Social Network Analysis and Mining (2022) 12:27
https://doi.org/10.1007/s13278-021-00818-z

ORIGINAL ARTICLE

Disrupting networks of hate: characterising hateful networks and removing critical nodes

Wafa Alorainy[1,3] · Pete Burnap[1] · Han Liu[4] · Matthew Williams[2] · Luca Giommoni[2]

Received: 3 June 2020 / Revised: 12 November 2020 / Accepted: 6 September 2021 / Published online: 31 January 2022

Abstract

Hateful individuals and groups have increasingly been using the Internet to express their ideas, spread their beliefs and recruit new members. Understanding the network characteristics of these hateful groups could help understand individuals' exposure to hate and derive intervention strategies to mitigate the dangers of such networks by disrupting communications. This article analyses two hateful followers' networks and three hateful retweet networks of Twitter users who post content subsequently classified by human annotators as containing hateful content. Our analysis shows similar connectivity characteristics between the hateful followers networks and likewise between the hateful retweet networks. The study shows that the hateful networks exhibit higher connectivity characteristics when compared to other "risky" networks, which can be seen as a risk in terms of the likelihood of exposure to, and propagation of, online hate. Three network performance metrics are used to quantify the hateful content exposure and contagion: giant component (GC) size, density and average shortest path. In order to efficiently identify nodes whose removal reduced the flow of hate in a network, we propose a range of structured node-removal strategies and test their effectiveness. Results show that removing users with a high degree is most effective in reducing the hateful followers network connectivity (GC, size and density), and therefore reducing the risk of exposure to cyberhate and stemming its propagation.

Keywords Network analysis · Cyberhate · Online hate · Hate diffusion · Hate prevention · Node removal · Network disruption

1 Introduction

Individuals and groups have increasingly used the Internet to express their ideas, spread their beliefs and recruit new members (Lee and Leets 2002). As with offline hate crime, online hate speech (or cyberhate) posted on social media has become a growing social problem. In 2016 and 2017, the UK's decision to leave the European Union, and a string of terror attacks, was followed by noticeable and unprecedented increases in cyberhate (Williams et al. 2019), with a rhetoric of invasion, threat and otherness (Alorainy et al. 2019). Some research suggests that the perpetrators of cyberhate have similar motivations to those who resort to violence offline (Williams et al. 2019; Awan 2014; Chan et al. 2016; Awan and Zempi 2017). Social psychologists have suggested that the perpetrators of hate crime may be influenced by their perception that certain groups pose a threat to them (Stephan and Stephan 2013), and (Glaser et al. 2002) suggest that racists often express their views more freely on the Internet than elsewhere. Thus, as online social media enables

✉ Wafa Alorainy
alorainyws@cardiff.ac.uk; waloraini@su.edu.sa

Pete Burnap
burnapp@cardiff.ac.uk

Han Liu
han.liu@szu.edu.cn

Matthew Williams
williamsm7@cardiff.ac.uk

Luca Giommoni
giommonil@cardiff.ac.uk

[1] School of Computer Science and Informatics, Cardiff University, Cardiff, UK

[2] School of Social Sciences, Cardiff University, Cardiff, UK

[3] College of Science and Humanities, Shaqra University, Shaqraa, Kingdom of Saudi Arabia

[4] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

individuals and groups to spread ideologies and even advocate hate crime, it is essential to study the online structure, communication and connectivity of online communities in order to determine users' exposure to hateful ideologies that could influence their own views and actions. Twitter, which has become an essential source of timely information, offers a unique opportunity to study social dynamics in online social networks in terms of (1) individuals' exposure to online hate and (2) individuals' role in propagating online hate among groups—more specifically, how we can disrupt the flow of cyberhate and reduce exposure to others, through targeted intervention in the flow of hate. Both of these insights directly respond to the UK's Online Harms whitepaper, which focuses the need to protect citizens online (Online 2020).

The detection of hate online has been widely discussed from the perspective of content analysis. However, the study of hateful networks on social media has received limited attention in the literature. A study of such networks could be valuable in the context of concern about exposure to, and contagion of, online hateful and offensive narratives in social media. On the Twitter platform, the hateful *followers' network* represents the user community directly exposed to hateful content. This network is a subset of users who directly receive information from each other. Furthermore, the hateful *retweet network* is a construct formed by users who propagate cyberhate to their own followers, thereby passing on hateful narratives from the people they follow—a form of cyberhate contagion. Several studies have applied Social Network Analysis (SNA) methods to Twitter hateful networks in order to use connectivity information as an indicator that a user is posting offensive content (Ribeiro et al. 2018; Al-garadi et al. 2016). Others have focused SNA analysis on the retweets network to measure diffusion (Sainudiin et al. 2019; Ribeiro et al. 2018). However, there is yet to be a study of *multiple* hateful networks with the aim of understanding whether there is evidence of similar of 'levels of friendship', and therefore a general exposure to the hate, nor similar levels of propagation behaviour and therefore general contagion effect.

It is important to note that people who are exposed to hateful content won't necessarily spread the hate. However, the exposure to hateful content potentially increases the risk of increasing the number of people adopting hateful ideologies.

Moreover, previous studies are yet to propose intervention methods to prevent cyberhate from spreading. Intervention methods could include the possibility of identifying contagion pathways in hateful networks and evaluating the reduction in exposure of the network's users to receiving hateful content, in the same way we might expect traditional offline virus spreading to be contained. The lack of such a study motivated us to undertake a baseline study that characterises several hateful networks extensively from multiple perspectives—namely exposure to cyberhate (in follower networks); diffusion of cyberhate (in retweet networks); and intervention methods for the curtailing and containment of cyberhate (through network pruning). To the best of our knowledge, this is one of the pioneer studies that have applied SNA to study (1) a range of hateful networks to compare and contrast baseline measures of connectivity and propagation across multiple hateful networks, and (2) evaluate interventions such as prevention strategies (Pastor-Satorras and Vespignani 2002) and disruption strategies (Xu and Chen 2008) that identify nodes in hateful networks which, if removed, would reduce the network connectivity (exposure reduction) and potentially diffuse the hate (contagion reduction).

The research presented in this paper comprises an analysis of Twitter networks containing anonymised accounts whose text was classified by a crowdsourced team of human annotators as containing cyberhate.

1.1 Contributions

- *C1* To the best of our knowledge, this is the first study to understand the connectivity characteristics of two hateful follower networks. The experiments show that the level of the connectivity of the hateful followers' networks is similar, and therefore have common levels of users' exposure to cyberhate. We also compared the hateful networks to another form of 'risky' network (i.e. a suicidal ideation network of similar size) and showed evidence of higher connectivity between the hateful users (higher exposure to the hateful content) compared to the suicidal users. They, however, have less reciprocated friendship behaviour than suicidal users (less connected around the topic).
- *C2* To the best of our knowledge, this is the first study to understand the communication characteristics of three hateful retweets networks. Experiments show a consistently and significantly greater reach of content (contagion), and greater degree of co-operation on the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the comparator "risky" network—suicidal ideation. Hateful content reaches more users in fewer hops.
- *C3* To the best of our knowledge, this is the first study to develop metrics that identify nodes within hateful networks (user accounts) whose removal is empirically shown to reduce the connectivity (largest component, density and average shortest path) in both the follower and retweet networks. Six node-removal strategies based on network connectivity were tested on three network metrics: giant component size, density and the average shortest path. Our experiments show that removing the nodes with the highest network *degree* has the highest impact on reducing the size of the largest component of

the hateful followers' and retweet networks. We demonstrate the rigour of these findings on different hateful networks.

The remainder of the paper is organised as follows. Section 2 describes the related work on this topic. Section 3 describes the methods of the present research, including data collection. Section 4 reports the results, and Sect. 5 presents the discussion and the conclusions.

2 Literature review

2.1 Hateful Networks' characterisation

Social network analysis, particularly on the Twitter platform, has been applied in a range of studies, e.g. student interaction (Stepanyan et al. 2010), quantifying influence on Twitter (Bakshy et al. 2011), political community structure and emotions (Cherepnalkoski and Mozetič 2016; Himelboim et al. 2016). In terms of online hate on the Web, Gerstenfeld et al. (2003) analysed 157 extremist sites and found links between most of these websites, and Zhou et al. (2005) also investigated Web communication and analysed the content and links of hate groups. In their research, they found that the main objective of these websites is to spread and promote ideas, such as white supremacists and Neo-Nazis. Moreover, Chau and Jennifer (2007) used the techniques of Social Network Analysis (SNA) to analyse hate groups on the Internet, formulating hypotheses around the specific features of each site. They showed that the network of bloggers in hate groups is decentralised. Also, they found that the number of "hate" bloggers increased steadily over a number of years.

Recently, Mathew et al. (2019) introduced a study that looked into the diffusion dynamics of posts made by hateful and non-hateful users on Gab. They collect a large dataset of 341K users with 21M posts and investigated the diffusion of the posts generated by hateful and non-hateful users. They observed that the content generated by the hateful users tended to spread faster, farther and reach a much wider audience when compared to the content generated by non-hateful users. Also, an important finding was that hateful users were far more densely connected among themselves compared to non-hateful users.

On Twitter, previous research has been aimed mainly at detecting hateful, offensive, abusive and aggressive speech on the platform using information about the network activity. Such studies analysed user network activity on Twitter to detect cyberhate by considering specific attributes of online activity using machine learning classifiers. An example is Chatzakou et al. (2017) who detected Twitter aggressors and bullies automatically, Ribeiro et al. (2018) who detected hateful users and Ting et al. (2013) who focused on hate group detection. Burnap et al. (2014) specifically looked at retweet virality following a terror attack—a likely trigger event for hateful responses—and found that sentiment expressed in the tweet was statistically significantly predictive of both size and survival of information flows of this nature. Wadhwa and Bhatia (2014) aimed to uncover/identify hidden radical groups in online social networks, providing evidence of the ability to discover subgroups. Ribeiro et al. (2017, 2018) aimed to define a user-centric view of hate speech by examining the difference between user activity patterns and network centrality measurements in the sampled graph. They discovered that hateful users were more central in the retweets network and therefore identifiable as key influencers within the network.

Their study prompted us to understand more about the formation of online hateful communities and how to intervene in an effective way to reduce the spread of hate. To the best of our knowledge, none of the above studies has compared network connectivity metrics across a number of networks to understand baseline metrics of users' exposure to hate and whether this metric is common between hateful networks. Existing work is also yet to investigate similarity in the propagation of hate (contagion) within multiple hateful networks.

2.2 Hateful content prevention

Removing nodes to reduce the network's connectivity (and therefore stem the flow of content) has been widely introduced in previous research. For example, these strategies are used for breaking complex networks (Cunha et al. 2015), the spread of computer viruses (Newman et al. 2002) and spam prevention (Colladon and Gloor 2019). Yip et al. (2012) examined the structural properties of the networks of personal interactions between cybercriminals in carding forums. They found that carding social networks are not scale-free as the degree distributions are log-normal, which has important implications for network disruption. It is widely accepted that scale-free networks are particularly resilient to random node removals, but highly vulnerable to targeted attacks due to there only being a small fraction of nodes possessing the majority of links. In their study, they did not use any node removal strategy, instead using the implication of the degree distributions characteristics. Petersen et al. (2011) examined removing the highest degree nodes based on the distance between the entire network's nodes. Their work proposed a node removal algorithm for a criminal network. As part of their study, they found that removing the high degree nodes had an impact on enlarging the distance between the criminals. Wiil et al. (2010) introduced a study that analysed the importance of links in terrorist networks. This study showed that removing nodes destabilised the network, noting that both the importance of nodes and links should be considered. All of these previous works have been examined on Web fora, which is structurally different to that of the Twitter platform (Kane et al. 2014).

For the Twitter platform, a node removal strategy was applied on political networks and showed that SNA metrics could be used to evidence impact on the network connectivity (Jürgens et al. 2011). Studies specifically aiming to reduce the connectivity in hateful networks by removing the nodes are rare, and no study yet exists looking at online social networks. An attempt by Xu and Chen (2008) found that terrorist networks on the Web were more vulnerable to attack on the bridges that connect different communities, than to attacks on their hubs. They applied two removal strategies on Web sites' networks: hub-based strategy and bridge based. To the best of our knowledge, no study yet exists that examines node removal strategies for hateful Twitter networks.

