Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube
Recommendation Algorithms

Josephine B. Schmitt[1], Diana Rieger[2], Olivia Rutkowski[1], & Julian Ernst[1]

1 Chair for Communication and Media Psychology, University of Cologne, Germany
2 Institute for Media and Communication Studies, University of Mannheim, Germany

In order to serve as an antidote to extremist messages, counter-messages (CM) are placed in the same online environment as extremist content. Often, they are even tagged with similar keywords. Given that automated algorithms may define putative relationships between videos based on mutual topics, CM can appear directly linked to extremist content. This poses severe challenges for prevention programs using CM. This study investigates the extent to which algorithms influence the interrelatedness of counter- and extremist messages. By means of two exemplary information network analyses based on YouTube videos of two CM campaigns, we demonstrate that CM are closely—or even directly—connected to extremist content. The results hint at the problematic role of algorithms for prevention campaigns.

Keywords: Information Network Analysis, YouTube, Counter-messages (CM), Algorithms, Extremist Messages, Selective Exposure.

doi:10.1093/joc/jqy029

Introduction

Facing the increasing threat by extremist actors worldwide, societal concerns are rising about the importance of the Internet as a distribution channel of extremist messages (i.e., propaganda, hate speech, conspiracy theories). Although the volume of extremist messages online cannot easily be quantified—among others due to their rapid appearance and disappearance (The Swedish Media Council, 2014)—Internet users may come across a large amount of this content (Costello, Hawdon, Ratliff, & Grantham, 2016; Rieger, Frischlich, & Bente, 2013). For instance, the potential of encountering extremist ideas online more than tripled between 2013 and 2015 (Kaakinen, Oksanen, & Räsänen, 2018). While in 2013 about 17% of Internet users aged 15 to 30 reported being exposed to extremist messages, in 2015, it was more than 60%. A German study

Corresponding author: Josephine B. Schmitt; e-mail: josephine.schmitt@uni-koeln.de

revealed that 81% of under-24-year-old online users had already experienced hate speech online (LFM NRW, 2016). A representative study conducted in 2016 among 14- to 19-year-olds showed that about 40% had been exposed to extremist content via video platforms such as YouTube (Reinemann, Nienierza, Riesmeyer, Fawzi, & Neumann, 2018). Ahmed and George (2017) demonstrated that not only overtly violent language or jihadi vocabulary generated search results in Google with extremist material, but benign, apolitical, and non-violent language also facilitated access to websites promoting violence and extremist ideologies. It is therefore important to investigate how closely extremist content is linked to other, inconspicuous content.

Traditionally, research explained the likelihood of encountering information through media with the selective exposure paradigm (Hart et al., 2009; Knobloch-Westerwick & Meng, 2009). According to the paradigm, we usually select attitude-consistent information. However, in digital media environments, recommendations by other entities, such as our friends or algorithms, also have an important influence on selection decisions (Courtois & Timmermans, 2018; Pariser, 2011). For instance, on YouTube, algorithms undertake an important organizing and gatekeeping function. They "do not merely transmit content, but filter it (…) thereby making the content more relevant to its potential consumers" (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015, p. 460). By doing this, they define putative relationships between videos and automatically link them based on similar catchphrases, catchphrases (for example, see Davidson et al., 2010), and provide endorsement for a relationship. Hence, it might be important to integrate the role of automatic recommendations into selection processes in the digital environment.

Although endorsements via social media can encourage people to engage with information they would usually ignore—such as articles from ideologically-incongruent sources (Messing & Westwood, 2014)—this may pose severe problems, especially for younger users. They lack the competence to critically evaluate and reflect online content, and to differentiate between "good" and "evil" sources (Sonck, Livingstone, Kuiper, & de Haan, 2011). When messages are endorsed due to algorithms, "they may be more difficult to discount, as they are unlikely to be seen as overtly partisan" (Bode & Vraga, 2015, p. 623). That is, in the context of extremist content, algorithmic "recommendation" could disguise ideological partisanship as they make content appear "related."

In order to serve as an "antidote" to extremist messages "on site" (Neumann, 2013, p. 7), prevention actors aim at spreading *anti*-extremist messages in the same environment in which extremist messages occur. This includes counter-message (CM) campaigns, that actively intend to counter extremist ideas—often distributed and promoted via social media channels such as YouTube. On YouTube, CM are often tagged with similar keywords as extremist messages (e.g., by the common keyword "Islam"), or they explicitly address extremist actors by their names (e.g., by referring to the so-called Islamic State [ISIS]) and, respectively, include extremist narratives in their titles. The potential linkage of extremist messages and CM may pose serious challenges for CM as prevention activities: An *anti*-extremist message—be it a

single video or a part of a larger campaign—could automatically (e.g., through similar meta-data) increase the likelihood of encountering extremist material (Zhou et al., 2016). Once accessing extremist content via YouTube, users are likely to be redirected to further extremist videos—"potentially leading to immersion in an extremist ideological bubble" (O'Callaghan et al., 2015, p. 473).

Against this background, it seems necessary to ask *to what extent* these algorithms have an impact on the interrelatedness of CM and extremist messages on YouTube, and *how* they affect the likelihood for users of CM to come across videos with extremist content. To answer these questions, we collected data of two exemplary CM campaigns, and of videos that are *related* to the campaign videos by YouTube recommendation algorithms. We built information networks and analyzed them regarding the interconnectedness of the videos within the respective network. However, before we elaborate on the role and potential effects of algorithms on the user and his or her choices, we will shed some light on the definitions of extremist messages as well as the concept that underlies CM, their aims and impact. Finally, we aim to contribute to theoretical considerations of integrating the role of recommendation algorithms into the selective exposure paradigm.

Extremist messages: definition, shapes and strategies

A single definition of extremism is difficult to formulate as the word *extreme* is often considered as an ideology being not "in the middle" of society. The definition of "the middle" is subject to cultural and societal norms and changes (Breton, Galeotti, Salmon, & Wintrobe, 2002; Sotlar, 2004). However, scholars agree upon a description of extremism as a desire to radically, and if necessary, forcefully and violently impose a political and/or religiously motivated ideology that has a "claim to totality in the sense of true interpretation" (Kemmesies, 2006, p. 11). Messages used to promote extremism appear in different "problematic" shapes, which may overlap each other, such as: (a) hate speech, (b) conspiracy theories, and (c) propaganda. According to Meibauer (2013), *hate speech* includes insults, abusive language and designations that devalue members of certain societal or demographic groups, as well as minorities (e.g., religious groups). *Conspiracy theories* can be defined as "a proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons, the conspirators, acting in secret" (Keeley, 1999, p. 116). *Propaganda* can be defined as "a systematic form of purposeful persuasion that attempts to influence the emotions, attitudes, opinions, and actions of specified target audiences for ideological, political or commercial purposes through the controlled transmission of one-sided messages" (Nelson, 1996, p. 232 f.). That is, the definition of what "extremist online material" is does not necessarily entail violence, unconstitutionalities or indictable topics.

In terms of the distribution of extremist messages, mainly right-wing and Islamist extremists use the Internet (Schmitt, Ernst, Frischlich, & Rieger, 2017). For spreading their beliefs, they predominantly rely on social media channels due to the ease of distributing ideas and messages rapidly to a large audience, and the

reachability of young audiences in particular (Gottfried & Shearer, 2016). By connecting their messages to terms relevant for younger age groups, borrowing marketing strategies from popular media culture such as games or music videos, using the "wolf-in-sheep's clothing" tactic, and non-violent communication strategies, extremists aim at addressing these younger users in particular (Jugendschutz.net, 2015a, 2015b; The Swedish Media Council, 2014). Their online messages might therefore not all be extremist as defined by the criminal code, but can be considered as "problematic" in the sense that they comprise hate speech, conspiracies and propaganda. In order to work against a potential influence of extremists, security agencies, civil education as well as youth prevention actors aim at providing a counter-voice online, for example through CM.

Counter-messages: definition, shapes and strategies

On a very general level, CM can be defined as *positive* messages directed against extremist ideologies, core elements of ideologies, or violent extremist behavior and are aimed at "helping people to see through the propaganda and misconception techniques of extremists" (Briggs & Feve, 2013, p. 9). Most of them target the public at large, known as primary prevention.[1] Although CM are delivered in an array of different formats (e.g., text, speech, pictures) many senders focus on audio-visuals; however, it seems to be much harder to find CM online than extremist messages - even when searching for relevant keywords (Rieger, Morten, & Frischlich, 2017).

