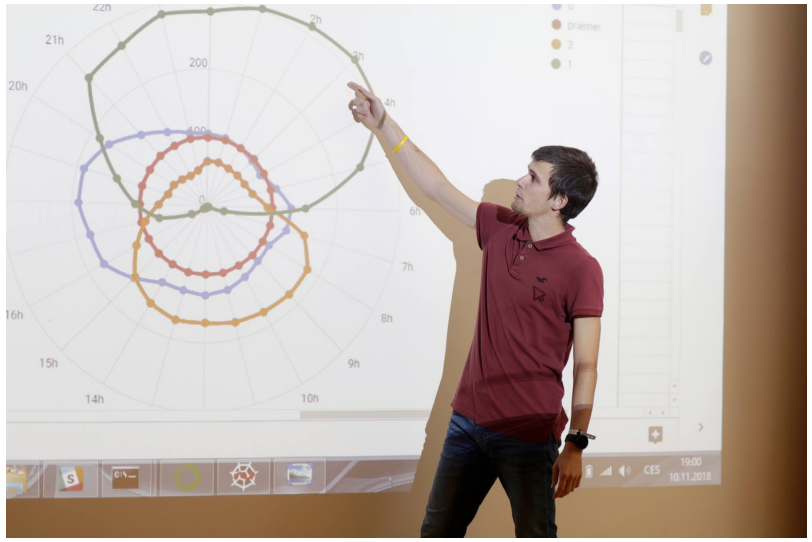
DATA SCIENCE IS A STATE OF MIND

MY BACKGROUND

- Masters in Applied Mathematics
- Former Data Science consultant at Teradata, Austria
- Co-founder of Knoyd, BaseCamp.ai, and Data Shift



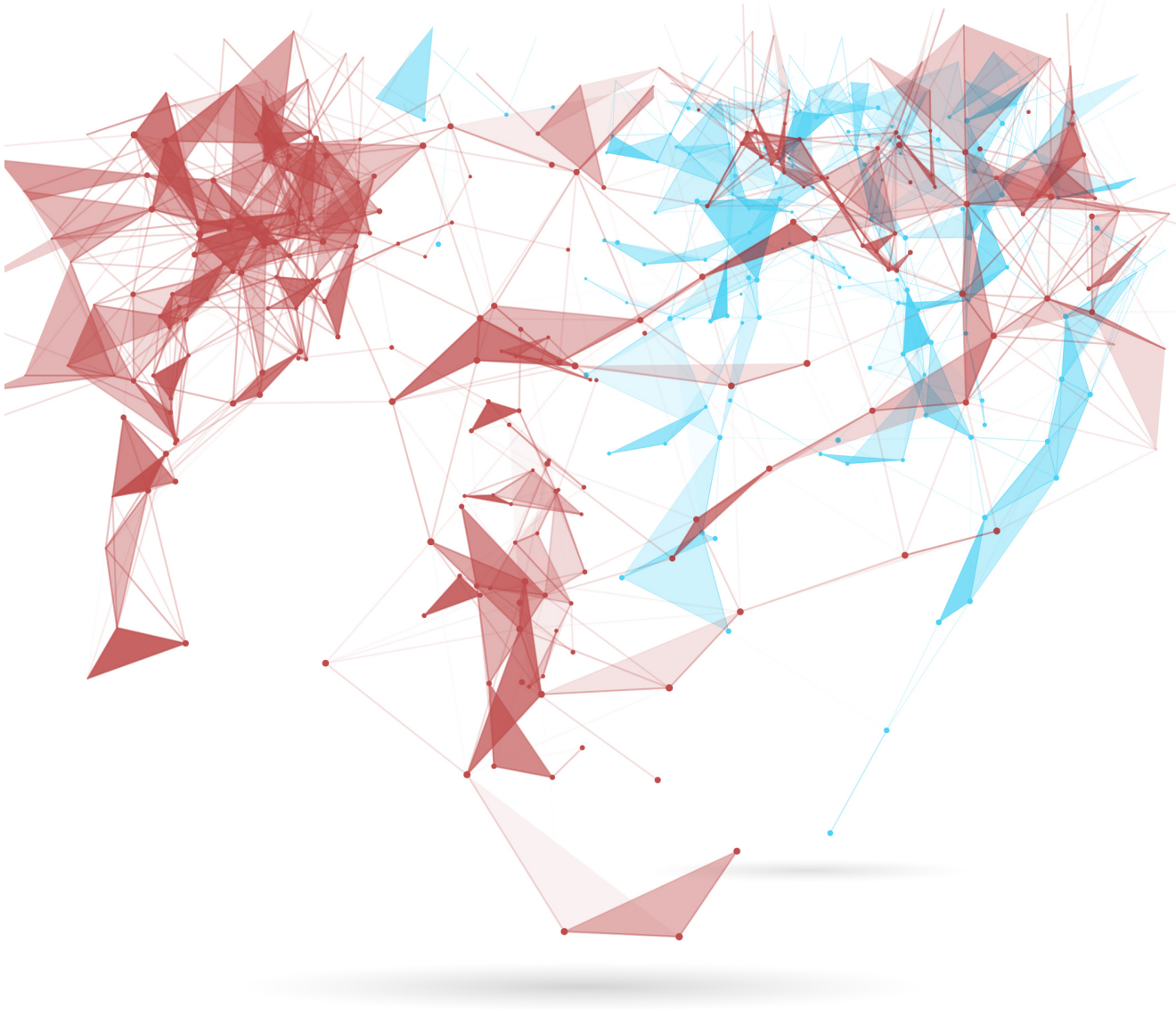
KNOYD

PROTOTYPING

INTEGRATION

STRATEGY

Prezly UniCredit poštová banka 365 SOLVED. Ulmon finAccel



CONTENTS

Google Trends | Compare

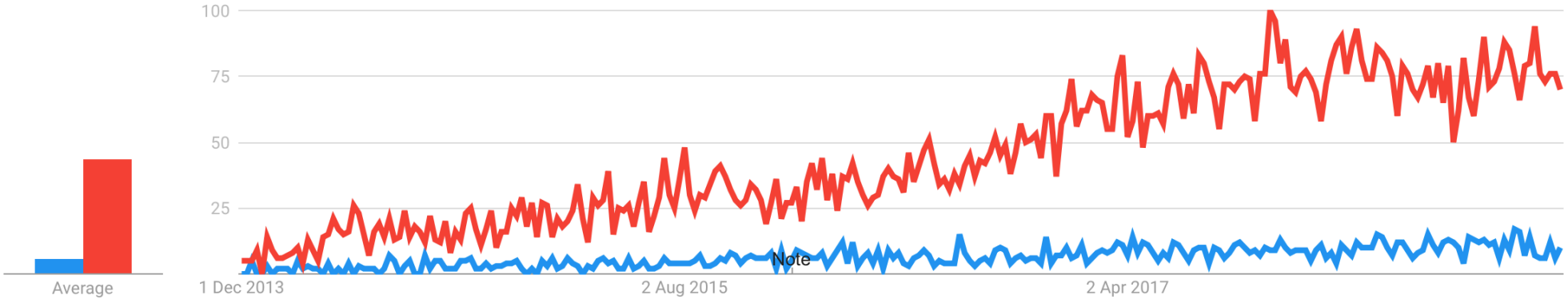
sas data science
Search term

R data science
Search term

+ Add comparison

Worldwide | Past 5 years | All categories | Web Search

Interest over time



Google Trends | Compare

sas data science
Search term

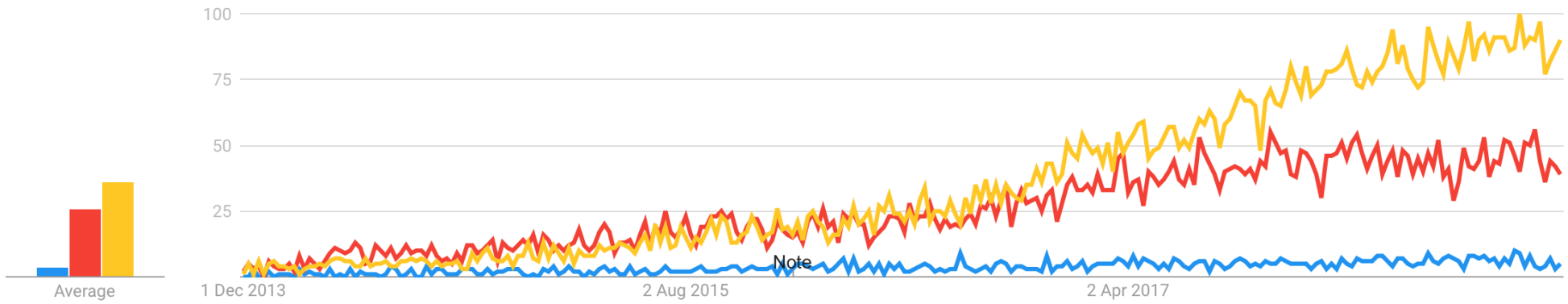
R data science
Search term

Python data science
Search term

+ Add comparison

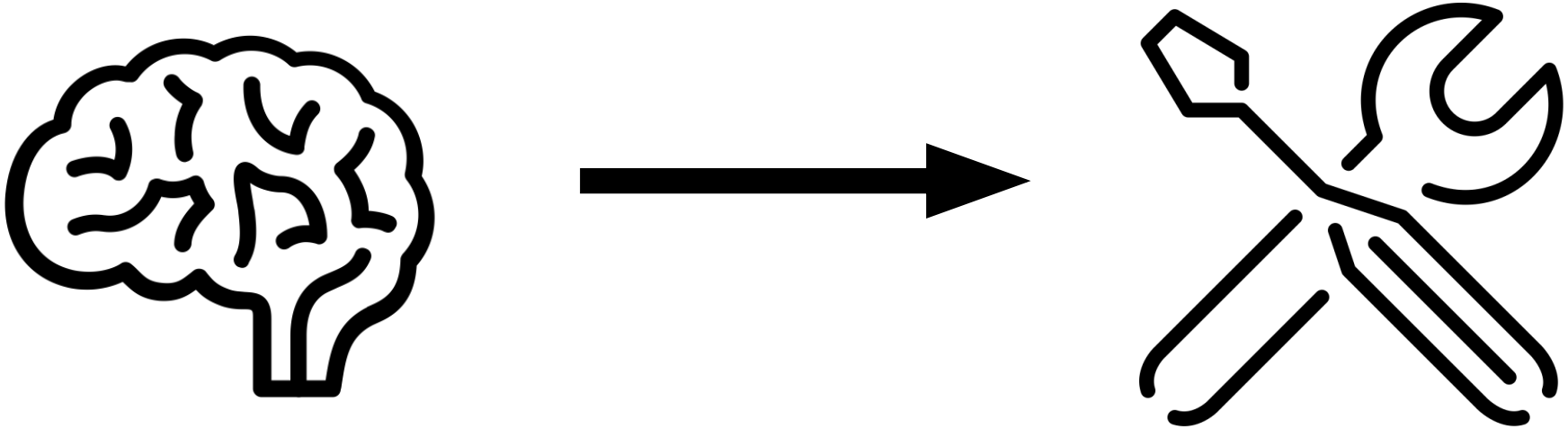
Worldwide | Past 5 years | All categories | Web Search

