8th largest cluster, labelled as "*cross correlation; google trend*", shows such a research direction. The most active citer of this cluster is by Heiberger (2015), and they applied Google query volumes to study the relationship between mass online behaviour and stock market movement.

Table 2. Top 10 clusters

Cluster ID	Size	Silhouette	Label (LLR)
0	44	0.99	internet stock message board; social network
1	41	0.549	social network; financial analysis
2	35	0.779	vector autoregression
3	32	0.763	automatic sentiment analysis
4	27	0.954	prototypical analysis; social network
5	26	0.932	textual information
6	26	0.829	forex rate prediction
7	24	0.811	cross correlation; google trend
8	22	0.997	rough set approach; macroeconomy
9	21	0.924	financial; data mining

Conclusion

In this paper, we conducted a bibliometric analysis to comprehensively investigate the research stream of literature related to financial market forecasting using online information. The 10 seminal articles based on local citations were listed, with Bollen, Mao et al. (2011) as the most influential article. Based on a co-citation network, 10 streams of research mainly focused on stock market forecasting based on online board and social media. Online search volumes (e.g., Google Trends) have also attracted much attention.

References

Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance*, 59(3), 1259-1294.

Bollen, J., H. Mao and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.

Da, Z., J. Engelberg and P. Gao (2011). In Search of Attention. *Journal of Finance*, 66(5), 1461-1499.

Engelberg, J. E. and C. A. Parsons (2011). The causal impact of media in financial markets. *The Journal of Finance*, 66(1), 67-97.

Heiberger, R. H. (2015). Collective attention and stock prices: Evidence from google trends data on standard and poor's 100. *PLoS ONE*, 10(8).

Smith, G. P. (2012). Google Internet search activity and volatility prediction in the market for foreign currency. *Finance Research Letters*, 9(2), 103-110.

Table 1 Top 10 seminal articles based on the LCS.

No.	Authors	Title	LCS	(%)
1	Bollen, J; Mao, HN; et al.	Twitter mood predicts the stock market	120	27.2
2	Antweiler W.; Frank M.Z.	Is all that talk just noise? The information content of Internet stock message boards	49	11.1
3	Da, Z; Engelberg, J; et al.	In Search of Attention	32	7.3
4	Tumarkin, R; Whitelaw, RF	News or noise? Internet postings and stock prices	24	5.4
5	Schumaker, RP; Chen, HC	Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System	21	4.8
6	Schumaker, RP; Chen, HC	A quantitative stock prediction system based on financial news	20	4.5
7	Hagenau, M; Liebmann, M; et al.	Automated news reading: Stock price prediction based on financial news using context-capturing features	19	4.3
8	Nassirtoussi, AK; Aghabozorgi, S; et al.	Text mining for market prediction: A systematic review	18	4.1
9	Schumaker, RP; Zhang, YL; et al.	Evaluating sentiment in financial news articles	16	3.6
10	Fung, GPC; Yu, JX; et al.	Stock prediction: Integrating text mining approach using real-time news	14	3.2

(%): percentage of local citations = LCS/the number of paper

The compound F^2-index as extension of the f^2-index in a dynamic perspective: An application in Corporate Governance research

Yves Fassin [1]

[1] *Yves.Fassin@ugent.be*
Ghent University, Department for Marketing, Innovation and Organisation, Tweekerkenstraat 2, B 9000 Gent (Belgium)

Introduction

The study of the status and evolution of a field focus on the main research question of exploratory nature: who are the most influential authors in the field? The study adopts an innovative methodology following the ghent-rating and f^2-index, based on categorization of articles in function of citations and positioning on the citation distribution.

Bibliometric studies on Corporate Governance

From an academic perspective, the research agenda has evolved in the last decades to develop corporate governance as a real discipline. Durisin and Puzone (2009), provide a first bibliometric study in the main specialized journal *Corporate Governance An International Review* (CGIR). They conclude to the maturation of a specific field with its distinct subfields.

The original search query for "corporate governance" in the Web of Science, restricted to the science categories business, management and ethics, leads to a selection of 3686 articles written by 1235 different authors; together 43351 citations (average is 11.8 citations per article).
Bibliometric studies traditionally start from two predominant methods: the list of the most productive authors in a field and the most influential authors.

Table 1. - Productivity

Researcher	n
FILATOTCHE	38
ZATTONI	38
AGUILERA	30
WRIGHT M	26
KUMAR P	26
JUDGE WQ	26
WESTPHAL	25
PENG MW	21
YOSHIKAWA	20

Table 2. Most - cited Authors

Researcher	tot cit
PENG MW	3271
WRIGHT M	3140
AGUILERA	3062
HOSKISSON	2996
WESTPHAL	2905
FILATOTCHE	2724
ZAJAC	2293
DALTON DR	1847
DAILY CM	1594

Table 3 h-index ranking

Researcher	h
FILATOTCHEY	26
WESTPHAL	19
AGUILERA	18
PENG MW	17
WRIGHT M	16
ZATTONI A	15
DALTON DR	11
YOSHIKAWA	11
HOSKISSON	10

The tables illustrate some similarities but also huge differences between those rankings and also between the h-index ranking.

The *gh*-rating anfd f^2-index methodology

I apply a recent methodology based on the ghent-rating that categorizes all articles in categories in function of their citation distribution (Fassin, 2018). The several grades (from AAA to E) depend from their relative position in the distribution of citations: all articles situated in the top 50 % (D), top 25 % (C), top 10 % (B), top 5 % (BB), the g-core (BA), the h-core (A) and the h^2-core (AAA) of the dataset.