Depending on the sender, the outlets and target groups, scopes and scales of CM campaigns as well as "typical users" differ largely. *Jigsaw's The Redirect Method*, for example, aims at redirecting users, susceptible to ISIS propaganda, who actively enter ISIS-related search terms, to CM which are "debunking recruitment narratives" (redirectmethod, 2016, p. 2). During an eight-week pilot study, The Redirect Method reached more than 300,000 individuals. However, in line with research on persuasion and attitude change, it can be assumed that this method is most effective in dissuading only those individuals who still have doubts regarding the ISIS narratives, but not those who are already convinced (Compton, 2013). In contrast, the campaign #NotInMyName by the U.K.-based *Active Change Foundation* targets the public. It reached its audience due to social media's snowball effect and broad mass media coverage. The *No Hate Speech Movement*, which is composed of national campaigns in over 40 countries, engages in various online and offline activities (e.g., YouTube videos, seminars, youth events), which makes it impossible to quantify the number of people reached. This campaign focuses on "the public at large and Internet users, with specific attention given to young users" (Council of Europe, 2017).

CM are released in order to function as an antidote to the potential effects of extremist messages. Nevertheless, research concerning the effectiveness of narrative persuasion and CM found mixed results (for an overview, see Braddock & Dillard, 2016; Frischlich, Rieger, Morten, & Bente, 2017; Hemmingsen & Castro, 2017).

Further, extremist messages and CM cannot be considered as independent from each other. Avoiding propaganda and fostering the distribution and effectiveness of CM at the same time is a challenging endeavor, since many CM rely on or deconstruct extremist ideologies and, thus, may repeat extremist narratives. Further, CM may evoke hateful or even extremist comments themselves (see e.g., Ernst et al., 2017), and campaigns such as The Redirect Method are based on the idea that after searching for ISIS-related terms, people are redirected to matching CM. These examples go hand-in-hand with the fact that in order to "improve" peoples' online experience and to provide a "customized experience" (Lazer, 2015, p. 1090), an increasing number of online applications (e.g., YouTube) rely on algorithms. This way, CM may also be technically linked to extremist content. Although algorithms act behind the scenes without being noticed, they are able to shape online users' realities and choices (Saurwein, Just, & Latzer, 2015).

Algorithmic selective exposure: how algorithms may shape realities and choices

In digital media environments, selective exposure does not solely refer to an individual choosing one piece of information over another, but also refers to selection due to automated algorithms. Algorithms, which may be defined as a "finite set of rules that gives a sequence of operations for solving a specific type of problem" (Knuth, 1997, p. 4)—are responsible for the type, diversity, and relevance of information we encounter on various online platforms. They rely on a complex interplay of computational decisions (e.g., based on the reputation of websites, keywords, location). These decisions sometimes reflect biases of programmers and data sets (Rainie & Anderson, 2017)—and human behaviors (Lazer, 2015; Saurwein et al., 2015). Algorithms can be quite helpful by attempting to predict which future choices the user will enjoy the most (Nguyen, Hui, Harper, Terveen, & Konstan, 2014). Based on previous user decisions, for instance, they "recommend" films or books a user may like, or restaurants close to the place the user is located at that moment.

However, this personalization may also have negative consequences for the information people encounter online (Helberger, Karppinen, & D'Acunto, 2018; Lazer, 2015; Saurwein et al., 2015), as it narrows the set and diversity of information over time (Nikolov, Oliveira, Flammini, & Menczer, 2015). Algorithms of mainstream search engines (e.g., Google) for example select the type and order of a certain subset of web pages. By doing this, they define the relevance or even the importance of information (Carlson, 2018; Hannak et al., 2013). On Facebook, people may never see all their friends' posts or their "liked" pages. Information is algorithmically evaluated, sorted and selected based on a calculated likelihood of interestingness (Lazer, 2015) due to the kind of network the people are part of (Bakshy, Messing, & Adamic, 2015). Moreover, algorithms filter information with regard to the political attitudes of the users (Bakshy et al., 2015; Flaxman, Goel, & Rao, 2016).

Having said that, there are serious concerns that this kind of personalized communication may have a negative impact on the public sphere and democratic

opinion-forming processes (e.g., Borgesius et al., 2016; Pariser, 2011). Being constantly confronted with a certain opinion could make individuals perceive this opinion as the majority opinion (Wojcieszak, 2009) and reduce their willingness to express dissenting opinions (Neubaum, 2016). Thus, a limitation of perspectives and ideas may foster polarization and adoption of more extreme attitudes (Stroud, 2010), as well as a misperception of facts about current events (Kull, Ramsay, & Lewis, 2003). Further, if this reduced set of opinions is prejudiced, it can contribute to an increased prejudice in society (Arendt, 2017), and to fragmentation (Bright, 2018). In the worst case, mass dissemination of extremist ideas could influence the societal discourse and silence moderate voices in the long-term. On the contrary, individuals with more extreme attitudes have more pronounced tendencies for selective exposure than people with moderate attitudes (Stroud, 2010). This is often attributed to the increased certainty they have in their beliefs (Wojcieszak, 2009). Likewise, Costello et al. (2016) recently demonstrated that, due to personalization of the online experience, people with a tendency toward anti-governmental attitudes are more likely to get in touch with extremist content.

While personalization on news sites seems not to be very common yet, it is an important feature of social media platforms such as YouTube (Borgesius et al., 2016). Adolescents, as the primary target group of propaganda, are heavy users of YouTube (Gottfried & Shearer, 2016). Besides extremist actors, various governmental and civil initiatives—which aim to counter extremist ideas by publishing CM videos—try to take advantage of this fact and use YouTube as an important distribution channel of their content. Hence, from a societal counter extremist perspective, it seems important to shed some light on the role automation may play for the coexistence of problematic extremist messages and CM on YouTube.

The role of algorithms on YouTube for the interconnectedness of videos

By personalizing "recommendations," algorithms essentially influence the selection—and finally also the potential reception—of content on YouTube, as they "help users find high quality videos related to their interests" (Davidson et al., 2010, p. 293). On YouTube, algorithms define relations between videos based on the interconnectedness of video producers, channels and videos, similar catchphrases, the user's own activity data and activity data of "similar" users (for an overview, see Davidson et al., 2010). The resulting recommendations of related videos may have an important impact on the user and his or her behavior on the platform. Based on an analysis of YouTube videos, Zhou, Khemmarat, and Gao (2010) found that, besides the search function, the related video recommendation was the main source responsible for video views. Figueiredo, Benevenuto, and Almeida (2011) support these findings demonstrating that YouTube's internal referrer system is a key mechanism through which users reach content.

Different consequences for users are conceivable concerning the potential interconnectedness of extremist messages and CM on YouTube. According to the selective exposure paradigm, automated algorithms provide users with the chance to

find videos that match themes and objects of their interest or attitudes. If they are interested in videos that challenge a certain (extremist) ideology or aspects of it, they may encounter more relevant—attitude-consistent—information. Moreover, as algorithms are based on mutual keywords, they may increase the likelihood for users to come across videos with a contrary message, just because of thematic congruence (e.g., tagged with "jihad"). In this way, algorithmic linkage could provide a broader set of attitudes and viewpoints (see Bode & Vraga, 2015). If an extremist message uses similar catchphrases (e.g., caliphate, jihad) as a CM video, the interconnectedness between both videos could be high. In this case, the positive possibility to foster manifold perspectives could turn into a scenario in which extremist perspectives are promoted. This does not seem to be unlikely, as there is a potential imbalance between the amount of propaganda material and CM (RAN, 2015). Producers of CM seem to be less active in producing and publishing content in comparison to extremist actors (Bartlett & Krasodomski-Jones, 2015). Against this background, we wonder which role algorithms play in the relationship between extremist messages (i.e., hate speech, propaganda, and conspiracy theories[2]) and CM on YouTube. Thus, we raise the following research questions:

RQ1: How are CM connected to videos promoting problematic extremist ideas?
RQ2: How close are CM related to other CM on YouTube?

Method

Data collection

In order to answer our research questions, we selected two exemplary CM campaigns that have been published within the last two years: (a) The campaign #WhatIS published by the German Federal Agency of Civic Education (Bundeszentrale für politische Bildung [bpb][3]) consists of eight videos tagged with #WhatIS. In these videos, popular German YouTubers explain selected concepts (e.g., caliphate, haram) that may arise in the context of public debates about Islam and Islamism. These videos were published on the respective YouTuber's channels. According to the bpb (2016), the campaign's aim is to counter distorted perceptions of Muslim life in Germany. One video, for example, explains and compares the different meanings of the concept "jihad", which is wrested from its complex religious context by both Islamist and right-wing extremists, (b) The campaign ExitUSA, run by the US-based non-profit organization *Life After Hate*, is part of an exit/outreach program. The program aims at helping individuals to leave white supremacist groups in the United States. Moreover, it provides support for former members of these groups. Within the campaign, four videos have been published on the YouTube channel ExitUSA. They were designed to "discredit far-right extremist groups, "sow the seeds of doubt" in far-right extremist individuals, and promote their exit program" (Silverman, Stewart, Amanullah, & Birdwell, 2016, p. 16).

