Spread of Hate Speech in Online Social Media

Binny Mathew
Indian Institute of Technology Kharagpur
Kharagpur, India
binnymathew@iitkgp.ac.in

Ritam Dutt
Indian Institute of Technology Kharagpur
Kharagpur, India
ritam.dutt@gmail.com

Pawan Goyal
Indian Institute of Technology Kharagpur
Kharagpur, India
pawang.iitk@gmail.com

Animesh Mukherjee
Indian Institute of Technology Kharagpur
Kharagpur, India
animeshm@gmail.com

ABSTRACT

Hate speech is considered to be one of the major issues currently plaguing the online social media. With online hate speech culminating in gruesome scenarios like the Rohingya genocide in Myanmar, anti-Muslim mob violence in Sri Lanka, and the Pittsburgh synagogue shooting, there is a dire need to understand the dynamics of user interaction that facilitate the spread of such hateful content. In this paper, we perform the first study that looks into the diffusion dynamics of the posts made by hateful and non-hateful users on **Gab**[1]. We collect a massive dataset of 341K users with 21M posts and investigate the diffusion of the posts generated by hateful and non-hateful users. We observe that the content generated by the hateful users tend to spread faster, farther and reach a much wider audience as compared to the content generated by normal users. We further analyze the hateful and non-hateful users on the basis of their account and network characteristics. An important finding is that the hateful users are far more densely connected among themselves. Overall, our study provides the first cross-sectional view of how hateful users diffuse hate content in online social media.

CCS CONCEPTS

• **Networks** → **Online social networks**; • **Human-centered computing** → *Social content sharing*; *Social media*; *Empirical studies in collaborative and social computing*; **Social network analysis**.

KEYWORDS

Hate Speech, Gab, Online Social Media, Information Diffusion, DeGroot Model

ACM Reference Format:
Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of Hate Speech in Online Social Media. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3292522.3326034

[1] gab.com

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326034

1 INTRODUCTION

The Internet is one of the greatest innovations of mankind which has brought together people from every race, religion, and nationality. Social media sites such as Twitter and Facebook have connected billions of people[2] and allowed them to share their ideas and opinions instantly. That being said, there are several ill consequences as well such as online harassment, trolling, cyber-bullying, and *hate speech.*

The rise of hate speech: Hate speech has recently received a lot of research attention with several works that focus on detecting hate speech in online social media [4, 10, 12, 21, 35]. Even though several government and social media sites are trying to curb all forms of hate speech, it is still plaguing our society. With hate crimes increasing in several states[3], there is an urgent need to have a better understanding of how the users spread hateful posts in online social media. Companies like Facebook have been accused for instigating anti-Muslim mob violence in Sri Lanka that left three people dead[4] and a United Nations report blamed them for playing a leading role in the possible genocide of the Rohingya community in Myanmar by spreading hate speech[5]. In response to the UN report, Facebook later banned several accounts belonging to Myanmar military officials[6] for spreading hate speech. In the recent Pittsburgh synagogue shooting[7], the sole suspect, *Robert Gregory Bowers*, maintained an account (@onedingo) on Gab[1] and posted his final message before the shooting[8]. Inspection of his Gab account shows months of antisemitc and racist posts that were endorsed by a lot of users on Gab.

The present work: In this paper, we perform the first study which looks into the diffusion dynamics of the posts by hateful users in Gab. We choose Gab for all our analysis. This choice is primarily motivated by the nature of Gab. Unlike other social media sites such as Twitter and Facebook, Gab promotes "free speech" and allows users to post content that may be hateful in nature without any fear of repercussion. This has led to the migration of several

[2] https://techcrunch.com/2018/07/25/facebook-2-5-billion-people
[3] http://www.aaiusa.org/unprecedented_increase_expected_in_upcoming_fbi_hate_crime_report
[4] https://www.theguardian.com/world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hate-speech
[5] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate
[6] https://www.reuters.com/article/us-myanmar-facebook/facebook-bans-myanmar-army-chief-others-in-unprecedented-move-idUSKCN1LC0R7
[7] https://en.wikipedia.org/wiki/Pittsburgh_synagogue_shooting
[8] https://www.independent.co.uk/news/world/americas/pittsburgh-synagogue-shooter-gab-robert-bowers-final-posts-online-comments-a8605721.html

Twitter users who were banned/suspended for violating its terms of service, namely for abusive and/or hateful behavior [45]. This provides a unique opportunity to study how the hateful content would spread in the online medium, if there were no restrictions.