3 Methods

3.1 Data collection

The study used data from two types of anti-religious content: Anti-Muslim and Anti-Semitic.

3.1.1 Anti-Muslim datasets

In order to collect and analyse hateful communication posted to Twitter, we first needed to identify accounts that were demonstrably posting hateful tweets. For the Anti-Muslim dataset, we collected data from Twitter around two 'trigger' incidents. The first was the murder of Lee Rigby, a solider based in Woolwich, London. Data collection lasted two weeks following the terrorist attack committed on May 23rd 2013; we named this data set 'Anti-Muslim 1'. Data were collected via the Twitter streaming Application Programming Interface (API), based on a manual inspection of the highest trending keyword following the event. The result was $N = 427{,}330$ tweets in this case. The second incident was the *#PunishAMuslimDay* event that took place on April 3rd 2018. The dataset was collected in the aftermath of a letter inciting others to commit violent and aggressive acts towards Muslims. We named this 'Anti-Muslim 2'. The collection spanned two weeks and resulted in $N = 919{,}854$ tweets.

A subsample of 2000 tweets from each dataset was chosen for a human annotation process. Human annotators were asked to label the offensive tweets using the crowd-sourced online service Crowdflower. Annotators were provided with each tweet and asked "Is this text offensive or antagonistic in terms of race, ethnicity or religion?" They were presented with a ternary set of classes: yes, no, undecided. The results from coders could then either be accepted or rejected on the basis of the level of agreement with other coders. We required at least four human annotations per tweet and retained only the annotated tweets for which at least three human annotators (75%) had agreed on the appropriate class as per related work (Thelwall et al. 2010; Burnap et al. 2017). The results of the annotation exercise produced a 'gold standard' dataset of 2000 tweets, with 973 and 1053 instances of offensive or antagonistic content tweets for Anti-Muslim 1 and Anti-Muslim 2 datasets, respectively. Our interest in these data was to flag the Twitter accounts of users posting hateful content. We searched the larger datasets for any duplicates of the annotated hateful tweets (tweets with the same text). This boosted the collection of hateful tweets to 2621 and 2097 tweets for Anti-Muslim 1 and Anti-Muslim 2, respectively. Finally, we extracted the distinct users in these collections, creating a list of 3502 and 8602 user accounts that were involved in creating or propagating hateful content for Anti-Muslim 1 and Anti-Muslim 2, respectively.

Each dataset was split into a followers' dataset (users who follow each other) and a retweet dataset (users who retweet each other). In the followers datasets, for each of the authors of the 3502 and 8602 Tweets classified as containing evidence of possible hateful speech, we retrieved Twitter profile information regarding the lists of followers (of the hateful users) and friends (users who the hateful users followed) so that we could identify the measures of connectivity between these types of user. This collection resulted in two sets of 2,018,950 followers and 1,942,614 friends for a list of 3502 distinct authors, and 3,855,37 followers and 4,977,47 friends for a list of 8602 distinct authors, respectively.

Next we used python tools[1] to generate two types of networks (see Sect. 3.2.1)—first type is (1) a followers' network (based on the followers dataset). This is a directed graph network in which each node has either a following relationship, a friend (followed) relationship or both. Algorithm 1 explains the steps for building a hateful followers network. It shows that we discarded any follower relationship of users who had not been shown to post hateful content. We extracted 1004 users and their 2644 followers who belonged to the Anti-Muslim 1 dataset. For Anti-Muslim 2, we extracted a total of 1073 and their 2895 followers. This led to two datasets of followers that contained the original users and their followers (those exposed to cyberhate)—one for Anti-Muslim 1 followers and the other for Anti-Muslim 2 followers.

The second type is (2) a retweet network. For the retweet datasets, two retweet networks were built: Anti-Muslim 1 with 1229 nodes and 2571 edges, and Anti-Muslim 2 with 5581 nodes and 16,338 edges. Each of these is a directed

[1] https://www.python.org/.

network having a *i* and *j* ∈ retweets' dataset for each node edged from *i* to *j*, indicating that *j* is a retweeter of a tweet posted by *i*.

Algorithm 1 Building the hateful Follower Network

```
Input: INPUT: H=$h_1, h_2, .., h_n$ ▷ users accounts who post hateful content
Output: OUTPUT: hateful Follower Network $H_{followers}$
for each $h_i \in$ H do
    Collect follower network $h_{fo}$ and friends network $h_{fr}$
    for each follow relation $h_i \longleftarrow x_i \in h_{fo}$ do
        if $x_i \in$ H then
            Return adjacency list $H_{followers}[h_i, x_i]$
        else
            discard the relation $h_i \longleftarrow x_i$
    end if
    end for
    for each friend relation $h_i \longleftarrow y_i \in h_{fr}$ do
        if $y_i \in$ H then
            Return adjacency list $H_{followers}[y_i, h_i]$
        else
            discard the relation $h_i \longleftarrow y_i$
    end if
    end for
end for
Return adjacency list $H_{followers}$
```

3.1.2 Anti-Semitic dataset

For the Anti-Semitic dataset, we collected using the COSMOS platform (Burnap et al. 2015).[2],[3] The data used for this analysis included tweets posted between 16/10/2015 and 21/10/2016 and were gathered in real time (this ensures that all tweets are collected). The raw dataset for the complete study period contained 31,282,472 tweets. Human annotation was used again, with workers asked "Is this text offensive or antagonistic related to a Jewish identity" with a yes/no label applied. As per common convention, the Crowdflower human coding task was reviewed and instances where agreement dropped below 75% were dropped from the training data (Thelwall et al. 2010; Burnap et al. 2017). This resulted in 372 anti-semitic tweets. As with the anti-muslim data, the annotated Anti-Semitic dataset was extended by adding duplicates and retweets of the original tweets from the larger data collection. This boosted the annotated dataset to 3874 tweets. For the Anti-Semitism dataset, we did not extract a followers' network because Twitter API did not recognise the relevant users' IDs (possibly removed by Twitter). We did build the retweet network for Anti-Semitism, which consisted of 2748 nodes and 5091 edges.

3.1.3 Comparative "risky" network dataset

Although Twitter networks of different size and nature inevitably show different characteristics, for the purposes of comparison between networks in this study (e.g. do hateful networks exhibit different characteristics to other risky networks?), we selected a similar size network from another "risky" category—one in which users in online social networks risk exposure to ideology and there is concern of the contagion of content. We found that the suicidal network published in Colombo et al. (2016) was similar to our networks from three perspectives: (1) comparable size, (2) similar data collection process (Twitter API), and (3) likely to spread content of concerning ideology i.e a "risky network". As with our hateful datasets, the suicidal network has two sub networks—the followers network and the retweet network. Both networks have similar sizes to our networks. The suicidal content was also labelled by human annotators in the original paper. For the suicidal network, the followers network contains 987 nodes and 2410 edges, whereas the retweet network contains 3209 nodes and 2211 edges.

3.2 General network characteristics

3.2.1 Metrics selection

We built social networks graphs from the datasets of followers and retweets. Then, we extracted metrics and compared them. As discussed in Pržulj (2007), the larger the number of common properties (metrics), the more likely it is that the two networks are similar. The metrics used in this study were selected as follows:

- *Giant component* The Giant Component (GC) is a connected component of a given graph that contains a finite fraction of the entire graph's nodes, e.g. a significant proportion of the nodes are connected in one GC. The GC of the networks was extracted using Depth-first Search and Linear graph algorithms (Tarjan 1972). From a hate spread perspective, the size of the GC is essential in that it reveals the maximum number of people who can be (directly or indirectly) reached by any other node in the same component. A large GC indicates high reachability because every node is reachable from almost every other.
- *Density* The ratio between the number of edges in the graph and the total number of possible edges. Measures how close the network is to complete. A complete graph has all possible edges and density equal to 1. The opposite, a graph with only a few edges, is a sparse

[2] A free software tool that allows researchers to connect directly to Twitter's streaming Application Programming Interface (API) to collect real-time social media posts by specifying keywords.

[3] https://cst.org.uk/public/data/file/4/2/antisemitic/Content/on/Twitter.pdf.

graph (Zykov 1990). High density indicates intimate, tightly knit networks, and ties between individuals in denser network are more likely to be strong.

- *Average degree metrics* are direct measures of how information travels throughout the network (Newman 2001). Average graph degree for a vertex is calculated as the number of links that end in that vertex. Also, we calculated the maximum value of the degree of the nodes over all graph vertices. Essentially, this metric is a measure of graph connectivity in terms of links/relations between nodes. This, in terms of followers degrees, means that users can directly consume (see, read) the content posted by other users. The spread of node degrees over a network is characterised by a distribution function, which is the probability that a randomly selected node has exactly k edges. The degree distribution has been calculated for the followers and retweet networks. In this directed case, we are interested in the out-degree, which represents the number of the users that someone follows (e.g. if A has an out-degree of 5, it means A follows five people). Also, the in-degree distribution is calculated, which represents the number of the followers that someone has (e.g. if A has in-degree of 5, it means there are five people that follow A). Higher out-degree values mean a wider exposure to different sources of hate propagators. Higher in-degree value refers to influential users (content creation hubs or conversational hubs) who can be responsible for hate creation and propagation. For the retweet network, out-degree represents the number of retweets (e.g. if A has out-degree of 5, it means they retweeted five tweets posted by five different users). Also, we interested in the in-degree distribution that shows the number of retweets that someone gained (e.g. if A has in-degree of 5, it means five users retweeted A’s tweet). High in-degree indicates high hate propagation, while high out-degree indicates to the level of diversity of the propagated content. Nodes with high out-degree centrality can exchange their opinion and build a conversation with others (Hanneman and Riddle 2005; Ishikawa et al. 2013).
- *Betweenness centrality* this metric is a measure of accessibility that is the number of times a node is crossed by shortest paths in the graph, which is useful for finding the individuals who influence the flow around a system.
- *Eigenvector centrality* The node with high eigenvector value is important as a connector for high information diffusion. Degree centrality measures the amount of connections a node has, but disregards the nodes to which these connections are established. Eigenvector centrality modifies this approach by giving a higher centrality score to those connections which are made with those nodes that are themselves central.
- *Average clustering coefficient* Firstly, we calculate the clustering coefficient for each node as the probability that two randomly chosen distinct neighbours of the given node are connected; this is also referred to as the local clustering coefficient for a node. Then, we average these values over all network nodes. The average clustering coefficient was calculated using the Matthieu Latapy algorithm (Latapy 2008). Clustering coefficient measures how some of the nodes can form dense groups in which each element has strong connections with the others. As a consequence, each piece of information posted by one of these nodes can rapidly spread within the groups but disseminates outside the group with more difficulty.
- *Reciprocity* the measure of the likelihood that nodes in a directed network are mutually linked. A higher value indicates many nodes have two-way links, reflecting high connectivity (high level of friendship) in the followers network and high cooperation for hate dissemination in retweet networks.
- *The average shortest path and diameter* is the average graph distance between all pairs of nodes. The diameter is the longest graph distance between any two nodes in the network (Albert and Barabási 2002). The Faster Algorithm for closeness centrality was used to extract the average shortest paths and the diameters (Brandes 2001). These metrics were chosen because they are direct measures of how information travels throughout the network. Followers paths represent links between a node and its neighbours, between them and their own networks, and so on. The shorter the length of the shortest path from a node to all others in the graph (and so their average), the easier the information can travel from a given node and spread over the network (Colombo et al. 2016).

3.3 Node removal strategy

In a theoretical study, Golub et al. (2007) found that the efficient diffusion of influence through a network is limited by the presence of highly influential, high degree nodes. Because these node are responsible for the robustness of the networks against Twitter suspension (Wei et al. 2015), the challenge for us is to identify these nodes within the hateful networks, whose removal would decrease the connectivity and reduce the flow of hateful content. As mentioned in Sect. 3.1, the followers’ networks are conceptually different from the retweets network, in that the former indicate exposure to hateful content, while the latter indicate the spread of content (contagion). Thus, our assumptions are that (1) removing the influential nodes from the followers’ networks would reduce the *exposure* to hateful content for others users (remove key content providers) and (2) removing the influential nodes from the retweets’ network would decrease the level of information propagation and *contagion*.