We aimed at obtaining robust results for both campaigns. Thus, we selected: (a) successful campaigns from two different countries—Germany and the United States,(b) that are cohesive and complete regarding their content, (c) that address two different topics: (1) countering right-wing extremism, and (2) Islamist extremism, and, (d) that focuses on different target groups. From a methodological perspective, the applied data collection tool was found to work more reliably with smaller amounts of videos to handle. Therefore, we decided to choose campaigns with a smaller range of videos (#WhatIS = eight videos; ExitUSA = four videos).

Videos of both campaigns were treated as seeds for data collection. We used the online tool YTDT Video Network (Rieder, 2015). This retrieves for each list of seeds a list of "related videos"[4] and their metadata (e.g., video ID, video title, URL) from YouTube's application programming interface (API) endpoint, to collect relevant network data (see also, Google Developers, 2017). Resources are first sorted based on their relevance, then in reversed chronological order based on the date they were created, their rating (highest to lowest), title (alphabetically) and view count (highest to lowest number of view counts; Rieder, 2017). Data collection took place on 10 March 2017. Browser history and cookies were deleted before crawling data, to reduce biased results due to the researchers' own search history. Table 1 gives an overview about descriptive data regarding the seed videos. Beginning with the eight seeds (#WhatIS) and, respectively, four seeds (ExitUSA), we collected data of "related" videos with a crawl depth[5] = 2.

Procedure

For each campaign, an information network was built; nodes represent videos, edges represent the relationships between videos. We visualized network data with the software Gephi (Version .9.1; Bastian, Heyman, & Jacomy, 2009). The network of the campaign #WhatIS consists of 11.954 nodes and 205.738 edges (directed), average degree[6] = 17.211, network density[7] = .001. The network of the ExitUSA campaign consists of 6.699 nodes and 99.377 edges, average degree = 14.385, network density = .002. ForceAtlas2 was used as visualization method. ForceAtlas2 is a force-directed layout that simulates a physical system to illustrate the spatial structure of the data. Thereby, "nodes repulse each other like charged particles, while edges attract their nodes, like springs (…). The position of a node cannot be interpreted on its own; it has to be compared to the others" (Jacomy, Venturini, Heymann, & Bastian, 2014, p. 2). Nodes connected by numerous edges are situated in the same region of the network; nodes with few relations to other nodes lie wider apart from each other.

To get an overview about the importance and influence of the seeds in the networks, we calculated the Eigenvector centrality (EC) (Bonacich, 1972). EC counts the number of nodes each node is connected to. Moreover, these nodes are weighted according to their centrality; in other words the centrality of a node is a function of the centrality (i.e., importance, well-connectedness) of their neighbors in the

Table 1 Overview About the Seeds' Descriptives for Both Campaigns

Nr.	Name	Dislikes	Likes	EC	Channel title	Number of views	Publication date	Cluster	Indegree[a]	Outdegree[b]
Network #WhatIS										
A	Info Islam: Was bedeutet KALIFAT? [What does "caliphate" mean?]	285	11,268	.1240	FlipFloid	143,264	12.11.2015	6	62	63
B	Islam und Wissen [Islam and knowledge]	336	7,740	.0347	FlipFloid	95,873	16.01.2016	6	24	64
C	Was bedeutet UMMA? [What does "umma" mean?]	226	4,909	.0491	Hatice Schmidt	63,633	12.10. 2015	8	34	65
D	Info Islam: Was bedeutet Dschahiliyya? [What does "jahiliyyah" mean?]	253	2,955	.0824	MrWissen2go	68,621	28.10.2015	1	45	64
E	Info Islam: Was bedeutet halal/haram? [What does "halal/haram" mean?]	86	1,918	.0310	mariemeimberg	26,688	10.01.2016	2	35	59
F	Infos Islam: Was bedeutet Gebiet des Krieges? [What does "dar al-harb" mean?]	65	657	.0437	LetsDenk	9,022	19.12.2015	9	32	65
G	Info Islam: Was bedeutet Bid'a? [What does "bid'a" mean?]	37	363	.0509	KWiNK	6,789	27.11.2015	1	39	59
H	Info Islam: Was bedeutet Dschihad [What does "jihad" mean?]	845	1,474	.1787	datteltäter	38,598	11.12.2015	2	86	60

Nr.	Name	Dislikes	Likes	EC	Channel title	Number of views	Publication date	Cluster	Indegree	Outdegree
Network #ExitUSA										
A	No judgment just help	0	0	.0017	Exit USA	1,021	07.10.2015	0	3	30
B	There is life after hate	4	15	.0047	Exit USA	5,052	06.10.2015	0	10	60
C	Oak creek	4	9	.0017	Exit USA	4,806	06.10.2015	0	3	39
D	The formers	11	17	.0053	Exit USA	7,253	06.10.2015	0	10	65

Note: [a]Number of incoming connections
[b]Number of outgoing connections.

network (Al-Taie & Kadry, 2017). Relative scores are assigned to all nodes in the network. The higher the value, the more the node is connected to other nodes in the network; thus, the more influential is the node.

With regard to the network #WhatIS seed A, a video by LeFloid, one of the most popular German YouTube channels, EC = .12, and seed H, a video by datteltäter, the first German-Muslim satire channel, EC = .18 are the most influential seeds beneath the seeds that built the basis of this network (see Table 1), indicating that these two seeds are the best connected seeds. The ExitUSA-seeds display very low values of EC, ranging from EC = .0017 to EC = .0053 (see Table 1).

Modularity, a measure of the quality of clustering (Newman & Girvan, 2004), was used to assess the videos' kind of interrelatedness in each network. Its values may vary between 0 and 1. A value $M < .4$ hints at a low separation among clusters, between $M = .4$ and $M = .6$ the level of separation may be considered as medium, $M \geq .6$ means a high distinctiveness of the clusters (Himelboim, Smith, & Shneiderman, 2013). Based on modularity measures, nodes and edges were organized into communities. Nodes *within* a community tend to share similar characteristics, information flows freely. Across communities, there is limited connectivity of the nodes in the network.

To analyze the content of the communities, we drew a randomized sample of 30%[8] of all of the respective community-associated videos. Next, we conducted a qualitative analysis of each video and its content individually. Available metadata associated with the videos—namely, their titles, descriptions, and associated keywords is usually limited to what the uploader provided. For example, it is not uncommon to find titles corresponding to file names, like IMG_0815, and the description is often left empty. In addition, problematic content such as hate speech or extremist propaganda is obviously not described and tagged as such. Thus, we manually inspected and screened each video in order to get a more accurate picture of the audio-visual material. Moreover, we categorized it into standard YouTube-categories (Entertainment, Gaming, HowTo & Style, Music, News & Politics, People & Blogs, Pets & Animals, Science & Technology and Sports), as recommended by prior studies (e.g., Filippova & Hall, 2011). Concerning problematic/extremist content, we defined and identified the following categories: conspiracy theories, hate speech, Islamist extremist propaganda (IE), and right-wing extremist propaganda (RE). We developed a categorization of this content based on previous research (e.g., Frankenberger, Glaser, Hofmann, & Schneider, 2015; Hepfer, 2016; Jugendschutz.net, 2015b; O'Callaghan et al., 2015; Table A in the supplementary material gives an overview of the categories). In order to address the problem of the "subjectivity of the coding process" and, as recommended by Elo et al. (2014), one researcher was responsible for the analysis and the other carefully followed up on the categorization and coding process. Divergent opinions were continuously discussed und resolved. The percentage frequency for each category in each community was determined.

Results

#WhatIS

We identified 11 communities (resolution = 5). A modularity value of M = .775 implies a high distinctiveness of the communities (Himelboim et al., 2013). The three largest communities accounted for about 64% of all collected videos (see Figure 1). Table 2 provides a brief characterization of the 11 communities. Table 3 gives a more detailed overview about the content of each community. The three largest communities include a large number of videos addressing extremist views: In Communities 2 and 7 we found a large percentage of videos (Cluster 2: 32.2%; Cluster 7: 31.4%) providing a very strict/radical understanding of Islam. Two seeds are part of the largest community (Community 2), two of the third largest

Figure 1 Communities within the network #WhatIS; 11 modularity classes, resolution = 5.