To this end, we crawl the Gab platform and acquire 21M posts by 341K users over a period of 20 Months (October, 2016 to June, 2018). Our analysis reveals that the posts by hateful users tend to spread faster, farther, and wider as compared to normal users.

Our **main contributions** are as follows.

- We perform the first study which looks into the diffusion dynamics of posts by hateful user accounts.
- We find that the hate users in our dataset (which constitute 0.67% of the total number of users) are very densely connected and are responsible for 26.80% of posts generated in Gab.
- We find that the posts of hate users tend to spread fast, farther, and reach a much wider audience as compared to the non-hateful users.

In summary, our analysis reveals that the posts by hateful users have a much higher spreading velocity. These posts receive a much larger audience and as well at a faster rate. As a case study, we also investigate the detailed account characteristics of *Robert Gregory Bowers*, the sole suspect of the Pittsburgh synagogue shooting[7].

2 DATASET

2.1 The Gab social network

Gab[1] is a social media platform launched in August 2016 known for promoting itself as the "Champion of free speech", but has been critised for being a shield for alt-right users [45]. The site is very similar to Twitter, but has very loose moderation policy. According to the Gab guidelines, the site does not allow illegal pornography and promotion of violence and terrorism[9]. All other forms of speech are allowed on Gab. The site allows users to read and write posts of upto 3,000 characters, called "gabs". The site employs an upvoting and downvoting mechanism for posts and allows categorization of posts into topics such as News, Sports, Politics etc.

2.2 Dataset collection

In order to understand the diffusion dynamics in Gab, we collect a massive dataset of posts and users by following the crawling methodology described in Zannettou et al. [45]. We use Gab's API to crawl the site using the well-known snowball strategy. We first obtain the data for the most popular user as returned by Gab's API and then collect the data for all their followers and followings. We collect different types of information as follows: 1) basic details about each user like username, score, account creation date; 2) all the posts of each user; 3) all the followers and followings for each users. This resulted in a massive dataset whose details are presented in Table 1. We have only collected the publicly available data posted in Gab and make no attempt to de-anonymize the users. We outline the procedure to distinguish between hateful and non-hateful users in the following section.

[9]https://gab.com/about/guidelines

Property	Value
Number of posts	21,207,961
Number of reply posts	6,601,521
Number of quote posts	2,085,828
Number of reposts	5,850,331
Number of posts with attachments	9,669,374
Number of user accounts	341,332
Average follower per account	62.56
Average following per account	60.93

Table 1: Description of the dataset.

2.3 Identifying hateful users

Gab has been at the center of several hate activity. With the recent Pittsburg shooting, and removal of the app from play store, it has become quite infamous. The volume of hateful content in Gab is 2.4 times higher than that of Twitter [45] which justifies our choice of Gab. We adopted a multi step approach to curate our dateset.

2.3.1 Lexicon based filtering. We created a lexicon[10] of 45 high-precision unigrams and bigrams that are often associated with hate like 'kike' (slur against Jews), 'paki' (slur against Muslims), 'beached whale' (slur against fat people). These hate words were initially selected from the Hatebase[11] and Urban dictionary[12]. Words such as 'banana', 'bubble' are present in Hatebase which could easily appear in benign context. In order to avoid ambiguity, we ran multiple iterations and carefully chose those keywords which were not ambiguous in Gab.

We leverage these high precision keywords to identify explicit hate posts based on their textual content. The total number of unique posts which have been identified explicitly as 'Hate' were 280,468 or 1.32% of the entire dataset. However, since posts need not necessarily contain solely textual information (45.59% of all posts include an attachment in the form of images, videos, and URLs), we resort to a diffusion based model of identifying hate users in the social network.