To efficiently identify nodes *v* whose removal reduces exposure and contagion within the network most, six structured heuristic node removal strategies were designed using different node centrality metrics. In addition to random node removal, these were the nodes with: the highest degree max_{deg}(v) (**degree -based strategy**)—see Algorithm 2; the highest in-degree max_{indeg}(v) (**in-degree -based strategy**)—see Algorithm 3; the highest out-degree max_{outdeg}(v) (**out-degree -based strategy**)—see Algorithm 4; the highest betweenness max_{bet}(v) (**betweenness -based strategy**)—see Algorithm 5; and the highest eigenvalue max_{eig}(v) (**eigenvalue -based strategy**)—see Algorithm 6.

Algorithm 2 Degree-based Node Removal

```
Input: INPUT: H=v$_1, v_2, .., v_h$
Output: OUTPUT: Giant component (GC), density (d)
 and average shortest path (l) of H
for each $v_i \in$ H do
    if v = $max_{deg}$ then
        remove $v$
        Return $GC(H), d(H), l(H)$
    else
        Return $null$
    end if
    end for Repeat until remove (H/10)
```

Algorithm 3 Indegree-based Node Removal

```
Input: INPUT: H=v$_1, v_2, .., v_h$
Output: OUTPUT: Giant component (GC), density (d)
 and average shortest path (l) of H
for each $v_i \in$ H do
    if v = $max_{indeg}$ then
        remove $v$
        Return $GC(H), d(H), l(H)$
    else
        Return $null$
    end if
    end for Repeat until remove (H/10)
```

Algorithm 4 Outdegree-based Node Removal

```
Input: INPUT: H=v$_1, v_2, .., v_h$
Output: OUTPUT: Giant component (GC), density (d)
 and average shortest path (l) of H
for each $v_i \in$ H do
    if v = $max_{outdeg}$ then
        remove $v$
        Return $GC(H), d(H), l(H)$
    else
        Return $null$
end if
end for
Repeat until remove (H/10)
```

Algorithm 5 Betweenness-based Node Removal

```
Input: INPUT: H=v$_1, v_2, .., v_h$
Output: OUTPUT: Giant component (GC), density (d)
 and average shortest path (l) of H
for each $v_i \in$ H do calculate betweenness centrality
    if v = $max_{btw}$ then
        remove $v$
        Return $GC(H), d(H), l(H)$
    else
        Return $null$
end if
end for
Repeat until remove (H/10)
```

Algorithm 6 Eigenvalue-based Node Removal

```
Input: INPUT: H=v$_1, v_2, .., v_h$
Output: OUTPUT: Giant component (GC), density (d)
 and average shortest path (l) of H
for each $v_i \in$ H do calculate eigenvalue centrality
    if v = $max_{egv}$ then
        remove $v$
        Return $GC(H), d(H), l(H)$
    else
        Return $null$
end if
    end for
Repeat until remove (H/10)
```

In addition, a random node removal strategy was used as a baseline to examine the performance of the five structured node removal strategies.

The closeness centrality metric was excluded because, in a highly-connected network, all the nodes would be shown with a similar score. It seemed perhaps more useful to use closeness to find the influencers in a single cluster rather than in an entire network.

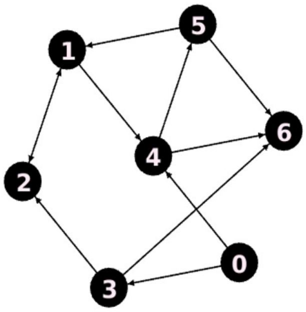


Fig. 1 Example graph that shows removing $node_3$ will increase the shortest path between $node_0$ and $node_2$

It is expected for the hateful network to be restructured after removing a portion of specific nodes. The impact of different node removal strategies was measured through changes in the networks' giant component (GC), the density, and the average shortest path metric of different networks. Network GC and network average shortest path have been widely used as an indicator of the network changes/failure/distribution (Xu and Chen 2008; Petersen et al. 2011; Boldi et al. 2011; Jürgens et al. 2011). Reducing the GC to a small connected component is a positive sign of the effectiveness of node removal strategy. In contrast, the increase in the average shortest path is a positive sign of the node removal strategy as it indicates the removal of the vital bridges (fundamental hubs) that "shorten" the distance between the nodes. Figure 1 shows a simple network containing numbered nodes as an example of a network being restructured after removing a node. The shortest path between $node_0$ to $node_2$ is 2, passing through $node_3$. When $node_3$ is removed, the shortest path between $node_0$ and $node_2$ will become 4 as the content needs to be passed via 3 nodes—$node_4$ then $node_5$, and then $node_1$, to reach its destination.

We also use the density metric because a previous study (Luarn and Chiu 2016) showed that network density is positively related to transmitter activity on social network sites. Moreover, on Twitter, the rate at which information is spread through a network was found to depend on its density (Lerman and Ghosh 2010).

For the degree-based strategy, we also specify whether the node should have high in-degree or high out-degree. According to Roland et al. (2017), in retweet networks, in- and out-degree centrality metrics capture the users' engagement with other users and the content of their posts, and they also form vital bridges. These metrics indicate the actual attention given to content and the action that users took to disseminate information. So, both in-degree and out-degree nodes could have an essential role in our networks.

Only the first 10% of the networks' nodes were removed, and the results of the metrics were recorded gradually for each 1% removed. Previous research results show that highly influential nodes are rare in social networks (Zhao et al. 2017), which is the reason we chose to only remove 10% in descending order of the centrality of the nodes. For instance, Otsuka and Tsugawa (2019), Gallos et al. (2005), Xu and Chen (2008) and Duijn et al. (2014) considered removing 4%–8%–10%. The results of the six removal strategies approximate the effects of different strategies that reflect the role of the node within the network. The steps of the degree-based strategy are simply explained in 2. In each round of node removal, we recalculated the metrics because according to Nie et al. (2015), Bellingeri et al. (2014), Cohen et al. (2000) and Iyer et al. (2013), this will provided more efficient deletion than the non-recalculated method. Node removal strategies were applied on the hateful networks (followers and retweet networks) and also applied on the suicidal network to show the similarities and differences in the role of the nodes within non-hateful networks. The fundamental differences between the degree-based, the betweenness-based and the eigenvalue-based strategies are that the degree-based method concentrates on reducing the total number of edges in the network as fast as possible, whereas the betweenness-based approach concentrates on removing as many edges in the shortest path as possible (Holme et al. 2002). The eigenvalue-based strategy aims to deconstruct the bridges between the highest impact nodes. Eigenvalue centrality was also used to measure the popularity and importance of a node in (non hateful) social networks by Newman (2008) and Bonacich (1972).

4 Results and discussion

Tables 1 and 2 show the graph metrics for the hateful followers' networks (Anti-Muslim 1 and Anti-Muslim 2), the hateful retweets networks (Anti-Muslim 1, Anti-Muslim 2 and Anti-Semitic), and the comparator suicide network.

Table 1 Graph metrics for the followers networks

Networks	Nodes	Edges	Giant component (%)	Density	Avg. deg.	Max. deg.	Avg clust.	Avg. sh.	Diameter	Reciprocity (%)
Anti-Muslim 1	1004	2644	60.7	0.0026	2.6	100	0.062	5.4	16	33.4
Anti-Muslim 2	1073	2895	66	0.0025	2.7	143	0.065	5.6	17	26.7
Suicidal	987	2410	50	0.0024	2.53	100	0.064	5.6	17	62

Table 2 Graph metrics for the retweet networks

Networks	Nodes	Edges	Giant component (%)	Density	Avg. deg.	Max. deg.	Avg clust.	Avg. sh.	Diameter	Reciprocity (%)
Anti-Muslim 1	1229	2571	69.2	0.0017	2.09	304	0.0097	5.2	21	18.89
Anti-Muslim 2	5581	16338	81.3	0.00054	2.3	1034	0.15	5.9	16	15.61
Anti-semitic	2748	5091	72.1	0.00067	1.85	522	0.029	6.3	14	12
Suicidal	3209	2211	31.3	0.00021	1.38	44	9.4E-03	5.05	13	0.9

4.1 Network characteristics

4.1.1 Follower graph: measure of hateful content exposure

Giant Component Table 1 shows that the hateful networks have a similar sized Giant Component (60.7% and 66%) while the suicidal network with a similar number of nodes and edges is smaller at 50%. The size of the giant component is the maximum number of people who can be exposed to/propagate hateful content. This suggests that the users in hateful networks are at similar levels of risk to exposure, with suicide networks as a comparator risky network at least 10% lower.

Hate Density In addition, Table 1 shows that hateful networks have a similar, and slightly higher density than the suicidal network by 0.0001. Despite the seemingly small numerical difference, this has an impact on the rate of information flow within the network. However, given the hateful networks have more nodes (higher number of users) than the suicidal one, it would be expected that they would have smaller density (Faust 2006). In reality, we see the hateful networks show slightly *higher* density values, suggesting that users in the hateful network are more interconnected than users in suicidal network. Highly interconnected users in a followers network mean increased potential of content exposure, which in turn increases the risk of a potential content propagation.

Average Degree The hateful followers' networks exhibited 2.6 and 2.7 average degrees, respectively—see Table 1. The hateful networks have a slightly higher average degree than the suicidal network which is of comparable size, while the max degree was slightly higher for the Anti-muslim 2 network. Generally, the expected average degree of a social network such as Twitter is around 3 (Chatfield and Brajawidagda 2012), which is similar to hateful followers' networks. It does not appear that hateful follower networks are significantly more connected than the comparator risky network or Twitter networks on average. The overall degree, in-degree and out-degree distributions are illustrated in Fig. 2, showing long-tail characteristics where the majority of users following a few numbers of the hateful account, between 1 and 10 (out-degree), and have very few followers between 1 and 10 (in-degree). This observation indicates the existence of hubs, i.e. a few nodes that are highly connected to other nodes, in the hateful and suicidal networks. The presence of large hubs results in a degree, in-degree and out-degree distribution with long tail.

However, the distribution of the hateful followers networks shows consistently similar in-degree distribution and have a higher "head" (maximum) out-degree appeared for degrees higher than 10. This suggests both hateful networks have fewer "influencers" (people with lots of followers) and more "superconsumers" (people who follow a lot of hateful posters). In contrast, the suicidal followers network shows slightly higher head for its in-degree distribution. This network has more influencers and less nodes who tend to follow a large number of similar users. Meaning that, the hateful followers network tends to be more vulnerable to hate exposure (Leskovec et al. 2007) than the suicidal follower network.

