The Journal of Social Media in Society
Spring 2025, Vol. 14, No. 1, Page 258-278
thejsms.org

The Role of Networks in Spreading Hate Speech on Twitter

Jari-Mikko Meriläinen
Department of Economics, Jyväskylä University School of Business and Economics, Jyväskylä, Finland
jari-mikko.jm.merilainen@jyu.fi

This study aims to investigate the role of networks in the dissemination of hate speech on Twitter. With a sample of participants engaging in hate speech on Twitter, we examine the influence of popular users and retweeting in the spread of hate speech, using a manually annotated tweet sample. We divide the sample of hate speech participants into two groups: the core and outer circle. The core refers to users frequently followed by other hate speech participants while the outer circle comprises users who engage in hate speech but are not regularly followed by other perpetrators. We programmatically investigate the dynamics of hate speech in this network. Our findings reveal that popular users play a dual role: actively sending hateful tweets on other users' threads and retweeting hateful tweets from outside the network to their followers. However, hate speech represents only a small portion of their overall Twitter activity, as they predominantly share non-hateful tweets. Consequently, their Twitter feeds remain, on average, primarily non-hateful.

Keywords: hate speech, networks, social media, online communication, Twitter

---

Even if hate speech is prevalent on Twitter, little is known about the behavior of the producers of such content. This study aims to fill this gap by analyzing a large sample of manually classified tweets to examine hate speech on Twitter. We focus on understanding the role of networks in the spread of hate speech. To accomplish this, we select a sample of Twitter users who have indulged in hate speech targeting Finnish government ministers, as previously identified by Meriläinen (2022). We aim to determine whether users followed by other individuals in this sample are prominent sources of hate speech, investigating whether hate speech created by highly followed users spreads throughout the network. Additionally, we explore whether popular users use this network to share hateful content with their followers by retweeting hate speech.

Despite the centrality of networking on social media platforms, research on the role of networks in hate speech has been limited. To our knowledge, only one study by Mathew et al. (2019) has addressed a similar issue. Their findings suggest that hateful content spreads more rapidly than general content on the Gab social network. Besides this, Stewart et al. (2023) examine the role of hate influencers in the propagation of hate on Telegram, indicating that influencers employ social media to mobilize members, share information, and spread misinformation.

Unlike Mathew et al. (2019), we employ a manual classification of tweets to distinguish between hateful and non-hateful content, enhancing the precision of the annotation. We divide the network into the core network and the outer circle, with the former representing the top 10% of users most frequently followed by other hate speech participants. In doing so, we assume that the most followed users are the most visible in this network. The outer circle consists of the residue, i.e., the nine other deciles of the sample users.

Our hypotheses are as follows: (i) the core produces hate speech that is redistributed (i.e., retweeted) by users in the outer circle, (ii) users in the outer circle produce hate speech that is likewise disseminated by the core users, (iii) hateful content created by core users is retweeted by other core users, and (iv) the core transmits hate speech from outside the network to both the core and the outer circle. We analyze tweet identification numbers and Twitter names programmatically to address these questions. As one of the pioneering studies focusing on the role of social media networks in hate speech, this research fills a significant gap in the current literature and offers new insights for social media and hate speech research.

This study is organized as follows: following the introduction, we present a literature review, followed by descriptions of the dataset and the definition of hate speech. Next, we present the results and conclude with a discussion of our findings.

LITERATURE REVIEW

As mentioned, only a few studies examine the role of networks in online hate speech. However, network structures and the different roles of social media users have been studied. For example, Recuero et al. (2019) collected Twitter data related to the corruption trial of former Brazilian President Lula. They studied the roles of Twitter users

in polarized political conversations. The authors selected key individuals based on their indegree, outdegree, and modularity within the network. They found that individuals with a high indegree, measured by a high number of retweets and mentions, act as activists with a clear political agenda, whereas users with a high outdegree, who retweet and mention various other users, are opinion leaders with a clear political position.

Åkerlund (2020) analyzed tweets with hashtags linked to Swedish far-right discourse on Twitter and showed that influential users stand out by creating original content and defining the conversation for others to engage in. Isa and Himelboim (2018) studied the patterns of information flow within the #FreeAJStaff movement on Twitter, finding that core movement actors and politicians played a mediating role in content distribution. However, Himelboim et al. (2017) used Twitter datasets to demonstrate that different conversation topics have different patterns of information flow. Political issues, for example, were often discussed in high-density and high-modularity topic-networks. Similarly, Himelboim and Han (2014) found that users who consistently tweeted about their cancer had different patterns of information flow, compared to those who tweeted about it sporadically.

Bruns and Highfield (2016) argued that one approach for recognizing the varying forms of online public communication is to unlock the traditional public sphere into a series of public sphericules and micro-publics, which co-exist, intersect, and overlap in various forms. Participating in one public is neither a prerequisite nor an implication that participation in another will result. Accordingly, Himelboim et al. (2013) suggested that Twitter users establish social ties based on not only political opinions but also a wide range of interests. However, they found that political content tends to be confined within like-minded clusters, resulting in a lack of cross-ideological interaction.