Table 2 Legend of Figure 1 (Percentage = Share of Videos/Nodes in the Network)

2	Diverse Problematic Content, IE Propaganda & News/Politics	31.60%
7	Diverse Problematic Content, IE Propaganda & People/Blogs	17.37%
1	Diverse Entertaining Content & Problematic Content	15.45%
6	People/Blogs & Diverse Entertaining Content	12.09%
9	Education & Diverse Entertaining Content	8.15%
8	Howto/Style & People/Blogs	7.51%
0	Music & Peoples/Blogs	4.98%
3	Education & Nonprofit/Activism	.82%
10	News/Politics, Nonprofit/Activism & IE Propaganda	.69%
4	Education & Peoples/Blogs	.69%
5	Music	.69%

communities (Community 1; see Table 3 and Figure 1). Nearly half of the videos in Community 2 can be considered as extremist messages (i.e., conspiracy theories, hate speech, IE/RE propaganda). Compared to that, Community 1 includes (as well as about 17% of extremist messages) a wide variety of entertaining videos. However, based on these results, the four seed-videos—D, E, G, and H—may be regarded as closely linked to extremist content (see Table 3 for a more detailed overview about the community composition).

The remaining four videos are part of Communities 6, 8, and 9. Community 6 mainly includes entertaining videos (i.e., for example comedy, music, let's play, film trailers). However, there are also some videos, which can be considered as extremist messages (i.e., conspiracy theories, hate speech)[9]. Community 8 mainly contains videos related to lifestyle topics. There is only one video which we considered as related to extremist Islamist ideas. Besides a large amount of entertaining and lifestyle videos, Community 9 contains a considerable amount of educational videos (27%; e.g., related to physics, psychology, philosophy), diverse entertainment as well as some CM (2.1%). There also are some videos with extremist content (i.e., conspiracy theories, RE propaganda). All other communities in the network mainly consist of entertainment/information-related videos, apart from Community 10 where we can find 12.5% of the videos being related to IE Propaganda. Thus, we can assume, regarding RQ1, that there is a great likelihood for users of videos D, E, G, and H to be exposed to extremist content—even Islamist and right-wing extremist propaganda. Compared to that, seeds A, B, C, and F have fewer connections to such content; given users rely on the recommendations by YouTube, they are less likely to come across extremist video material. Concerning RQ2, we found that other CM—apart from those, which served as seeds—are underrepresented. Community 0 contains the highest share of CM (8.9%). In Communities 1 and 2, the seeds are connected to other CM; however, the number is very small. In Community 2, most of them have been published on the channel datteltäter, on which one of the seed videos has also been published.

Table 3 Overview of the Composition of the 11 Communities (clusters) in the Network of the #WhatIS Campaign

	(in %)										
	C0 (n = 177)	C1 (n = 555)	C2 (n = 1131)	C3 (n = 30)	C4 (n = 24)	C5 (n = 24)	C6 (n = 432)	C7 (n = 621)	C8 (n = 270)	C9 (n = 291)	C10 (n = 24)
Problematic/extremist content											
Conspiracy theories		8.0	2.3				1.2	8.0		2.8	
Hate speech		.2	3.3				.3	5.2			
RE Propaganda	.9	8.5	7.9					2.5		1.1	
IE Propaganda			32.2					31.4	.4		12.5
Counter-messages											
Counter-messages	8.9	.4	1.8				.3		.4	2.1	
Other											
Comedy	5.5	13.1	5.1	7.7			10.0	5.0	1.9	1.8	
Education	2.1	14.4	3.1	46.7	75.0		.3	3.9		27.0	4.1
Entertainment	7.5	6.5	10.7			4.2	12.0	3.5		27.6	
Film & Animation		2.0					10.2	.9		7.6	
Gaming		1.3					10.5			.9	
Howto & Style	8.2	.9					2.8	.9	84.0	5.9	
Music	18.1	2.7	4.4	4.3		91.7	6.8	.2		4.2	
News & Politics	5.6	18.4	15.1	4.3		4.1	3.3	8.4		4.9	66.7
Nonprofits & Activism			.4	2.7				1.9		1.5	
People & Blogs	30.0	19.8	13.4	16.3	25.0		40.8	28.1	13.0	11.8	16.7
Pets & Animals							0.7				
Science & Technology	2.8						.3			.8	
Sports		4.0	.2								
Travel & Events	11.4		.1				.5		.3		

Note: The table is based on samples of 30% we randomly drew for each cluster. Values are rounded.

Thus, it can be concluded that it is unlikely to encounter other videos countering extremist ideas by using videos of the #WhatIS campaign.

ExitUSA

We identified 25 video communities (resolution = 5). A modularity value of M = .860 implies a high distinctiveness of the communities. The three largest communities accounted for about 50% of all collected videos (Figure 2). Table 4 provides a brief characterization of the clusters of the network. Table 5, in turn, gives a more detailed overview.

All four-seed videos are part of the largest community (Community 0; see Table 4 and Figure 2). This community contains a wide range of topics. Besides a large amount of entertaining videos dealing with topics related to lifestyle, celebrities, music and film, there are various videos containing (political) information and expressions of political opinions (without extremist content; see also Table 5). Moreover, results show that about 4.6% of the videos in Community 0 can be considered as extremist propaganda (4.5% RE propaganda, 0.1% IE propaganda). The remaining communities in the network mainly contained entertaining videos, ranging from topics related to animals, lifestyle topics, and music, as well as educational content. We found only four other communities with extremist content: In Community 7, we identified 79.2% of the videos dealing with conspiracy theories; Community 11—albeit rather underrepresented compared with other topics in this community—covers videos containing hate speech (13%); whereas, about half of

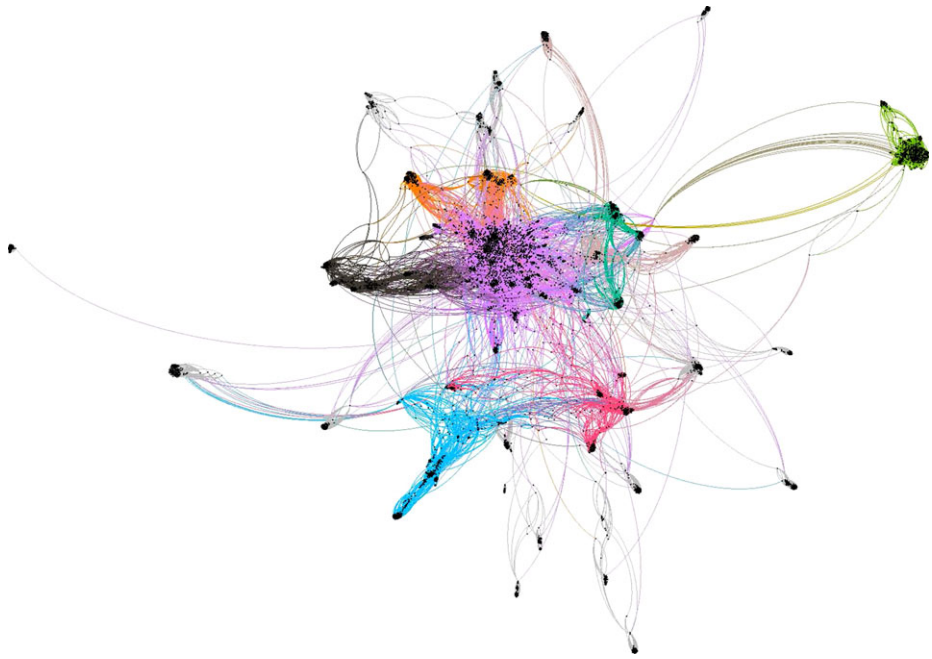

Figure 2 Communities within the network ExitUSA; 25 modularity classes, resolution = 5.

Table 4 Legend of Figure 2 (Percentage = Share of Videos/Nodes in the Network)

0	Entertainment, News & People/Blogs	38.08%
2	People & Blogs (I)	8.27%
14	Film & Animation, Gaming, & Music	6.90%
18	Diverse Problematic Content & People/Blogs	6.78%
6	News, Politics, & People/Blogs	5.48%
22	Music (I)	5.34%
21	Education & People/Blogs	4.64%
19	Music & Entertainment	3.39%
11	Entertainment, Education, Hate Speech	2.72%
16	Entertainment	2.07%
13	News & Politics/Nonprofit & Activism	1.88%
4	Music (II)	1.42%
7	Conspiracy Theories	1.31%
9	Music (III)	1.22%
10	People & Blogs (II)	1.19%
15	Gaming & People/Blogs	1.16%
23	Entertainment & Gaming	1.07%
24	Pets & Animals	1.04%
8	Music & Nonprofit/Activism	1.03%
3	Entertainment, People, & Travel/Events	.97%
20	Education	.96%
5	Travel & Events	.94%
12	Entertainment	.90%
17	News & Politics	.85%
1	Nonprofit & Activism	.37%

the videos in Community 18 can be regarded as extremist content. We categorized more than 18% of the videos as RE propaganda, 14.2% IE propaganda, 5.3% as hate speech, and 10.5% as conspiracy theories. Thus, with regard to RQ1, we can conclude that—although the relative amount of extremist content is rather small and the seeds cannot be considered as very well-connected based on their EC (see Table 1)—there is a certain likelihood for users to encounter extremist messages. Other CMs—apart from those that served as seeds for the network—are underrepresented. Only Community 0 contains further CMs. Although, these CMs are directly related to the seed videos, CMs are in stark competition with a much higher amount of extremist content. Thus, with regard to RQ2 it can be concluded that—based on the mere share of videos—it is quite unlikely to encounter other videos countering extremist ideas by using a video of the ExitUSA campaign.