Average Clustering Coefficient The average clustering coefficient of the Anti-Muslim 1 and Anti-Muslim 2 followers' networks was 0.062 and 0.065, respectively. Clustering coefficient values for the hateful followers' networks were similar to the suicidal followers network. Even though the average clustering coefficient of the hateful networks is similar to that of the suicidal network, their distribution showed that there are some differences. Empirically, nodes with higher degree(K_i) have a lower *local* clustering coefficient on average; thus, the *local* clustering coefficient (C_i) decreases with increasing degree (Myers et al. 2014). The metric quantifies how close the neighbours are to being a complete graph (a clique). The distribution of clustering coefficients of the hateful followers networks and the comparator network is shown in Fig. 3. For the hateful followers networks, several nodes with $(K_i) \geq 30$ have (C_i) greater than 0.2 while for the suicidal follower network all (C_i) of nodes that have $(K_i) \geq 30$ do not exceed 0.15. The probability of a node's neighbours being also connected (densely connected neighbours) is higher for the hateful network than the suicidal network. Whether these nodes are "hate consumers" (out-degree edges) or "hate influencers" (in-degree edges), both cases exhibited densely connected neighbours. This behaviour has not been found in the suicidal followers network, providing further evidence of tight connectivity

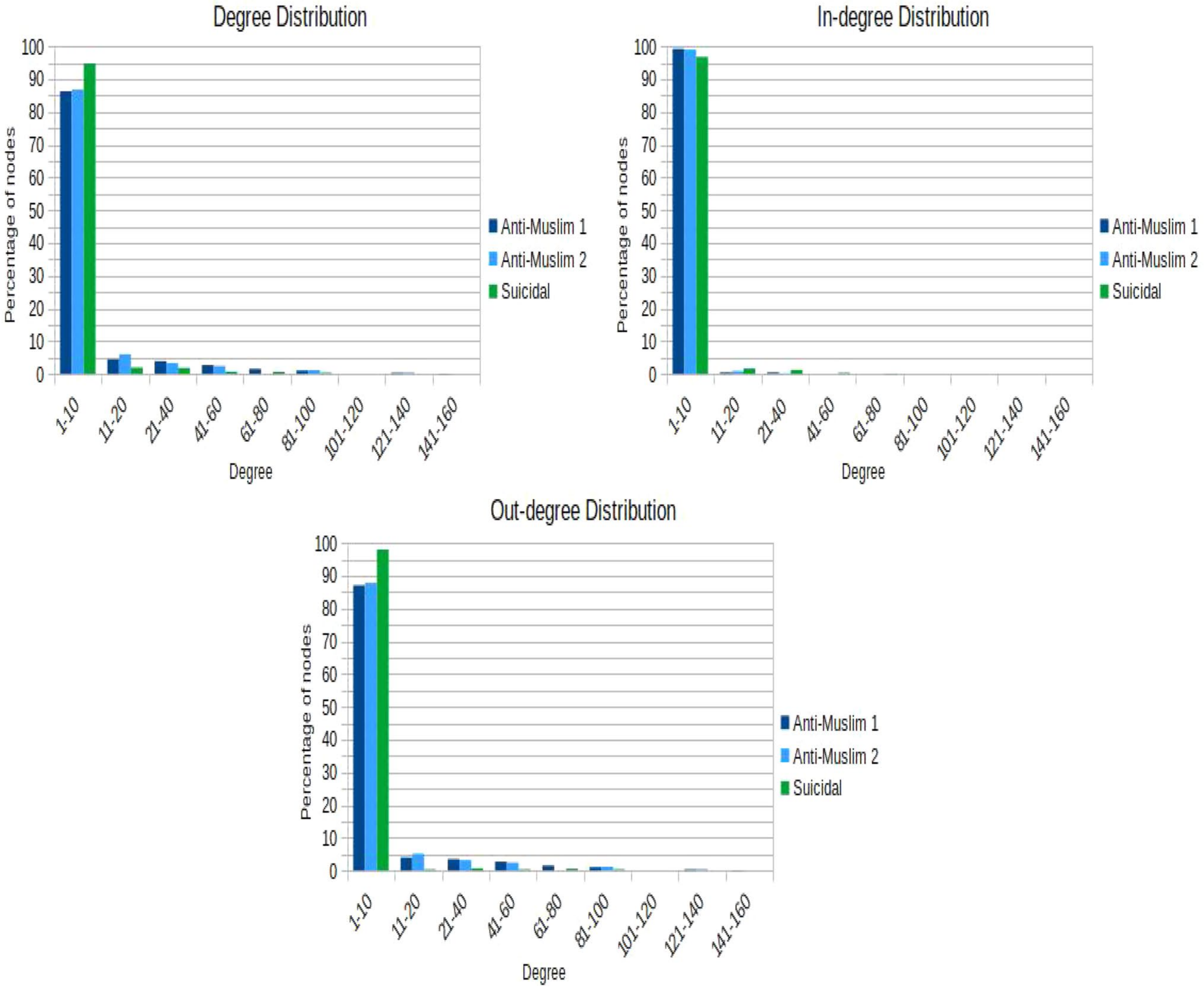


Fig. 2 Degree, in-degree and out-degree distributions for the followers networks

between the hateful users, which means increasing the risk of the hateful content exposure.

Reciprocity Table 1 shows that around 33% and 26% of the Anti-Muslim 1 and Anti-Muslim 2 followers' edges were reciprocal. These percentages are significantly lower than those recorded for the suicidal network, which was 62%. The presence of the reciprocity in the follower network means that people with common interests, known as "homophily", are exposed to each other's content (Scott 1988; Rao and Bandyopadhyay 1987; Paul et al. 2018). Thus, about the third of the hateful accounts are exposed to each other's content, while more than the half of the suicidal users do so. This suggests that the suicidal users form a more cohesive community based around reciprocal follower relationships (Pelaprat and Brown 2012; Putnam 2000). However, research shows reciprocity has a connection with emotional distress which is significantly associated with the suicidal users (Mueller and Abrutyn 2015). Thus, a study that investigates the incentive for reciprocal behaviour for different "risky" followers networks is required for clearly understanding that behaviour.

Diameter

Table 1 shows that Anti-Muslim 1 and Anti-Muslim 2 recorded similar diameters (16 and 17) and average shortest paths (5.4 and 5.6). The maximum diameter is susceptible to outliers (Palmer et al. 2001), and the average shortest path is a more rational measurement than the diameter, because the diameter decreases when edges are added, while the latter may remain unchanged. Thus, we focused on the average shortest path for characterising the hateful network. The average shortest paths are for the largest connected component (Giant Component) (Lehmann and Ahn 2018). For example, in Anti-Muslim 1, between 5 and 6 steps are needed (5.4 avg. sh. path) to reach up to 60% of people who

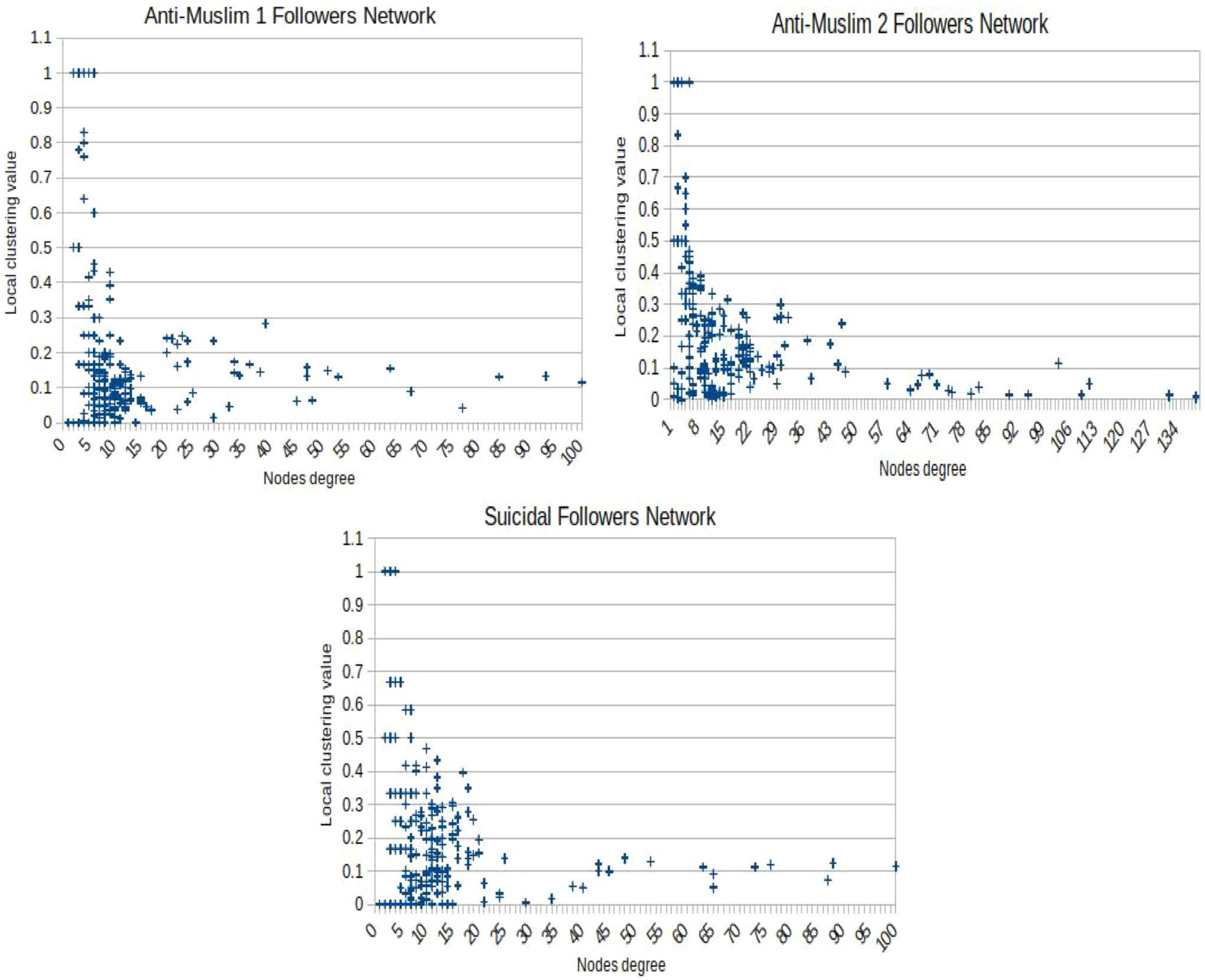


Fig. 3 Distribution of the local clustering coefficient per degree for the followers networks

belong to the Anti-Muslim 1 followers network (Giant Component). The bigger both metrics are, the easier the content will flow through the network. The metrics are similar for all three networks, though for suicide the Giant Component is smaller so less people are reached.

Although the average path length should actually decrease with small sized networks (Leskovec et al. 2005), the average shortest path for the hateful followers' networks runs in line with that reported in a public retweets Konect dataset (5.45), which depicts a much more extensive Twitter network of online communications, with three million nodes and over ten million edges (Kunegis 2013). This provided some evidence that the hateful followers' networks exhibit data flow properties resembling large-scale communication followers' networks, albeit in a very small-scale network.

4.2 Retweets graphs: measure of the hateful content contagion

Giant Component Table 2 shows that more than 69%, 81% and 72% of the nodes in the hateful retweets networks exist in the largest (Giant) component. While some fluctuations exist between the hateful networks, the percentage in the comparator suicidal retweet network measures only 31.3%, which is significantly lower. These percentages reflect the percentage of users who are part of a connected community. The results suggest that there is a consistently and significantly greater reach of content (contagion) in the hateful networks.