2.3.2 Extraction of hateful users. Using the high precision lexicon would miss out on several users who might be hateful in nature but are not selected as they did not post any content with words from our lexicon (like using images and videos). In order to capture such obscure hate users, we leverage the methodology used by Ribeiro et al. [33]. We enumerate the steps of our methodology below.

- We identify the initial set of hateful users as those who have written at least 10 posts, with at least one hateful keyword in each of them. This results in a set of 2,769 hateful users.
- We create a repost network where nodes represent the users and edge-weights denote posting and reposting frequency. We convert the repost network into a belief network by reversing the edges in the original network and normalizing the edge weights between 0 and 1. We explain this further in the subsequent section.

[10]The lexicons: www.github.com/binny-mathew/Spread_Hate_Speech_WebSci19
[11]https://www.hatebase.org
[12]https://www.urbandictionary.com

- We then run a diffusion process based on the DeGroot's learning model [19] on the belief network. We assign an initial belief value of 1 to the 2,769 users identified earlier and 0 to all the other users. The diffusion model aims to identify users who did not explicitly use any of the hateful keywords, yet have a high potential of being a hateful user due to homophily.
- We observe the belief values of all the users in the network after *five* iterations of the diffusion process and divide the users into four strata, [0, .25), [.25, .50), [.50, .75) and [.75, 1] according to their associated belief.

We define users whose belief values lie within [.75, 1] as hateful and those whose belief values lie within [0, .25] as non-hateful with the additional constraint that each of these users should have at least *five* posts. We do so since it is difficult to judge a person on the basis of a single post. We thus obtain a set of 2,290 hateful users and 58,803 non-hateful users, which comprises ∼ 0.67% and ∼ 17.23% of the entire dataset. We refer to the set of hateful and non-hateful users as KH (read 'Known hateful user') and NH (read 'Not hateful user') respectively henceforth.

2.3.3 DeGroot's model of information diffusion. : We illustrate a repost network with three users (A, B, C) in Figure 1a. An edge-weight of 9 from B to A denotes that user B has reposted 9 posts of A while a self loop of A of weight 17 denotes that A has posted 17 times. We convert the repost network into a diffusion network as shown in 1b by reversing the edges, with the edge-weights normalized. The edge weights are normalized by dividing the edge weight from C to A in the original network by the sum of the edge weights originating from C (including self loops). For example, user C in Figure 1a has reposted A 5 times and has posted 10 times. Thus the value of edge weights from A to C is $\frac{5}{15}$ or 0.33 and the weight of the self-loop at C is $\frac{10}{15}$ or 0.67 as shown in Figure 1b. The normalized edge-weight is a measure of the user's belief being influenced by her neighbors. Let us denote the belief of A, B and C at the time instant i as b_A^i, b_B^i, b_C^i respectively. The belief of user C at time instant $i+1$ can be written as

$$b_C^{i+1} = 0.33 \times b_A^i + 0.67 \times b_C^i \quad (1)$$

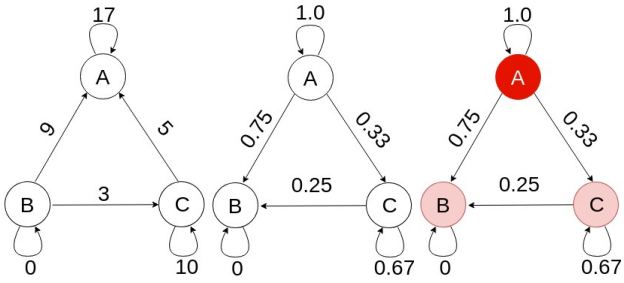
Thus belief propagation takes place in an iterative fashion using the DeGroot's model. If we consider the initial beliefs of A, B and C to be 1, 0 and 0 respectively, their corresponding beliefs at time instant 1 would be 1, 0.75 and 0.33 as demonstrated in Figure 1c.