Concerning hate speech, Mathew et al. (2019) is the only study we are aware of regarding the role of networks. They used a large sample from Gab to show that content generated by hateful users spreads faster and reaches a wider audience, compared to content generated by normal users. The network of hateful users was found to be almost seventeen times denser than that of normal users, with higher reciprocity values. Additionally, hateful users created nearly 25 percent of the content on Gab, despite comprising only 0.67 percent of the user base. Van Sant et al. (2012) revealed that a small

group of users predominantly created hate speech targeting Finnish ministers on Twitter. However, Meriläinen (2022) found that hate speech tends to be fragmented, rather than centralized.

Ribeiro et al. (2018) concluded that hateful users on Twitter tend to be power users, tweeting more frequently and at shorter intervals. They also follow other users more and favorite more tweets, compared to normal users. Hateful users are densely connected, with 41 percent of their retweets being to other hateful users. However, they have fewer followers than normal users, and on average, their accounts are younger than those of normal users. Meriläinen (2022) found that Twitter accounts associated with hate speech tend to be more recent and short-lived than other users' accounts. ElSherief et al. (2018) observed that hate instigators' Twitter accounts are younger than those of hate speech targets or general users. They also argued that participation in hate speech and being more visible and popular are related.

Studies that specifically focus on hate speech perpetrators are lacking. Siegel (2022) provided a review of such studies. According to Costello and Hawdon (2018), users who produce hateful material online are more likely to be men than women. Certain social networking sites, such as Reddit, Tumblr, and messaging boards, are associated with the dissemination of hateful content. Users closer to online communities or spending more time in hate-populated areas are more likely to produce hateful material. Generally, spending more time online reduces the probability of producing hateful material.

Zych et al. (2023) conducted an online survey involving individuals aged 15–25 from Finland, South Korea, the USA, and Spain. They found that impulsivity and involvement in online identity bubbles were linked to more cyberaggression in all four countries. Women were found to engage in less offensive and threatening messaging, displaying less aggression online than men. Additionally, older users exhibited less aggressive online behavior than younger users. Belonging to an offline network, such as family, friendship groups, school, or work communities, decreased aggressive online behavior. Blaya and Audrin (2019) analyzed survey data from 12–20-year-olds in France and revealed that cyberhate perpetration is strongly correlated with the amount of time spent online, racism, positive attitudes towards violence, and belonging to a deviant youth group. Further, exposure to cyberhate is associated with producing cyberhate.

It is noteworthy to mention that several studies have suggested that online hate speech increased during the COVID-19 pandemic. For example, Kim and Kesari (2021) showed that the prevalence of anti-Asian hate speech increased after Donald Trump called COVID-19 the Chinese virus. Using a large sample of tweets from Italian and British radical right parties and party leaders, Caiani et al. (2021) argued that the pandemic gave the radical right a political opportunity to spread hate speech and attack their political adversaries more broadly. These authors suggested that the radical right, in both the UK and Italy, tends to instrumentalize the pandemic to attack their traditional enemies.

DEFINITION OF A HATEFUL TWEET, DATA, AND NETWORK CHARACTERISTICS

Definition of a Hateful Tweet and Manual Annotation

Defining a hateful tweet is not a straightforward task. Moreover, as argued by Vidgen and Derczynski (2020), abuse is subjective. Consequently, hate speech is defined variously (MacAvaney et al., 2019). Castaño-Pulgarín et al. (2021) concluded that online hate speech could be categorized as religious, racist, political, and gendered. Additionally, online hate speech is often triggered by terrorism. We follow a classification similar to that of Southern and Harmer (2021), based on Papacharissi (2004). We consider a tweet as hateful if it involves racism, sexism, or stereotyping, as well as name-calling and other insulting behavior such as questioning one's intelligence. Even indirect threats of violence are categorized as hateful tweets. However, unlike Southern and Harmer (2021), we do not consider swearing hateful.

Our definition of hate speech is broader than that of the Council of Europe (2016), which defines hate speech as racism, xenophobia, antisemitism, and similar forms of intolerance. Using a similar definition would exclude a significant amount of hateful content. The hate speech perpetrators included in our sample were initially identified in a study focusing on hate speech targeting the government ministers of Finland. This speech was often politically motivated, which may introduce a bias towards political hatred, rather than racial hatred and intolerance. Additionally, due to the overlap of the sampling period with the COVID-19 restrictions, some of the hateful content targeted public health officials.

Various studies have used different approaches for automatically classifying messages as hateful and non-hateful[1]. These approaches include lexicon-based methods (Lingiardi et al., 2020), rule-based methods (Gorrell et al., 2020), and machine learning algorithms (Burnap & Williams, 2016; Vidgen & Yasseri, 2020; Kettunen & Paukkeri, 2021). However, the accuracy of these methods has been criticized by several authors. For example, Kwarteng et al. (2022) argued that automated detection tools lack sensitivity to context. Manual annotation, where tweets are classified by human readers, overcomes this problem, as the context is known. Southern and Harmer (2021) suggested that a tweet may be uncivil without being hateful, leading to false positives. Further, MacAvaney et al. (2019) argued that limitations in data availability for training and testing pose challenges to automatic detection. Arango et al. (2022) suggested that even state-of-the-art models that perform well on their original test sets might not generalize well to other datasets.