Discussion

CM are ascribed a huge potential in the context of extremism prevention, radicalization intervention and online youth protection (e.g., Szmania & Fincher, 2017).

Table 5 Overview of the Composition of the Communities (clusters) in the Network of the ExitUSA (Part I) Campaign

(in %)	C0 (n = 757)	C1 (n = 9)	C2 (n = 164)	C3 (n = 20)	C4 (n = 30)	C5 (n = 30)	C6 (n = 111)	C7 (n = 24)	C8 (n = 21)	C9 (n = 24)	C10 (n = 24)	C11 (n = 54)	C12 (n = 18)
Problematic/extremist content													
Conspiracy theories	1.1							79.2					
Hate speech	1.3											13.0	
RE Propaganda	4.5						.9						
IE Propaganda	.1												
Counter-messages													
Counter-messages	3.4												
Other													
Autos & Vehicles				4.8									
Comedy	2.5						1.8					24.1	
Education	3.1						.9						
Entertainment	18.5		9.1	28.6	10.0		9.9	4.2	9.5	8.3	8.3	35.2	94.4
Film & Animation	7.5			4.8			1.8						5.6
Gaming	.7		.6										
Howto & Style	1.6		4.8	4.8		5.6							
Music	6.7				90.0		3.6		42.9	91.7		16.7	
News & Politics	14.0	11.1					49.6	12.5				1.9	
Nonprofits & Activism	5.9	88.9					4.5	4.2	38.1				
People & Blogs	27.4		85.5	14.3			22.5		9.5		87.5	9.3	
Pets & Animals	.4												
Sports	.3						.9						
Travel & Events	.9			42.9		94.4	3.6				4.2		

(in %)	C13 (n = 39)	C14 (n = 138)	C15 (n = 24)	C16 (n = 42)	C17 (n = 18)	C18 (n = 136)	C19 (n = 69)	C20 (n = 18)	C21 (n = 93)	C22 (n = 109)	C23 (n = 21)	C24 (n = 21)
Problematic/extremist content												
Conspiracy theories						10.4						
Hate speech						5.2						

(*Continued*)

Table 5 Continued

(in %)	C13 (n = 39)	C14 (n = 138)	C15 (n = 24)	C16 (n = 42)	C17 (n = 18)	C18 (n = 136)	C19 (n = 69)	C20 (n = 18)	C21 (n = 93)	C22 (n = 109)	C23 (n = 21)	C24 (n = 21)
RE Propaganda						18.5						
IE Propaganda						14.1						
Counter-messages												
Counter-messages												
Other												
Autos & Vehicles		.7										
Comedy		.7				7.4			2.2			
Education	5.1					7.4		100.0	32.3			
Entertainment	2.6	5.1	4.2	100.0		1.5	31.9		16.1	3.7	68.7	
Film & Animation	2.6	25.6	4.2				2.9					
Gaming		28.3	75.0			7.4					33.3	
Howto & Style	5.1	1.5							4.3			
Music		37.0				2.2	63.8		1.1	96.3		
News & Politics	43.6	.7			100.0	2.2						
Nonprofits & Activism	41.0		4.2						17.2			
People & Blogs		.7	12.5			43.7	1.5		25.8			
Pets & Animals									1.1			100.0
Sports												
Travel & Events												

Note: Table is based on samples of 30% we randomly drew for each cluster. Values are rounded.

However, the content that people encounter online no longer solely depends on their individual selections, but also on algorithms that feature certain content, for instance because of mutual keywords, (Zhou et al., 2016). Thus, all attempts at using social media to spread CM might run the risk of guiding people to extremist, or at least problematic material, due to a certain thematic congruence. This relationship might limit or prevent an individual's exposure to only attitude-consistent information (e.g., Garrett, 2009) and increase the engagement with ideologically incongruent sources (e.g., Messing & Westwood, 2014). Although the latter might be useful in order to extend or respectively overcome individual filter bubbles, in the case of extremism prevention, it might increase the risk of encountering and promoting anti-democratic and extremist ideas.

The present study aimed at investigating the interconnectedness of YouTube videos, which target *countering* extremist ideas, and videos, which *seek to promote* them. We built information networks based on data sets of two exemplary CM campaigns and: (a) videos that YouTube considers to be directly "related" to them (Crawl Depth 1, "first click"), and (b) videos that are "related" to the latter (Crawl Depth 2, "second click") collected by means of the YTDT Video Network tool (Rieder, 2015). We regionally grouped each network's videos regarding their interrelatedness. Based on a content analysis of the resulting communities, we demonstrated that extremist content—however, rather the subtle, non-indictable content—*might* be closely connected with CM.

Both campaigns differ regarding the amount and diversity of (extremist) content to which they may relate. These differences may be due to a structural difference: While the videos of the campaign #WhatIS are part of the participating YouTubers' channels[10]—and, therefore, connected to a more diverse set of videos—the videos of ExitUSA are all published on the same channel. We further find the CM videos by ExitUSA grouped in *one* community. They are mainly connected with diverse entertainment and information-related videos and, only to a lesser extent, with extremist content (i.e., mainly right-wing propaganda). Nevertheless, even though a large number of the extremist videos is not part of the same community as the ExitUSA CM seeds—and, thereby, more loosely connected with them—people can easily be confronted with them within two clicks via the YouTube recommendations.

Presumably, due to the distribution of videos of #WhatIS over different communities, they have a higher variability in the type of videos they are connected to—also with different kinds of extremist messages. We found a remarkable number of connections of the seeds with IE propaganda videos. This can be explained by the thematic overlap of the keywords and topics used (e.g., "jihad"), which especially hint at the risks and challenges of YouTube's recommendation system for users and the problematic role of automated algorithms in the context of CM campaigns. Since algorithms can produce relations or endorsements from CM to extremist material, they make it more difficult to discount problematic content and could prevent extremist material to be seen as overtly partisan (Bode & Vraga, 2015). This poses serious challenges for political communication, democratic opinion-forming

and the society as a whole. From a theoretical lens, selections based on recommendation algorithms could lead to different effects than those typically found in selective exposure research (e.g., Knobloch-Westerwick & Meng, 2009) since not only attitude-consistent information is presented but also opposing content (if it is, for instance, connected through common keywords).

From a (optimistic) countering violent extremism (CVE) respectively preventing violent extremism (PVE) perspective, one could also argue that the interrelatedness could work the other way around: The search or selection of extremist messages may lead to finding more CM on YouTube. Due to the extensive publication activities of extremist actors, there is an unfavorable imbalance to the disadvantage of CM: There are far more active (and effective) extremist actors/organizations than organizations/actors who publish CM (Berger, 2016). This is also mirrored in our networks: Other CM than those, which served as seeds are underrepresented. This imbalance seems to be even more problematic as the YouTube relevance-algorithms do not necessarily rely on popularity metrics (e.g., views, likes), but seem to feature channels with high activity in terms of video publication (Rieder, Matamoros-Fernández, & Coromina, 2018). Thus—although we practically lack this information in our study—it seems likely that extremist content, which we found to be "related" to CM, published on very "active" channels might appear further up in YouTube's list of recommended videos. This, in turn, may increase the likelihood of getting in contact with extremist content, as users often tend to rely on the ranking provided by search engines—independently from the senders' trustworthiness (Kammerer & Gerjets, 2014). On YouTube, automated recommendations even were found to be the main reason to click on a video (Figueiredo et al., 2011; Zhou et al., 2010). Moreover, CM as prevention or countering activities are not effective per se. They can even be problematic, mainly because: (a) they can often be considered as user-generated content, thus, journalistic quality requirements cannot be applied, (b) they address the same problematic topics as extremist messages by sometimes even repeating the problematic arguments, and (c) many of them tend to use humor or satire as means, which run the risk of not being understood by everyone or even to evoke reactance (see Rutkowski, Schötz, & Morten, 2017). We did not include actual user data; thus, we do not have any information on the actual audience, the information flow and the videos' effects.