Density Table 2 shows that the densities of all the hateful retweet networks are slightly higher (by 0.00033) than the suicidal network. Although that Anti-Muslim 2 recorded the lowest density of all the hateful retweet networks, its density

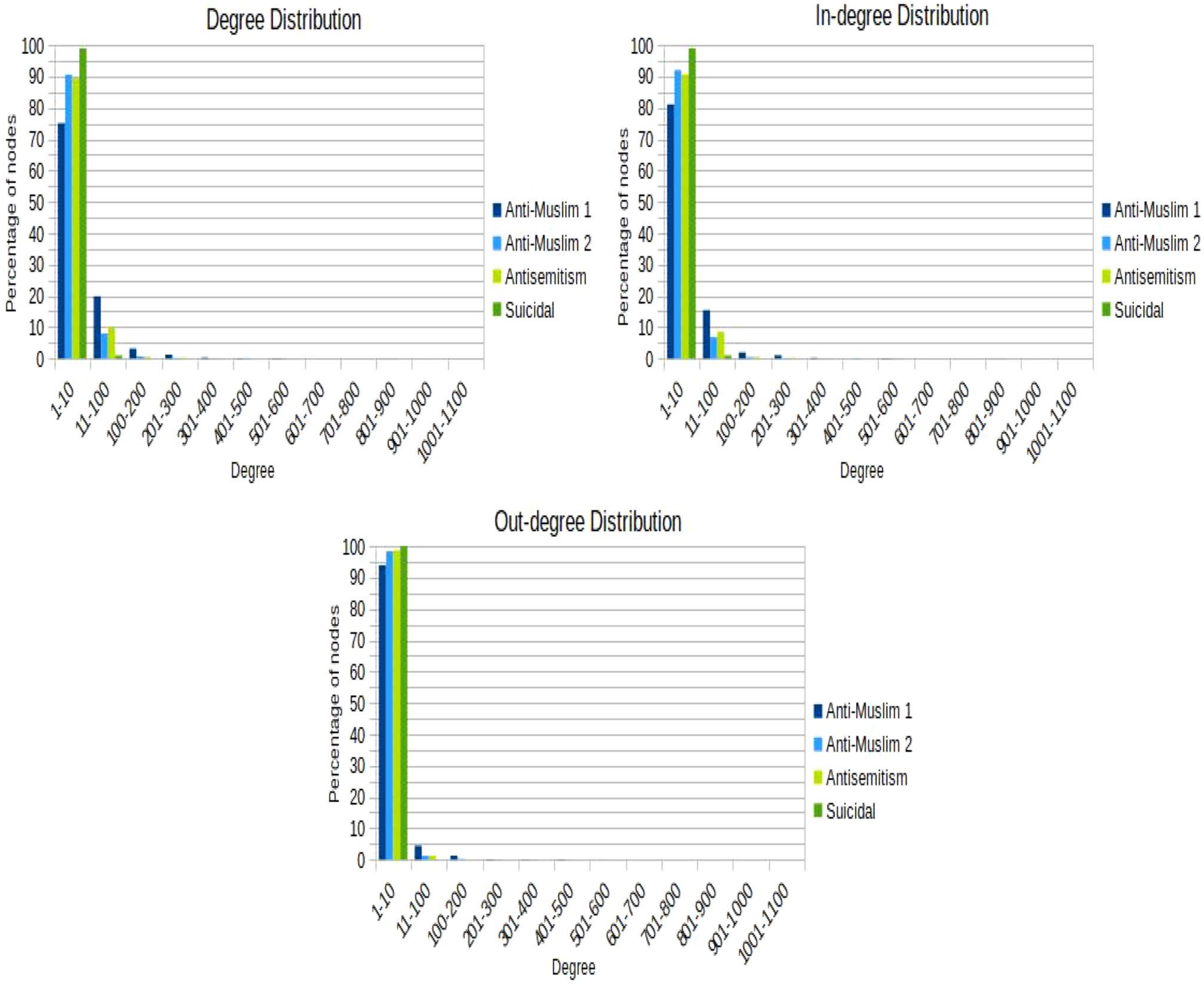


Fig. 4 Degree, in-degree and out-degree distributions of the retweet networks

is double than the suicidal retweet network. This means that the suicidal network has a smaller number of node connections than the hateful networks. Broadly speaking, the densities of all the retweet networks are low compared to the densities of the follower networks. However, we would expect this for retweets as a high density would require all Twitter users retweeting each other, which is unlikely.

Average Degree Table 2 shows the average degree of the Anti-Muslim networks (2.09, 2.3) is slightly higher than the Anti-Semitic network (1.85), with all hateful networks higher than the comparison suicide network (1.38). The consistently higher average degree indicates more nodes were reachable on average and increased propagation of hateful content through the network. The maximum degree of the hateful content is far higher than that of the suicide network for all three hateful retweet networks. The overall degree, in-degree and out-degree distributions are illustrated in Fig. 4, showing a propriety of scale-free networks with existence of fewer nodes in the network with higher levels of retweets and many other nodes with fewer retweets. Moreover, for all the retweet networks in this study, in-degree distribution shows higher "head" than out-degree distribution for degrees larger than 10, indicating that popular users responsible for creating information cascades exist in the hateful and the suicidal retweet networks. In contrast, the suicidal retweet network shows an absence of nodes with high out-degree (degrees larger than 10), compared to the number of out-degree edges in the hateful retweet networks. Out-degree suggests a considerable number of the hateful users engaged significantly with hateful conversation by retweeting other users' hateful messages, while the suicidal network do not exhibit this behaviour. This suggests that we would see more co-operation on the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the suicide network.

Average Cluster Coefficient Table 2 shows that the average clustering coefficient is low for all networks, suggesting a lack of coherent sub-groups within the overall networks. In addition, it is not feasible to show here the degree distribution of the *local* clustering coefficient as all the network do not exhibit similar average clustering coefficient.

Reciprocity Table 2 shows that the reciprocity among the hate networks (12–18%) is higher than the comparator suicide network (0.9%) and fairly consistent. While the density and avg. clustering co-efficients do not suggest significant smaller 'organised' sub groups exits (where users all retweet each other), there is clearly a consistent level of reciprocal retweeting in the context of hateful content that would appear to be higher than a comparator network. Although the relationship between these retweeters is not necessarily a friendship link, there is a noticeable level of co-operation.

Diameter Table 2 shows there is a consistent level of information flow among the retweet networks, with the Anti-Muslim network exhibiting a slightly higher diameter, but the Anti-Semitic network having a slightly higher average shortest path. However, we should consider the size of the Giant Component as the average shortest path is calculated for this largest sub-graph and not for the entire network. Table 2 shows that the hateful networks have a consistently sized (69–81%) and significantly larger Giant Component than the suicidal network (31.3%). This means while the same number of steps is needed to reach the largest cluster, the hateful content reach is consistently much larger than the suicide content—with between 37.9 and 50% more users reached by retweeting.

4.3 Node removal strategies

4.3.1 Followers networks

Figure 5 shows the six removal strategies that were applied on the followers networks. Of the six removal strategies tested, the random removal strategies were the least effective. This is inline with Jahanpour and Chen (2013), Crucitti et al. (2003) and Wang (2003), who found that small-world networks have strong resilience against a random-based node removal strategy. In general, targeting the highest degree nodes resulted in the greatest decrease in the size of the largest component (GC) and the density for all the hateful followers networks.

For both hateful follower networks, 75–83% of the largest component (GC) was disconnected, and this is also true for the suicidal network. In contrast, using other strategies, saw a reduction of only a 55–75% in GC size. From a network analysis perspective, the highly linked nodes act as a tie that connect global bridges.[4] Practically, their removal disconnects a community into two communities.

Moreover, we observed that node removal based on out-degree (people who follow a lot of others) was more effective than in-degree for the hateful networks. This is inline with the finding in Sect. 4.1.1 that the maximum out-degree was higher than the maximum in-degree for the hateful follower networks. Again, we could observe that people who follow a number of different hateful accounts are connecting different communities more than influential people.

The degree-based strategy also reduced the density of the hateful networks by 47–76% for Anti-Muslim 1 and 2, respectively, whereas other strategies recorded reduction not more than 43%.

Applying the eigenvalue-based strategy was found to be the most effective strategy for increasing the average shortest path of the hateful follower networks and also for the suicidal network. Eigenvalue-based strategy elongated the average shortest path of Anti-Muslim 1 follower network from 5.4 to 6.4, Anti-Muslim 2 from 5.6 to 13.8 and also from 5.05 to 9.5 for the suicidal network. This means that the average steps that are needed for delivering content to all the users in the biggest connected community are increased. For example, in the Anti-Muslim 2 follower network a hateful tweet would potentially reach all the nodes in the largest community (GC) within 5.6 steps. By removing nodes that are connected to highly linked nodes (high eigenvalue nodes), we saw an increase in the number the steps that are needed to reach the majority of the nodes in the largest community to 13 steps—essentially obstructing of the information flow. Nevertheless, the eigenvalue-based strategy was the least effective strategy for reducing the size of the GC all the hateful networks. The eigenvalue-based strategy reduced the GC of Anti-Muslim 1 and 2 followers networks by 63% and 57% which is less than the reduction made by degree-based strategy. Our explanation for this is that nodes with high eigenvalues could be connected to important local bridges (Huang et al. 2019). Granovetter (1977) stated that when the local bridge is removed, the nodes on either side of the bridge become reachable only via very long paths. Thus, while the eigenvalue-based strategy had less impact on reducing the size of the largest component, it resulted in a sparse largest component and therefore elongated the path needed to reach all users in the network.

4.3.2 Retweet networks

Figure 6 shows the six removal strategies that were applied on the retweets networks.. We observed that of the six

[4] A bridge is a direct tie between nodes that would otherwise be in disconnected components of the graph.

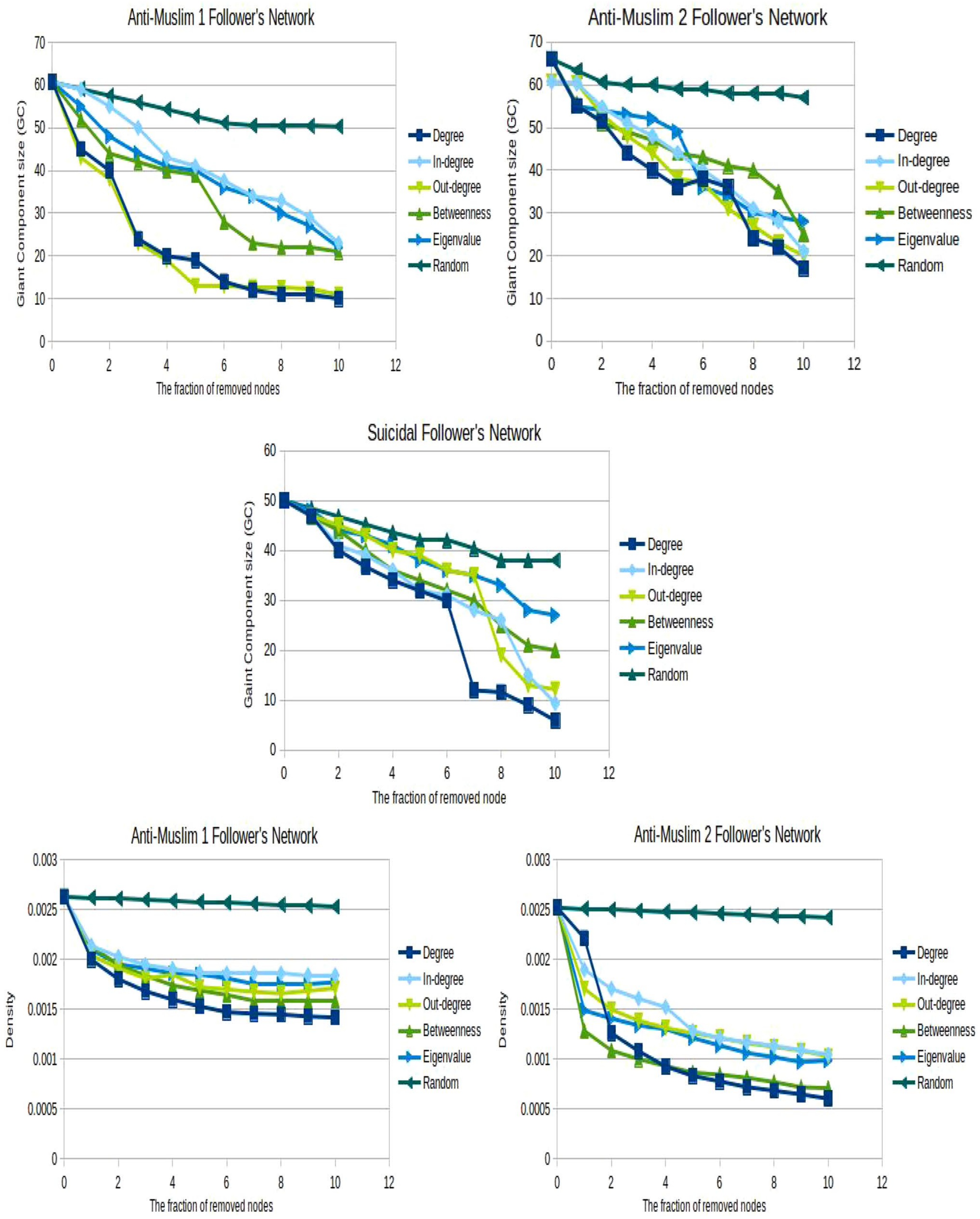


Fig. 5 Impact of different removal strategies on the level of the giant component size, the density and the average shortest path of the hateful and suicidal followers networks