Interest over time

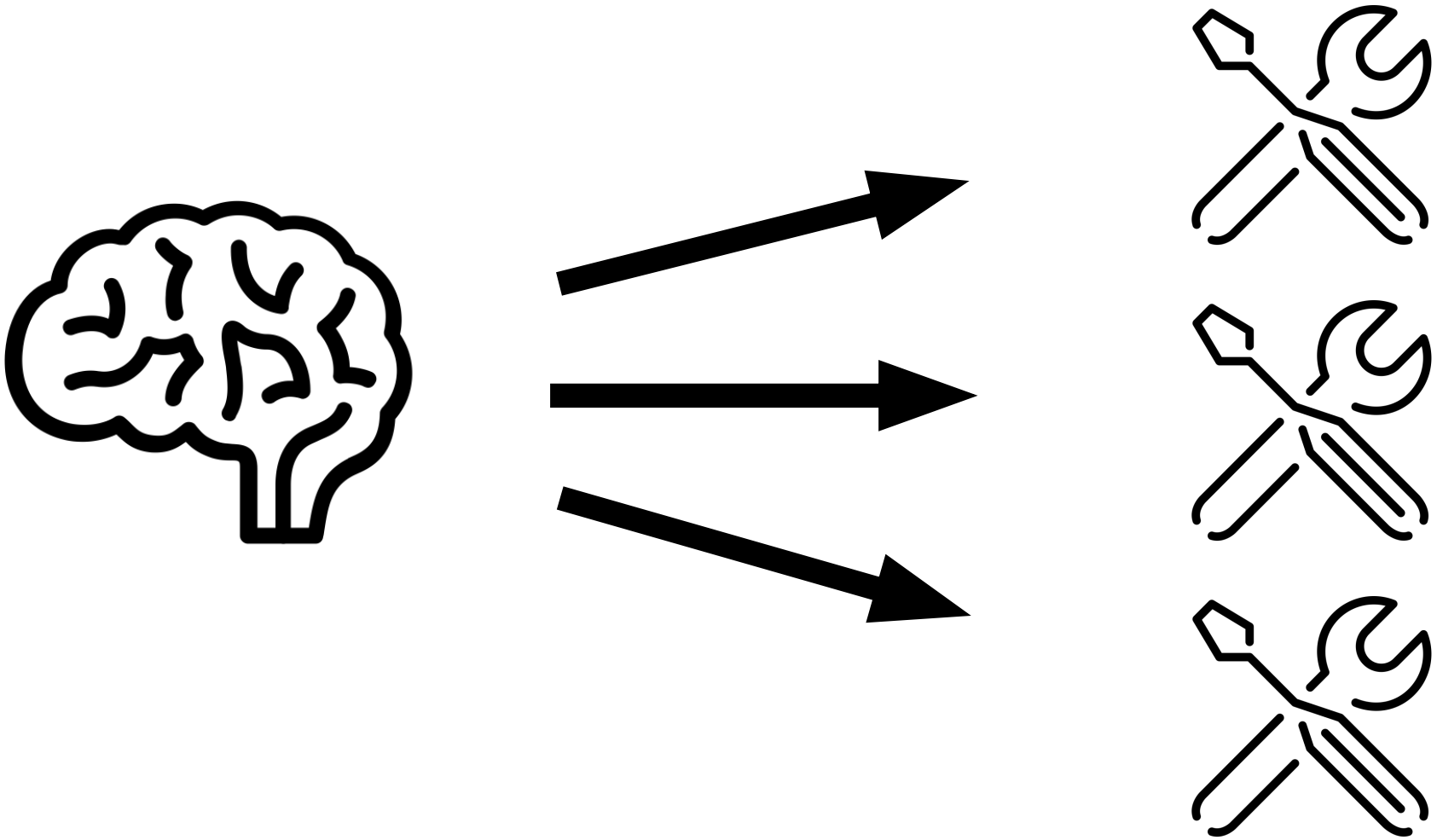


EVOLUTION OF PRODUCTIVITY

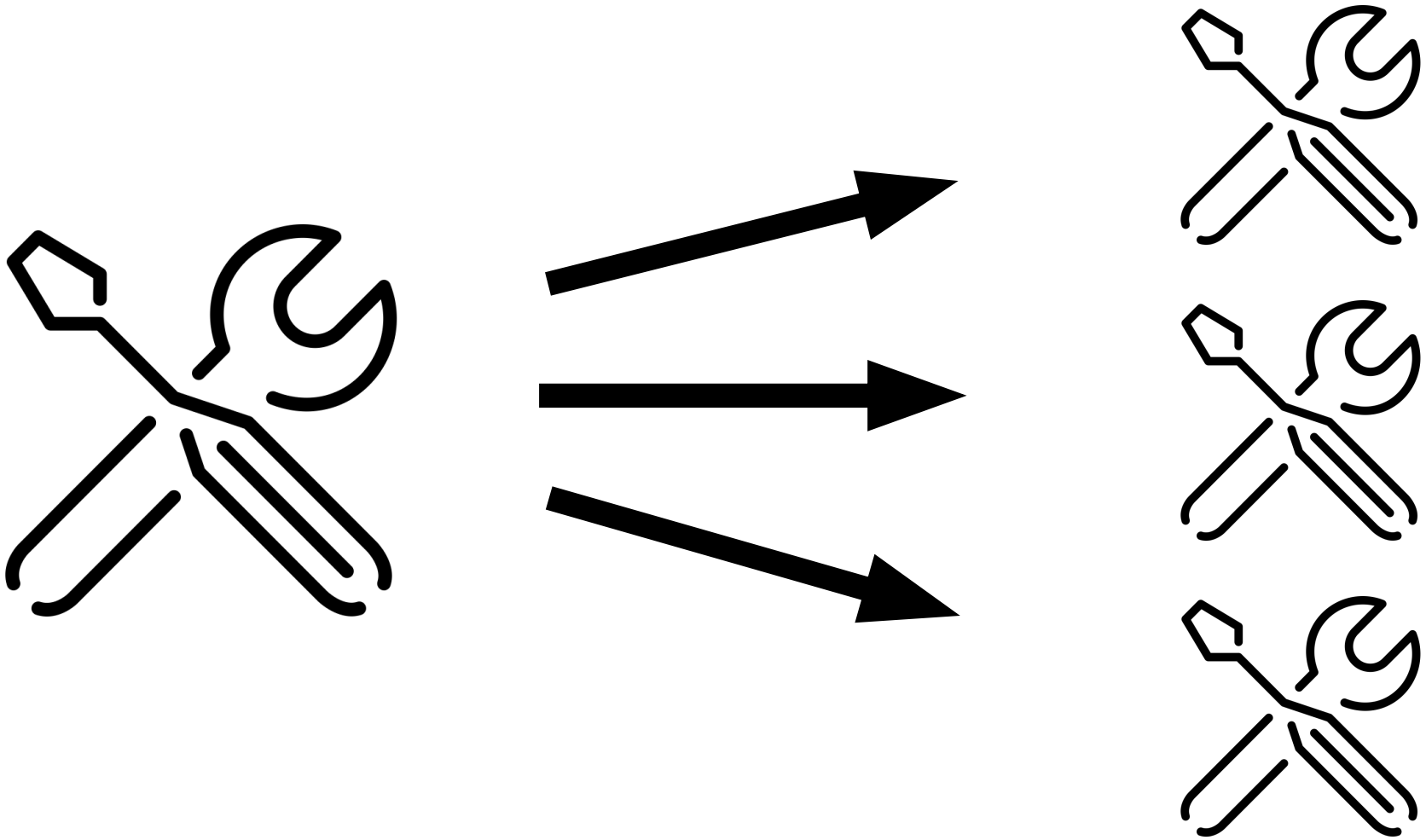
MAN MAKES TOOLS



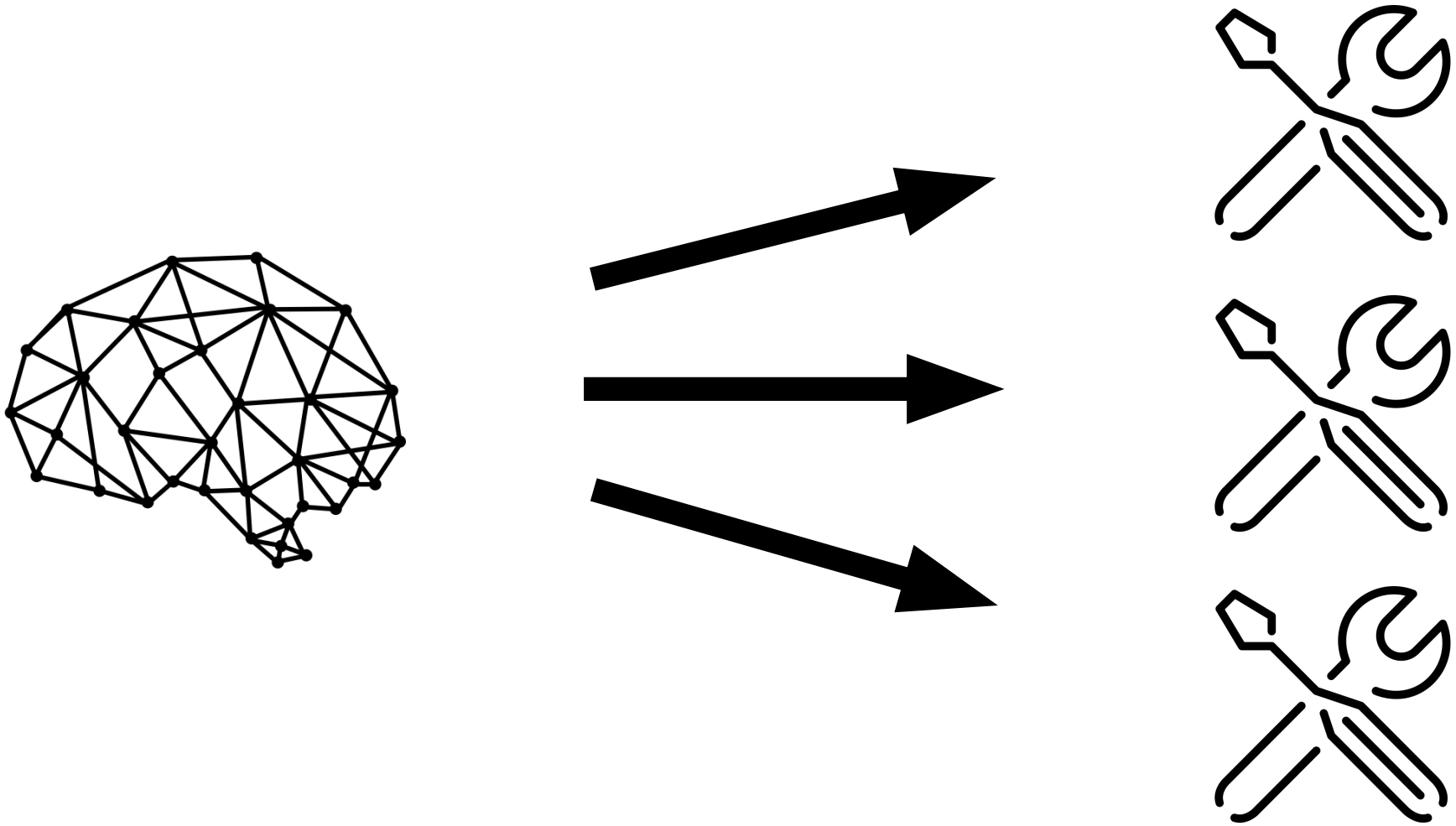
MAN MAKES MORE TOOLS



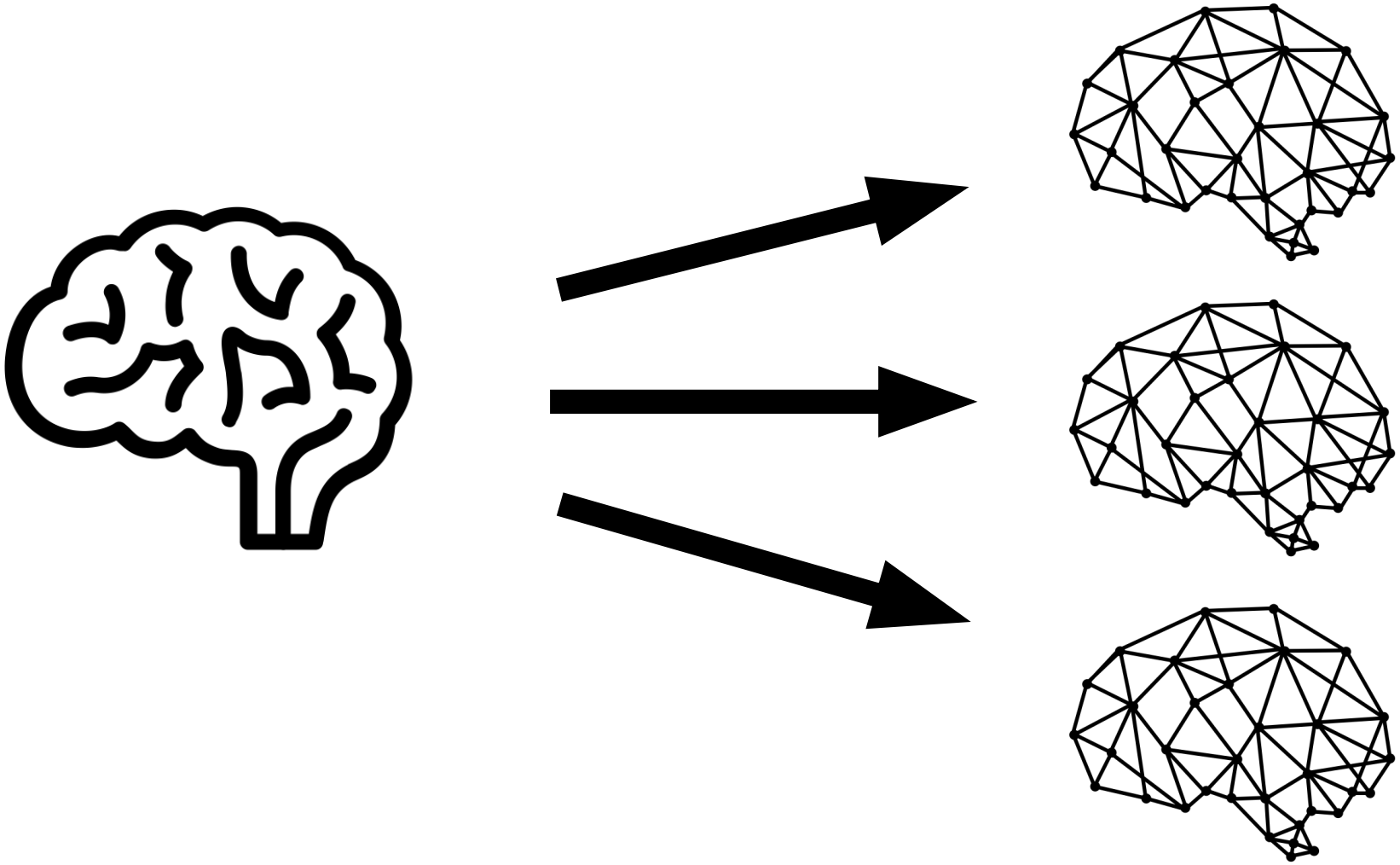
TOOLS MAKE TOOLS



MACHINE MAKES TOOLS

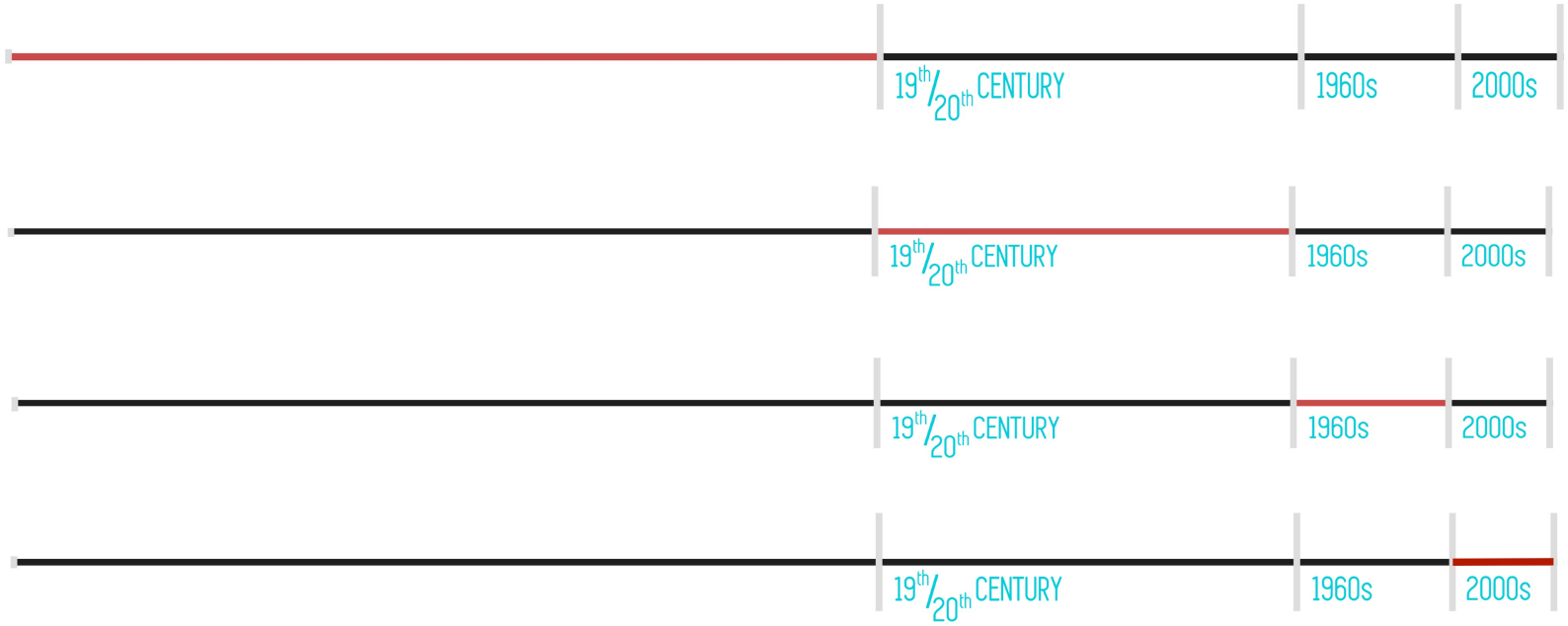


MAN MAKES MEN



WHY SHOULD YOU CARE?

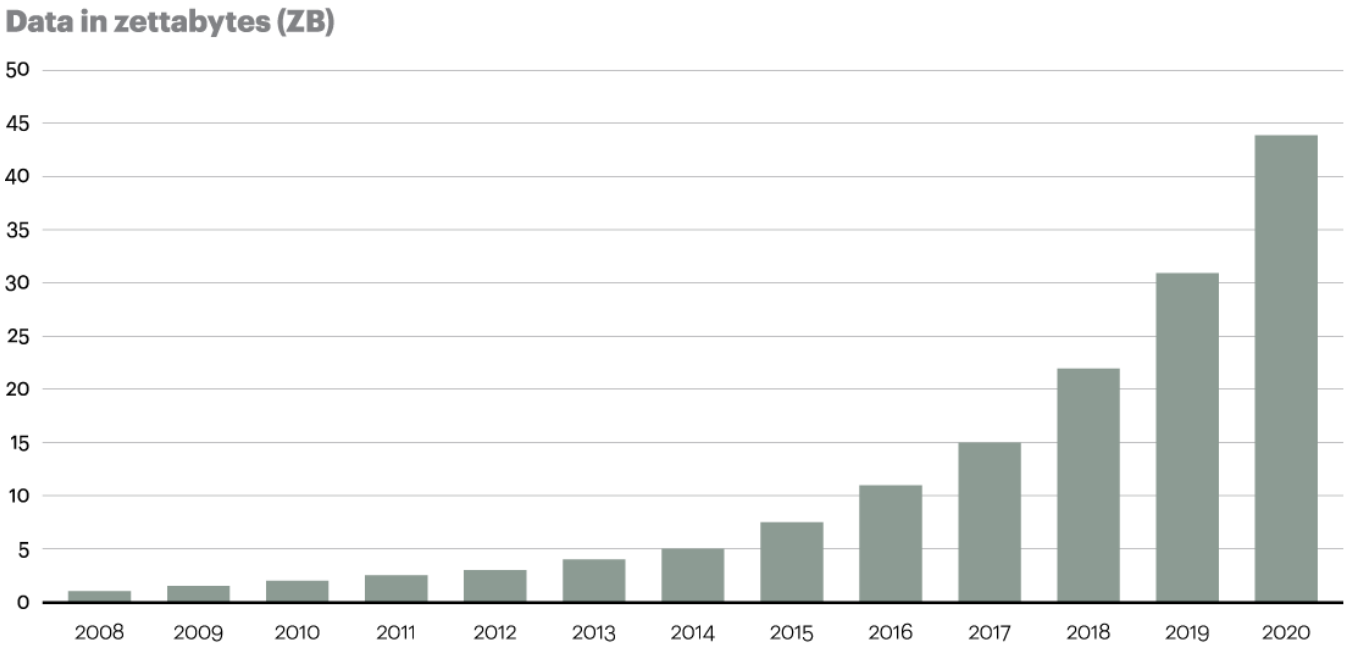
SPEED OF DEVELOPMENT



AMOUNT OF DATA

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020



Source: Oracle, 2012

Internet live stats

JOB MARKET

"… the sexiest job of 21st century."

Relative growth of # of Data Science jobs

DATA VS. NON-DATA MANAGMENT

MANUAL VS. AUTOMATION

INTUITION VS. VALIDATION

BBDO | Optimizely

REACTIVE VS. PROACTIVE

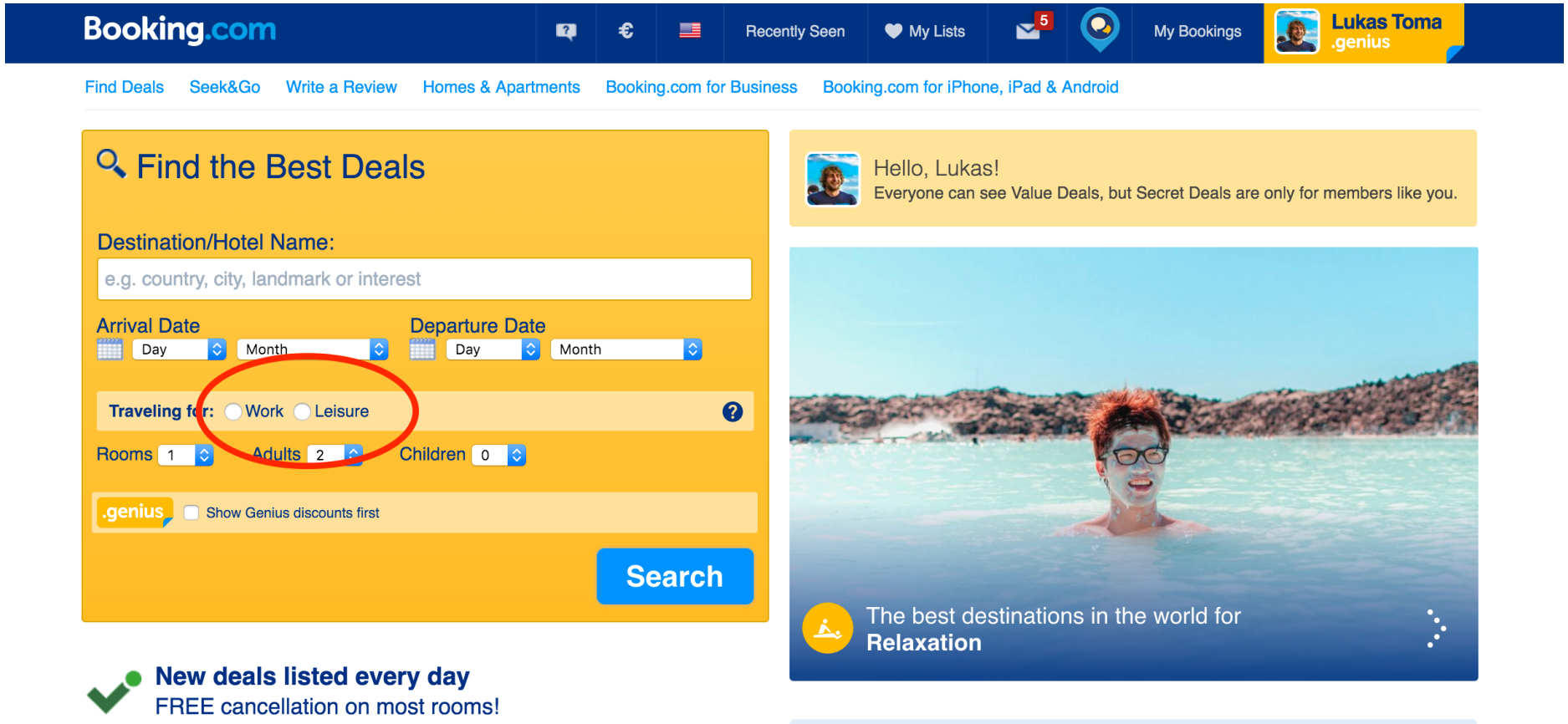
Slovenská elektrizačná prenosová sústava, a. s.