Future research could benefit from a multi-methodological approach. By combining these network data with survey data and/or data of actual video users, scholars could get deeper insights into audience flows and how possible interconnections between videos are perceived by online users—for instance, how they relate to the emergence of boomerang effects or, in case of more biased users, to hostile media effects. Further, apart from this homogeneous network data, it is also conceivable to include data from other sources such as Twitter or Facebook in order to get a broader picture of interrelations of extremist and counter messages. This is even more relevant as research on audience fragmentation provides evidence for audience duplication across media outlets (Webster & Ksiazek, 2012). Moreover, this kind of research

should be extended to further CM campaigns in order to analyze possible differences in credibility, trustworthiness, and authenticity of different CM senders. In the context of CVE/PVE, there is a need to use credible voices (Cherney, 2016). Especially in times of a "crisis in trust" in political institutions (see e.g., Foster & Frieden, 2017, p. 511), it seems necessary to shed light on the effectiveness of state-led CM campaigns as well as campaigns published by actors from civil society, social media influencers, and so forth.

Limitations and future perspectives

Although the present research makes a noteworthy contribution to research on the interrelatedness of extremist messages and CM, it is important to mention some more limitations and come up with implications for future studies. First, our analysis is limited to two specific CM campaigns collected at one point in time. Although we found comparable patterns in both networks, the resulting networks are unique and exemplary concerning these aspects. Future studies could aim at accumulating data at different points in time in order to compare resulting networks (e.g., Courtois & Timmermans, 2018). Moreover, data was collected to a crawl depth of 2, meaning that only those "related" videos were included, which were directly related to the seeds (first level, "first click") as well as those videos that were related to the "related" videos on the first level ("second click"). That is, further linkages between videos were not taken into account. Thus, although we found comparable patterns, the results of this study are not generalizable to CM campaigns in general or overall linkages between videos. Here again, studies including a longitudinal perspective could address this limitation.

Using the YTDT Video network tool, we have to rely on what YouTube defines as "related" and provides via the data API (Google Developers, 2017). This fact does not only influence the research presented here but also the "average" user's daily usage of YouTube and the "recommendations" he or she receives. Although many users know about the existence of algorithms that influence behaviors and selections, the concrete underlying mechanisms are subject to the utmost secrecy.

Further, the present networks are based on data of a "blank prototype" of a user, meaning that browser history and cookies were deleted before data collection (nevertheless, the influence of metadata on the computer, on which data was collected, probably cannot be entirely avoided). Of course, a "regular" user would not encounter the same conditions. This procedure was chosen due to two—in our view—even more confounding considerations. First, our browser history as reseachers in the field might have a very specific pattern. It is not representative for the "average user" as we might have, for example, looked for extremist content (e.g., for scientific purposes) more often. This higher frequency could imply even more ease of finding this type of content when starting with a CM. Second, the individual browser history would have created a highly idiosyncratic precondition. This would thus have created a unique and researcher-biased condition that would not have been representative either. Future studies could focus on simulating different "user types,"

with their particular conditions in terms of browser history and cookies and how this influences the interrelatedness of extremist messages and CMs.

Although our results indicate a certain likelihood of encountering extremist content when starting from one of the videos that are part of the CM campaigns, results differ between the networks. Some videos are related to extremist content, whereas others are not. Therefore, we cannot argue that algorithms are problematic per se. Future research should focus on the investigation of concrete factors that lead to a connection of CM and extremist videos on YouTube (e.g., the role of common keywords). Further, it should take the social media structures and dynamics into account. In this regard, research should also shed light on the role of user behavior and the behavior of presumably "similar" users, as these are important criteria upon which algorithms are based.

Practical and theoretical implications

Needless to say, the found and assumed relationships and consequences, discovered thus far, are mainly based on the idea of "average" and unobtrusive social media users. However, it seems plausible to assume that younger users especially are not aware of the existence of algorithms, and if they are, they may not be able to oversee the power of these algorithms in influencing their media selection and recommendation. From a theoretical perspective, the results of the present study underline the importance to add an "algorithmic dimension" to selective exposure research as well as to media effects. Theoretical ideas on the filter bubble (Pariser, 2011) could be informed by the possibility of algorithms to increase the likelihood of finding attitude-consistent information as well as the potential to connect unrelated or even contradictory information through keyword linkage. The current study adds to this discussion by providing information on the potentials of recommendation algorithms to connect problematic material to users' YouTube fare.

Concerning potential effects, theories suggest bigger effects for attitude-consistent media fare. Although the potential effects of unintentional exposure to online political communications are rather small, especially when the message is inconsistent with the recipient's prior attitudes (Bowyer, Kahne, & Middaugh, 2017), social media's distinctive features such as "recommendations" of "related" content may be interpreted as social cues (i.e., others seem to like that, so I could like that too!; Messing & Westwood, 2014). This could facilitate access to and acceptance of rather unexpected messages. Taking the example of the effects of extremist messages, although it is nearly impossible to quantify the degree to which extremist narratives influence radicalization processes, research underlines the *potential* of narrative persuasion and the psychological appeal of themes inherent in extremist narratives (Braddock & Dillard, 2016; Braddock & Horgan, 2016). Especially, online propaganda seems to be able to consolidate preexisting extremist beliefs (von Behr, Reding, Edwards, & Gribbon, 2013; Wojcieszak, 2009). Thus, the discussed potential interrelatedness can be even more dangerous for users that are susceptible to extremist narratives (Ribeau, Eisner, & Nivette, 2017). Relatedly, research on the

hostile media effect found that people with strong preexisting attitudes are more likely to perceive alternative viewpoints as biased and hence filter them out (Kim, 2011). Like this, algorithmic linkage to extremist content could contribute to polarization processes, foster extreme attitudes (see e.g., Bright, 2018) and even make a positive effect of CMs more unlikely—especially for already susceptible individuals (Valkenburg & Peter, 2013). Although the current study did not test these assumptions, future studies should more directly address these concerns and investigate the effects of algorithmic recommendation, for instance regarding the consequences of heterogeneous networks (see e.g., Huckfeldt, Morehouse Mendez, & Osborn, 2004).

With regard to a more practical perspective, this study has an important implication. Pedagogical guidance to frame online CM campaigns is strongly needed: Based on the present analysis and previous work on the role of social media features (Ernst et al., 2017), we conclude that the exposure to CMs may be tainted with risks. Thus, online and offline PVE and CVE efforts *should be combined* in order to successfully counter the negative effects of extremist messages (see above). Besides strengthening peoples' social cognitive resilience to violent extremism (see e.g., Aly, Taylor & Karnovsky, 2014), they should aim to develop a comprehensive knowledge and deeper understanding of the functional principles of social media and foster a critical understanding of manipulating messages themselves as well as the role of the Internet as a distribution channel (Rieger et al., 2017).

Notes

1 Primary prevention describes "organized programs for reducing the incidence (rate of new cases) of a disorder in a defined population" (Caplan & Caplan, 2000, p.131). In the case of countering violent extremism (CVE) respectively preventing violent extremism (PVE) measures, primary prevention means mainly prevention of news cases of radicalization especially within youths.

2 Videos containing explicitly violent content contradict the rules of YouTube and have usually been deleted by the platform. Thus, they are not part of the sample.

3 The *German Federal Agency of Civic Education* is a public institution pursuing the provision of "[…] citizenship education and information on political issues for all people in Germany" (Bundeszentrale für politische Bildung, 2012).

4 On YouTube, videos are regarded as "related" based on the interconnectedness of video producers, channels, and videos, similar catchphrases, the user's own activity data and activity data of "similar" users as well as co-visitation counts (for a detailed overview, see Davidson et al., 2010).

5 Crawl depth (CD) specifies how far from the seeds the script should go. Crawl depth = 0 will get only the relations between the seeds; CD = 1 determines the relation between the seeds and directly-related videos (recommendations on first level, see also red frame in Figure 1), CD = 2 collects videos on the second level—in other words, videos that are related to the collected videos on the first level (CD = 1).

6 Average number of relationships (edges) between the nodes in the network; graphs were treated as directed.

7 Ratio of the number of links present to the maximum number of links possible. Values may vary between 0 and 1. Density values are high, if nodes are highly interconnected with one another.
8 We decided for 30% to get a sample size small enough to work with (i.e., to conduct a qualitative categorization), but balanced and big enough to be representative for the total cluster. Codes and categorization of the videos can be provided upon request.
9 As YouTube aims at combatting overtly violent extremist and terrorist content (YouTube, 2017), these messages may be regarded as rather subtle extremist content, which is not indictable.
10 The eight relevant videos were published on seven different channels.

References

Ahmed, M., & George, F. L. (2017). *A war of keywords: How extremists are exploiting the internet and what to do about it*. London: Tony Blair Institute for Global Change.

Al-Taie, M. Z., & Kadry, S. (2017). *Python for graph and network analysis*. Cham: Springer International Publishing.