The biggest disadvantage of machine learning algorithms is that they learn from a manually classified sample of messages, thus lacking an information advantage over manual annotation. Consequently, their accuracy can at best be as precise as manual annotation. Hence, similar to Southern and Harmer (2021) and Åkerlund (2020), we manually classify tweets into hateful and non-hateful categories by reading them one by one. All the tweets were analyzed by a single researcher.

Southern and Harmer (2021) argued that a machine learning model can produce both false positives and false negatives. Tweets may contain uncivil expressions but may still not be hateful, leading to false positives being thrown up by a machine learning model. Additionally, abusive comments may not be directed at the person being replied to, as suggested by Gorrell et al. (2020).

Concerning false negatives, machine learning techniques may not correctly identify othering and other more subtle forms of abuse. For example, Burnap and Williams (2015) argued that a hateful message can contain othering phrases like "send them home," while Southern and Harmer (2021) pointed out that gendered hate, such as "get back to the kitchen," may be overlooked by automated software. MacAvaney et al. (2019) found that the subtleties in language make automated hate speech detection challenging.

[1]For a review on the use of automatic hate speech detection, see Fortuna and Nunes (2018). Balaji et al. (2021) reviewed the use of machine learning algorithms for social media analysis.

The content of hateful tweets often includes racism. For example, the tweet "*Ei ollut rahtihametta tuohon aikaan*" (They did not have a cargo skirt back then) is a racist and stereotyping tweet targeting the Roma minority of Finland. Likewise, the tweet "*Tämä sompassi [sic] on kova yrittäjä.*" (This Som-panzee tries hard) is a racist tweet targeting Somalians. Islamophobia is also present, as seen in the tweet "*Islam on väkivaltainen maailmanvalta*" (Islam is a violent global power). Many hateful tweets oppose immigration in general, such as "*Sama tarvis tehdä matuinvaasiota vastaan.*" (The same should be done against the invasion of immigrant-invaders) and "*Pentti Heinonen huomasi taistelleensa turhaan, sillä Suomi on miehitetty.*" ([War veteran] Pentti Heinonen learned that he fought for nothing because Finland is occupied).

Hate speech often targets politicians, as shown in the tweets "*Missä sinun aivot ovat?*" (Where are your brains?), "*Ei kansa noita akkoja äänestänyt tuonne, kyllä ne varjoista nostettiin...*" (The people did not vote those bitches there, they were raised from the shadows…" and "*Linnaan sinä tuholaismuija kuulut taustakuiskaajinesi.*" (You vermin bitch belong to a prison together with your prompters). In addition, due to the impact of the COVID-19 restrictions and vaccination campaigns, hateful tweets sometimes targeted pandemic officials, such as "*Maailman vaarallisin massamurhaaja Fauci.*" (The world's most dangerous mass murderer Fauci) and "*Vedä vittu päähäsi sekopää runkku*" (Pull a cunt over your head you nutcase wanker), the latter having been sent to a Finnish pandemic official.

Data

The sample consists of a total of 34,542 tweets. A portion of these has been manually classified as hateful or non-hateful. Specifically, these were tweeted by users who are among the top ten most followed hate speech perpetrators in the sample of 269 users. There are 28 users in the top decile. The sample period is the week from January 21, 2022, to January 28, 2022. Therefore, the tweet sample includes all tweets sent by these users during this period. Consequently, tweets are not filtered by location, language, or similar criteria. We used the Twitter API to collect the sample. The top decile users sent 11,138 tweets, accounting for approximately one-third of the total tweet sample. Importantly, this sample includes only public messages and excludes private messages sent through Twitter. Further, Rogers (2020) argued that the 'deplatforming' of extremists

by Twitter and other leading social media platforms has led to their migration to alternative platforms like Telegram. Therefore, using Twitter data alone may not provide a complete picture of online hate speech.

As mentioned in the introduction, hate speech perpetrators were identified by Meriläinen (2022), in whose study, the number of users sending hateful tweets was approximately 300. In January 2022, we could find 269 of these users on Twitter. Thus, in our sample, the number of users engaging in hate speech is 269. For this study, we specifically focus on the top decile of these 269 users in terms of followers. We consider these users the most popular within the sample of hate tweet perpetrators. Of the 269 users, 255 follow at least one other hate speech participant. Therefore, a group of users within this sample does not follow any other sample user.

Fig. 1 displays the average number of followers within the sample network by decile in terms of followers. The figure demonstrates that many hate speech participants do not have numerous followers within this network. Approximately half the users have fewer than ten followers in the sample. However, on average, users in the highest decile are followed by over seventy other hate speech participants.