2.4 Quality of the labels

We evaluate the quality of the final dataset of hateful and non-hateful accounts through human judgment. We ask four annotators to determine if a given account is hateful or non-hateful as per their perception. The annotators consisted of three undergraduate students with major in Computer Science and one PhD student in Social Computing. Since Gab does not have any policy for hate speech, we use the guidelines defined by Twitter[13] for this task. We provide the annotators with a class balanced random sample of 200 user accounts[14]. Each account was evaluated by two independent

[13]https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[14]We have used a random sample of 200 accounts per class to keep the monetary cost manageable



(a) Repost network **(b) Belief network** **(c) Belief diffusion**

Figure 1: Description of the DeGroot's model for information diffusion in a toy network.

annotators. We follow the definition used by ElSherief et al. [14] to identify a post as hateful. The authors define hate speech as a "*direct and serious attack on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease*".

We observe that the two annotators found 86.9% and 93.2% of the hate accounts from our sample as hateful, yielding a substantial high Cohen's κ score of 0.69. Likewise 92.2% and 99.4% of the non-hateful accounts from our sample were adjudged to be non-hateful yielding a very high κ score of 0.87. These results show that the dataset generated by our method is of high quality with minimal noise.

3 USER CHARACTERISTICS

We first try to understand the characteristic differences between the KH and NH users identified by our method.

3.1 Account characteristics

Here we analyze the differences in the account characteristics of hateful and non-hateful users. The different account characteristics include the number of posts, followers and followings (normalized over time) and the number of likes, dislikes, replies, reposts (normalized over the number of posts) of the KH and NH users. The normalization over time is done by dividing the account characteristic (say number of posts) of a user by the number of days elapsed from the first post of the user to the date the last post was crawled. We report the mean and median details of these characteristics in Table 2. We measure the statistical significance between the two distributions using the two sample K-S test and observe that each of the account characteristics are significantly different (p-value<0.001). The inordinate difference in the mean and median values between NH and KH can be attributed to the prolific activity of hateful users. The raw quantity of posts generated by the KH and NH amount to 26.80% (5.68M) and 45.53% (9.65M) of all posts, respectively. This implies that as small as *0.67% of the users generated 26.80% of all the content in Gab.* Some of the striking observations from the table are that the normalized number of followers of the KH users is more than double the number for the NH users. Although the normalized number of likes for the KH users is larger than that of the NH users, what is more notable is that the normalized number of dislikes

for the KH users is more than double compared to the NH users. This indicates that there is a (possibly growing) mass in Gab (albeit small) who have built an overall opposition against hate speech and we believe that this could be attributed to the rising influence of counter speech and the corresponding counter speakers. Typically, the posts from the KH users receive double the number of replies and reposts compared to the NH users.

Feature	Mean KH	Mean NH	Median KH	Median NH
post	6.899	0.633	1.813	0.075
follower	1.651	0.639	0.629	0.138
following	1.658	0.881	0.299	0.052
like	2.628	1.533	1.490	0.678
dislike	0.124	0.051	0.038	0.000
score	2.576	1.733	1.453	0.875
reply	0.222	0.116	0.162	0.000
repost	0.401	0.224	0.146	0.000
F:F	4.732	5.219	1.545	1.611

Table 2: Account characteristics of the hateful and the non-hateful users. Hateful users generate more popular content and also posts frequently. All the differences in account characteristics are significant (p-value<0.001 and marked in different color), except F:F(Follower/Following).