Aly, A., Taylor, E., & Karnovsky, S. (2014). Moral disengagement and building resilience to violent extremism: An education intervention. *Studies in Conflict and Terrorism, 37*, 369–385. doi:10.1080/1057610X.2014.879379

Arendt, F. (2017). Impulsive facial-threat perceptions after exposure to stereotypic crime news. *Communication Research, 44*, 793–816. doi:10.1177/0093650214565919

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*, 1130–1132. doi:10.1126/science.aaa1160

Bartlett, J., & Krasodomski-Jones, A. (2015). *Counter speech. Examining content that challenges extremism online*. Retrieved from https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Bastian, M., Heyman, S., & Jacomy, M. (2009). *Gephi*. Retrieved from https://gephi.org/

von Behr, I., Reding, A., Edwards, C., & Gribbon, L. (2013). *Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism*. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR453/RAND_RR453.pdf

Berger, J. M. (2016). *Making CVE work: A focused approach based on process disruption*. Retrieved from https://www.icct.nl/wp-content/uploads/2016/05/J.-.pdf

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication, 65*, 619–638. doi:10.1111/jcom.12166

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology, 2*, 113–120. doi:10.1080/0022250X.1972.9989806

Borgesius, J. F. Z., Trilling, D., Moeller, J., Bodó, B., Vreese, D., Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review. Journal of Internet Regulation, 5*. doi:10.14763/2016.1.401

Bowyer, B. T., Kahne, J. E., & Middaugh, E. (2017). Youth comprehension of political messages in YouTube videos. *New Media & Society, 19*, 522–541. doi:10.1177/1461444815611593

Braddock, K., & Dillard, J. P. (2016). Meta-analytic evidence for the persuasive effect of narratives on beliefs, attitudes, intentions, and behaviors. *Communication Monographs, 83*, 446–467. doi:10.1080/03637751.2015.1128555

Braddock, K., & Horgan, J. (2016). Towards a guide for constructing and disseminating counternarratives to reduce support for terrorism. *Studies in Conflict & Terrorism, 39*, 381–404. doi:10.1080/1057610X.2015.1116277

Breton, A., Galeotti, G., Salmon P., & Wintrobe, R. (2002). *Political extremism and rationality*. Cambridge: Cambridge University Press.

Briggs, R., & Feve, S. (2013). *Review of programs to counter narratives of violent extremism: What works and what are the implications for government?* Retrieved from https://www.counterextremism.org/resources/details/id/444/review-of-programs-to-counter-narratives-of-violent-extremism-what-works-and-what-are-the-implications-for-government

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication, 23*, 17–33. doi:10.1093/jcmc/zmx002

Bundeszentrale für politische Bildung. (2012). *Strengthening democracy—Fostering a civil society. The federal agency for civic education: Our mission and activities*. Retrieved from http://www.bpb.de/die-bpb/138853/our-mission-and-activities

Bundeszentrale für politische Bildung. (2016). *Begriffswelten Islam* [*Concepts of Islam*]. Retrieved from http://www.bpb.de/lernen/digitale-bildung/medienpaedagogik/213243/webvideoformate-begriffswelten-islam

Caplan, G., & Caplan, R. B. (2000). The future of primary prevention. *The Journal of Primary Prevention, 21*, 131–136, doi:10.1023/A:1007062631504

Carlson, M. (2018). Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. *New Media & Society, 20*, 1755–1772. doi:10.1177/1461444817706684

Cherney, A. (2016). Designing and implementing programmes to tackle radicalization and violent extremism: lessons from criminology. *Dynamics of Asymmetric Conflicts, 9*, 82–94. doi:10.1080/17467586.2016.1267865

Compton, J. (2013). Inoculation theory. In J. P. Dillard & L. Shen (Eds.), *The Sage handbook of persuasion: Developments in theory and practice,* Vol. 2 (pp. 220–237). Thousand Oaks, CA: Sage.

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior, 63*, 311–320. doi:.1016/j.chb.2016.05.033

Council of Europe. (2017). *No hate speech youth campaign*. Retrieved from https://www.coe.int/en/web/no-hate-campaign/about-the-campaigns

Courtois, C., & Timmermans, E. (2018). Cracking the tinder code: An experience sampling approach to the dynamics and impact of platform governing algorithms. *Journal of Computer-Mediated Communication, 23*, 1–16. doi:10.1093/jcmc/zmx001

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., … Sampath, D. (2010). The YouTube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 293–296). New York: ACM. doi:10.1145/1864708.1864770

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative content analysis: A focus on trustworthiness. *Sage Open, 4*(1), 1–10. doi:10.1177/2158244014522633

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., & Roth, H.-J. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization, 10*, 1–49.

Figueiredo, F., Benevenuto, F., & Almeida, J. M. (2011). The tube over time: Characterizing popularity growth of YouTube videos. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 745–754). New York: ACM. doi:10.1145/1935826.1935925

Filippova, K., & Hall, K. B. (2011). Improved video categorization from text metadata and user comments. In *Proceedings of the 34th International ACM SIGIR Conference on Research and development in Information Retrieval* (pp. 835–842). New York: ACM, SIGIR'11.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly, 80*, 298–320. doi:10.1093/poq/nfw006

Foster, C., & Frieden, J. (2017). Crisis of trust: Socio-economic determinants of Europeans' confidence in government. *European Union Politics, 18*, 522–535. doi:10.1177/1465116517723499

Frankenberger, P., Glaser, S., Hofmann, I., & Schneider, C. (2015). Islamismus im Internet. Propaganda—Verstöße—Gegenstrategien [*Islamism on the Internet. Propaganda—Violations—Counter Strategies*]. Retrieved from http://www.hass-im-netz.info/fileadmin/dateien/pk2015/Islamismus_im_Internet.pdf

Frischlich, L., Rieger, D., Morten, A. & Bente, G. (Eds.), (2017). Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand [*Videos against extremism? Counternarratives put to the test*]. Wiesbaden: Bundeskriminalamt.

Garrett, R. K. (2009). Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of Communication, 59*, 676–699. doi:10.1111/j.1460-2466.2009.01452.x

Google Developers. (2017). *Search: list: YouTube Data API*. Retrieved from https://developers.google.com/youtube/v3/docs/search/list

Gottfried, J., & Shearer, E. (2016). *News use across social media platforms 2016*. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/13/2016/05/PJ_2016.05.26_social-media-and-news_FINAL-1.pdf

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 527–538). New York: ACM. doi:10.1145/2488388.2488435

Hart, W., Albarracin, D., Eagly, A. H., Lindberg, M., Lee, K. H., & Brechan, I. (2009). Feeling validated vs. being correct: A meta-analysis of exposure to information. *Psychological Bulletin, 135*, 555–588. doi:10.1037/a0015701

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society, 21*, 191–207. doi:10.1080/1369118X.2016.1271900

Hemmingsen, A.-S., & Castro, K. I. (2017). *The trouble with counter-narratives*. Retrieved from http://pure.diis.dk/ws/files/784884/DIIS_RP_2017_1.pdf

Hepfer, K. (2016). *Verschwörungstheorien: Eine philosophische Kritik der Unvernunft* [*Conspiracy theories: Philosophical criticism on irrationality*]. Bonn: Bundeszentrale für politische Bildung.

Himelboim, I., Smith, M., & Shneiderman, B. (2013). Tweeting apart: Applying network analysis to detect selective exposure clusters on Twitter. *Communication Methods and Measures, 7*, 195–223. doi:10.1080/19312458.2013.813922

Huckfeldt, R., Morehouse Mendez, J., & Osborn, T. (2004). Disagreement, ambivalence, and engagement: The political consequences of heterogeneous networks. *Political Psychology, 25*, 65–95. doi:10.1111/j.1467-9221.2004.00357.x

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One, 9*, e98679. doi:10.1371/journal.pone.0098679

Jugendschutz.net. (2015a). *Kinder als Instrument dschihadistischer Propaganda* [*Children as instrument of jihadist propaganda*]. Retrieved from https://www.jugendschutz.net/fileadmin/download/pdf/IS_Kinder_2015.pdf

Jugendschutz.net. (2015b). *Rechtsextremismus online beobachten und nachhaltig bekämpfen* [*Observing right-wing extremism online and countering it in a sustained manner*]. Retrieved from http://www.hass-im-netz.info/fileadmin/dateien/PM2015/bericht2014.pdf

Kaakinen, M., Oksanen, A., & Räsänen, P. (2018). Did the risk of exposure to online hate increase after the November 2015 Paris Attacks? A group relations approach. *Computers in Human Behavior, 78*, 90–97. doi:10.1016/j.chb.2017.09.022.