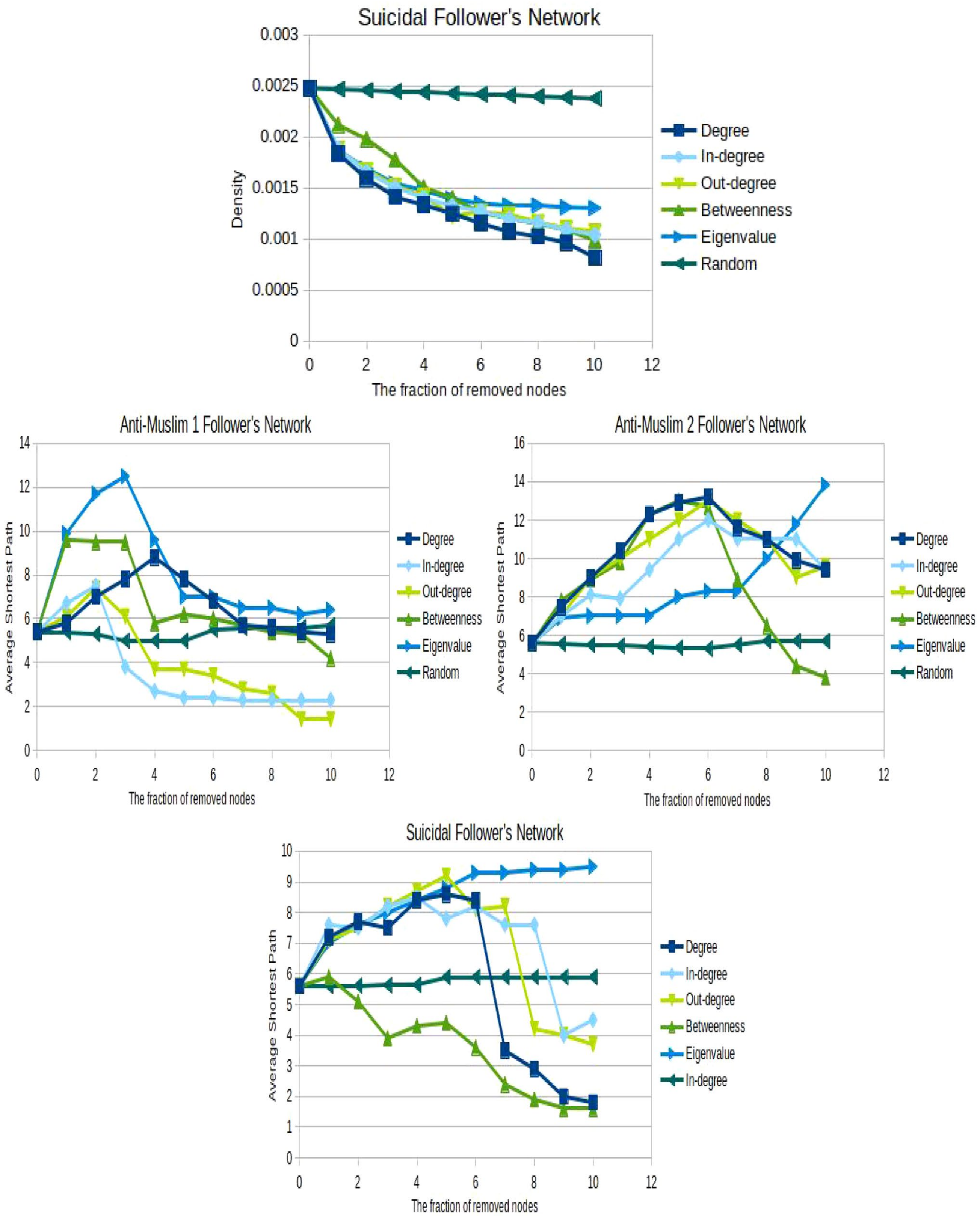
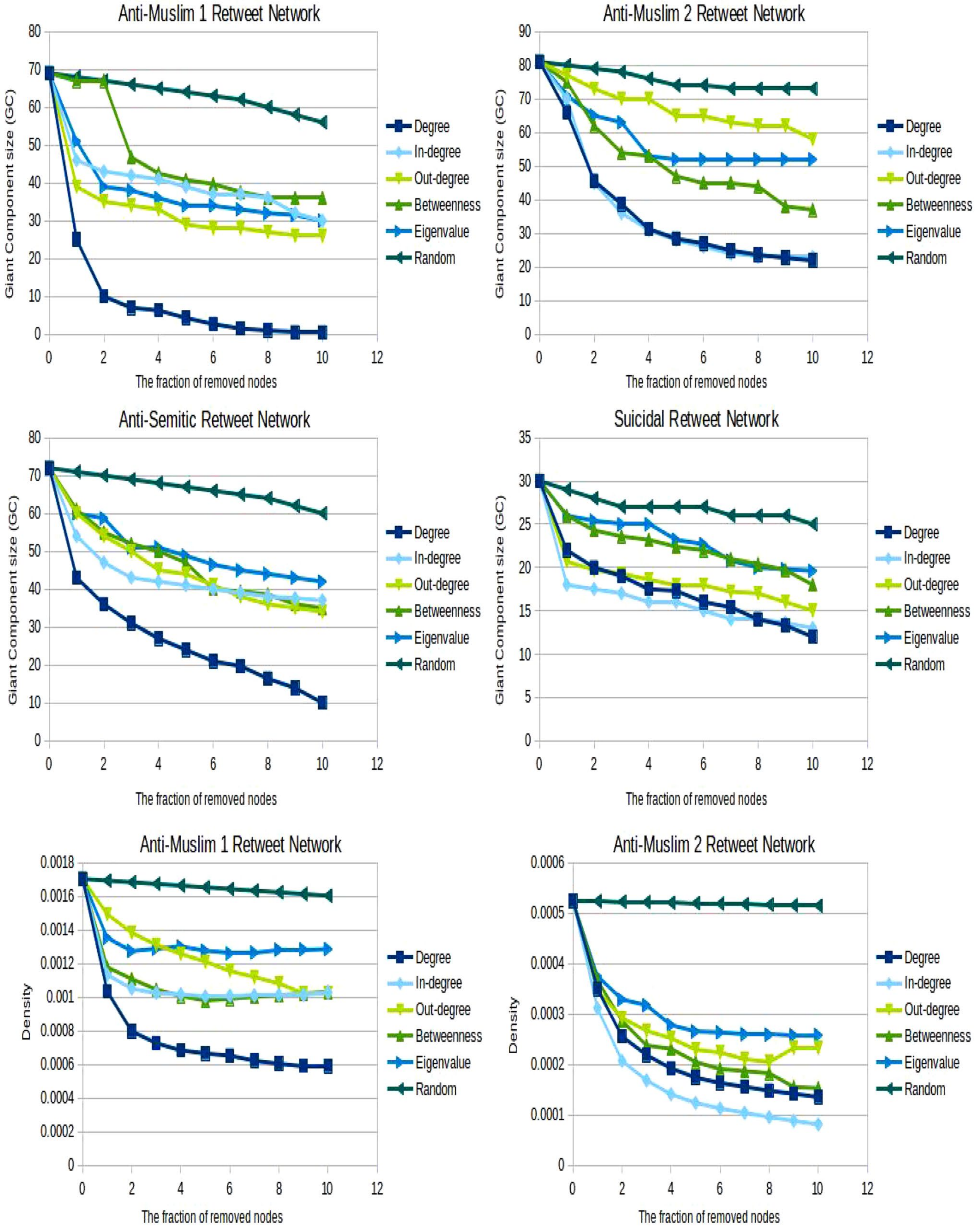


Fig. 5 (continued)


Anti-Muslim 1 Retweet Network

Anti-Muslim 2 Retweet Network

Anti-Semitic Retweet Network

Suicidal Retweet Network

Anti-Muslim 1 Retweet Network

Anti-Muslim 2 Retweet Network

Fig. 6 Impact of different removal strategies on the level of the giant component size, the density and the average shortest path of the hateful and suicidal retweets networks

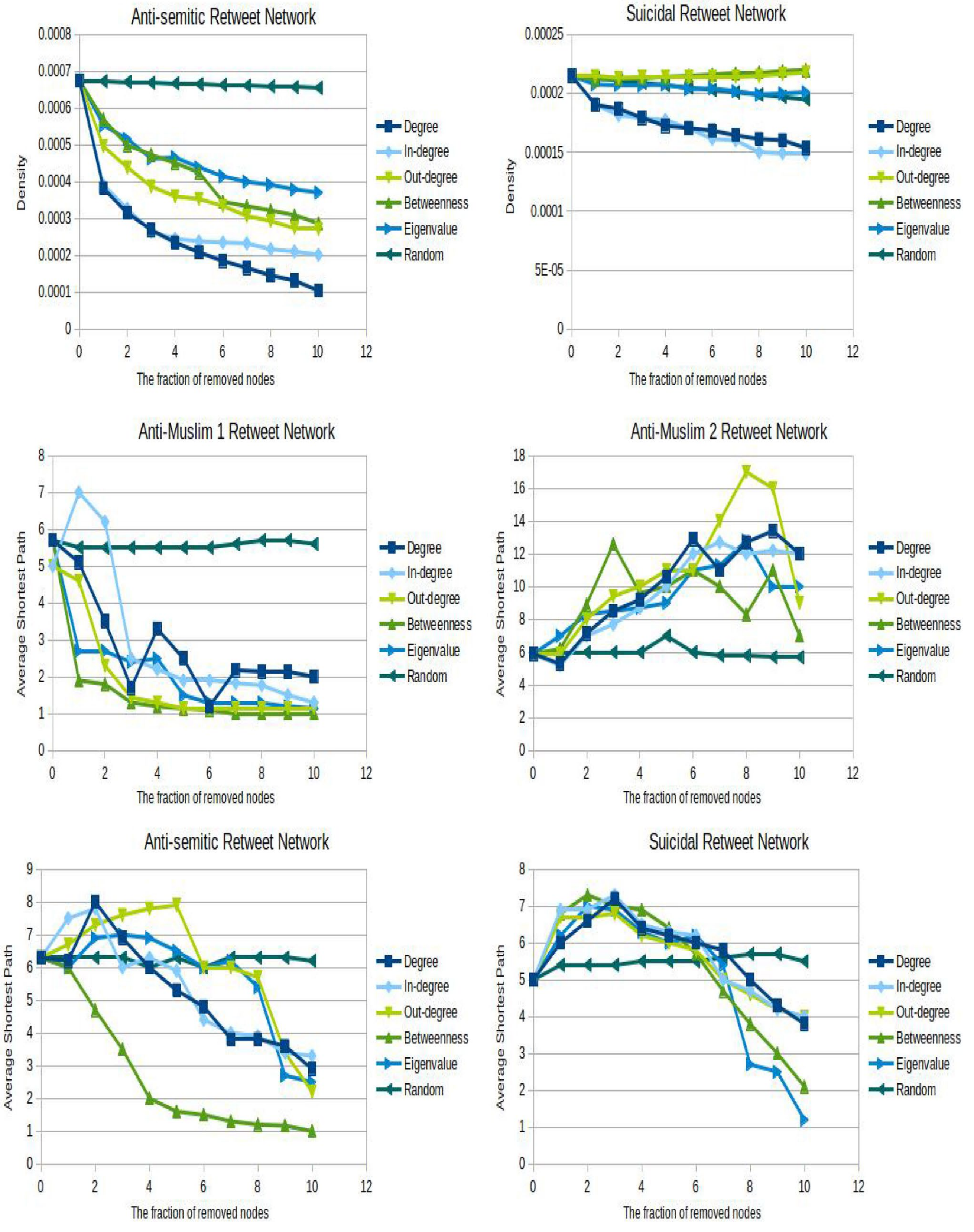


Fig. 6 (continued)

removal strategies tested, the random removal strategy showed the least impact on reducing the size of GC, the density and the average shortest path. In general, targeting the highest degree nodes resulted in the greatest decrease in the size of the largest component (GC) and the density for all the hateful retweets networks. A degree-based strategy reduced the size of GC by 94%, 85% and 75% for Anti-Muslim 1, Anti-Semitic and Anti-Muslim 2, respectively, while for the same fractions of nodes removed other strategies led to a reduction of between 60 and 70% reduction in GC size. Moreover, the degree-based strategy reduced the densities of the hateful networks by 65, 83 and 80% for Anti-Muslim 1, Anti-Semitic and Anti-Muslim 2, respectively. In contrast, other strategies recorded performances lower than the degree-based strategy by approximately 20%. This is in line with the results of previous studies showing that the most connected people (hubs) are the key players, being responsible for the most massive scale of the spreading process (Albert et al. 2000; Cohen et al. 2001).