3.2 Network characteristics

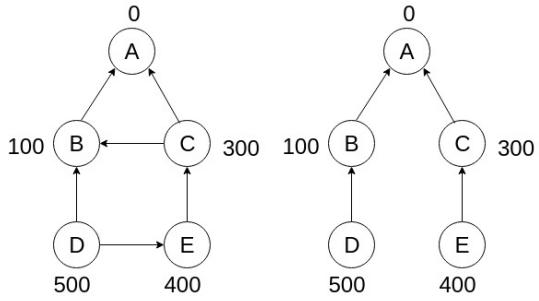
In this section, we investigate the network characteristics of the KH and the NH users on the basis of their follower-following relationship. We construct a subgraph over the entire network with nodes being the set of KH and NH users and edges representing the follower-following relationship between these users only. This subgraph so formed has 61.1K nodes and 7.56 M edges. We observe that the network of KH users (2.29K nodes, 156.1K edges) is ≈ 16.74 times more dense than the NH users (58.8K nodes, 6.15M edges). The KH users also demonstrate higher reciprocity values (35.00%) as opposed to the NH users (32.75%) with (p-value~0.0). Moreover, an NH user is 5.4 times more likely to follow a KH user than a KH user following a NH user, inkling at the higher popularity of KH users. It is also 20.675 times more likely that a KH user will follow another KH user than a NH one. This indicates strong cohesiveness among the KH users. Typically, the KH users seem to operate in closed groups or clans which is a well-known property of extremist networks [31].

4 DIFFUSION DYNAMICS OF POSTS

In this section, we observe the diffusion of information throughout the network and analyze the differences in diffusion of posts generated by the KH users and those generated by the NH users.

4.1 Model description

We refer to the path traced by a post as it is reposted by other users as a cascade and the original user as the root user. Since it is not possible to trace the exact influence path, i.e., the user who influenced the reposting, we leverage the social network connections (followers and friends) as means of information diffusion and influence similar to Taxidou and Fischer [38]. In all the models, an edge is formed between two users if there exists a follower-following relationship between the users. We deploy the Least Recent Influencer Model (LRIF) [5] to observe the information diffusion. Previous research [2, 38] have also used such models to study the diffusion of information in online social media. In the LRIF model, users are influenced by the first exposure to a message even if they do not act immediately. Essentially, the model seeks to avoid exhaustive search by converting the network into a directed acyclic graph, thereby, reducing the time complexity. We illustrate the DAG generated by the LRIF models in Figure 2. The sample network is shown in Figure 2a comprising five users. A directed edge between any two users (say from B to A) specifies the follower-following relationship (B follows A). The number beside each user specifies the time of reposting, with A being the root user. The DAG generated by the LRIF model is shown in Figure 2b.



(a) Repost graph **(b) LRIF model**

Figure 2: DAG generated by the LRIF model on a sample repost network. The numbers indicate the time in seconds of reposting. The links are formed between User C and A since A posted earlier than B.

4.2 Characteristic cascade parameters

In order to characterize the cascades generated by KH and NH users, we employ the following features as used in Vosoughi et al. [39].

- **Size** represents the number of nodes in the DAG which are reachable from the root user. It corresponds to the total number of unique users involved in the cascade of the post.
- **Depth** is the length of the largest path from the root node of the cascade. The depth of a cascade, D, with n nodes is defined as

$$D = \max(d_i), 0 \leq i \leq n \tag{2}$$

  where d_i is the depth of node i.
- **Average depth** is the average path length of all nodes reachable from the root user. For a cascade with n nodes, we define its average depth (AD) as

$$AD = \frac{1}{n-1} \sum_{i=1}^{n} d_i \tag{3}$$

  where d_i is the depth of the node i.

	Posts				Attachments				Topics			
	KH		NH		KH		NH		KT		NT	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Size	1.28	447	1.21	602	1.34	447	1.23	455	1.68	237	1.51	252
Depth	0.13	7	0.09	7	0.16	7	0.11	7	0.30	11	0.24	6
Breadth	1.13	275	1.10	533	1.15	275	1.11	391	1.30	162	1.24	189
Average depth	0.11	4.82	0.08	4.53	0.14	4.82	0.10	4.53	0.26	4.52	0.22	3.74
Structural virality	0.13	5.46	0.09	5.07	0.16	5.46	0.11	5.07	0.31	6.10	0.25	4.89

Table 3: Diffusion characteristics of posts of the KH and the NH users. The minimum value for all the characteristics were same: a post with no repost.

- **Breadth** is the maximum no. of nodes present at any particular depth in the DAG.

$$B = \max(b_i), 0 \leq i \leq d \tag{4}$$

where b_i denotes the breadth of the cascade at depth i and d denotes the maximum depth of the cascade.