GRIDCURE

EXPERT-ONLY VS. AUGMENTED

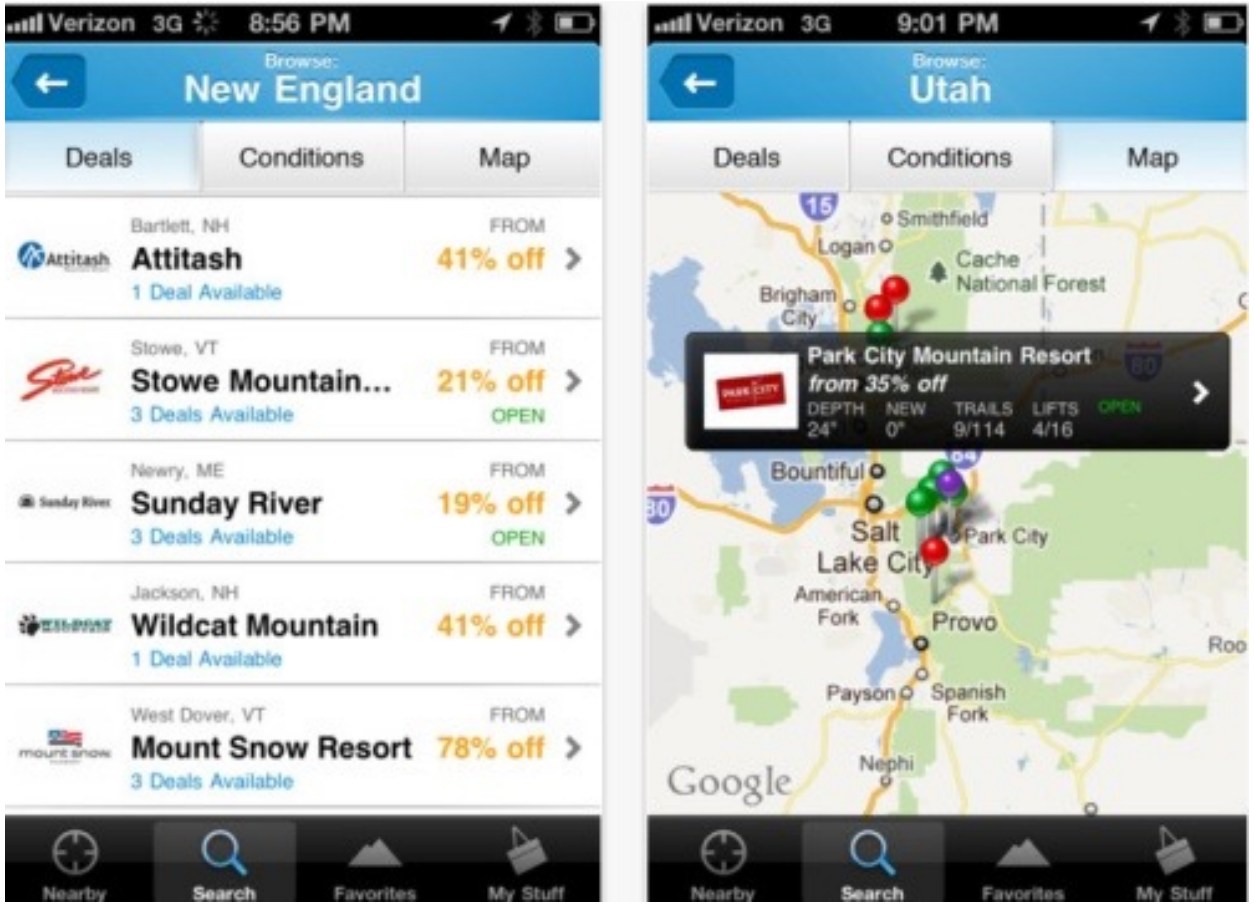
AREAS OF APPLICATION

Classification



Regression

liftopia



Source: http://www.datapine.com/blog/big-data-and-ski-resorts/

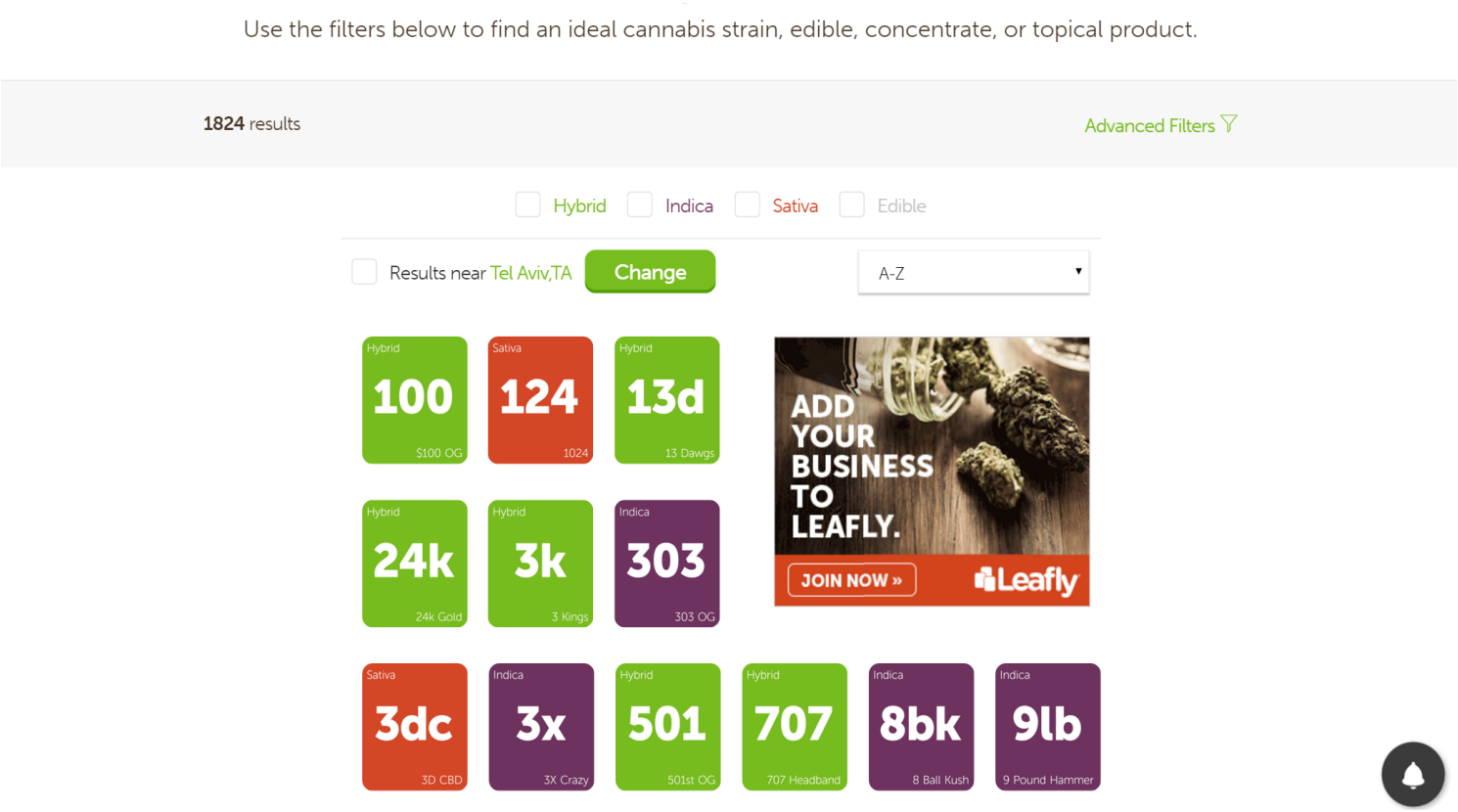
Clustering

Source: https://2.bp.blogspot.com/-KS6-whrl9pM/Vr3fT6e51qI/AAAAAAAABfA/rVvPkY-N_14/s1600/neurokiff-tablettte.jpg

Recommender Systems

Source: http://www.slideshare.net/Hadoop_Summit/shah-june27-425pmroom210av2

Recommender Systems



Source: Nina Rabinowitz, Marijuana: Turn another leaf

Social Network Analysis

Barack Obama campaign

Text Analytics

Source: https://hsto.org/getpro/habr/post_images/8e4/768/568/8e47685686001174765c7b00bb4b8b44.png

Text Analytics

AlchemyAPI™ different text analytics functions

Google reinventing human resources

Deep Learning

[Image captioning]

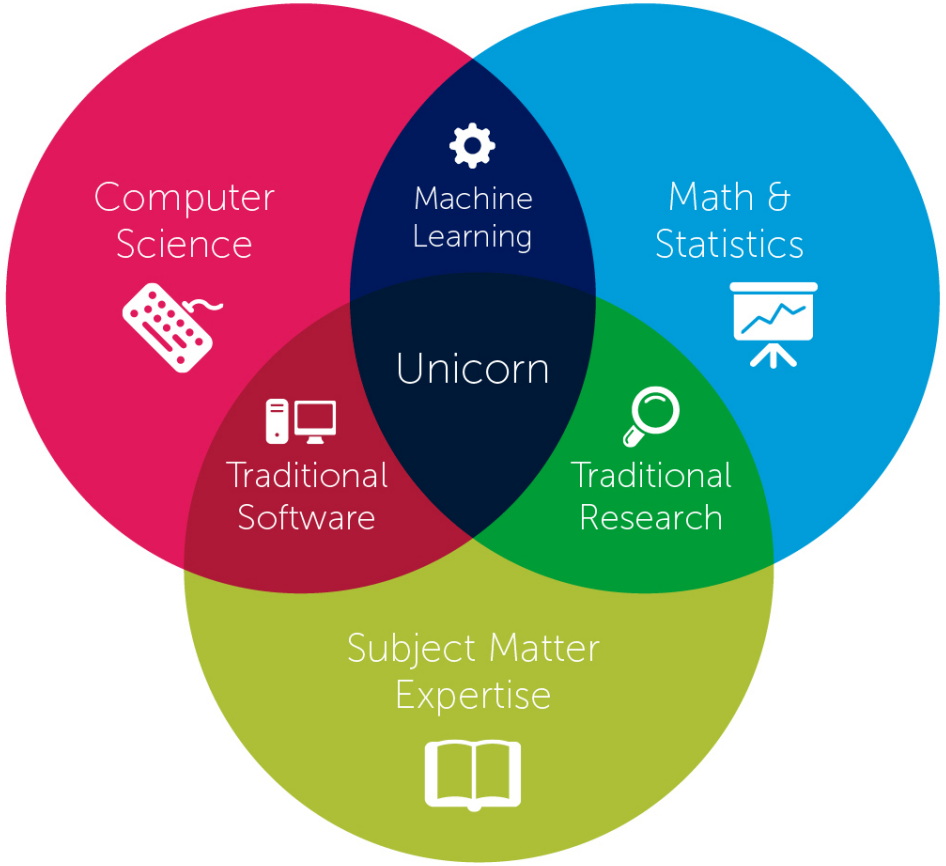
Data Visualization

Aging of the US population

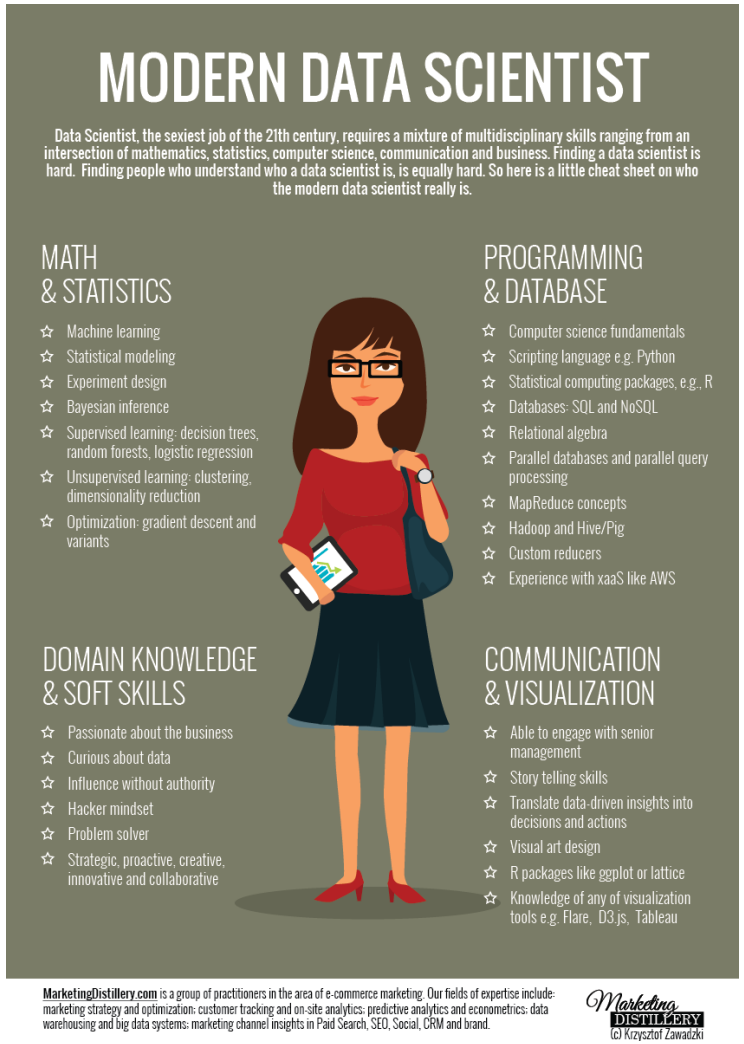
How parents' income predicts college chances of children

KINDS OF DATA SCIENCE

UNICORN DATA SCIENTIST



TYPES OF DATA SCIENTIST



- Research Scientist
- Data Engineer
- Visualization Expert
- Data Manager

Source: http://www.marketingdistillery.com/wp-content/uploads/2014/11/mds_f.png

SOCIAL NETWORK ANALYSIS

MU Kosice

November 2018

Section 1: Introduction

INTRODUCTION

- Immediately associated with Facebook, Linkedin or Twitter
- Not limited to these highly visible instances
- Example of less traditional social networks
    - The set of individuals with whom you make regular money transactions
    - The set of people with whom you are regularly on the phone
    - The set of people with whom you regularly exchange gifts (eshops)
    - The set of people with whom you share media content
- “Social network analysis is fast emerging as an important discipline for predicting and influencing consumer behavior” - Teradata magazine

INTRODUCTION - HISTORY

- The paper written by Leonhard Euler on the Seven Bridges of Königsberg and published in 1736 is regarded as the first paper in the history of graph theory
- SNA has origins in Graph theory and Social science
- Social network analysis (SNA) is the process of investigating social structures through the use of network and graph theories
- So when did Social Network theory and data analysis start?

INTRODUCTION - HISTORY

- Definitely, it has been around for a while, but the majority of social scientists cite Milgram's small world experiment.
- In the 60's (long before the Internet), Milgram asked people from the Boston area to send a letter to a person they didn't know. They had to route the message to a personal acquaintance that was more likely than the sender to know the target person. It turned out, that an average number of intermediaries was close to 6 "degrees of separation"((Backstrom et al., 2012)
- The 6 degrees of separation hypothesis was further supported in 2008 on Facebook data (Blackstrom et al., 2012). It was also tested on MSN messenger and the average path length turned to be 6.6 (Aggarwal et al. 2012). However, a more recent study in 2011 showed that Facebook now has 4 degrees of separation (Blackstrom et al., 2012).
- So our world becomes smaller as people become more connected!

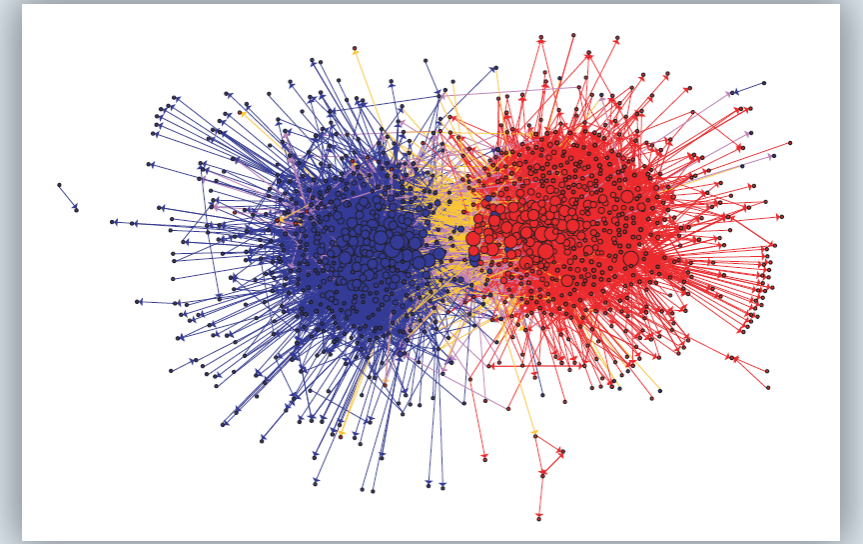
BACKGROUND - GRAPH ANALYSIS

- Graph theory provides a set of abstract concepts and methods for the analysis of graphs.
- G = (V, E) —> comprising a set V of vertices or nodes together with a set E of edges or lines
- V and E are usually taken to be finite
    - many of the well-known results are not true (or are rather different) for infinite graphs because many of the arguments fail in the infinite case.

BACKGROUND - SOCIAL SCIENCE

- Studying society from a network perspective is to study individuals as embedded in a network of relations and seek explanations for social behavior in the structure of these networks rather than in the individuals alone

A visualization of US bloggers shows clearly how they tend to link predominantly to blogs supporting the same party, forming two distinct clusters (Adamic and Glance, 2005)



Section 2: List of Use Cases

USE CASES

- SNA is used to identify criminal and terrorist networks and then identify key players in these networks
- Social Network Sites, like Facebook, use basic elements of SNA to identify and recommend potential friends
- SNA helps to reduce churn in telco companies. Myth about SNA: By saving the important people (influencers) you can save his circles as well.

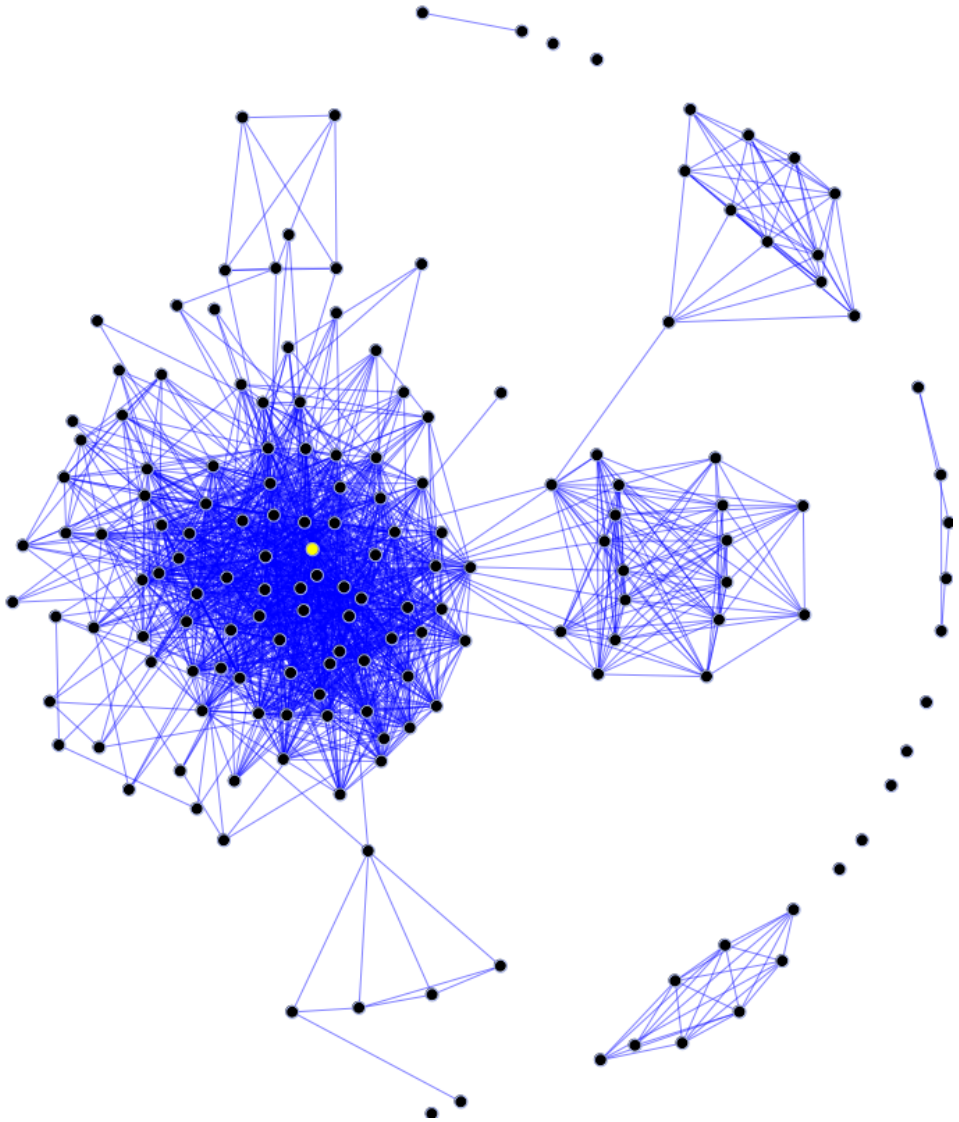
USE CASES

- Cross-sell and up-sell: Ability to contact influential people can provide additional profit opportunities across and beyond the entire contact circle.
- Computer Scientists use social network analysis to study webpages or Internet traffic
- In Life sciences is the use of network analysis to study food chains in different ecosystems

Section 3: Theory and Basic Concepts

TERMINOLOGY

- network = graph
- nodes, people = vertices
- links, connections = edges
- communities = clusters



BASIC CONCEPTS

Networks

Tie Strength

Key Players

Cohesion

How to represent various social networks

How to identify strong/weak ties in the network

How to identify key/central nodes in network

Measures of overall network structure

- Basic Concepts are based on presentation from Cheliotis, Giorgos - Social Network Analysis (SNA)

RELATIONS AS NETWORKS

Can we study their interactions as a network?

Communication

Anne: Jim, tell the Murrays they're invited

Jim: Mary, you and your dad should come for dinner!

Jim: Mr. Murray, you should both come for dinner

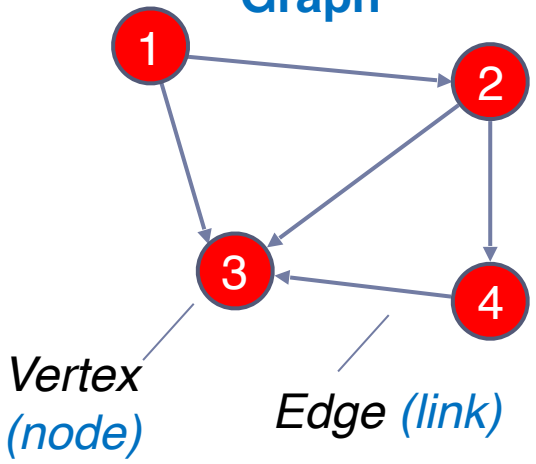
Anne: Mary, did Jim tell you about the dinner? You must come.