The effectiveness of the degree-based strategy was also best performing in the reduction in density, which is very clear in Anti-Muslim 1 and Anti-Semitic retweet networks. For Anti-Muslim 2, it is still best performing, but we observed that in-degree-based strategy is very close. We observe that an in-degree-based strategy is slightly more effective strategy for reducing the density of suicidal retweet network.

For the hateful retweet networks, however, there is no single strategy that has a significant impact on increasing the average shortest path. We observe that the average shortest path gets longer at the beginning due to the removal of nodes, but when the percentage of removed nodes becomes very large, existing shortest paths start getting shorter. Crucitti et al. (2003) observed that the average shortest path increases rapidly when the most connected nodes are eliminated and then remains unchanged. Indeed, Boldi and Vigna (2012) and Boldi et al. (2013) suggested that this measure is not useful when networks get significantly disconnected, which is inline with our finding. Nevertheless, this metric is widely used as an indicator for nodes removal efficiency in the previous studies (Jürgens et al. 2011; Kane et al. 2014; Wiil et al. 2010; Xu and Chen 2008). Also, we examined this metric for characterising the hateful followers and retweet networks and it showed interesting results regarding to the networks connectivity for both followers and retweet networks, which motivated us to consider it as an indicator of the removal strategies efficiency.

5 Conclusion

In this article, we have analysed the graph characteristics of three Twitter datasets of users who have posted tweets that human annotators agreed should be classified as containing evidence of hateful content. For the purposes of the research, we referred to these networks as hateful networks. We conducted a range of social network analysis experiments by investigating the social graphs of the followers and retweets of hateful users. Six metrics were applied on the hateful networks to examine the similarity and the differences between them, and we compared the results with another ’risky’ network—suicidal ideation language.

For the hateful followers networks, we found users within the hateful networks are at similar levels in terms of risk of exposure to hateful content, with suicide networks as a comparator risky network at least 10% lower. The size of the largest component and density are slightly higher for both two hateful followers networks than that of the suicide network by over 10% and 0.0001 for the largest component and density, respectively. This suggests a higher exposure to, and potential virality of, hateful content. For the average degree, it does not appear that hateful follower networks are significantly more connected than the comparator risky network or Twitter networks on average.

Interestingly, both hateful networks have fewer “influencers” (people with lots of followers), and more “superconsumers” (people who follow a lot of hateful posters). In contrast, the suicidal followers network shows a slightly higher volume of “influencers”, and less nodes who tend to follow a large number of similar users. These findings suggest that hateful follower networks tend to be more vulnerable to hate exposure (Leskovec et al. 2007) than the suicidal follower network. Moreover, the clustering coefficient metric revealed that the probability of a node’s neighbours being also connected (densely connected neighbours) is higher for the hateful networks than the suicidal network.

Comparing the average shortest path of the followers networks with the relative largest component size suggested that the hateful followers networks are more connected than the suicidal followers network. Five steps are needed to reach more than 60% of the users of the hateful follower networks while same number of steps can reach only a half of the suicidal users. In addition, the average shortest path of the hateful networks is similar to a more extensive Twitter network, suggesting that the hateful followers’ networks exhibit data flow properties resembling large-scale communication followers’ networks, albeit in a very small-scale network.

Regarding the social reciprocity or social exchange, the hateful follower networks recorded similar levels of reciprocity (users who follow each other) and were lower than the reciprocated behaviour for the suicidal followers network. In general, Twitter shows a lower level of reciprocity of 22.1% (Kwak et al. 2010). This suggests that the hateful networks are above the Twitter average and reasonably similar between networks—meaning more of the hateful users follow each other than average Twitter users—but below that of another ideological interest group—suicidal users. This could be interpreted as there being a 'community of interest' around hate on Twitter, but less connected than that of suicide. This reflects the danger of exposure to hateful ideologies among the offensive users who are considered information consumers and potential information propagators. Thus, a study that investigates the incentive for reciprocal behaviour for different "risky" followers networks is required for clearly understanding that behaviour.

For the hateful retweets networks, we observed several structural similarities and also differences between the hateful retweet network in terms of social network metrics. The hateful retweet networks recorded sizes of the largest component higher than 69%, densities higher than 0.0005 and reciprocities higher than 12%. All these metrics are higher than those recorded for the suicidal retweets network. This provides evidence that the hateful retweeters exhibit consistently high levels of information propagation behaviour, with significantly greater reach of content contagion in the hateful retweet networks than the comparator 'risky' network. Also, comparing the value of the average shortest path within the largest component showed that the hateful retweet networks needed between 5 and 6 steps to reach between 70 and 80% of the users, while 5 steps reached only 30% of the suicidal users. This provides further evidence of the high reachability (contagion) of the propagated message among the hateful users. This means while the same number of steps is needed to reach the largest cluster, the hateful content reach is consistently much larger than the suicide content—with between 37.9 and 50% more users reached by retweeting.

Regarding the reciprocity, the hateful retweets networks show significantly higher reciprocity than the suicidal retweet's network. This suggests that we would see more co-operation on the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the suicide network (Sparrowe et al. 2001). The danger associated with this finding is that the high level of reciprocity among hateful retweets networks could help in building a collaborative network which eventually raises the level of cooperation in hate propagation. The density and average clustering co-efficients do not suggest significant smaller 'organised' sub groups exits (where users all retweet each other), but there is clearly a consistent level of reciprocal retweeting in the context of hateful content that would appear to be higher than a comparator network.

Finally, we developed a range of node-removal strategies targeting users depending on their role in the network. A simulation of six node-removal strategies indicated that targeting nodes with the highest degree is more effective in reducing the largest component size of the hateful followers and retweets networks compared to the other strategies, thus limiting the exposure and transmission of hateful content. For both hateful follower networks, 75–83% of the largest component (GC) was disconnected, and this is also true for the suicidal network. In contrast, using other strategies, there is saw reduction of only a 55–75%. For the retweet networks, the degree-based strategy reduced the size of GC by 94%, 85% and 75% for Anti-Muslim 1, Anti-Semitic and Anti-Muslim 2, respectively, while for the same fractions of nodes removed other strategies led to a reduction of between 60 and 70% reduction in GC size. Moreover, the degree-based strategy reduced the densities of the hateful networks by 65, 83 and 80%. In contrast, other strategies recorded performances lower than the degree-based strategy by approximately 20%

Targeting nodes that highly follow or retweet other hateful users also had a signifciant effect on reducing the largest component and essentially obstructing the hateful information flow (except for Anti-Muslim 2 retweet network), suggesting that nodes with higher out-degree—the "super-consumer" and "super-retweeters" of the hateful content—constitute indispensable bridges responsible for connecting different clusters in a network. We might assume that removing 'producers'—influential nodes (with high in-degree), would be more effective at reducing spread, but the findings suggest that removing the 'spreaders' is actually more effective. This suggests that targeting such people (e.g. by suspending them) would more likely reduce the spread of hate. The bridging role also makes high out-degree users good targets for using counter-speech to reduce the propagation of hateful ideologies (Mathew et al. 2018), or for content moderation to reduce hate speech on mainstream platforms (Kiesler et al. 2012).

Broadly speaking, social networks are resistant to node removal, and the results of the strategies were not inevitable. A study showed that nodes of high degree are not actually so relevant for the global structure of the network (Boldi et al. 2013). Also, we noticed that some strategies are effective on specific metrics, and some are not, suggesting that combining different node removal strategies might have a significant impact on the entire network.

Indeed, our work is not without limitation, although we characterise different hateful networks in detail, we contended with a scarcity of baselines datasets. For future work, we suggest that more extensive hateful datasets could be constructed and compared to our baseline study, to establish consistency of findings across a larger range of hateful networks, and also study differences over time. Where possible, this may also include other online social networks with open APIs. It may also include comparison to different "risky" networks (e.g. harassment and cyberbullying). Furthermore, the cost of human annotation limits the size of the annotated posts. In future, we may see the use of machine classifiers for the detecting of cyberhate applied to larger datasets, with sub-sampling to validate performance. Current approaches are continually improving in performance up to 96% accuracy (Alorainy et al. 2018; Liu et al. 2019a, b).

To conclude, we aimed to understand the network characteristics of hateful online social networks to help understand individuals' exposure to hate and derive intervention strategies to mitigate the dangers of such networks by disrupting communications. The findings provide some evidence that shows consistent metrics between hateful follower and retweet networks, many of which show higher risk of content exposure and contagion that a comparator 'risky' network of similar size and connectivity. Our intervention strategies in particular have been highly effective in disrupting hateful networks—reducing the spread of content and elongating pathways between users. This would be a useful basis from which to inform policy around network disruption and where /who to introduce counter speech. The empirical measurements in this study also provide a baseline which can be followed by future studies to help inform the development of a body of knowledge around hateful networks and intervention strategies.

6 Future work

While we have identified some interesting and promising results, future research is needed in order to develop further knowledge. Our analysis could be extended to measure the indirect influence of a node to help decision-makers to decide the priority of removing a wider range of nodes. Node removing may cause an impact on the nodes that are directly connected to it, leading to *direct influence*, e.g. loss of a "bridge" leads to a fragmented network. We have measured this kind of influence in this article. It is also possible that node removal may have a further cascading effect on nodes that are not connected directly to the removed node, leading to *indirect influence*, e.g. the increase in the propagation length due to node removal (Liu et al. 2012).

In future, we could envisage the development of an algorithm that provides each node in the hateful network with a score, which combines its direct and indirect influence. The direct influence score may be calculated as a function of the number of users who are directly connected to a node (degree of a node), while the indirect influence score may use the concept of the personal network or 'ego' network for each node in the hateful network, and an "n-step neighbourhood" to include influence on indirectly connected nodes.

Acknowledgements This work was partly supported by the Economic and Social Research Council grant: 'Centre for Cyberhate Research and Policy: Real-Time Scalable Methods and Infrastructure for Modelling the Spread of Cyberhate on Social Media' (Grant Number: ES/P010695/1).