- **Structural virality** as defined by Goel et al. [18], is the average distance between all pairs of nodes in the DAG, assuming the DAG to be a tree. It is simply the Weiner index.

$$SV = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \tag{5}$$

where d_{ij} represents the length of the shortest path between nodes i and j.

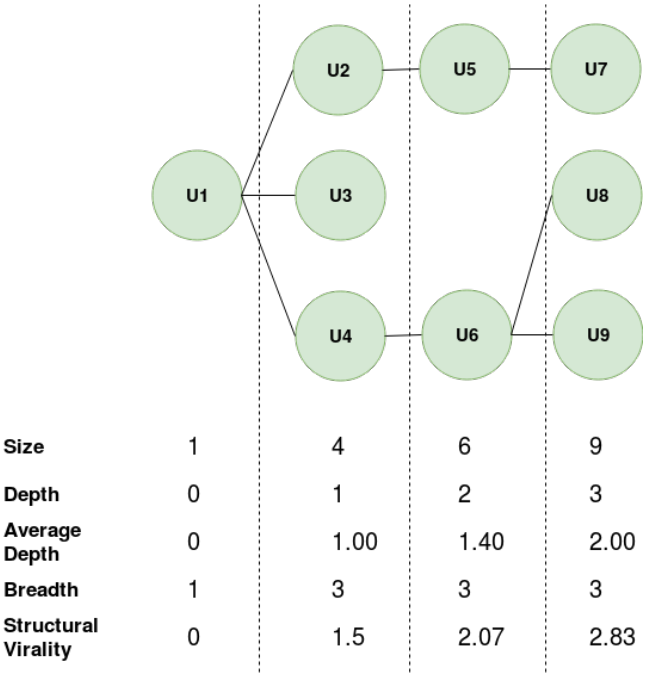


Figure 3: Repost path between a set of Gab users. At each level we show the various cascade properties.

In Figure 3, we show an example cascade in which we show the different values of the above measures at each level of the cascade.

4.3 Experiments on varied nature of posts

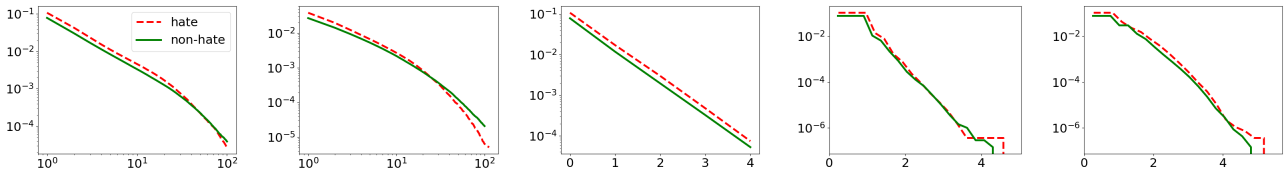
All subsequent evaluation is carried out on the DAG generated by the LRIF model. KH users had 2.73M posts and NH users had 6.87M posts which we considered as the root posts for our cascade. We do not include the posts of KH and NH users which are 'quotes' or 'replies', since such posts might not represent the user's actual opinion. We also observe the diffusion characteristics for posts having attachments (images or media content) separately since such posts are hypothesized to be more viral. The supposed virality is attributed to the appeal of an image/ meme over plain textual information. Finally, in order to observe the topic perspective, we look into posts which have been posted in topics. We report the mean and max score of the cascade features for the different experiments in Table 3. Note that we did not report min values since they were the same for the KH and NH users for all the cascade features.

4.4 Characteristic differences in cascades of the KH and NH users

4.4.1 General cascade parameters. The mean size (number of unique users) of a cascade is larger for posts of KH users than NH users as observed from Table 3. Figure 4a shows the Complementary Cumulative Distribution Function (CCDF) of a cascade's size for both KH and NH users. We observe that almost 90% of the posts do not get reposted for both KH and NH users. Although the maximum size of the NH's cascade is larger, the cascade's size is significantly larger for KH users especially for the initial stages. Thus, the posts of KH users have a larger audience.