John: Mary, are you hungry?

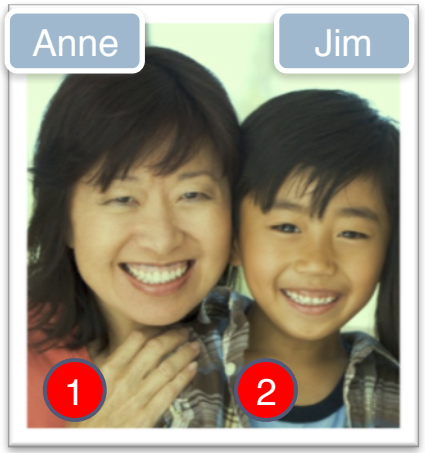
…

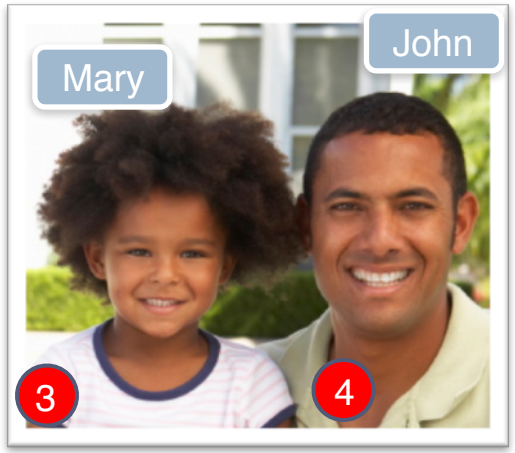
Who is who in the graph?

Graph



RELATIONS AS NETWORKS





Can we study their interactions as a network?

Communication

Anne: Jim, tell the Murrays they're invited

Jim: Mary, you and your dad should come for dinner!

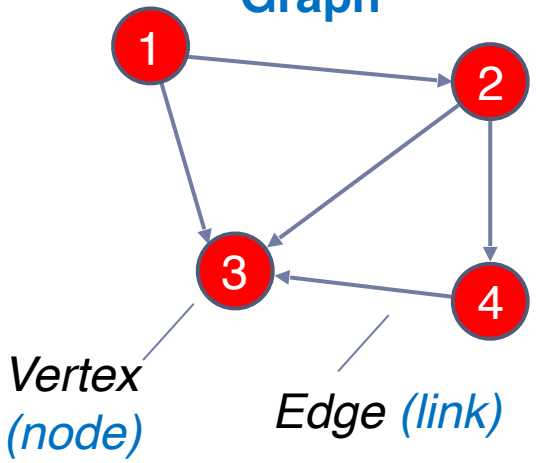
Jim: Mr. Murray, you should both come for dinner

Anne: Mary, did Jim tell you about the dinner? You must come.

John: Mary, are you hungry?

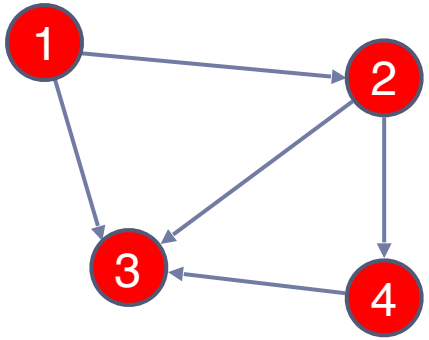
…

Graph



DIRECTED GRAPH - SIMPLE EXAMPLE

Directed Graph



Edge List

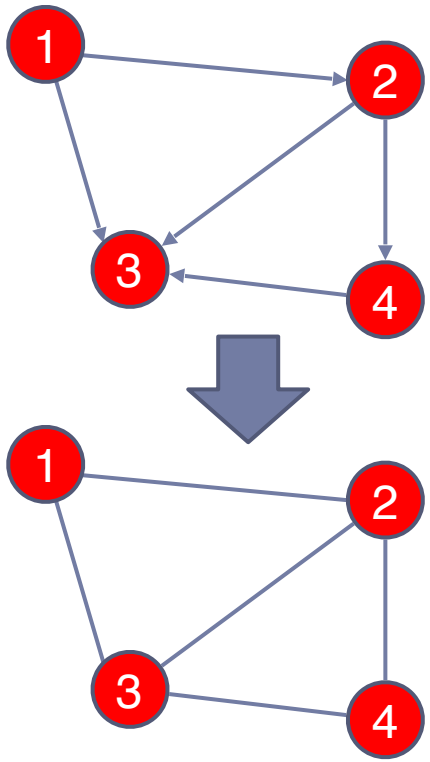
Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

Adjacency matrix

Vertex	1	2	3	4
1	-	1	1	0
2	0	-	1	1
3	0	0	-	0
4	0	0	1	-

UNDIRECTED GRAPH - SIMPLE EXAMPLE

Directed Graph
(who contacts whom)



Undirected Graph
(who knows whom)

Edge list remain the same

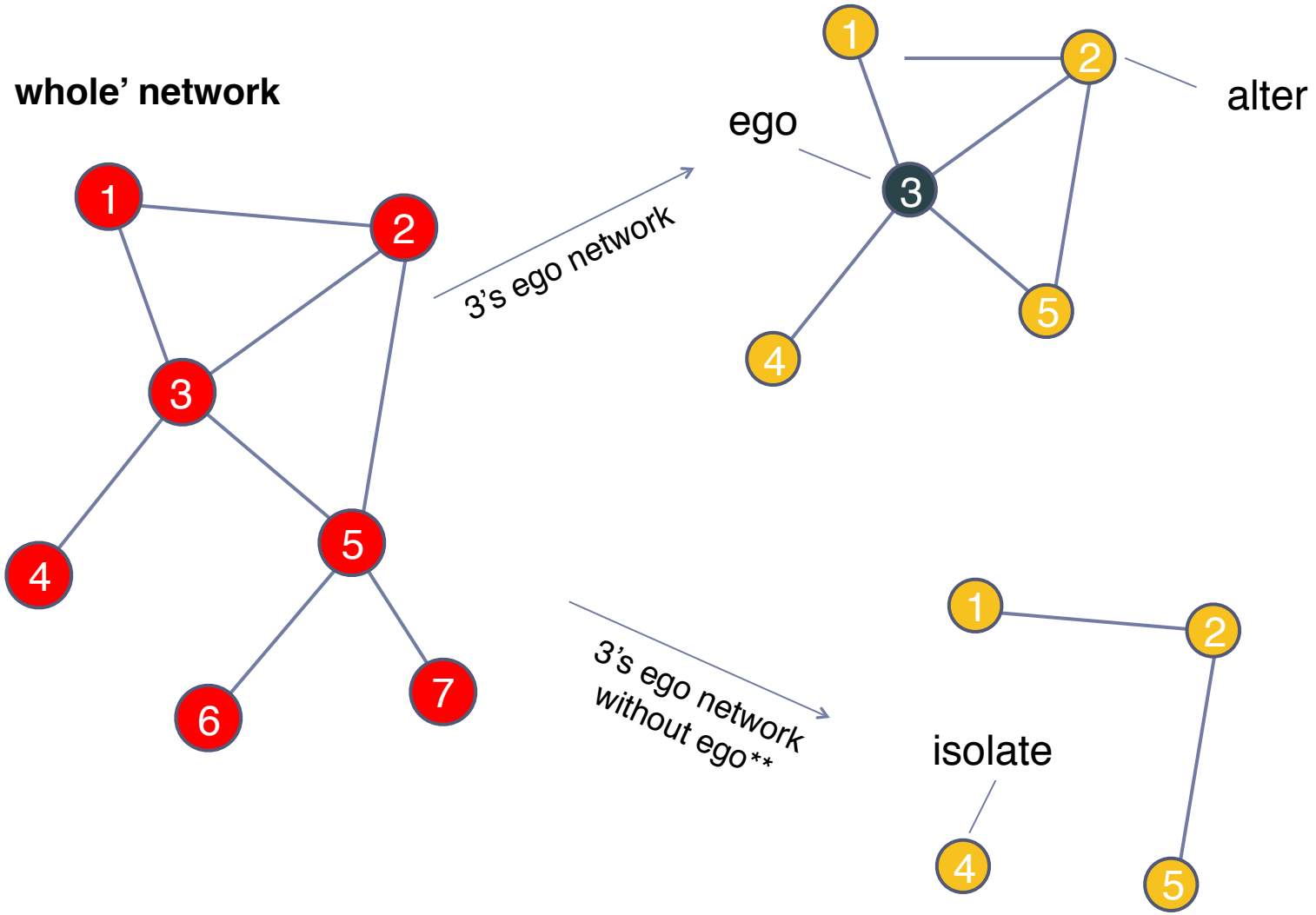
Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

But interpretation is different now

Adjacency matrix becomes symmetric

Vertex	1	2	3	4
1	-	1	1	0
2	1	-	1	1
3	1	1	-	1
4	0	1	1	-

WHOLE AND EGO NETWORKS



BASIC CONCEPTS

Networks

Tie Strength

Key Players

Cohesion

How to represent various social networks

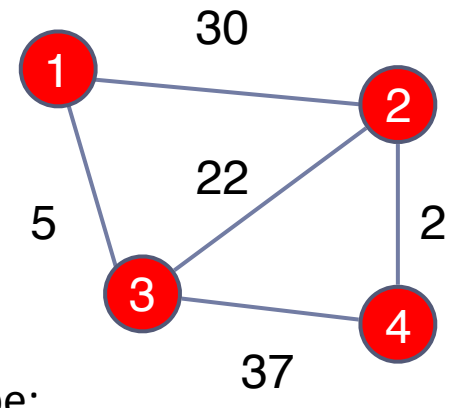
How to identify strong/weak ties in the network

How to identify key/central nodes in network

Measures of overall network structure

ADDING WEIGHTS TO EDGES

Weights represent strength of the link



Weights could be:

- Frequency of interaction in period of observation
- Number of items exchanged in period
- Individual perceptions of strength of relationship
- Costs in communication or exchange, e.g. distance
- Combinations of these

Edge list: add column of weights

Vertex	Vertex	Weight
1	2	30
1	3	5
2	3	22
2	4	2
3	4	37

Adjacency matrix: add weights instead of 1

Vertex	1	2	3	4
1	-	30	5	0
2	30	-	22	2
3	5	22	-	37
4	0	2	37	-

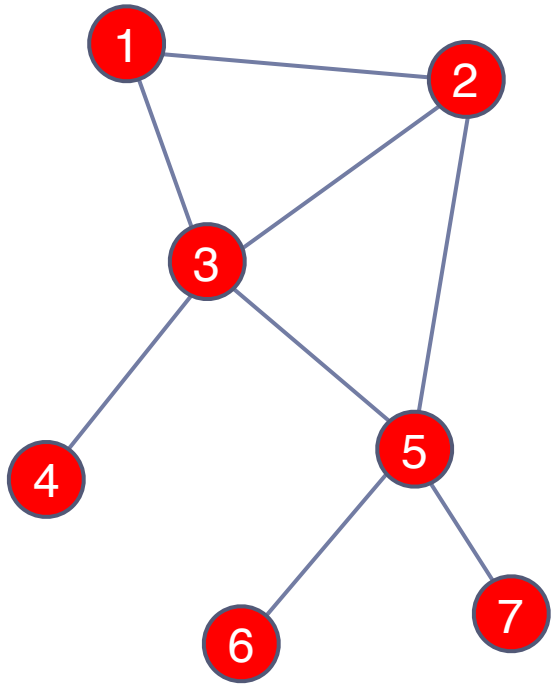
BASIC CONCEPTS

Networks	How to represent various social networks
Tie Strength	How to identify strong/weak ties in the network
<u>Key Players</u>	<u>How to identify key/central nodes in network</u>
Cohesion	Measures of overall network structure

DEGREE CENTRALITY

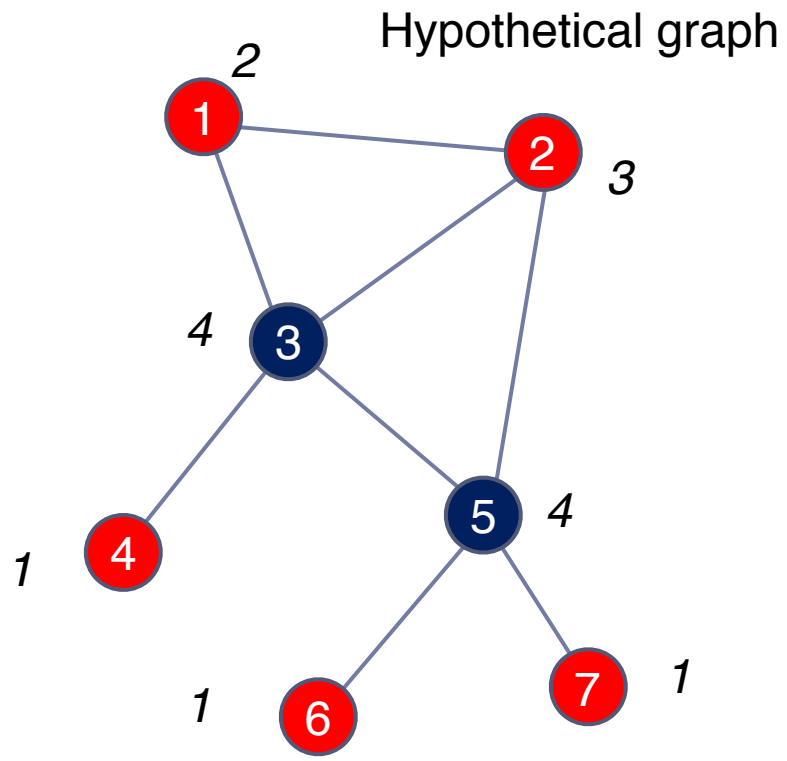
- A node's (in-) or (out-)degree is the number of links that lead into or out of the node
- In an undirected graph they are of course identical
- Often used as measure of a node's degree of connectedness and hence also influence and/or popularity
- Useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate 'neighborhood'

Hypothetical graph



DEGREE CENTRALITY

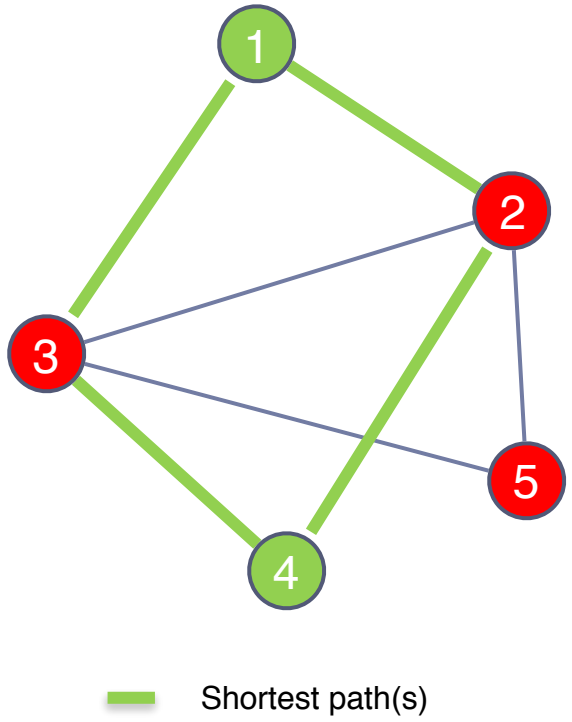
- A node's (in-) or (out-)degree is the number of links that lead into or out of the node
- In an undirected graph they are of course identical
- Often used as measure of a node's degree of connectedness and hence also influence and/or popularity
- Useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate 'neighborhood'



PATHS AND SHORTEST PATHS

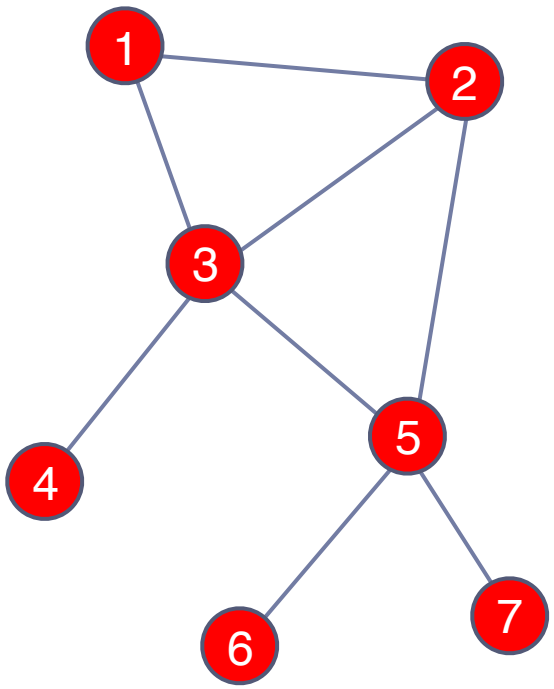
- A path between two nodes is any sequence of non-repeating nodes that connects the two nodes
- The shortest path between two nodes is the path that connects the two nodes with the shortest number of edges (also called the distance between the nodes)
- In the example to the right, between nodes 1 and 4 there are two shortest paths of length 2: {1,2,4} and {1,3,4}
- Other, longer paths between the two nodes are {1,2,3,4}, {1,3,2,4}, {1,2,5,3,4} and {1,3,5,2,4} (the longest paths)
- Shorter paths are desirable when speed of communication or exchange is desired (often the case in many studies, but sometimes not, e.g. in networks that spread disease)

Hypothetical graph



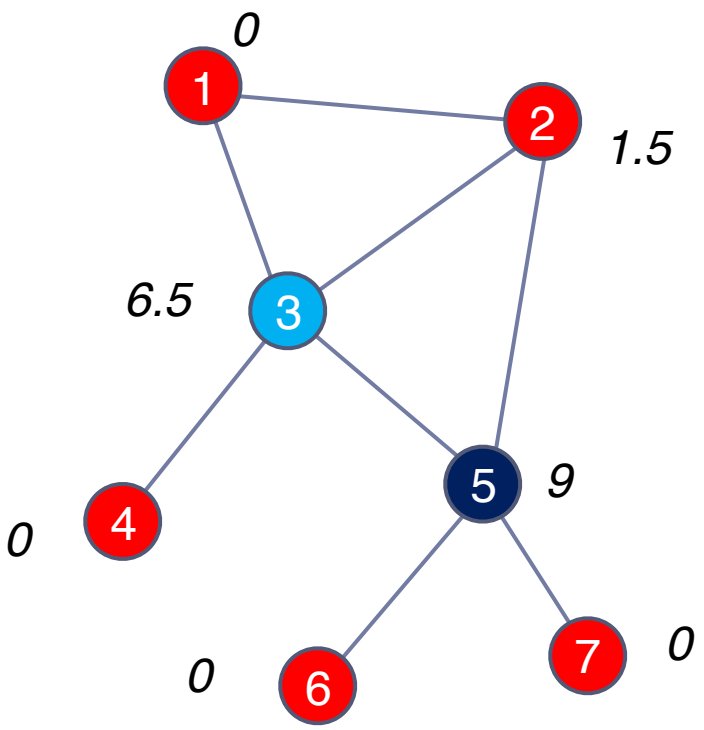
BETWEENNESS CENTRALITY

- For a given node v, calculate the number of shortest paths between nodes i and j that pass through v, and divide by all shortest paths between nodes i and j
- Sum the above values for all node pairs i,j
- Sometimes normalized such that the highest value is 1or that the sum of all betweenness centralities in the network is 1
- Shows which nodes are more likely to be in communication paths between other nodes
- Also useful in determining points where the network would break apart (think who would be cut off if nodes 3 or 5 would disappear)