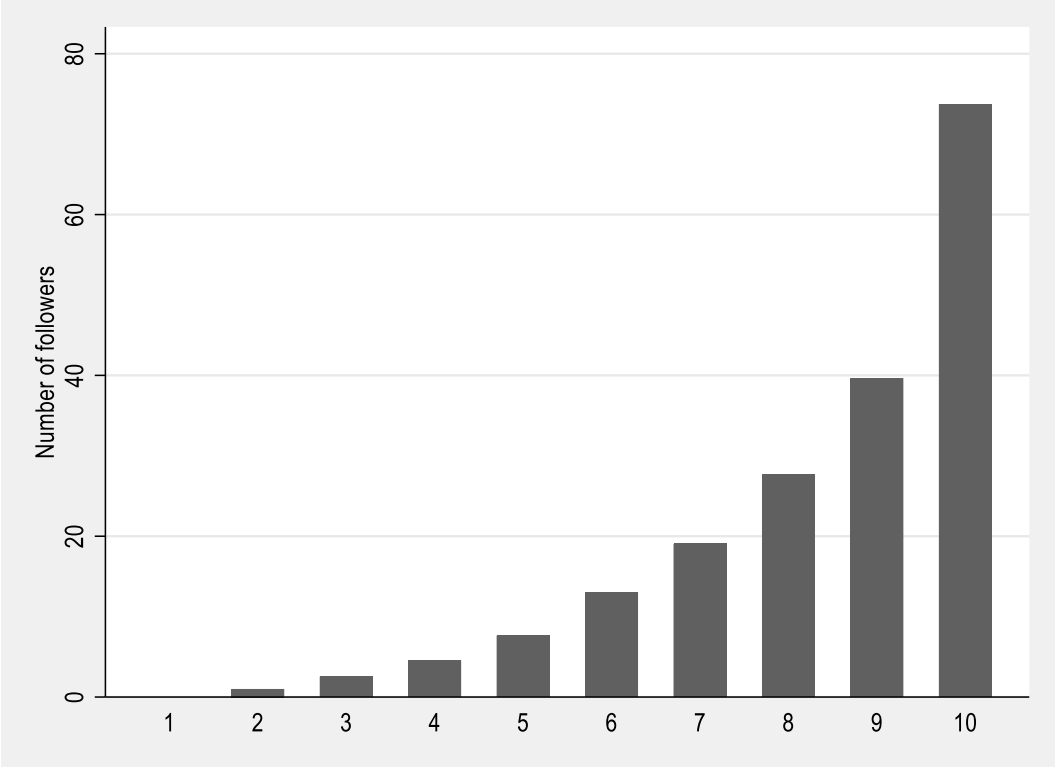

Figure 1. Number of In-sample Followers by Decile. The bars illustrate the number of followers, grouped by decile, with followers consisting of other hate speech perpetrators.

Table 1 shows the relative frequency of hate speech by the core user. We anonymized the dataset for ethical reasons[2]. The table is arranged by the number of followers the users have in the sample. On average, about seven percent of tweets, i.e., every fourteenth tweet, sent by the core users are hateful. However, there are large differences between the core users. The highest relative frequency is 37 percent, and the lowest relative frequencies are zeros. Thus, these two users did not send any hateful tweets during the sample period. Further, some of them did not actively participate in discussions. Hence, their number of observations is quite low.

In contrast, some users were very active on Twitter during the sample period. User A is the most followed of the core users, and about every sixth of her tweets is hateful. User N sent the most hateful tweets in absolute terms. The rightmost column shows the number of edges the users have in this network. This figure is visibly correlated with the number of followers (the linear correlation coefficient is 0.72). The difference is that edges include both directions, i.e., followers and those that the users follow. Thus, User B is not only frequently followed in this network because he also follows other users in this sample.

[2]Burnap et al. (2017) recommended that sensitive social media content should not be published without consent. Fiesler and Profeser (2018) found that most users feel uncomfortable when their tweets are used in a research study.

Table 1

The Core Users, Their Relative Hate Speech Frequency, and the Number of Followers Among the Hate Speech Participants

	Relative hate speech frequency	n	Number of hateful tweets	Followers	Edges
User A	0.138	254	35	114	239
User B	0.21	243	51	108	368
User C	0.02	406	8	106	113
User D	0.103	783	81	97	208
User E	0.009	439	4	86	126
User F	0.046	302	14	86	187
User G	0.043	282	12	84	166
User H	0.075	398	30	82	155
User I	0.074	472	35	77	153
User J	0.085	106	9	72	102
User K	0.168	125	21	70	157
User L	0.029	485	14	69	119
User M	0.052	786	41	64	129
User N	0.045	1,907	86	61	144
User O	0.344	160	55	59	85
User P	0.062	694	43	56	169
User Q	0.089	661	59	56	87
User R	0.007	286	2	56	57
User S	0.076	184	14	56	84
User T	0.369	122	45	56	114
User U	0.094	469	44	54	157
User V	0.111	117	13	53	108
User W	0	35	0	52	52
User X	0.025	755	19	51	107
User Y	0.03	464	14	50	90
User Z	0.026	39	1	49	114
User AA	0.18	61	11	48	116
User AB	0	7	0	48	55
Mean	0.069	394.4	27.2	68.6	134.3

Notes. Relative hate speech frequency shows the share of hateful tweets out of the total tweets sent by the user. The number of followers is the number of other hate speech participants following the core user. n = the number of observations. Edges is the number of edges, i.e., connections, the individuals have in the sample.