Availability of data and materials The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

Online harms white paper (2020). https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

Al-garadi MA, Varathan KD, Ravana SD (2016) Cybercrime detection in online communications: the experimental case of cyberbullying detection in the twitter network. Comput Hum Behav 63:433–443

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47

Albert R, Jeong H, Barabási A-L (2000) Error and attack tolerance of complex networks. Nature 406(6794):378

Alorainy W, Burnap P, Liu H, Williams M (2018) The enemy among us: Detecting hate speech with threats based'othering'language embeddings. arXiv:1801.07495

Alorainy W, Burnap P, Liu H, Williams ML (2019) The enemy among us detecting cyber hate speech with threats-based othering language embeddings. ACM Trans Web (TWEB) 13(3):1–26

Awan I (2014) Islamophobia and twitter: a typology of online hate against Muslims on social media. Policy Internet 6(2):133–150

Awan I, Zempi I (2017) 'I will blow your face off'—virtual and physical world anti-Muslim hate crime. Br J Criminol 57(2):362–380

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the* *fourth ACM international conference on Web search and data mining*, pp 65–74. ACM

Bellingeri M, Cassi D, Vincenzi S (2014) Efficiency of attack strategies on complex model and real-world networks. Phys A Stat Mech Appl 414:174–180

Boldi P, Vigna S (2012) Four degrees of separation, really. In: 2012 IEEE/ACM international conference on advances in social networks analysis and mining, pp 1222–1227. IEEE

Boldi P, Rosa M, Vigna S (2011) Hyperanf: approximating the neighbourhood function of very large graphs on a budget. In: Proceedings of the 20th international conference on World wide web. ACM, pp 625–634

Boldi P, Rosa M, Vigna S (2013) Robustness of social and web graphs to node removal. Soc Netw Anal Min 3(4):829–842

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Sociol 2(1):113–120

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163–177

Burnap P, Williams ML, Sloan L, Rana O, Housley W, Edwards A, Knight V, Procter R, Voss A (2014) Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. Soc Netw Anal Min 4(1):206

Burnap P, Rana OF, Avis N, Williams M, Housley W, Edwards A, Morgan J, Sloan L (2015) Detecting tension in online communities with computational twitter analysis. Technol Forecast Soc Change 95:96–108

Burnap P, Colombo G, Amery R, Hodorog A, Scourfield J (2017) Multi-class machine classification of suicide-related communication on twitter. Online Soc Netw Media 2:32–44

Chan J, Ghose A, Seamans R (2016) The internet and racial hate crime: offline spillovers from online access. MIS Q 40(2):381–403

Chatfield A, Brajawidagda U (2012) Twitter tsunami early warning network: a social network analysis of twitter information flows

Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Detecting aggressors and bullies on twitter. In: Proceedings of the 26th international conference on world wide web companion. International World Wide Web Conferences Steering Committee, pp 767–768

Chau M, Jennifer X (2007) Mining communities and their relationships in blogs: a study of online hate groups. Int J Hum-Comput Stud 65(1):57–70

Cherepnalkoski D, Mozetič I (2016) Retweet networks of the European parliament: evaluation of the community structure. Appl Netw Sci 1(1):2

Cohen R, Erez K, Ben-Avraham D, Havlin S (2000) Resilience of the internet to random breakdowns. Phys Rev Lett 85(21):4626

Cohen R, Erez K, Ben-Avraham D, Havlin S (2001) Breakdown of the internet under intentional attack. Phys Rev Lett 86(16):3682

Colladon AF, Gloor PA (2019) Measuring the impact of spammers on e-mail and twitter networks. Int J Inf Manag 48:254–262

Colombo GB, Burnap P, Hodorog A, Scourfield J (2016) Analysing the connectivity and communication of suicidal users on twitter. Comput Commun 73:291–300

Crucitti P, Latora V, Marchiori M, Rapisarda A (2003) Efficiency of scale-free networks: error and attack tolerance. Phys A Stat Mech Appl 320:622–642

da Cunha BR, Gonzalez-Avella JC, Goncalves S (2015) Fast fragmentation of networks using module-based attacks. PLoS ONE 10(11):e0142824

Duijn PAC, Kashirin V, Sloot PMA (2014) The relative ineffectiveness of criminal network disruption. Sci Rep 4:4238

Faust K (2006) Comparing social networks: size, density, and local structure. Metodoloski zvezki 3(2):185

Gallos LK, Cohen R, Argyrakis P, Bunde A, Havlin S (2005) Stability and topology of scale-free networks under attack and defense strategies. Phys Rev Lett 94(18):188701

Gerstenfeld PB, Grant DR, Chiang C-P (2003) Hate online: a content analysis of extremist internet sites. Anal Soc Issues Public Policy 3(1):29–44

Glaser J, Dixit J, Green DP (2002) Studying hate crime with the internet: what makes racists advocate racial violence? J Soc Issues 58(1):177–193

Golub B, Jackson MO (2007) Nαıve learning in social networks: convergence, influence, and the wisdom of crowds. http://www.stanford.edu/acksonm/naivelearning.pdf

Granovetter MS (1977) The strength of weak ties. In: Social networks. Elsevier, pp 347–367

Hanneman RA, Riddle M (2005) Introduction to social network methods

Himelboim I, Sweetser KD, Tinkham SF, Cameron K, Danelo M, West K (2016) Valence-based homophily on twitter: network analysis of emotions and political talk in the 2012 presidential election. New Media Soc 18(7):1382–1400

Holme P, Kim BJ, Yoon CN, Han SK (2002) Attack vulnerability of complex networks. Phys Rev E 65(5):056109

Huang C-Y, Yu-Hsiang F, Tsai Y-S et al (2019) Beyond bond links in complex networks: local bridges, global bridges and silk links. Phys A Stat Mech Appl 536:121027

Ishikawa Y, Li J, Wang W, Zhang R, Zhang W (2013) Web technologies and applications: 15th Asia-Pacific web conference. APWeb 2013, Sydney, Australia, April 4–6, 2013, Proceedings, volume 7808. Springer

Iyer S, Killingback T, Sundaram B, Wang Z (2013) Attack robustness and centrality of complex networks. PLoS ONE 8(4):e59613

Jahanpour E, Chen X (2013) Analysis of complex network performance and heuristic node removal strategies. Commun Nonlinear Sci Numer Simul 18(12):3458–3468

Jürgens P, Jungherr A, Schoen H (2011) Small worlds with a difference: new gatekeepers and the filtering of political information on twitter. In: Proceedings of the 3rd international web science conference, pp 1–5

Kane GC, Alavi M, Labianca G (2014) What's different about social media networks? A framework and research Agenda. MIS Q 38(1):275–304

Kiesler S, Kraut R, Resnick P, Kittur A (2012) Regulating behavior in online communities. Building successful online communities: evidence-based social design

Kunegis J (2013) Konect: the koblenz network collection. In: Proceedings of the 22nd international conference on world wide web. ACM, pp 1343–1350

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, pp 591–600. ACM

Latapy M (2008) Main-memory triangle computations for very large (sparse (power-law)) graphs. Theor Comput Sci 407(1–3):458–473

Lee E, Leets L (2002) Persuasive storytelling by hate groups online: examining its effects on adolescents. Am Behav Sci 45(6):927–957

Lehmann S, Ahn Y-Y (2018) Complex spreading phenomena in social systems. Springer, Berlin

Lerman K, Ghosh R (2010) Information contagion: An empirical study of the spread of news on digg and twitter social networks. In: Fourth international AAAI conference on weblogs and social media

Leskovec Jure, Kleinberg Jon, Faloutsos Christos (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM

Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. ACM Trans Knowl Discov Data (TKDD) 1(1):2

Liu L, Tang J, Han J, Yang S (2012) Learning influence from heterogeneous social networks. Data Min Knowl Discov 25(3):511–544

Liu H, Burnap P, Alorainy W, Williams ML (2019a) Fuzzy multi-task learning for hate speech type identification. In: The world wide web conference, pp 3006–3012

Liu H, Burnap P, Alorainy W, Williams ML (2019b) A fuzzy approach to text classification with two-stage training for ambiguous instances. IEEE Trans Comput Soc Syst 6(2):227–240

Luarn P, Chiu Y-P (2016) Influence of network density on information diffusion on social network sites: the mediating effects of transmitter activity. Inf Dev 32(3):389–397

Mathew B, Kumar N, Goyal P, Mukherjee A et al (2018) Analyzing the hate and counter speech accounts on twitter. arXiv:1812.02712

Mathew B, Dutt R, Goyal P, Mukherjee A (2019) Spread of hate speech in online social media. In: Proceedings of the 10th ACM conference on web science, pp 173–182

Mueller AS, Abrutyn S (2015) Suicidal disclosures among friends: using social network data to understand suicide contagion. J Health Soc Behav 56(1):131–148

Myers SA, Sharma A, Gupta P, Lin J (2014) Information network or social network?: The structure of the twitter follow graph. In: Proceedings of the 23rd international conference on world wide web. ACM, pp 493–498

Newman MEJ (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E 64(1):016132

Newman MEJ (2008) The mathematics of networks. New Palgrave Encycl Econ 2(2008):1–12

Newman MEJ, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66(3):035101

Nie T, Guo Z, Zhao K, Zhe-Ming L (2015) New attack strategies for complex networks. Phys A Stat Mech Appl 424:248–253

Otsuka M, Tsugawa S (2019) Robustness of network attack strategies against node sampling and link errors. PLoS ONE 14(9):e0221885

Palmer CR, Siganos G, Faloutsos M, Faloutsos C, Gibbons PB (2001) The connectivity and fault-tolerance of the internet topology

Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. Phys Rev E 65(3):036104

Paul I, Khattar A, Kumaraguru P, Gupta M, Chopra S (2018) Elites tweet? Characterizing the twitter verified user network. arXiv:1812.09710

Pelaprat E, Brown B (2012) Reciprocity: understanding online social relations. First Monday 17(10)

Petersen RR, Rhodes CJ, Wiil UK (2011) Node removal in criminal networks. In: 2011 European intelligence and security informatics conference. IEEE, pp 360–365

Pržulj N (2007) Biological network comparison using graphlet degree distribution. Bioinformatics 23(2):e177–e183

Putnam RD (2000) Bowling alone: the collapse and revival of American community. Simon and Schuster, New York

Rao AR, Bandyopadhyay S (1987) Measures of reciprocity in a social network. Sankhyā Indian J Stat Ser A 141–188

Ribeiro MH, Calais PH, Santos YA, Almeida VAF, Meira Jr W (2017) “Like sheep among wolves”: characterizing hateful users on twitter. arXiv:1801.00317

Ribeiro MH, Calais PH, Santos YA, Almeida VAF, Meira Jr W (2018) Characterizing and detecting hateful users on twitter. arXiv:1803.08977

Roland D, Spurr J, Cabrera D (2017) Preliminary evidence for the emergence of a health care online community of practice: using a netnographic framework for twitter hashtag analytics. J Med Internet Res 19(7):e252

Sainudiin R, Yogeeswaran K, Nash K, Sahioun R (2019) Characterizing the twitter network of prominent politicians and SPLC-defined hate groups in the 2016 us presidential election. Soc Netw Anal Min 9(1):34

Scott J (1988) Social network analysis. Sociology 22(1):109–127

Sparrowe RT, Liden RC, Wayne SJ, Kraimer ML (2001) Social networks and the performance of individuals and groups. Acad Manag J 44(2):316–325

Stepanyan K, Borau K, Ullrich C (2010) A social network analysis perspective on student interaction within the twitter microblogging environment. In: 2010 10th IEEE international conference on advanced learning technologies. IEEE, pp 70–72

Stephan WS, Stephan CW (2013) An integrated threat theory of prejudice. In: Reducing prejudice and discrimination, pp 33–56. Psychology Press

Tarjan R (1972) Depth-first search and linear graph algorithms. SIAM J Comput 1(2):146–160

Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 61(12):2544–2558

Ting IH, Wang SL, Chi H-M, Wu J-S (2013) Content matters: a study of hate groups detection based on social networks analysis and web mining. In: 2013 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2013). IEEE, pp 1196–1201

Wadhwa P, Bhatia MPS (2014) Discovering hidden networks in on-line social networks. Int J Intell Syst Appl 6(5):44–54

Wang XF (2003) Complex networks: small-world, scale-free and beyond. IEEE Circuits Syst Mag 3(1):6–20

Wei W, Joseph K, Liu H, Carley KM (2015) The fragility of twitter social networks against suspended users. In: 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 9–16

Wiil UK, Gniadek J, Memon N (2010) Measuring link importance in terrorist networks. In: 2010 international conference on advances in social networks analysis and mining. IEEE, pp 225–232

Williams ML, Burnap P, Javed A, Liu H, Ozalp S (2019) Hate in the machine: anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. Br J Criminol 60(1):93–117

Xu J, Chen H (2008) The topology of dark networks. Commun ACM 51(10):58–65

Yip M, Shadbolt N, Webber C (2012) Structural analysis of online criminal social networks. In: 2012 IEEE international conference on intelligence and security informatics. IEEE, pp 60–65

Zhao X, Liu F, Wang J, Li T et al (2017) Evaluating influential nodes in social networks by local centrality with a coefficient. ISPRS Int J Geo-Inf 6(2):35

Zhou Y, Reid E, Qin J, Chen H, Lai G (2005) Us domestic extremist groups on the web: link and content analysis. IEEE Intell Syst 20(5):44–51

Zykov AA (1990) Fundamentals of graph theory. BCS Associates Moscow