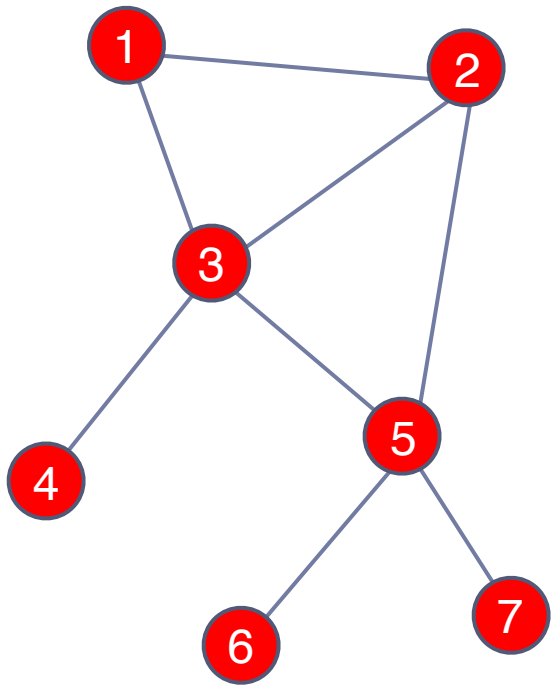
BETWEENNESS CENTRALITY

- For a given node v, calculate the number of shortest paths between nodes i and j that pass through v, and divide by all shortest paths between nodes i and j
- Sum the above values for all node pairs i,j
- Sometimes normalized such that the highest value is 1or that the sum of all betweenness centralities in the network is 1
- Shows which nodes are more likely to be in communication paths between other nodes
- Also useful in determining points where the network would break apart (think who would be cut off if nodes 3 or 5 would disappear)



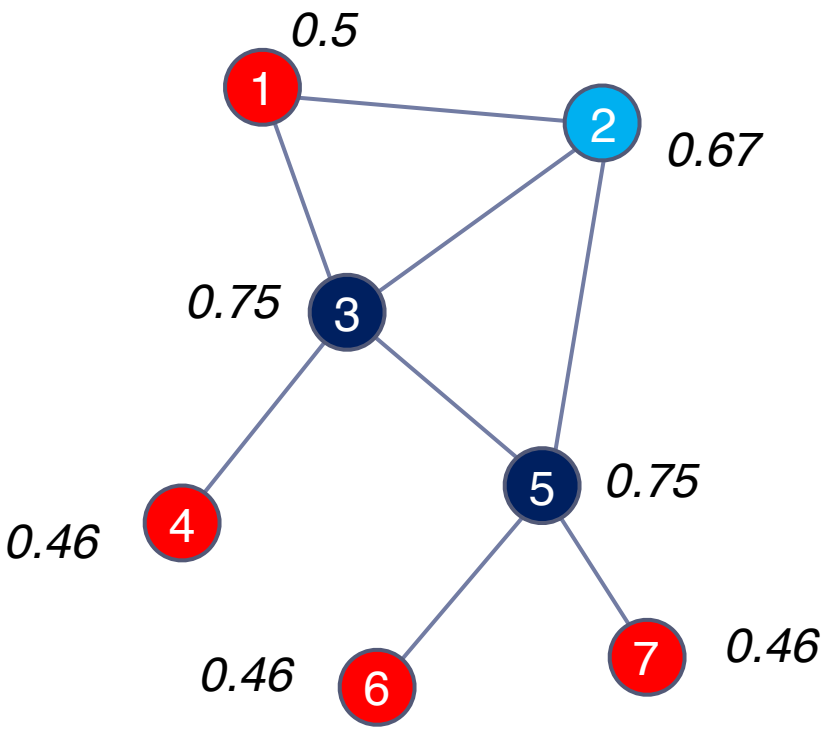
CLOSENESS CENTRALITY

- Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)
- Take the reciprocal of the above value so that higher values are ‘better’ (indicate higher closeness) like in other measures of centrality
- It is a measure of reach, i.e. the speed with which information can reach other nodes from a given starting node



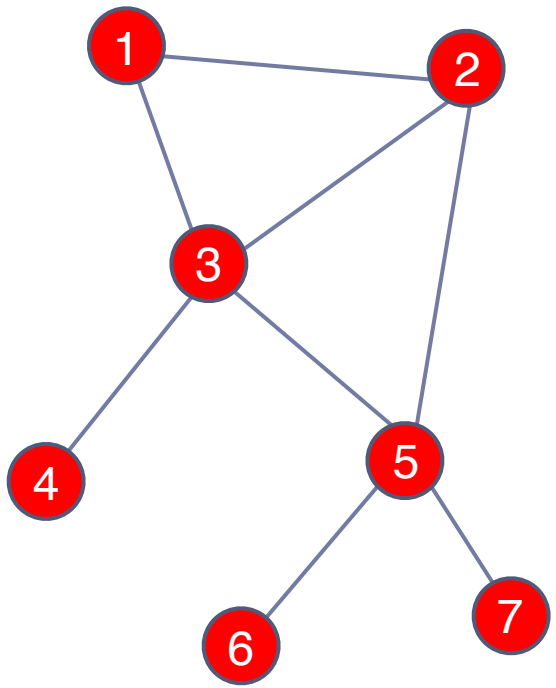
CLOSENESS CENTRALITY

- Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)
- Take the reciprocal of the above value so that higher values are ‘better’ (indicate higher closeness) like in other measures of centrality
- It is a measure of reach, i.e. the speed with which information can reach other nodes from a given starting node



EIGENVECTOR CENTRALITY

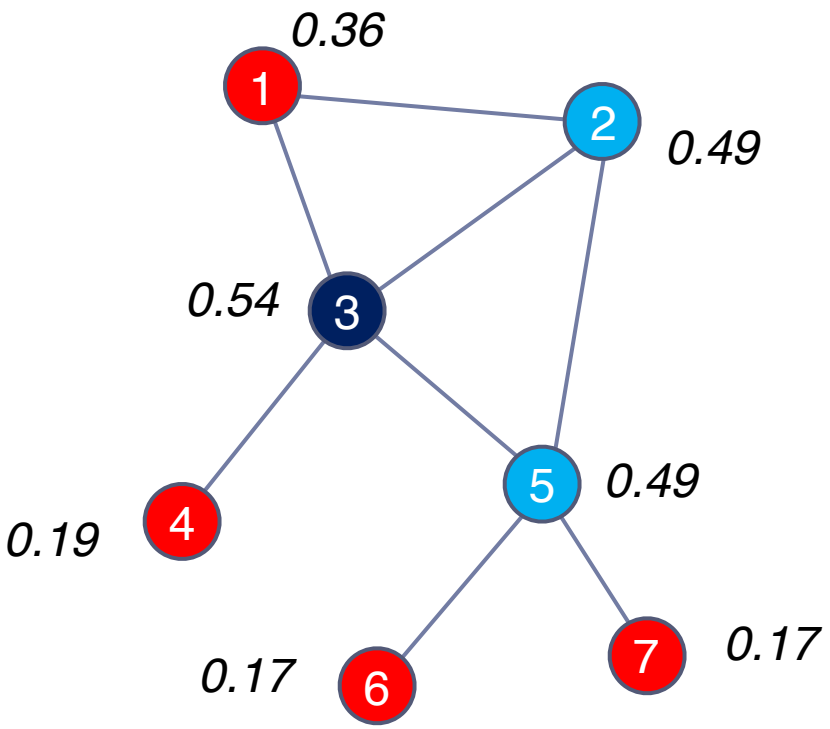
- A node's eigenvector centrality is a measure of the influence of a node in a network
- In other words, a node with a high eigenvector centrality is connected to other nodes with high eigenvector centrality
- This is similar to how Google ranks web pages: links from highly linked-to pages count more
- Useful in determining who is connected to the most connected nodes



$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

EIGENVECTOR CENTRALITY

- A node's eigenvector centrality is a measure of the influence of a node in a network
- In other words, a node with a high eigenvector centrality is connected to other nodes with high eigenvector centrality
- This is similar to how Google ranks web pages: links from highly linked-to pages count more
- Useful in determining who is connected to the most connected nodes



$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

INTERPRETATION OF MEASURES

Degree

- In network of music collaborations: how many people has this person collaborated with?
- How many people can this person reach directly?
- How many friends do you have?

INTERPRETATION OF MEASURES

Betweenness

- How likely is this person to be the most direct route between two people in the network?
- In network of spies: who is the spy though whom most of the confidential information is likely to flow?

INTERPRETATION OF MEASURES

Closeness

- In network of sexual relations: how fast will an STD spread from this person to the rest of the network?
- How fast can this person reach everyone in the network?

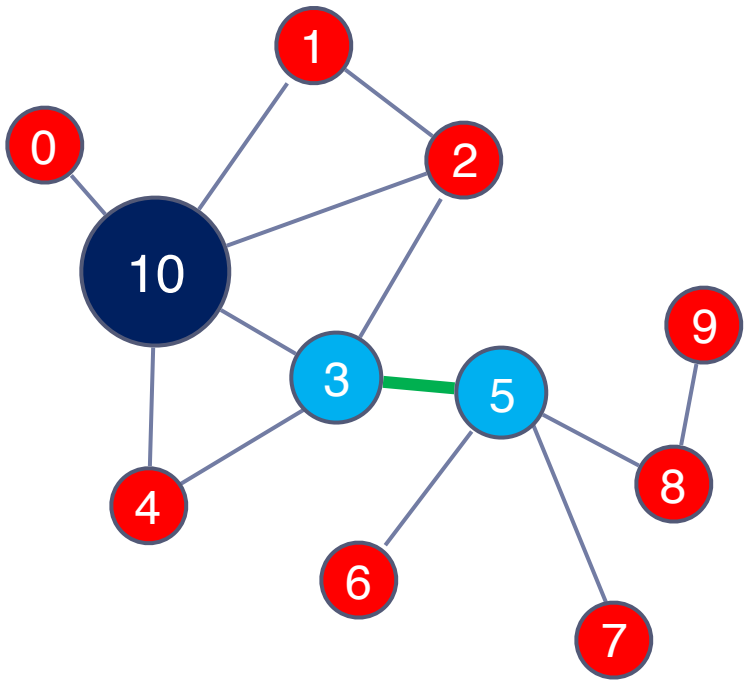
INTERPRETATION OF MEASURES

Eigenvector

- How well is this person connected to other well-connected people?
- In network of paper citations: who is the author that is most cited by other well-cited authors?

IDENTIFICATION OF KEY PLAYERS

- In the network to the right, node 10 is the most central according to degree centrality
- But nodes 3 and 5 together will reach more nodes
- Moreover the tie between them is critical; if severed, the network will break into two isolated sub-networks
- It follows that other things being equal, players 3 and 5 together are more ‘key’ to this network than 10
- Thinking about sets of key players is helpful!

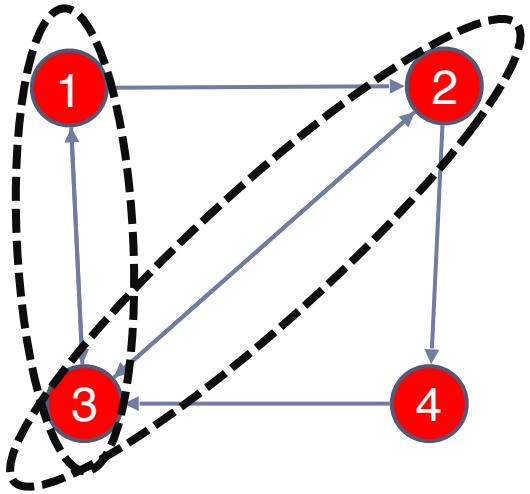


BASIC CONCEPTS

Networks	How to represent various social networks
Tie Strength	How to identify strong/weak ties in the network
Key Players	How to identify key/central nodes in network
<u>Cohesion</u>	<u>Measures of overall network structure</u>

RECIPROCITY

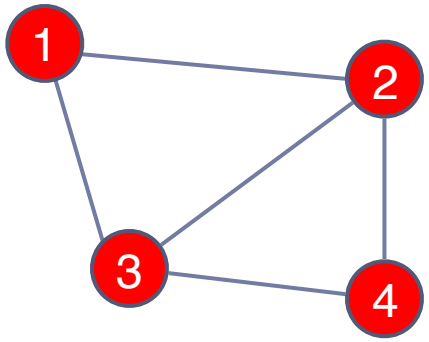
- The ratio of the number of relations which are reciprocated (i.e. there is an edge in both directions) over the total number of relations in the network
- …where two vertices are said to be related if there is at least one edge between them
- In the example to the right this would be 2/5=0.4 (whether this is considered high or low depends on the context)
- A useful indicator of the degree of mutuality and reciprocal exchange in a network, which relate to social cohesion
- Only makes sense in directed graphs

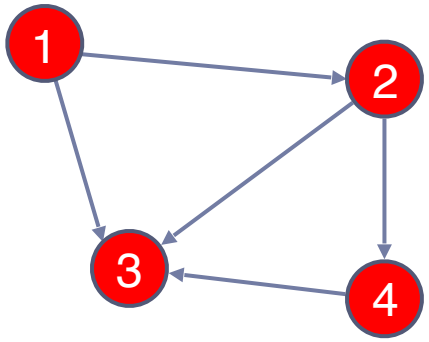


Reciprocity for network = 0.4

DENSITY

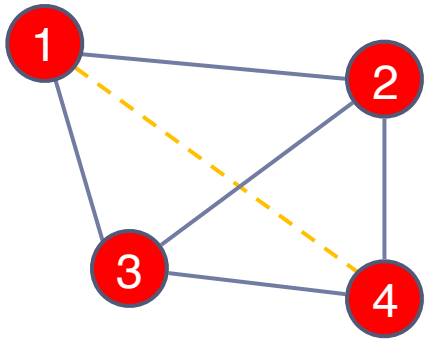
- A network's density is the ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes (which is n(n-1)/2, where n is the number of vertices, for an undirected graph)
- In the example network to the right density=5/6=0.83 (i.e. it is a fairly dense network; opposite would be a sparse network)
- It is a common measure of how well connected a network is (in other words, how closely knit it is) – a perfectly connected network is called a clique and has density=1
- A directed graph will have half the density of its undirected equivalent, because there are twice as many possible edges, i.e. n(n-1)
- Density is useful in comparing networks against each other, or in doing the same for different regions within a single network



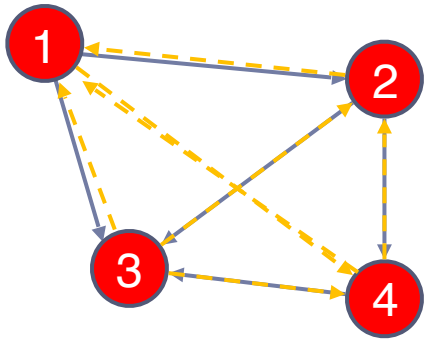


DENSITY

- A network's density is the ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes (which is n(n-1)/2, where n is the number of vertices, for an undirected graph)
- In the example network to the right density=5/6=0.83 (i.e. it is a fairly dense network; opposite would be a sparse network)
- It is a common measure of how well connected a network is (in other words, how closely knit it is) – a perfectly connected network is called a clique and has density=1
- A directed graph will have half the density of its undirected equivalent, because there are twice as many possible edges, i.e. n(n-1)
- Density is useful in comparing networks against each other, or in doing the same for different regions within a single network



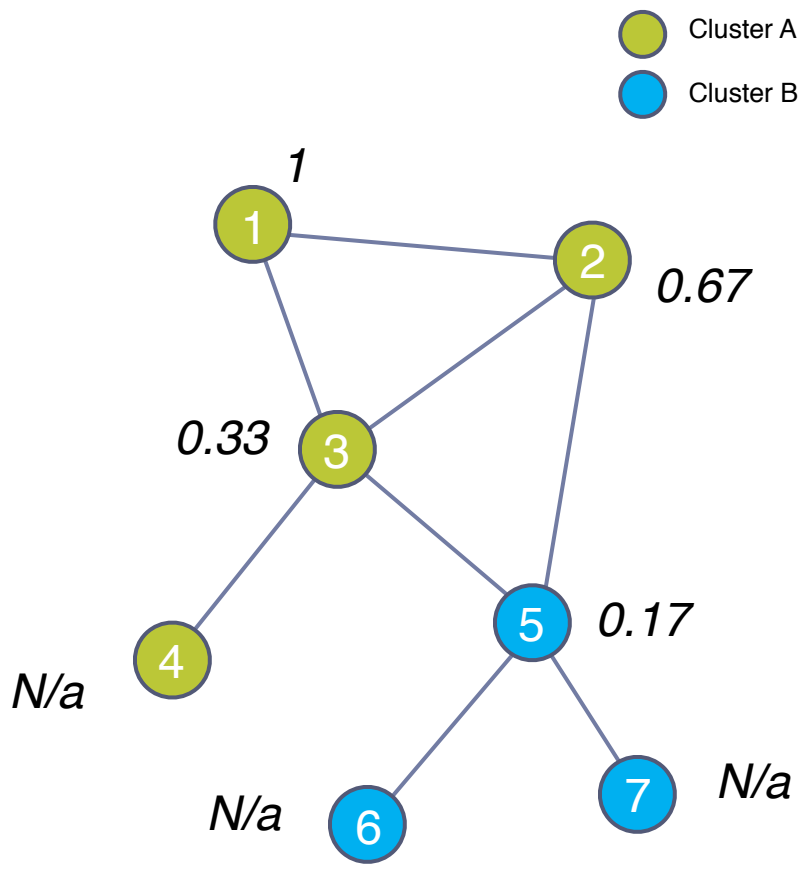
density = 5/6 = 0.83



density = 5/12 = 0.42

CLUSTERING

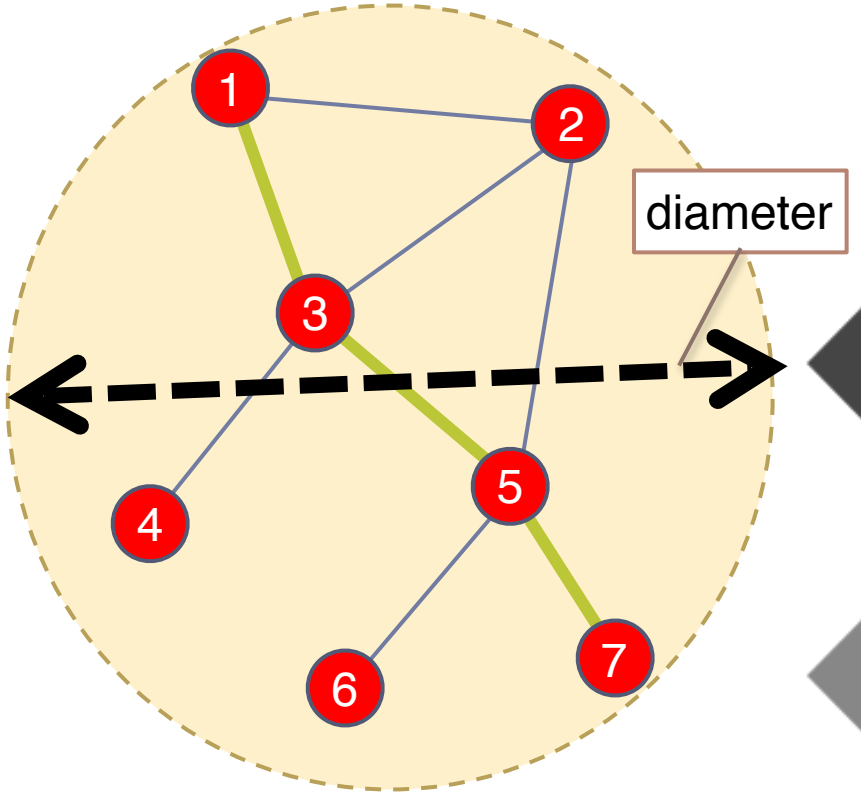
- A node's clustering coefficient is the number of closed triplets in the node's neighborhood over the total number of triplets in the neighborhood. It is also known as transitivity.
- E.g., node 1 to the right has a value of 1 because it is only connected to 2 and 3, and these nodes are also connected to one another (i.e. the only triplet in the neighborhood of 1 is closed). We say that nodes 1,2, and 3 form a clique.
- Clustering algorithms identify clusters or 'communities' within networks based on network structure and specific clustering criteria (example shown to the right with two clusters is based on edge betweenness, an equivalent for edges of the betweenness centrality presented earlier for nodes)