Network characteristics

We built an edge list of the 269 hate speech participants, which includes all the following/follower relations between these 269 users. Thus, the hate speech participants represent the nodes in the network, and the connections between them are the edges. As explained, there are 28 users in the top decile. Therefore, the rest of the sample, i.e., the outer circle, contains 241 users. Table 2 shows the network characteristics for the top

decile, the nine other deciles, and the full sample. The figures suggest that the users in the top decile are relatively well connected with each other. The edge density, i.e., the share of connections out of the total possible connections, is 67 percent.

Moreover, the reciprocity is 83 percent, which means that if user A follows user B, the latter has an 83 percent likelihood of following the former. Similarly, the likelihood of following is 84 percent if the two users share a connection. This is shown by the transitivity parameter.

The figures are perceptibly lower for the sample containing the nine other deciles. The edge density is 4.1 percent in this group. Compared to the edge density of the top decile, this is a low result. Nonetheless, reciprocity is 65 percent, indicating that users are likely to follow each other in the outer circle if there is a connection between them. Transitivity is quite low at 21 percent. To conclude, the core is denser, and the mean distance is shorter than that of the outer circle. In addition, users are more likely to follow each other.

Regarding the degree centralization figures, the results do not support a star-shaped network for the core or the outer circle. This can be interpreted from the low scores for the three types of degree centralization. A value of 1 would indicate a star-shaped network, i.e., that a single node interacts with all the other nodes, but the others are tied only to the first node. As for a value of 0, it would indicate that degree centrality is evenly dispersed in the network. For example, a circle-shaped network, i.e., A follows B, B follows C, C follows D, etc., has a centrality value of 0. Similarly, if all the nodes follow all the other nodes in the network, the value is 0. Since all the figures of both the core and the outer circle are quite low, we conclude that there are no central nodes in these networks. However, out-degree centralization is always higher than in-degree centralization.

The results are different when the figures are calculated for the full network containing all the hate speech perpetrators. Considering both directions, the degree centralization measure is 0.53. Further, the out-degree measure is considerably higher at 0.64 than the in-degree measure at 0.43. This indicates that the network includes users who follow many other users. Thus, the star-shaped form of the network is caused mainly by users following many other users in the same network, and not someone being followed by many others.

Table 2
Network Characteristics

	Core	Outer circle	Core and outer circle
Edge density	0.67	0.04	0.09
Mean distance	1.34	2.79	2.22
Reciprocity	0.83	0.65	0.70
Transitivity	0.84	0.21	0.39
Degree centralization: all	0.22	0.21	0.53
Degree centralization: in-degree	0.18	0.12	0.43
Degree centralization: out-degree	0.29	0.29	0.64

RESULTS

Table 3 shows the number of hateful and non-hateful tweets by tweet type. First, only a small fraction (about 7 percent) of tweets by the core users are hateful. This indicates that these users use Twitter primarily for purposes other than creating or spreading hateful content. Further, the rightmost column shows that the users in the top decile retweet very often because almost half of their updates are retweets. However, hate speech occurs most often as replies to other Twitter users, but the difference from retweeting is small. Quoting a hateful tweet as a reply to another user is rare. Quoting a non-hateful tweet as a reply is rare as well. Moreover, the users in the top decile are more likely to publish independent hateful tweets than quoted hateful tweets. To sum up, these users actively participate in discussions and send hateful tweets as replies to other users. Besides this, they retweet hateful tweets. However, most (93 percent) of their tweets are not hateful.

Table 3
Descriptive Statistics by Tweet Type for Tweets Sent by the Core Users

	Hateful		Non-hateful		Total	
	#	%	#	%	#	%
Independent	67	8.80 %	940	9.14 %	1007	9.12 %
Quoted tweet	62	8.15 %	415	4.04 %	477	4.32 %
Quoted reply	8	1.05 %	152	1.48 %	160	1.45 %
Reply	316	41.52 %	3587	34.89 %	3903	35.35 %
Retweet	308	40.47 %	5187	50.45 %	5495	49.76 %
Total	761	100.00 %	10281	100.00 %	11042	100.00 %

Notes. An independent tweet is a tweet published by the user in the user's own timeline. A quoted tweet is a retweet with a comment.

Table 4 shows the means and standard deviations for retweet counts of hateful and non-hateful tweets by the users in the top decile. The figures suggest that independent tweets published by these users are not frequently retweeted. On average, their tweets are retweeted only 2.3 times. Further, even the maximum number of retweets for a hateful independent tweet is low, at 14. Moreover, their non-hateful independent tweets are retweeted more often than their hateful independent tweets. Similarly, quoted tweets, quoted tweets sent as a reply, and replies to other users are not often retweeted. There is a striking difference in the number of retweets for retweets. Nonetheless, this figure is the total number of retweets for the tweet on Twitter. It is even higher for non-hateful tweets, which can be interpreted such that these users have retweeted popular non-hateful tweets that circulate on Twitter. The maximum retweet count for a retweet is more than 100,000. Resulting from such outliers, the standard deviations for retweet counts are large.