The mean breadth of a cascade is also larger for posts generated by KH users implying that such posts spread wider (farther amongst a user's followers) than those generated by NH users. The CCDF of a cascade's breadth 4b exhibits similar characteristics as a cascade's size. Only the top 0.1% of the NH user's cascade had more breadth as compared to the KH user's cascade.

The mean depth, mean average depth and mean structural virality of a cascade are also significantly larger for posts generated by KH users. Not only does it imply that such posts diffuse deeper into the network but they are also more viral [18]. Moreover, as the CCDF for depth, avg-depth and virality (Figures 4c, 4d and 4e



(a) CCDF of size (b) CCDF of breadth (c) CCDF of depth (d) CCDF of average depth (e) CCDF of virality

Figure 4: Different diffusion dynamics of the posts by hate and non-hate users using LRIF model. The cascade properties namely size, breadth, depth, average depth and virality are larger for the posts of hateful users.

respectively) depicts, these properties remain consistently larger for the KH users throughout their entire distribution.

4.4.2 Posts with attachment. In order to understand the diffusion dynamics of posts with attachment, we consider only those posts which have an attachment. The attachment can be images, videos or urls. From Table 3, it is observed that posts with attachments have a larger mean size, breadth, depth, average depth and structural virality implying that such posts have a greater outreach, diffuse wider, deeper and are more viral. This agrees with our hypothesis that attachments with memes and images are more instrumental in information diffusion than textual content. The different characteristics manifest as significant (p-value < 0.01) according to the KS-test for posts and attachments.

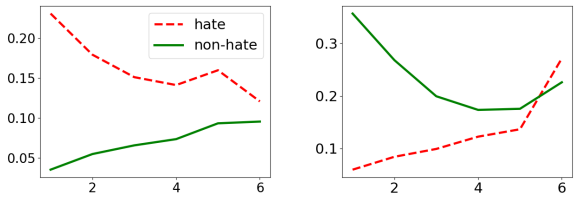
4.4.3 Posts in topics. Topics represent sub-communities in Gab catered to a certain cause or serving a niche interest. We attempt to compare topics having a higher proportion of hateful content with those having a lower proportion of hateful content. We consider topics which have at least 100 users and at least 500 posts to ensure that our cascades formed are well represented. We then rank the topics in decreasing fraction of hateful content and take the top 250 topics as hateful topics (HT) and the bottom 250 topics as non-hateful topics (NT). We show some of the top 10 instances of hateful and non-hateful topics in Table 4. We observe that HT consists of topics which aim to promote hate speech in the community.

The cascade properties of the posts in the HT and NT topics are summarized in Table 3. It is evident that community involvement has increased the cascade properties significantly.

HT	Jews Are The Synagogue Of Satan, The Black Race SUCKS, Street Shitter, Israel Holocaust Remembrance Day
NT	Xenoblade Chronicles 2(Spoilers), 2018 memes to amuse you, What's Going On?, Landscape, Classic Cars and Trucks,

Table 4: Prominent hateful and non-hateful topics of Gab.

4.4.4 Early adopters in a cascade. Figures 5a and 5b illustrate the proportion of hateful and non-hateful propagators at each depth. It is evident that the hateful users are early adopters in the cascades of KH users, exhibiting strong degree of homophily. The reverse also holds true for non-hateful users who are the early adopters in the cascades of NH users. The change in monotonicity of the curves in both the diagrams after depth 4 can be attributed to the small number of cascades whose depth exceeded 4 levels (0.0065% and 0.0057% of KH and NH users respectively). These are fast cascades where the information was propagated by a larger fraction of hateful users in KH posts and larger fraction of non-hateful users in NH posts.



(a) % of KH propagators (b) % of NH propagators

Figure 5: The proportion of hateful and non-hateful users who have reposted the root user across different depths. Here, the X-axis represents the depth of the cascade while the Y-axis represent the proportion of users. Hate users are early propagators for the posts of hateful users while non-hateful users are the early propagators for the posts of non-hateful users.