Network clustering coefficient = 0.375
(3 nodes in each triangle x 2 triangles = 6 closed triplets divided by 16 total)

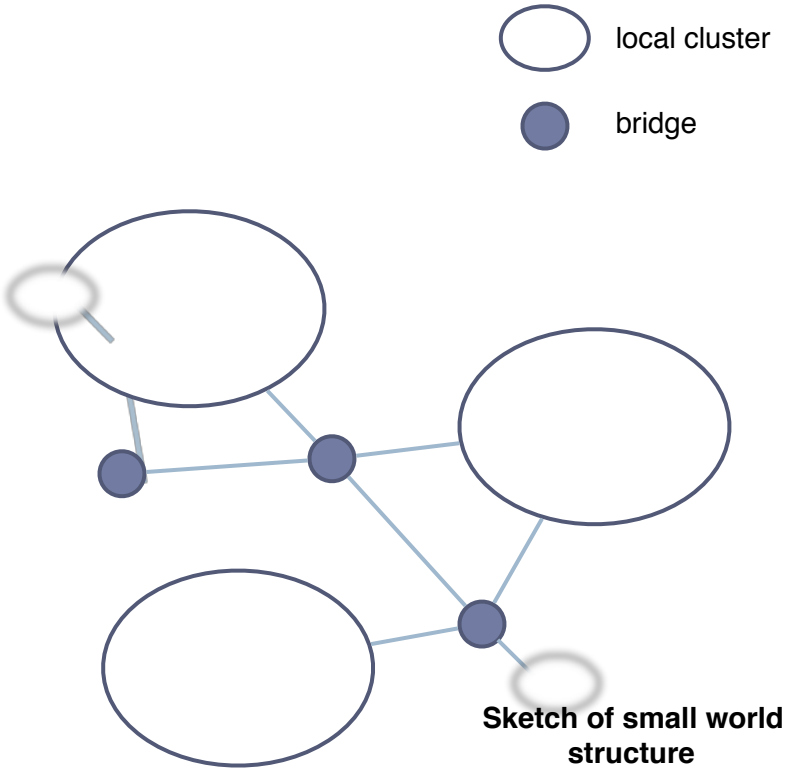
AVERAGE AND LONGEST DISTANCE

- The longest shortest path (distance) between any two nodes in a network is called the network's diameter
- The diameter of the network on the right is 3; it is a useful measure of the reach of the network (as opposed to looking only at the total number of vertices or edges)
- It also indicates how long it will take at most to reach any node in the network (sparser networks will generally have greater diameters)
- The average of all shortest paths in a network is also interesting because it indicates how far apart any two nodes will be on average (average distance)



SMALL WORDS

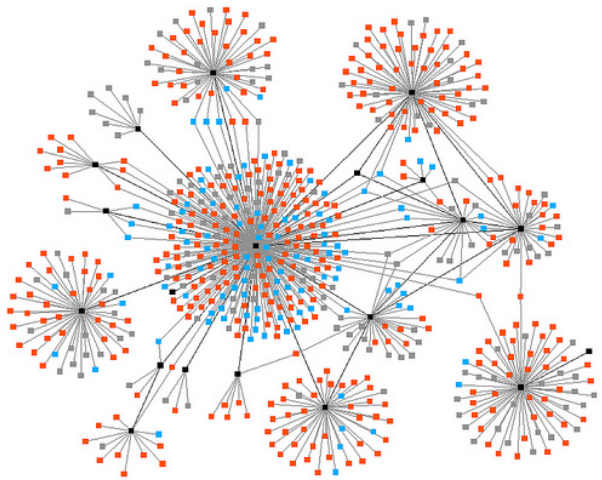
- A small world is a network that looks almost random but exhibits a significantly high clustering coefficient (nodes tend to cluster locally) and a relatively short average path length (nodes can be reached in a few steps)
- It is a very common structure in social networks because of transitivity in strong social ties and the ability of weak ties to reach across clusters
- Such a network will have many clusters but also many bridges between clusters that help shorten the average distance between nodes



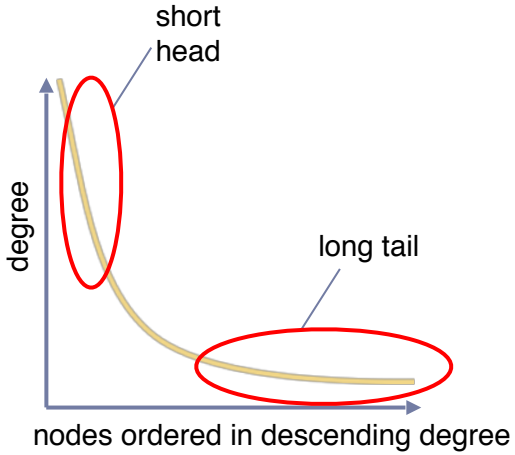
Sketch of small world structure

PREFERENTIAL ATTACHMENT

- A property of some networks, where, during their evolution and growth in time, a the great majority of new edges are to nodes with an already high degree; the degree of these nodes thus increases disproportionately, compared to most other nodes in the network
  - The result is a network with few very highly connected nodes and many nodes with a low degree
  - Such networks are said to exhibit a long-tailed degree distribution
  - And they tend to have a small-world structure!



Example of network with preferential attachment



Sketch of long-tailed degree distribution

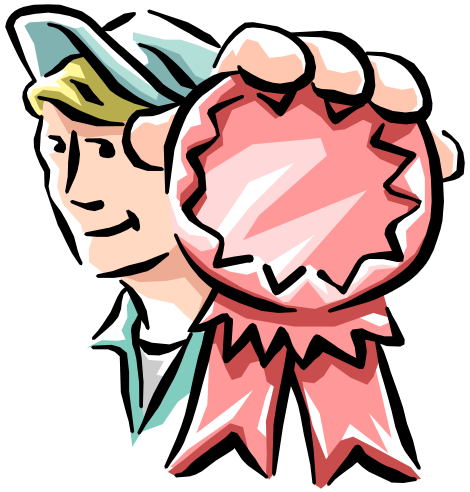
PREFERENTIAL ATTACHMENT - REASONS

Popularity

- We want to be associated with popular people, ideas, items, thus further increasing their popularity, irrespective of any objective, measurable characteristics

Also known as
'the rich get richer'

PREFERENTIAL ATTACHMENT - REASONS



Quality

- We evaluate people and everything else based on objective quality criteria, so higher quality nodes will naturally attract more attention, faster

Also known as
'the good get better'

PREFERENTIAL ATTACHMENT - REASONS

Mixed model

- Among nodes of similar attributes, those that reach critical mass first will become ‘stars’ with many friends and followers (‘halo effect’)

May be impossible to predict who will become a star, even if quality matters

Section 4: Regular Expressions

WHAT ARE REGULAR EXPRESSIONS?

- Definition: A Regular expression is a pattern describing a certain amount of text.

- A regular expression, often called a pattern, is an expression that describes a set of strings.
  - Wikipedia

WHAT ARE REGULAR EXPRESSIONS?

- Regular expressions allow matching and manipulation of textual data.
- Use:
  - Matching/Finding
  - Doing something with matched text
  - Validation of data
  - Case insensitive matching
  - Parsing data (ex: html)
  - Converting data into diff. form etc.

GRAMMAR OF REGEX

RE = one or more non-empty ‘*branches*‘ separated by ‘|’

Branch = one or more ‘*pieces*’

Piece = *atom* followed by quantifier

Quantifier = ‘*,+,?’ or ‘*bound*’

Bound = *atom{n}, atom{n,}, atom*{m, n}

Atom = (RE) or

() or

‘^,$,’ or

\ followed by `^.[$()|*+?{\’ or

any-char or

‘*bracket expression*’

Bracket Expression = is a list of characters enclosed in `[ ]'

REGULAR EXPRESSIONS EXAMPLES

Characters	Regular Expression
t	. t [a-z]
1	. 1 [0-9]
text	[a-z]+
asdgf	.+ .*
^	Matches string at the beginning of the text
$	Matches string at the end of the text
?	Matches 0 or 1 time
.	match any character
*	Matches 0 or more times
+	Matches 1 or more times

REGULAR EXPRESSIONS- SUMMARY

- Everyone should know basics of regular expression
  - What it is
  - Simple examples
- No need to remember all patterns - use cheat sheet!

Section 5: Use Case I

SNA in Telecommunications

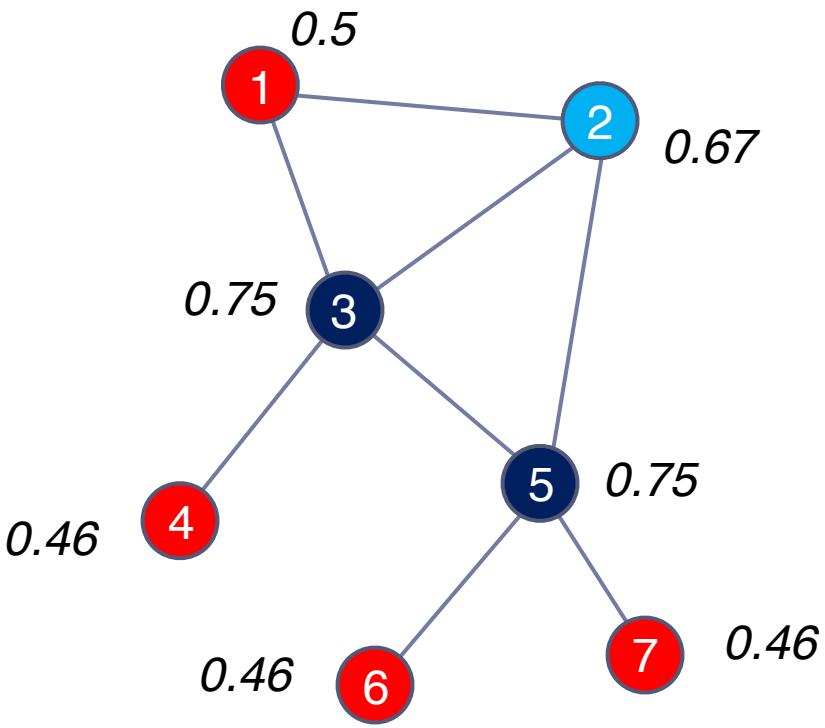
INTRODUCTION

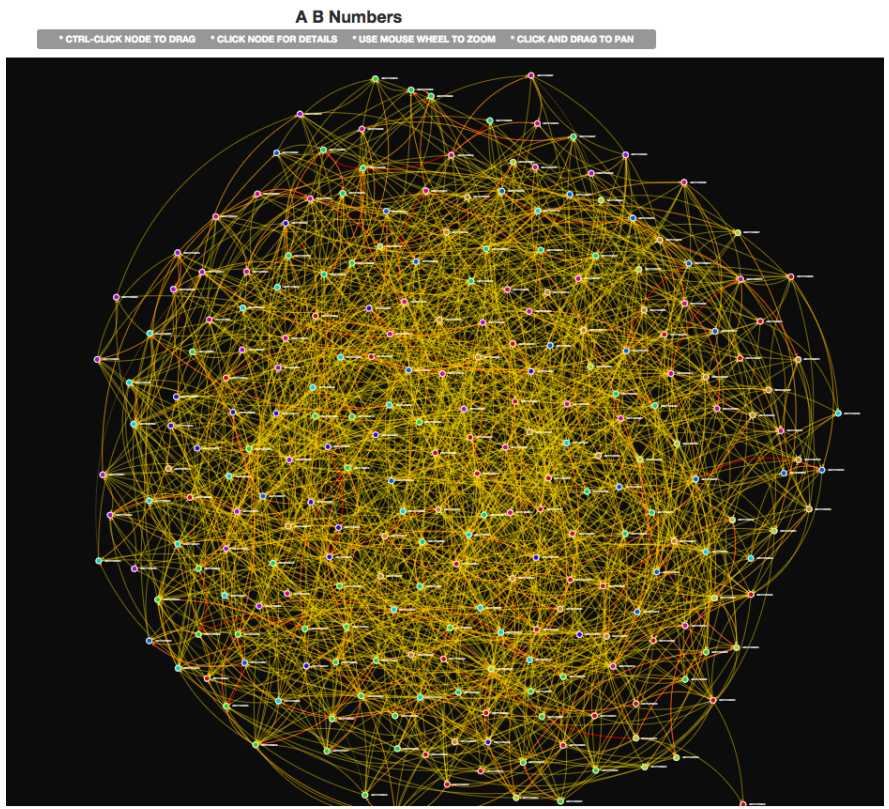
- SNA in Telcos is based on CDR data (call data records)

call_id	call_dt	call_tm	call_duration	call_price	A number	B number	type	direction	A prefix	B prefix
12	2016-10-01	08:20:23	23	0.0	904246821	902487123	A	O		
245	2016-10-03	13:43:52	12	0.23	902654289	918786534	B	I		
321	2016-10-07	23:22:21	123		902654289	904246821	A	O		
231	2016-10-23	19:54:09	345	0.45	902654289	908765432	A	O		
221	2016-10-15	11:10:00	32	1.4	908765432	664529751	C	O		43
17	2016-10-11	12:38:37	9	0.32	911654789	908765432	B	O		
789	2016-10-10	02:34:09	0	0.12	911654789	904246821	A	O		
753	2016-10-08	14:41:33	98		904246821	911654789	D	O		
537	2016-10-27	17:21:24	10	0.89	908234876	911654789	B	I		
98	2016-10-22	18:22:21	22	0.65	908234876	664356980	C	I		43

INTRODUCTION

- You cannot rely just on the visualization





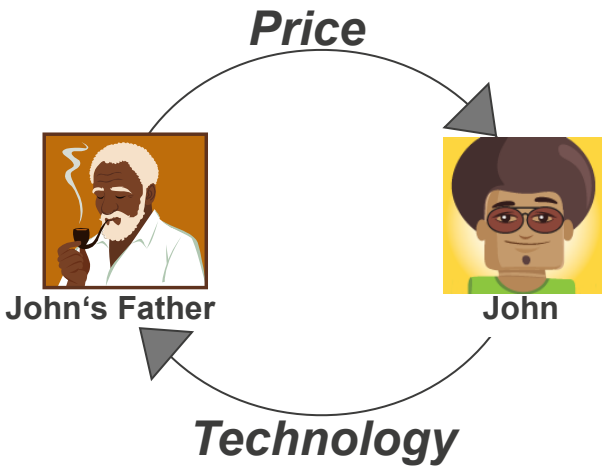
INTRODUCTION

Identify the social network

- Who contacts whom?
- How often?
- How long?
- Both directions?
- On net, off net

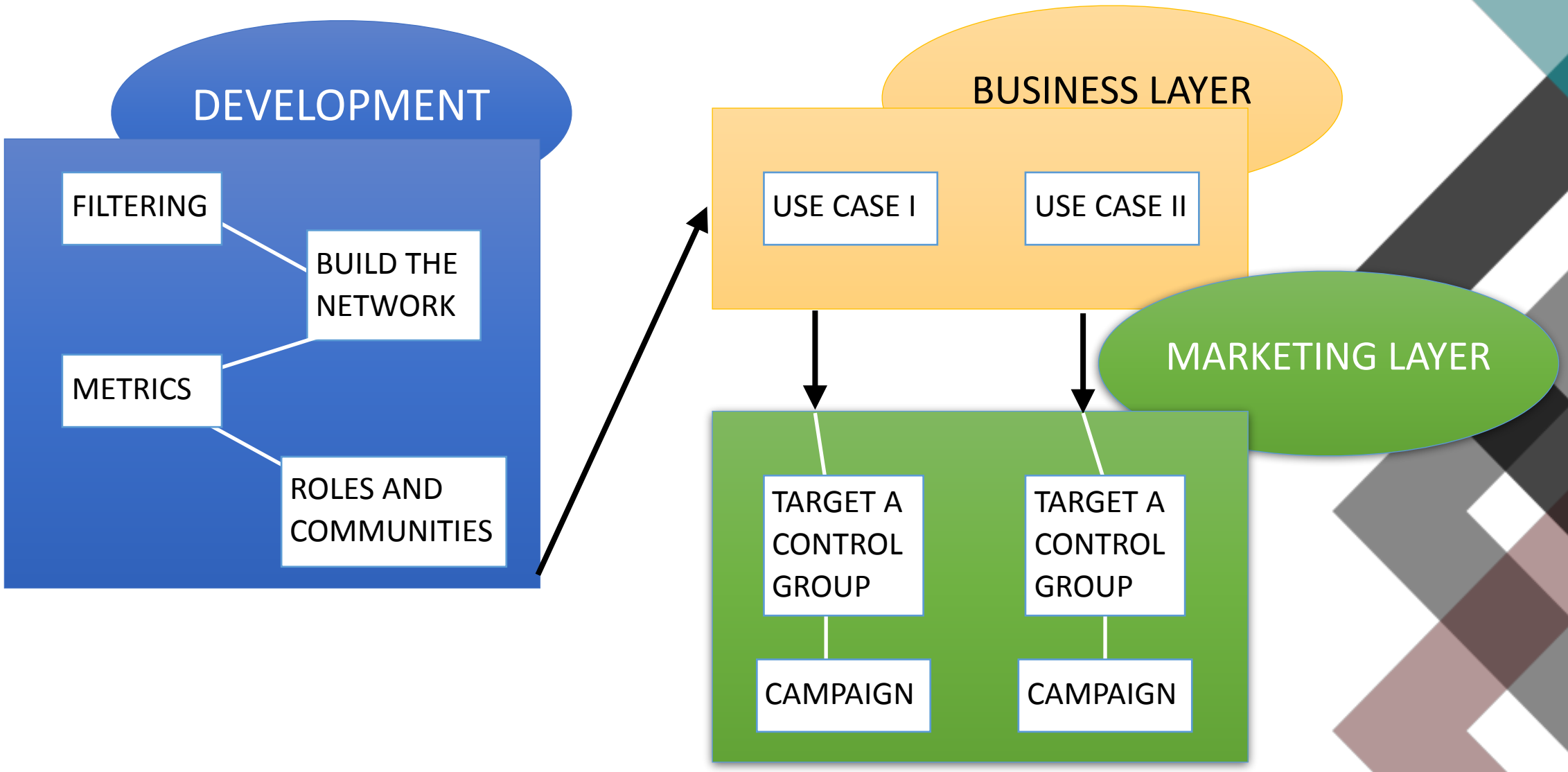
Identify Influencers for each topic

- Who influences whom how much on purchases?
- Who influences whom how much on churn?