Table 4
Descriptive Statistics for Retweet Counts by Tweet Type for Tweets by the Core Users

Hateful tweets	Mean	Std. Dev.	Freq.	Max
Independent	1.7	2.6	67	14
Quoted tweet	1.7	5.0	62	32
Quoted reply	0.9	2.5	8	7
Reply	0.4	1.1	316	11
Retweet	58.7	382.5	308	4564
All hateful tweets	24.2	244.8	761	4564
Non-hateful tweets	Mean	Std. Dev.	Freq.	Max
Independent	2.4	8.4	940	80
Quoted tweet	2.1	6.3	415	81
Quoted reply	0.4	1.3	152	14
Reply	0.2	1.4	3,587	49
Retweet	190.2	1871.5	5,187	101,093
All non-hateful tweets	96.4	1332.6	10,281	101,093
Full sample	Mean	Std. Dev.	Freq.	Max
Independent	2.3	8.1	1007	80
Quoted tweet	2.0	6.1	477	81
Quoted reply	0.4	1.4	160	14
Reply	0.3	1.4	3903	49
Retweet	182.9	1820.7	5495	101,093
All tweets	91.4	1287.6	11,042	101,093

Notes. An independent tweet is one published by the user in the user's own timeline. A quoted tweet is a retweet with comment.

Table 5 displays the statistics for retweets by the core users. In total, there are 5495 retweets. We programmatically compared the sample usernames to the usernames in the core's retweets. The figures suggest that most retweets (4779 tweets) are tweets from outside the network. This means that the core users transmit tweets from outside the network to their followers.

Further, the core users retweet much more from the core (296 tweets) than from the outer circle (150 tweets). Despite this, the number of hateful retweets from the core is low because about five percent (16 tweets) of tweets retweeted from the core are hateful. This is higher at about 17 percent for the tweets retweeted from the outer circle. Finally, approximately five percent (270 tweets) of retweets from external sources are deemed hateful. Consequently, we conclude that despite core users occasionally retweeting hateful content, they share non-hateful material far more frequently.

Table 5 also reveals similar statistics for quoted tweets. Quoted tweets are retweets with comments. As the figures in Table 3 already indicate, quoting a tweet is much rarer than retweeting. Accordingly, the number of quoted hateful tweets that originate from the core is very low, at 6 tweets. Likewise, the number of hateful tweets quoting tweets from the outer circle is minimal.

Similar to retweets, the core quotes tweets from outside the network. However, the number of these tweets is much lower than that of retweets. Reading these 62 tweets reveals hateful comments directed at other users' tweets. Please note that the tweet identification numbers were programmatically compared to the identification numbers of the quoted tweets. Consequently, it is possible that a user in the outer circle quoted the tweet before the core user. Therefore, quoted tweets originating from outside the network may have been previously tweeted by a user or users in the outer circle.

Table 5
Hateful and Non-hateful Retweets by the Core Users

	Retweets from the core	Retweets from the outer circle	Retweets from outside	Total
Hateful	16	22	270	308
Non-hateful	280	128	4779	5187
Total	296	150	5049	5495
	Quoted tweets from the core	Quoted tweets from the outer circle	Quoted tweets from outside	Total
Hateful	6	7	49	62
Non-hateful	46	0	369	415
Total	52	7	418	477

Notes. A quoted tweet is a retweet with a comment.

Table 6 presents the statistics for retweets from the core by the outer circle. The data reveal that approximately three percent of their retweets originate from the core users. These figures were generated programmatically by comparing the usernames of retweets with those of the core users. Additionally, about ten percent of the retweets in our classified sample of tweets are identified as hateful. We identified this by comparing the tweet identification numbers of the outer circle's retweets to the identification numbers of the identified hateful tweets. Considering the presence of 28 core users, this averages to approximately one tweet per core user. We consider this a relatively low number. Consequently, we infer that users in the outer circle do not frequently share hateful content created by the core users.

We also investigated whether the hateful tweets retweeted by the core users spread within the outer circle as retweets. This was accomplished by comparing the identification numbers of the core users' hateful retweets to the identification numbers of the outer circle's retweets. The results indicate that the 293 hateful retweets were collectively retweeted 517 times by users in the outer circle. It is possible that some of these retweets were seen by users in the outer circle before the core users retweeted them. Unfortunately, the Twitter API does not allow for examination of this issue because the identification data include only the number of the original tweet, not that of the retweet. Last, users in the outer circle did not share hateful content created by the core as quoted tweets. These results are not reported because they are zeroes. Naturally, they are available upon request.

Table 6
Hateful and Non-hateful Retweets from the Core by the Outer Circle

Outer circle	Retweets from the core
Hateful retweets	31
Non-hateful retweets	256
Total retweets from the core by the outer circle	287
Total retweets by the outer circle	9273
Hateful retweets by the core	308
Same tweets retweeted by the outer circle	517

DISCUSSION

This study used a large sample of manually classified tweets to examine the role of networks in hate speech on Twitter. We employed a previously identified sample of hate speech perpetrators to investigate whether popular users play a key role in the dissemination of hate speech. Our sample comprised all the tweets by the 269 hate speech participants during one week in January 2022. We categorized these users into two groups: the core, representing the most followed decile by users in this network, and the outer circle, comprising the other nine deciles of the sample users. They participate in hate speech but are not frequently followed by others in this network. This study extends those of Åkerlund (2020) and Isa and Himelboim (2018), who studied the role of influential actors in social media information flow, as well as Mathew et al. (2019), who investigated the role of networks in hate speech.