Kammerer, Y., & Gerjets, P. (2014). The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *International Journal of Human-Computer Interaction, 30*, 177–192. doi:10.1080/10447318.2013.846790

Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy, 96*, 109–126. doi:10.2139/ssrn.1084585

Kemmesies, U. (2006). *Terrorismus und Extremismus—der Zukunft auf der Spur. Forschungsstand zum Phänomenfeld des islamischen Extremismus und Terrorismus* [*Terrorism and extremism—On the trace for the future. State of research on the phenomenon field of Islamic extremism and terrorism*]. München: Luchterhand.

Kim, K. (2011). Public understanding of the politics of global warming in the news media: The hostile media approach. *Public Understanding of Science, 20*, 690–705. doi:10.1177/0963662510372313

Knobloch-Westerwick, S., & Meng, J. (2009). Looking the other way: Selective exposure to attitude-consistent and counter-attitudinal political information. *Communication Research, 36*, 426–448. doi:10.1177/0093650209333030

Knuth, D. E. (1997). *The art of computer programming: Fundamental algorithms* (3rd ed.). Bonston, MA: Addison-Wesley.

Kull, S., Ramsay, C., & Lewis, E. (2003). Misperceptions, the media, and the Iraq War. *Political Science Quarterly, 118*, 569–598. doi:10.1002/j.1538-165X.2003.tb00406.x

Lazer, D. (2015). The rise of the social algorithm. *Science, 348*, 1090–1091. doi:10.1126/science.aab1422

LFM NRW. (2016). *Ethik im Netz—Hate Speech* [*Ethics in the Internet—Hate speech*]. Retrieved from http://www.lfm-nrw.de/fileadmin/user_upload/lfm-nrw/Service/

Veranstaltungen_und_Preise/Medienversammlung/2016/EthikimNetz_Hate_Speech-PP.pdf

Meibauer, J. (2013). Hassrede—von der Sprache zur Politik. [Hate speech—from language to politics]. In J. Meibauer (Ed.), Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion [*Hate Speech. Interdisciplinary contributions to a recent discussion*], (pp. 1–16). Gießen: Gießener Elektronische Bibliothek.

Messing, S., & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research, 41*, 1042–1063. doi:10.1177/0093650212466406

Nelson, R. A. (1996). *A chronology and glossary of propaganda in the United States*. Westport, CT and London: Greenwood Press.

Neubaum, G. (2016). *Monitoring and expressing opinions on social networking sites-empirical investigations based on the spiral of silence theory*. Retrieved from http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-42707/Neubaum_Diss.pdf

Neumann, P. R. (2013). Radikalisierung, Deradikalisierung und Extremismus [Radicalization, de-radicalization and extremism]. *Aus Politik und Zeitgeschichte, 63* (29–31), 3–10.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 026113.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 677–686). New York: ACM. doi:10.1145/2566486.2568012

Nikolov, D., Oliveira, D. F. M., Flammini, A., & Menczer, F. (2015). Measuring online social bubbles. *PeerJ Computer Science, 1*, e38. doi:10.7717/peerj-cs.38

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (white) rabbit hole the extreme right and online recommender systems. *Social Science Computer Review, 33*, 459–478. doi:10.1177/0894439314555329

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Penguin.

Rainie, L., & Anderson, J. (2017). *Code-dependent: Pros and cons of the algorithm age*. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/14/2017/02/08181534/PI_2017.02.08_Algorithms_FINAL.pdf

RAN. (2015). *Counter narratives and alternative narratives*. Retrieved from https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/networks/radicalisation_awareness_network/ran-papers/docs/issue_paper_cn_oct2015_en.pdf

redirectmethod. (2016). *The redirect method. A blueprint for bypassing extremism*. Retrieved from https://redirectmethod.org/

Reinemann, C., Nienierza, A., Riesmeyer, C., Fawzi, N., & Neumann, K. (2018). Verdeckter Extremismus, offener Hass? [*Conveiled extremism, open hate?*]. Baden-Baden: Nomos.

Ribeau, D., Eisner, M. & Nivette, A. (2017). *Können gewaltbereite extremistische Einstellungen vorausgesagt werden?* [*Can we predict violent extremist attitudes?*] Retrieved from http://www.media.uzh.ch/dam/jcr:41381576-3db2-4b9a-bb04-6464c538be16/Forschungsmemo.pdf

Rieder, B. (2015). *YTDT video network*. Retrieved from https://tools.digitalmethods.net/netvizz/youtube/mod_videos_net.php

Rieder, B. (2017). *YouTube-Data-Tools*. Retrieved from https://github.com/bernorieder/YouTube-Data-Tools/blob/master/mod_videos_net.php

Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (2018). From ranking algorithms to "ranking cultures": Investigating the modulation of visibility in YouTube search results. *Journal of Research into New Media Technologies, 24*, 50–68. doi:10.1177/1354856517736982

Rieger, D., Ernst, J., Schmitt, J. B., Vorderer, P., Bente, G., & Roth, H.-J. (2017). Medienpädagogik gegen Extremismus? Propaganda und Gegenentwürfe im Internet. [Media literacy education against extremism? Propaganda and counter-messages online] *merz | medien + erziehung, 3*, 27–35.

Rieger, D., Frischlich, L., & Bente, G. (2013). *Propaganda 2.0. Psychological effects of right-wing and Islamic extremist Internet videos*. Cologne: Luchterhand.

Rieger, D., Morten, A., & Frischlich, L. (2017). Verbreitung und Inszenierung [Dissemination and production]. In L. Frischlich, D. Rieger, A. Morten, & G. Bente (Eds.), Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand [*Videos against extremism? Counternarratives put to the test*] (pp. 47–80). Wiesbaden: Bundeskriminalamt.

Rutkowski, O., Schötz, R., & Morten A. (2017). Subjektives Erleben [Subjective experience]. In L. Frischlich, D. Rieger, A. Morten, & G. Bente (Eds.), Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand [*Videos against extremism? Counternarratives put to the test*] (pp. 47–80). Wiesbaden: Bundeskriminalamt.

Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: options and limitations. *Digital Policy, Regulation and Governance, 17*, 35–49. doi:10.1108/info-05-2015-0025

Schmitt, J. B., Ernst, J., Frischlich, L. &. Rieger, D. (2017). Rechtsextreme und islamistische Propaganda im Internet: Methoden, Auswirkungen und Präventionsmöglichkeiten. [Right-wing and Islamist online-propaganda: Methods, effects and prevention]. In R. Altenhof, S. Bunk, & M. Piepenschneider. (Hrsg.), Politischer Extremismus im Vergleich [*Political extremism by comparison*] (pp. 171–208). Berlin-Münster-Wien-Zürich-London: LIT Verlag Dr. W. Hopf.

Silverman, T., Stewart, C. J., Amanullah, Z., & Birdwell, J. (2016). *The impact of counter-narratives. Insights from a year-long cross-platform pilot study of counter-narrative curation, targeting, evaluation and impact.* Retrieved from https://www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE_1.pdf

Sonck, N., Livingstone, S., Kuiper, E., & de Haan, J. (2011). *Digital literacy and safety skills*. Retrieved from http://eprints.lse.ac.uk/33733/1/Digital%20literacy%20and%20safety%20skills%20%28lsero%29.pdf

Sotlar, A. (2004). *Some problems with definition and perception of extremism within society*. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/Mesko/208033.pdf

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication, 60*, 556–576. doi:10.1111/j.1460-2466.2010.01497.x

Szmania, S., & Fincher, P. (2017). Countering violent extremism online and offline. *Criminology & Public Policy, 16*, 119–125. doi:10.1111/1745-9133.12267

The Swedish media council. (2014). *Pro-violence and anti democratic messages on the Internet*. Retrieved from https://www.statensmedierad.se/publikationer/publicationsinenglish/proviolenceandantidemocraticmessagesontheinternet.605.html

Valkenburg, P. M., & Peter, J. (2013). The differential susceptibility to media effects model. *Journal of Communication, 63*, 221–243. doi:10.1111/jcom.12024

Webster, J. G., & Ksiazek, T. B. (2012). The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of Communication, 62*, 39–56. doi:10.1111/j.1460.2466.2011.01616.x

Wojcieszak, M. (2009). "Carrying online participation offline"—Mobilization by radical online groups and politically dissimilar offline ties. *Journal of Communication, 59*, 564–586. doi:10.1111/j.1460-2466.2009.01436.x

YouTube. (2017). *An update on our commitment to fight violent extremist content online*. Retrieved from https://youtube.googleblog.com/2017/10/an-update-on-our-commitment-to-fight.html

Zhou, R., Khemmarat, S., & Gao, L. (2010). The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement* (pp. 404–410). New York: ACM. doi:10.1145/1879141.18791

Zhou, R., Khemmarat, S., Gao, L., Wan, J., Zhang, J., Yin, Y., & Yu, J. (2016). Boosting video popularity through keyword suggestion and recommendation systems. *Neurocomputing, 205*, 529–541. doi:10.1016/j.neucom.2016.05.002