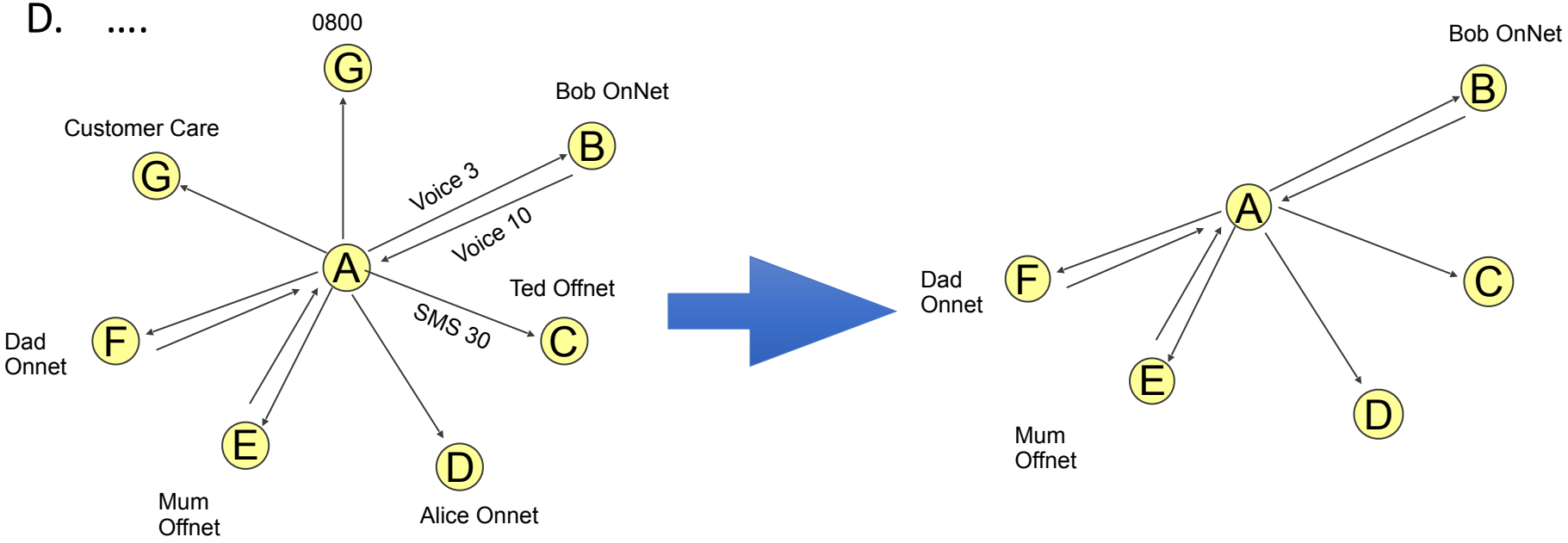
There is no "general" influencer

ARCHITECTURE OF SNA PROCESS



FILTERING THE EDGES

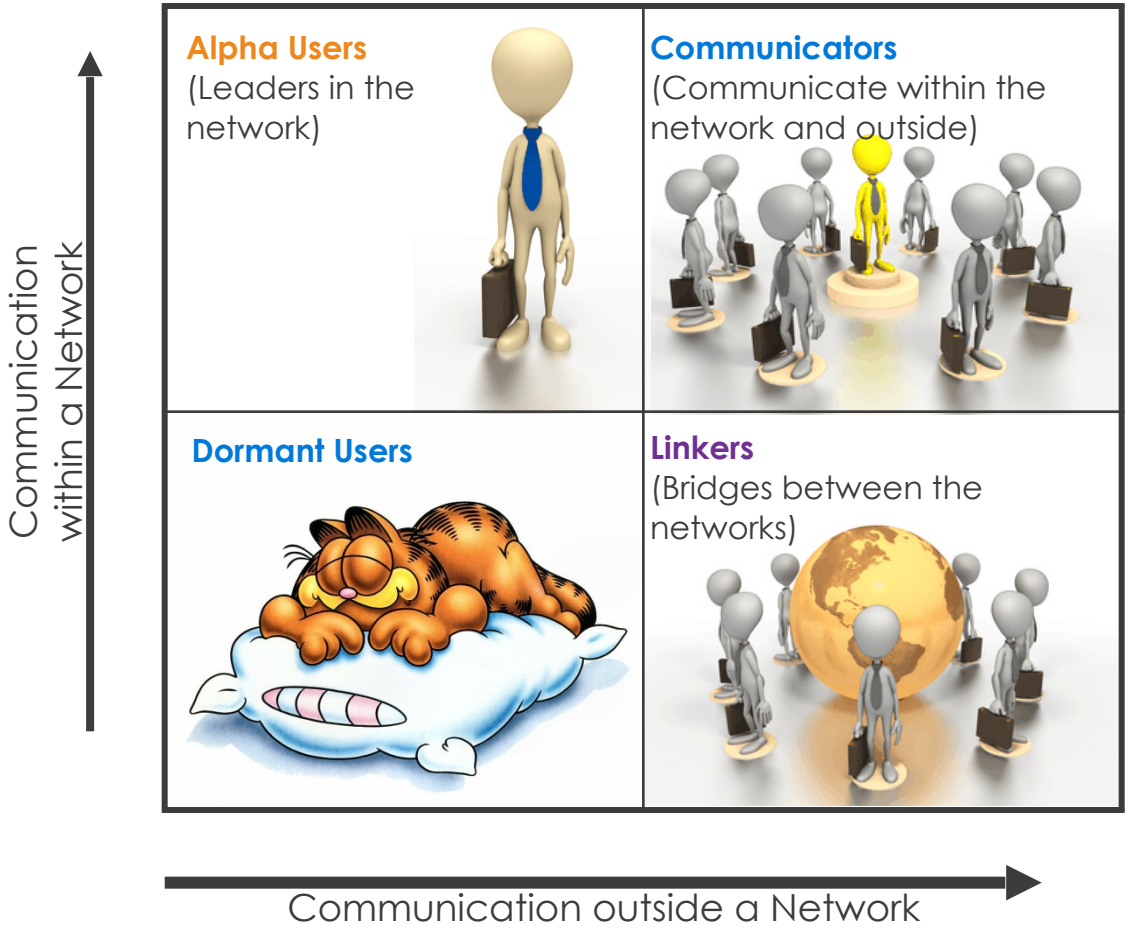
- Filtering away the irrelevant nodes and edges
  - A. Automatic call numbers
  - B. Special non-human gateways
  - C. Emergency calls
  - D. ....



COMPUTE METRICS

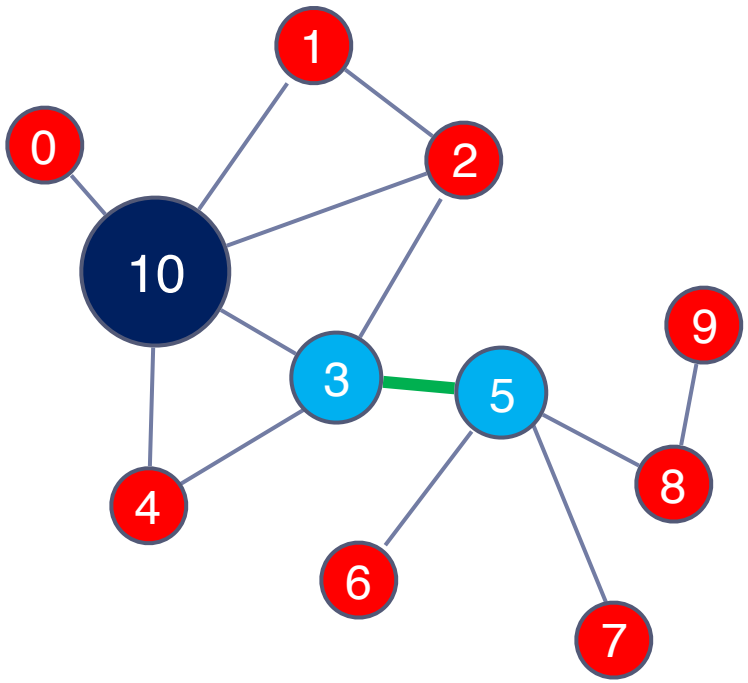
- Computing many nodes and edges metrics
- Node metrics
  - Degree
  - Betweeness
  - ……
- Edges metrics
  - Strength
  - Reciprocal
  - short calls at night
  - ……

ROLES AND COMMUNITIES



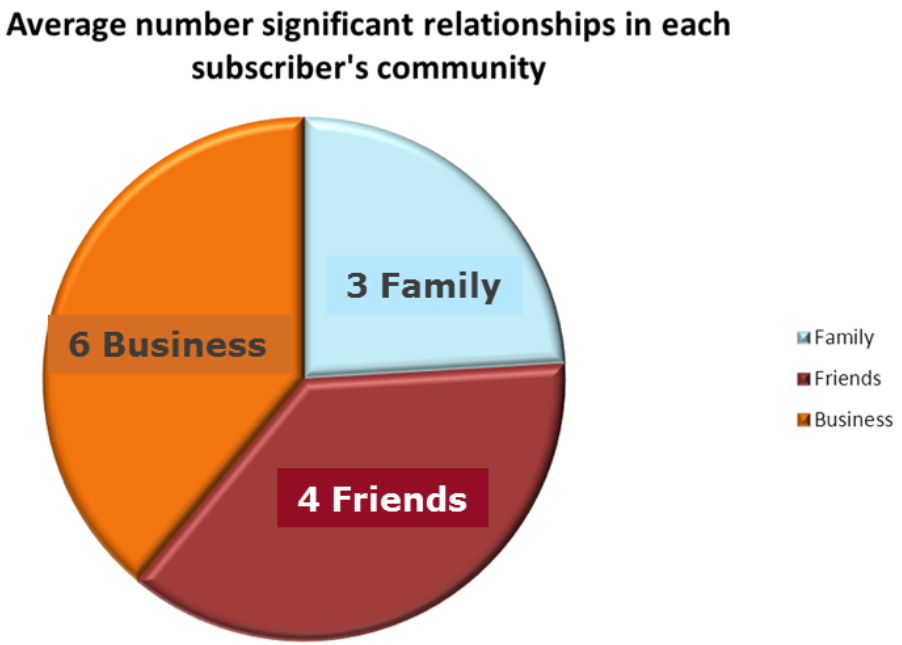
ROLES AND COMMUNITIES

- Alpha users
- Dormant users
- Communicators
- Linkers



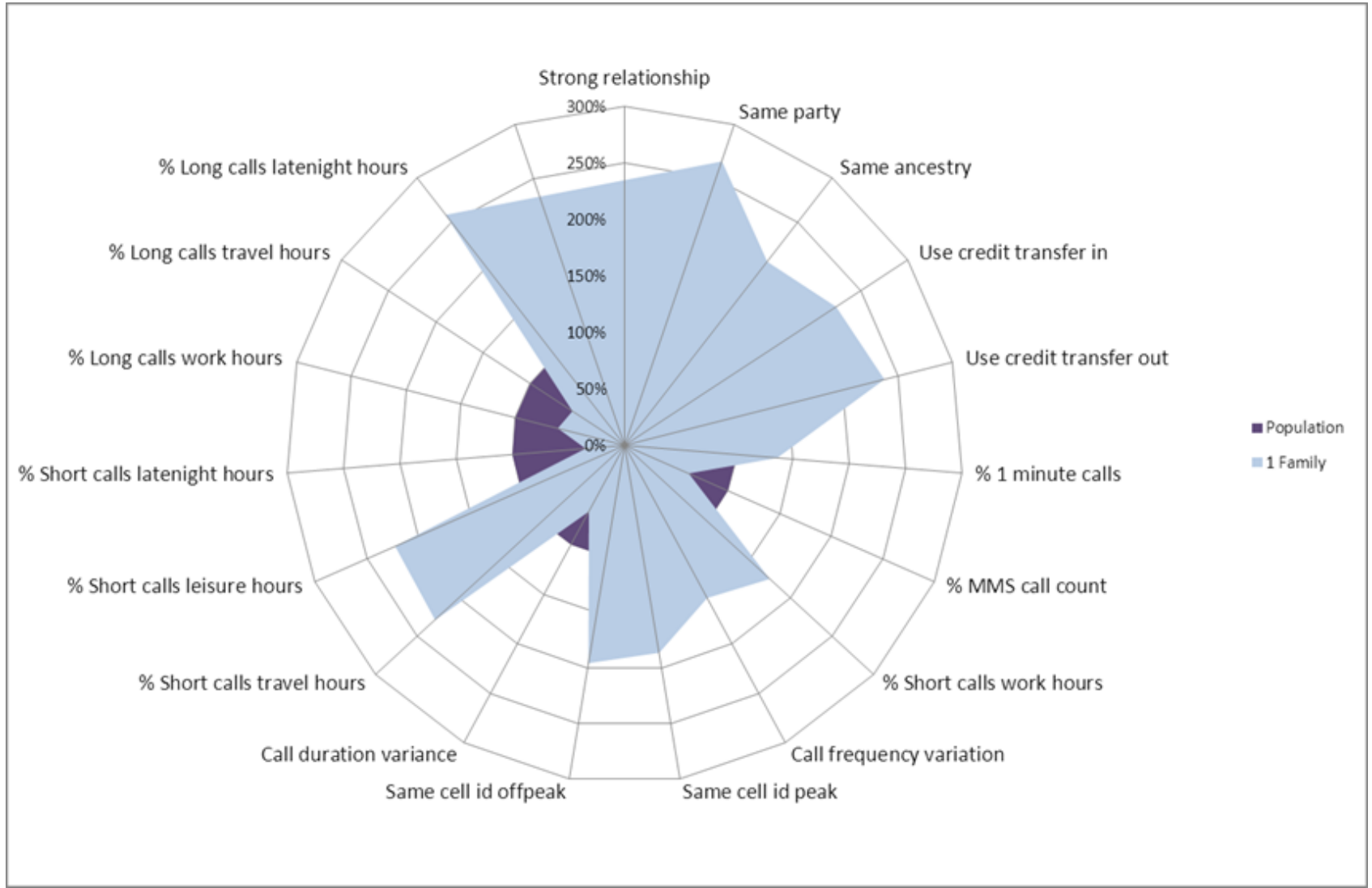
ROLES AND COMMUNITIES

- Non-trivial relationships between people identified as:
  - Family
  - Friends
  - Business
- SNA experience was used to develop hypotheses and analyses about how types of calling patterns are related to the different relationships
- Each customer is profiled in terms of his relationships with the people he calls and who call him



ROLES AND COMMUNITIES

- What makes up a ‘Family’?



USE CASE I - MULTI SIM DETECTION

- Using the SNA for Multi SIM detection, companies are able to identify 4 types of clients
  1. customers, who bought a new competitive SIM-card and going to churn
  2. customers, who bought additional SIM-card and become permanent Multi-SIM users
  3. existing Multi-SIM users
  4. customers, who have a main SIM card with competition, however recently bought a new onnet SIM

USE CASE I - MULTI SIM DETECTION

- Multi SIM Detection Opportunities:
  - Reduction of the loss of revenue caused by customers churn, through early identification of customers, who started using competitors' SIM cards
  - Reduction of the loss of revenue by identifying customers, who use competitors' SIM cards on a regular basis
  - Revenue acquisition by identifying customers, who started using onnet SIM-card in addition to their main SIM-card of competitors network
  - Reduction of customer acquisition expenses through determining internal churn.

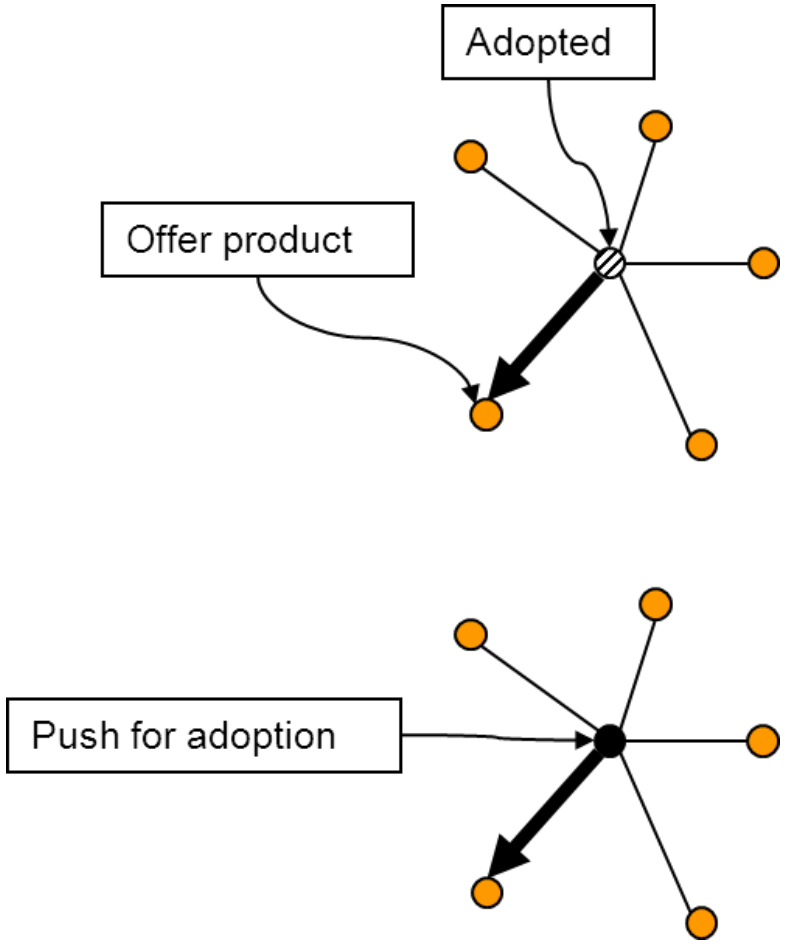
USE CASE II - CROSS SELL

Leverage the Collateral Adoption

- **Reactive**
- Identify subscribers whose affinity for products has increased due to adoption of product in their friends & family community

Identify Influencers among Friends & Family

- **Proactive**
- Identify subscribers who, should they adopt, would push a few friends and family to do the same



Section 6: Live Demo - Similarities between flags of countries

SNA BETWEEN FLAGS OF COUNTRIES

- Download flags in png format from https://www.gosquared.com/resources/2400-flags
- Similar flags are taken as connected

Section 7: Use Case II

WHO DO WE INTERACT WITH?

- A case study about how to use the social network analysis method to analyze the interactions between groups of people online

REFERENDUM IN SLOVAKIA 2015

- Do you agree that the term marriage can describe only a coexistence of a man and a woman?
- Do you agree that pairs or groups of people of the same gender are not allowed to adopt children?
- Do you agree that schools are not allowed to demand the participation of a child in the course of sexual education, if the parent or the child itself disagree?

REFERENDUM IN SLOVAKIA 2015

- Source of the data:
    - Public debate online
    - Several Facebook groups on both sides
    - Articles and discussion forums

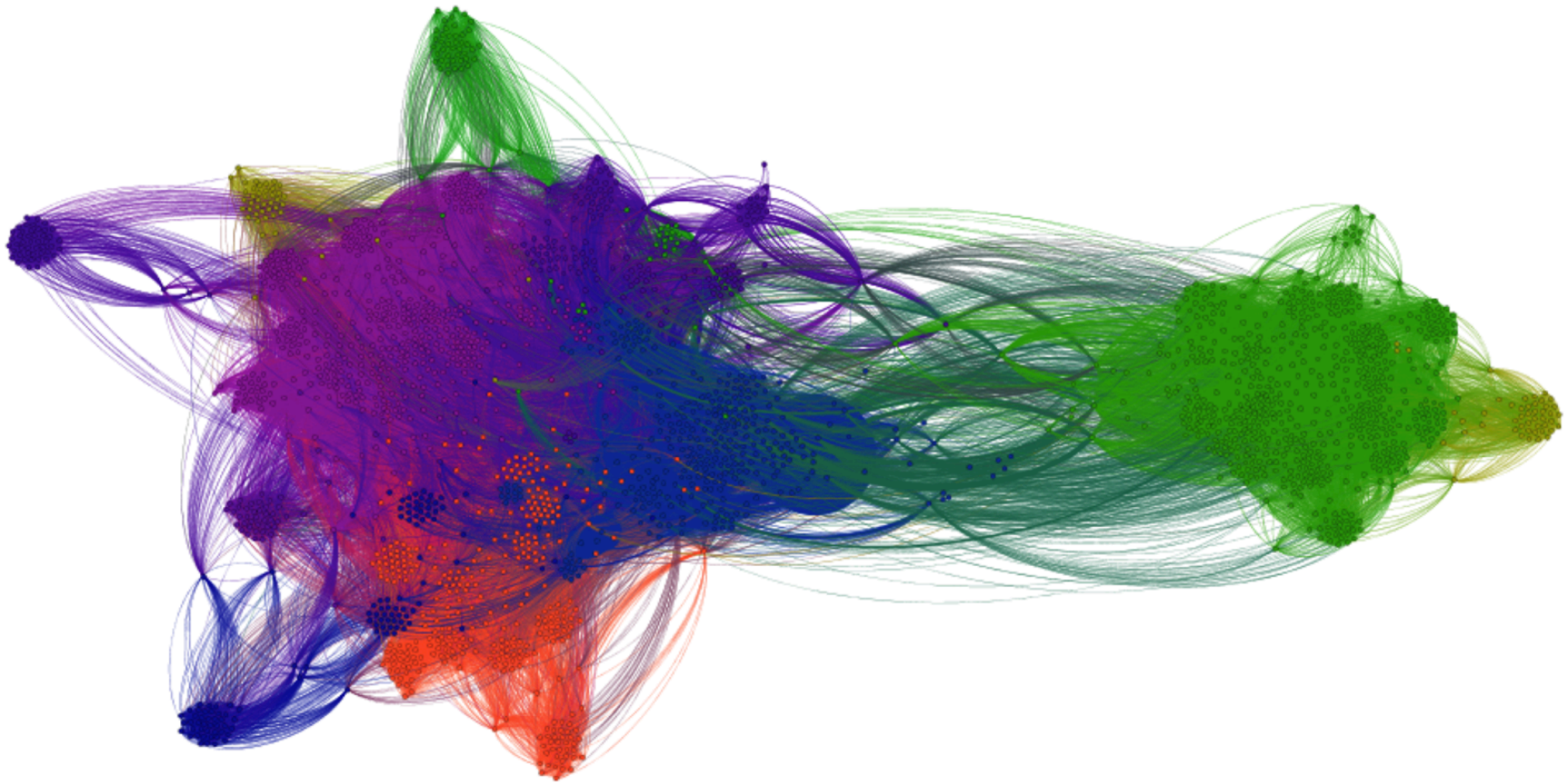
RESEARCH QUESTION

- To what extend did the members of opposite opinion groups in Slovak 2015 referendum interact with each other and among themselves in the public debate in social media?