The results suggest that the most followed hate speech participants play a dual role in hate speech. They actively send hateful tweets in other users' threads and, additionally, they retweet hateful tweets from outside the network, distributing them to their Twitter followers. To a much lesser extent, the core users also create hateful content as quoted tweets, with most of it related to sources outside the network. Consequently, we conclude that Twitter networks have a role in hate speech because prominent hate users leverage their visibility to disseminate hateful content. Hateful content independently created by these users has a much smaller volume than retweeted hateful content and hateful content created as replies. However, users in the outer circle do not often retweet the hateful tweets created by the core users. This implies that the network serves to amplify the visibility of hateful tweets but is not used to spread them further.

Further, network analysis revealed that the network has a star shape. However, the out-degree network is much closer to a star than the in-degree network. This implies that no users are being followed at the center of the star; instead, there are users following many other users in the center. These users have the potential to retweet content shared or created by the core users. Nonetheless, our results do not suggest that hateful content created by the core users or the outer circle is frequently retweeted.

It is important to highlight that the core users primarily create and retweet non-hateful content, producing many times more non-hateful content than hateful content. Consequently, the Twitter feeds of these users are, on average, non-hateful since only a small fraction of their tweets are hateful. We find this problematic for preventing hateful speech because recognizing users engaged in hate speech, whether automatically or manually, is challenging. This difficulty arises because hateful content is dispersed among non-hateful content. Moreover, this supports the suggestions of Meriläinen (2022) and Ayo et al. (2021), who argue that hate speech is a fragmented phenomenon. Whether the predominantly non-hateful flow of tweets is an intentional strategy for concealing hate speech or simply an outcome of the core users' personalities is a question we are unable to answer. Moreover, during the reading process, we noticed that meme photos can sometimes be abusive, which means a machine learning algorithm analyzing only text would miss abusive messages. Perhaps this is done intentionally, to prevent automated hate speech detection.

The tweet sample covered one week, introducing the possibility of selection bias in our results. Additionally, the hate speech perpetrators were originally identified in a study focusing on hate speech targeting politicians, which could introduce bias toward political hate speech into our findings. Further, the tweets were classified by a single author. Since abuse is subjective, this may cause bias in our results. As a suggestion for further research, we propose a similar study using a different sample.

References

Åkerlund, M. (2020). The importance of influential users in (re) producing Swedish far-right discourse on Twitter. *European Journal of Communication*, *35*(6), 613–628. doi:10.1177/0267323120940909

Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, *105*, 101584. doi:10.1016/j.is.2020.101584

Ayo, F. E., Folorunso, O., Ibharalu, F. T., Osinuga, I. A., & Abayomi-Alli, A. (2021). A probabilistic clustering model for hate speech classification in twitter. *Expert Systems with Applications*, *173*, 114762. doi:10.1016/j.eswa.2021.114762

Balaji, T. K., Annavarapu, C. S. R., & Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, *40*, 100395. doi:10.1016/j.cosrev.2021.100395

Blaya, C., & Audrin, C. (2019, June). Toward an understanding of the characteristics of secondary school cyberhate perpetrators. *Frontiers in Education*, *Vol. 4*, p. 46. Frontiers Media SA. doi:10.3389/feduc.2019.00046

Bruns, A., & Highfield, T. (2015). Is Habermas on Twitter?: Social Media and the Public Sphere. In *The Routledge Companion to Social Media and Politics*, 56–73. Routledge.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242. doi:10.1002/poi3.85

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*, 1–15. doi:10.1140/epjds/s13688-016-0072-6

Burnap, P., Sloan, L., & Williams, M. L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*(6), 1149–1168. doi:10.1177/0038038517708140

Caiani, M., Carlotti, B., & Padoan, E. (2021). Online hate speech and the radical right in times of pandemic: The Italian and English cases. *Javnost—The Public*, *28*(2), 202–218. doi:10.1080/13183222.2021.1922191

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. doi:10.1016/j.avb.2021.101608

Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender*, *5*(1), 55–60. doi:10.1089/vio.2017.0048

Council of Europe (2016) Recommendations and declarations of the Committee of Ministers of the Council of Europe in the field of media and information society. Available at: https://go.coe.int/URzjs [last accessed 24 October 2023]

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018, June). Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, *Vol. 12*, No. 1. doi:10.1609/icwsm.v12i1.15038

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society*, *4*(1), 2056305118763366. doi:10.1177/2056305118763366

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30. doi:10.1145/3232676

Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., & Bontcheva, K. (2020). Which politicians receive abuse? Four factors illuminated in the UK general election 2019. *EPJ Data Science*, *9*(1), 18. doi:10.1140/epjds/s13688-020-00236-9