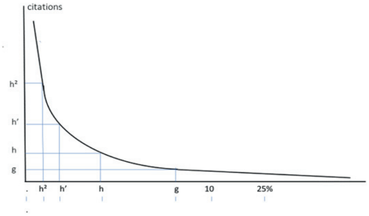


Figure 1: **Citation distribution curve – *gh*-rating (Fassin, 2018)**

Table 4 shows the number of articles in each category for the corporate governance field, and for the recent years of the field (last 5 years 2013-2017).

Table 4. Publications in corporate governance

Category	AAA	AA	A	BA	BB	B	C	Z
n	h^2	h/2	h	g	5%	10%	25%	100%
CG field	16	42	83	208	184	369	922	3686
CG rec	7	14	27	74	86	172	430	1722

Table 5 provides the amount of citations for the successive percentiles for the corporate governance field and for the recent field.

Table 5. Thresholds per category

Category		AAA	AA	A	BA	BB	B	C	D
	max	h^2	h/2	h	g	5	10	25	50
CG field	896	264	141	83	45	49	28	11	2
CG rec	153	57	41	27	15	13	9	3	1

The author's fame-index or f^2-index (Fassin, 2018) is calculated as the weighted sum of the articles within the author's h^2-core with the weighting factor determined by the field percentile categories. Applied to a simplified categorization of a researcher's articles (into A, B, C and R categories), the f^2-index or fame-index is thus defined as

$$f^2 = 2a + b + c/2 + r/4 + 2aaa \qquad \text{with}$$

aaa the number of articles in the h^2-core of the dataset
a the number of articles in the h-core
b the number of articles in the 10%-decile
c the number of articles in the 25%-decile
r the number of articles not in the 25%-decile
all limited to the author's recent h^2-core.

Table 6. The f^2-index calculation and distribution.

Researcher	h2	n	25	10	h	h^2	f^2
PENG MW	9	21	17	11	8	3	23.5
WESTPHAL	10	25	21	17	8	1	21
AGUILERA	9	30	17	13	7	2	20.5
ZAJAC	9	11	9	9	8	1	19
FILATOTCHEV	9	38	30	15	7	1	18.5
HOSKISSON	8	13	9	8	4	2	17
WRIGHT M	8	26	18	8	4	2	16.5

The dynamism aspect and recent contribution

Most rankings are static; also the f^2- classification. In order to have a better view on the evolution, a similar analysis can be executed on the data of the recent years: the f^{2}'-index is calculated on the basis of the publications in the last (full) 5 years.

$$f^{2}{}' = a' + b' \quad \text{within } h^{2}{}'\text{-core}$$

with a' the number of articles in the recent h-core and b' the number of articles in the 10%-decile, limited to the author's recent h^{2}'-core. This f^{2}'-classification in table 7 includes the active scholars and identifies and the newcomers in the field.

Table 7.Distribution and recent f^{2}'-index (5 years)

Researcher	h^{2}'	n'	10%'	h'	f^{2}'
PENG MW	5	8	8	3	8
AGUILERA	4	13	7	4	8
FILATOTCHEV	4	9	5	3	7
VAN ESSEN	4	6	4	2	6
ZATTONI	4	27	6	1	5

The combined F^2-index

In order to present a more dynamic description, we introduce a compound F^2-index, calculated by the sum of the overall f^2-index, the 5 years f^{2}'-index and the number of highly cited papers, within certain limits (increase limited to 50% if $f^2 > 6$)..

Table 8. f^2-index and combined F^2-index

Researcher	f^2	f^{2}'	HCF	F^2
PENG MW	23.5	8	4	35.25
AGUILERA	20.5	8	4	30.75
FILATOTCHEV	18.5	7	2	27.5
WESTPHAL	21	3		24
ZAJAC	19	2.5		21.5

The comparison between the f^2 and the compound F^2-classifications (table 9) shows the progress of the individual researchers compared to their peers. The data reveal the future trends.

Table 9. Ranking with various criteria

Researcher	r cit	r n	r h	r f^{3}'	r f^2	r F^2
PENG MW	1	8	4	1	1	1
AGUILERA	3	3	3	1	**3**	**2**
FILATOTCHEV	6	1	1	3	**5**	**3**
WESTPHAL	5	7	2	20	2	4
ZAJAC	7	20	11	50	4	5
HOSKISSON	4	13	9	20	6	6
WRIGHT M	2	4	5	T250	7	7
ZATTONI	22	2	6	5	**15**	**14**
VAN ESSEN	T75	22	19	4	**27**	**19**

Contribution to bibliometrics

The implementation of the f^2-index to this particular field illustrates the prudent selectivity and discriminative power of the method. The classification following the f^2-index mitigates between the classical ranking (productivity, citations, h-index)

The extension of the f^2-index to the compound F^2-index presents an innovate tool to examine dynamism in citation analysis. Especially, the difference in ranking between the F^2 The classification on the basis of the compound F^2-index can be applied to estimate the future trends and to identify the rising stars.

References

Durisin, B. & Puzone, F., 2009. Maturation of Corporate Governance Research, 1993–2007: An Assessment. *Corporate Governance: An International Review*, 17, 266-291

Fassin, Y. (2018). A new qualitative rating system for scientific publications and a fame Index for academics. *Journal of the Association for Information Science and Technology*, 69(11):1396–1399.

Hirsch, J.E. 2005. 'An index to quantify an individual's scientific research output'. *Proceedings of the National Academy of Sciences*, 102:16569-16572.

Kosmulski, M. 2006. 'A new Hirsch-type index saves time and works equally well as the original h-index'. ISSI Newsletter, 2(3): 4-6.

Author Index

Printed in Italy, August 2019