DATA

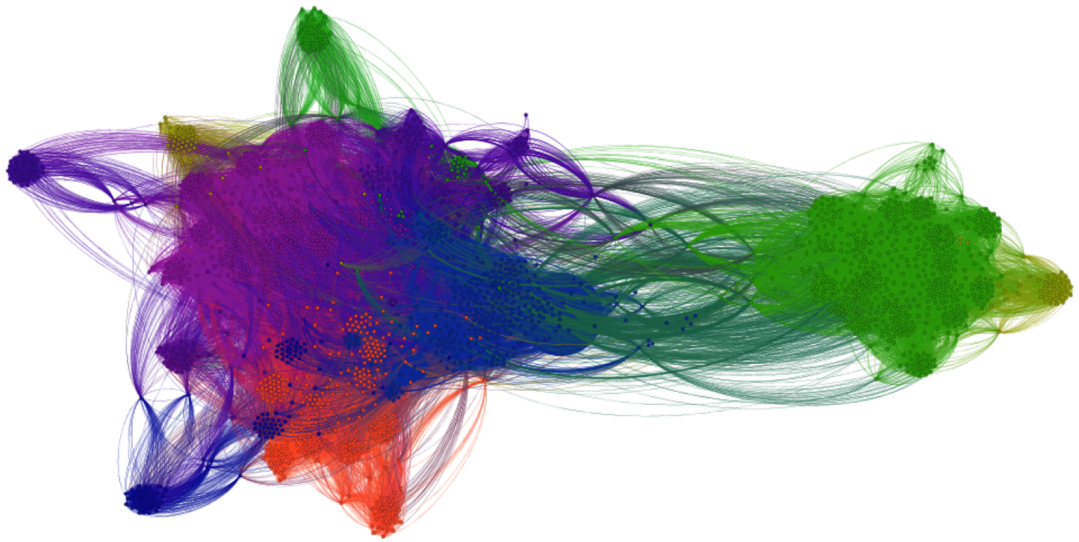
- period November 1, 2014 till February 13, 2015 (1 week after the referendum took place)
- The data included the links – “edges” between users or pages, in the form of “likes”, “shares” or “comments”
- The final dataset included 79,160 edges.

RESULTING SOCIAL NETWORK



RESULTING SOCIAL NETWORK - INTERPRETATION

- Different colours represents different Facebook Groups
- People are more connected inside the group
- Nodes are User Profiles
- Only few nodes represent people who connect different pages



Section 8: Live Demo - Amazon - Products bought together

FREE SNA DATASETS

- More awesome SNA datasets for practice are at https://snap.stanford.edu/data/

AMAZON DATASETS

- Network was collected by crawling Amazon website.
- It is based on Customers Who Bought This Item Also Bought feature of the Amazon website.
- If a product i is frequently co-purchased with product j, the graph contains a directed edge from i to j.

AMAZON DATASETS

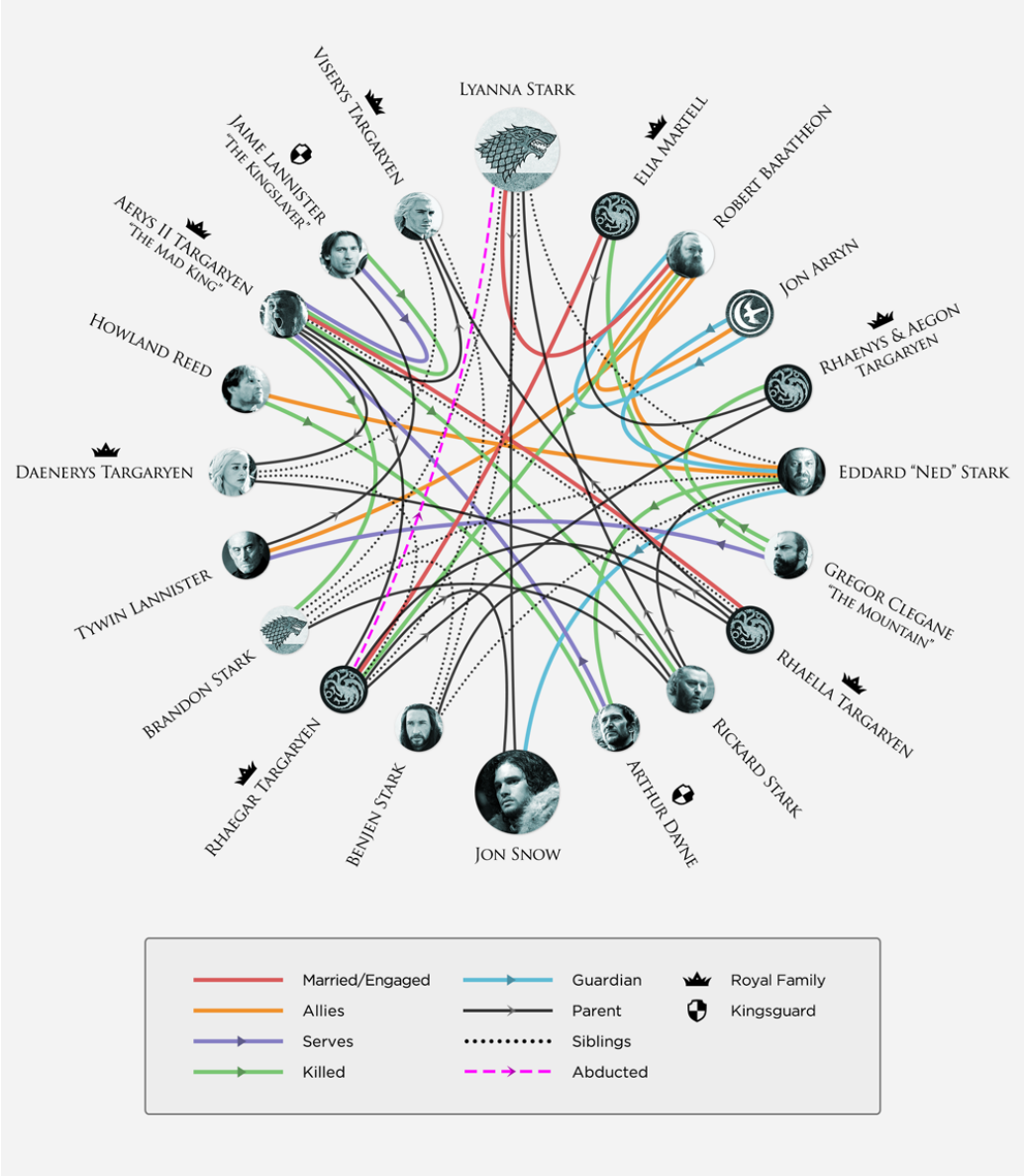
DATASET Statistics	
Nodes	262111
Edges	1234877
Nodes in largest WCC	262111 (1.000)
Edges in largest WCC	1234877 (1.000)
Nodes in largest SCC	241761 (0.922)
Edges in largest SCC	1131217 (0.916)
Average clustering coefficient	0.4198
Number of triangles	717719
Fraction of closed triangles	0.09339
Diameter (longest shortest path)	32
90-percentile effective diameter	11

AMAZON EXAMPLE

- Live demo on social network of Amazon products

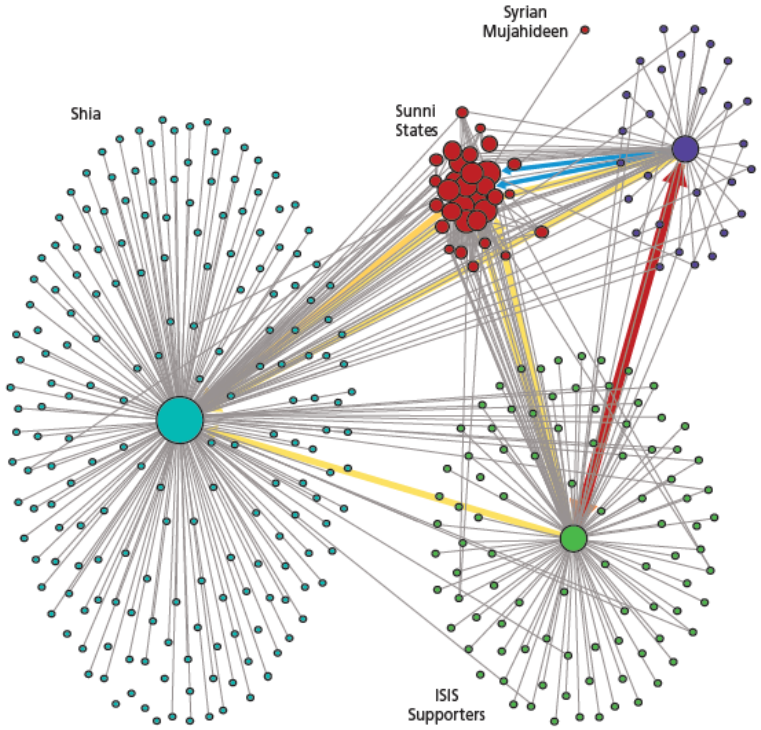
Section 9: More Examples

GAME OF THRONES EXAMPLE



ISIS ON TWITTER EXAMPLE

Figure 3.2
Community of Communities (Metacommunities) Network

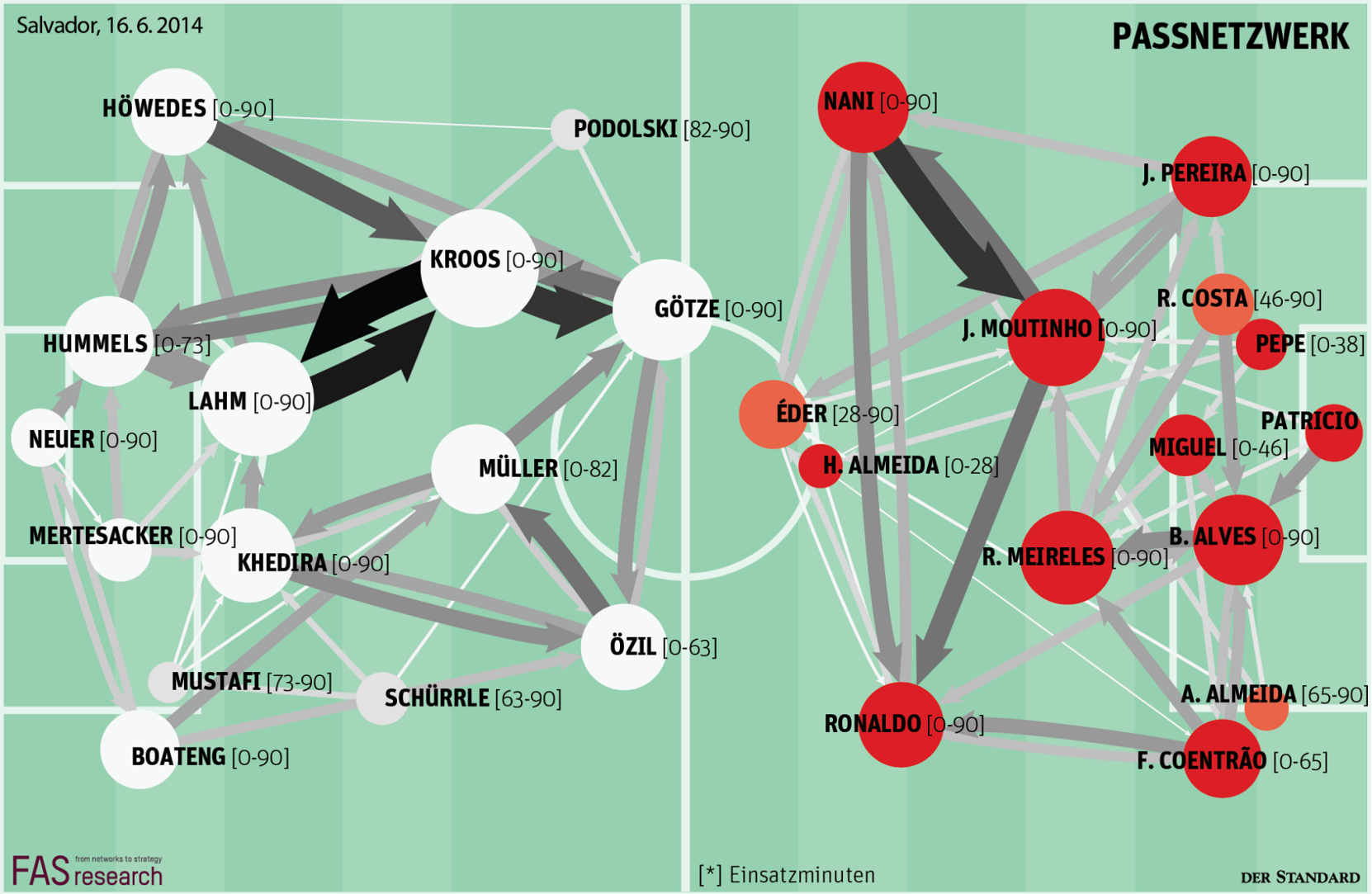


Full Article: http://www.rand.org/pubs/research_reports/RR1328.html

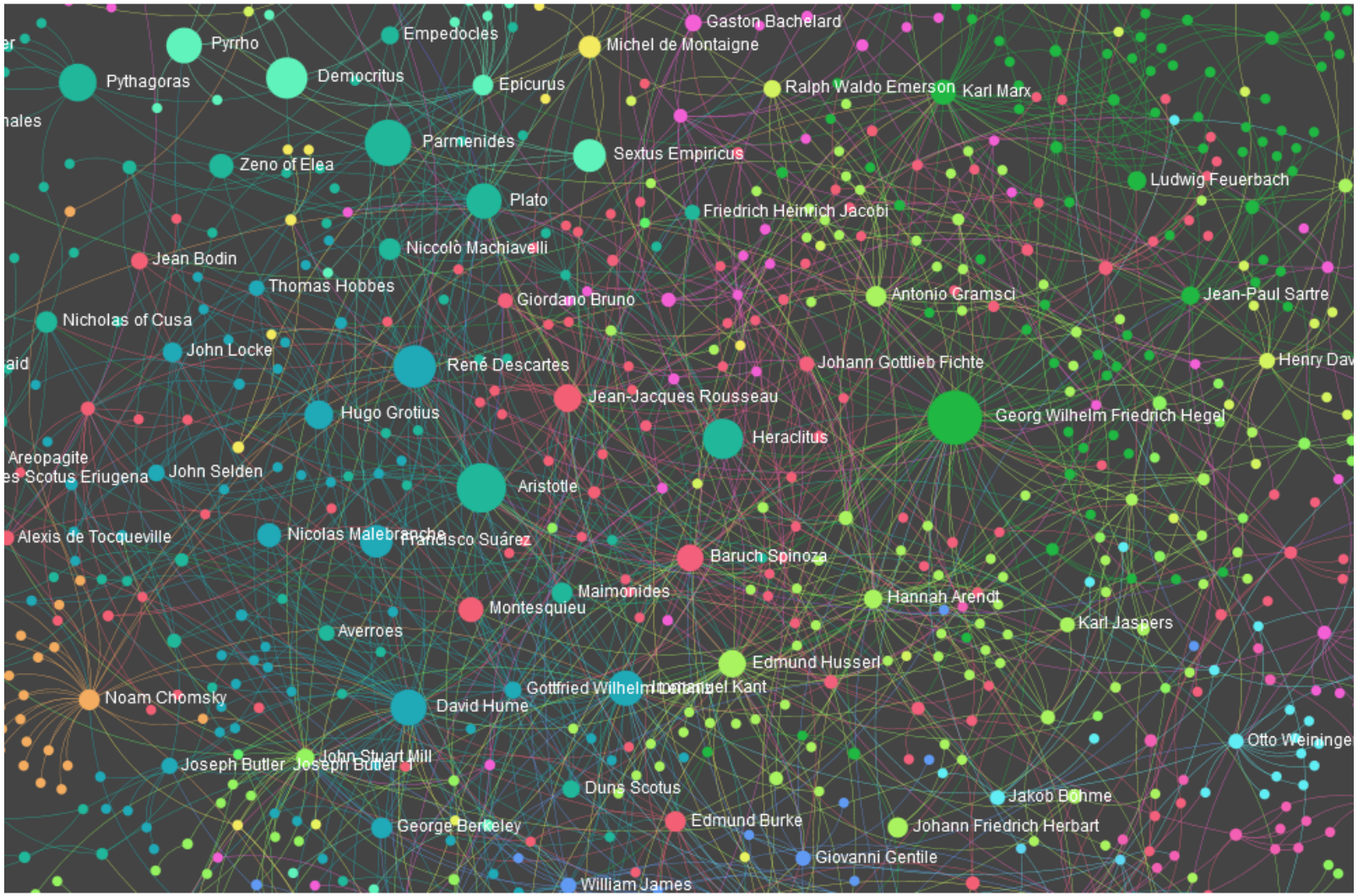
KREMLIN'S TWITTER BOT CAMPAIGN

Full article: https://globalvoices.org/2015/04/02/analyzing-kremlin-twitter-bots/

SNA ON PASSES IN FOOTBALL



SNA ON RELATIONS IN PHILOSOPHY



Interactive version: http://www.designandanalytics.com/philosophers-gephi/

Section 10: Live Demo - Analyzing Enron dataset

ENRON

- Enron: was an American energy, commodities, and services company based in Houston, Texas.
- in 2000, $111 billion revenue
- America’s Most Innovative Company for 6 years in a row (1996-2001)
- At the end of 2001, it was revealed that reported financial condition was sustained by systematic, and creatively planned accounting fraud

ENRON DATASET

- After the investigation, the emails and information collected were deemed to be used for historical research and academic purposes
- One of the only publicly available mass collections of real emails (typically bound by numerous privacy and legal restrictions)
- Expanded corpus, containing over 1.7 million messages, is now available on Amazon S3 for easy access
- We will use random sample with 20000 messages

ENRON DATASET

- Data are in unstructured form
- Data parsing and cleaning is required before SNA methods are applied

SUMMARY

- Important stuff to remember:
  I. Concepts of Social Network Analysis
  II. R packages for Social Network Analysis
  III. SQL (not related to SNA)
  IV. Identify calling circle of one particular person (could be use to target additional people for investigation)
  V. What is regex

Thanks for your attention

Juraj Kapasny