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-mediated Communication*, *18*(2), 154–174. doi:10.1111/jcc4.12001

Himelboim, I., & Han, J. Y. (2014). Cancer talk on twitter: community structure and information sources in breast and prostate cancer social networks. *Journal of Health Communication*, *19*(2), 210–225. doi:10.1080/10810730.2013.811321

Himelboim, I., Smith, M. A., Rainie, L., Shneiderman, B., & Espina, C. (2017). Classifying Twitter topic-networks using social network analysis. *Social Media + Society*, *3*(1), 2056305117691545. doi:10.1177/2056305117691545

Isa, D., & Himelboim, I. (2018). A social networks approach to online social movement: Social mediators and mediated content in #FreeAJstaff twitter network. *Social Media + Society*, *4*(1), 2056305118760807. doi:10.1177/2056305118760807

Kettunen, L., & Paukkeri, M. S. (2021). Tekoälyn hyödyntäminen vihapuheen seurannassa [*Utilising artificial intelligence for monitoring hate speech*]. http://urn.fi/URN:ISBN:978-952-259-893-6 [last accessed 30 October 2023].

Kim, J. Y., & Kesari, A. (2021). Misinformation and hate speech: The case of anti-Asian hate speech during the COVID-19 pandemic. *Journal of Online Trust and Safety*, *1*(1). doi:10.54501/jots.v1i1.13

Kwarteng, J., Perfumi, S. C., Farrell, T., Third, A., & Fernandez, M. (2022). Misogynoir: challenges in detecting intersectional hate. *Social Network Analysis and Mining*, *12*(1), 166. doi:10.1007/s13278-022-01008-1

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, *39*(7), 711–721. doi:10.1080/0144929X.2019.1607903

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS ONE*, *14*(8), e0221152. doi:10.1371/journal.pone.0221152

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019, June). Spread of hate speech in online social media. *Proceedings of the 10th ACM conference on web science*, 173–182. doi:10.1145/3292522.3326034

Meriläinen (2022). Marinin hallituksen ministereihin Twitterissä suunnattu vihapuhe (*Hate speech targeting the ministers of Sanna Marin's cabinet in Twitter*). *Prologi 18*(1). doi:10.33352/prlg.122025

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. doi:10.1177/1461444804041444

Recuero, R., Zago, G., & Soares, F. (2019). Using social network analysis and social capital to identify user roles on polarized political conversations on Twitter. *Social Media + Society*, *5*(2), 2056305119848745. doi:10.1177/2056305119848745

Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr, W. (2018). Characterizing and detecting hateful users on twitter. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. *12*, No. 1). doi:10.1609/icwsm.v12i1.15057

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. doi:10.1177/0267323120922066

Siegel, A. (2020). Online Hate Speech. In N. Persily & J. Tucker (Eds.), *Social Media and Democracy: The State of the Field, Prospects for Reform* (SSRC Anxieties of Democracy, pp. 56–88). Cambridge: Cambridge University Press.

Southern, R., & Harmer, E. (2021). Twitter, incivility and "everyday" gendered othering: An analysis of tweets sent to UK members of Parliament. *Social Science Computer Review*, *39*(2), s. 259–275. doi:10.1177/0894439319865519

Stewart, N. K., Al-Rawi, A., Celestini, C., & Worku, N. (2023). Hate influencers' mediation of hate on Telegram: "We Declare War Against the Anti-White System". *Social Media + Society*, *9*(2), 20563051231177915. doi:10.1177/20563051231177915

Van Sant, K., Fredheim, R., & Bergmanis-Korats, G. (2021). Abuse of power: Coordinated online harassment of Finnish government ministers. Riga: NATO Strategic Communications Centre of Excellence. Direct link (last accessed 11 Nov 2021): https://stratcomcoe.org/publications/abuse-of-power-coordinated-online-harassment-of-finnish-government-ministers/5

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos ONE*, *15*(12), e0243300. doi:10.1371/journal.pone.0243300

Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, *17*(1), 66–78. doi:10.1080/19331681.2019.1702607

Zych, I., Kaakinen, M., Savolainen, I., Sirola, A., Paek, H. J., & Oksanen, A. (2023). The role of impulsivity, social relations online and offline, and compulsive Internet use in cyberaggression: A four-country study. *New Media & Society*, *25*(1), 181–198. doi:10.1177/14614448211009459

Funding and Acknowledgements

This work was supported by the OP Research Foundation and the Foundation for Economic Education. The author declares no conflicts of interest. The author is deeply grateful for the anonymous reviewer's time and effort in reviewing the manuscript. Their valuable comments and constructive feedback have significantly enhanced the quality and clarity of this study.

Data availability statement

The data and codes supporting the findings of this study are openly available in the Harvard DataVerse at https://doi.org/10.7910/DVN/HQMVWK. The data has been anonymized, and all Twitter data has been removed due to copyright reasons.

Online Connection

Jari-Mikko Meriläinen: https://www.linkedin.com/in/jari-mikko-meril%C3%A4inen-929904bb
X handle: @JariMikko