4.4.5 Dynamics of the cascade properties. We also explore the different dynamic properties of the cascade parameters. In particular, we investigated the temporal aspects as well as the relationship between the parameters of the cascade.

Temporal aspects: The temporal aspects of information diffusion, namely the evolution of different cascade parameters over time are illustrated for both KH and NH cascades in terms of size, breadth, depth, average depth and structural virality via the Figures 6a, 6b, 6c, 6d and 6e respectively. For all such diagrams, the x-axis represents the specific characteristic (like size or depth) and the y-axis represent the time taken in thousand seconds. It is quite evident that the time taken for the KH cascades to reach a particular value is lower in the initial stages implying that KH cascades are significantly faster initially. This can be attributed to the high proportion of KH users as early propagators.

Dynamic relationship between diffusion parameters: Next, we observe the dynamic relationship between the cascade parameters. In Figure 7a, we observe that the cascade of KH users reach a higher depth for almost all the values of breadth. We note similar results for other relations such as size vs avg. depth (Figure 7b),

size vs vepth (Figure 7d), and size vs virality (Figure 7e). In case of size vs breadth (Figure 7c), we observe that this does not hold. The posts of KH users seem to have smaller breath for almost every size of the cascade as compared to the posts of NH users.

4.4.6 Summary.

- The posts of KH users diffuse significantly farther, wider, deeper and faster than the NH ones.
- Posts having attachments tend to be more viral.
- KH users are more proactive and cohesive. This observation is based on their fast repost rate and the high proportion of them being early propagators.

5 CASE STUDY: THE PITTSBURG SHOOTING

In the aftermath of the Pittsburg synagogue shooting[7], the Gab website owners were forced to shutdown the site temporarily for a week[15]. The reason behind the decision to ban the website arose from Robert Bowers' history of posting antisemitic messages on Gab (under the username @onedingo). Bowers allegedly killed eleven people at a Pittsburgh synagogue with a gun on October 27, 2018.

We illustrate the account characteristics of @onedingo that were present in our dataset in Table 5. We observe that all the characteristics of @onedingo are close to the characteristics of the KH users shown in Table 2. We also manually inspected into the user's posts and found several hateful instances such as the following.

> Kikes are enemy number one. Dealing with anything after will be a relative piece of cake. I will not fire on someone who is shooting my enemy.

Moreover, 40.9% of onedingo's followings were KH users, also 25.4% of his followers were KH users.

We observed the cascade properties of onedingo and found that it aligns more with the non-hateful users. Thus, if one only looks at the cascade properties it would be very difficult to ascertain the vindictive nature of this user. The user successfully managed to camouflage himself and portray a non-hateful behavior in its message cascading patterns. However, at a micro level, a closer observation of the posts made by the user, reveals that several of the posts of onedingo talk about killing and genocide of Jews. We would need models that can differentiate between different intensities of hate speech for obtaining clearer insights in such nuanced cases. We plan to take this work up as an immediate future work.

6 ADDITIONAL OBSERVATIONS

In this section we put forward certain additional observations that we believe are necessary to render completion to this research.

6.1 Influential users

As posts of KH users have higher virality as compared to the NH users, we wanted to know how much of it was due to user popularity. To get a better understanding of the popularity, we analyze the users based on two different criteria: 1) the number of followers; 2) the user PageRank.

[15]https://www.technadu.com/godaddy-forces-gab-shut-down-temporarily/46040

Account characteristics			Cascade characteristics	
Property	Value	Normalized value	Property	Mean value
post	206	1.355	size	1.158
follower	212	1.395	depth	0.105
following	232	1.526	breadth	1.052
like	568	2.757	average depth	0.087
dislike	2	0.01	virality	0.078
score	566	2.748		
reply	113	0.549		
repost	114	0.553		
F:F	0.91379	- -		

Table 5: Description of onedingo's characteristics.

We compute the PageRank on the follower/following network of the Gab users and rank them according to their scores. We take the top k users in the PageRank score and compute the percentage of users that are tagged as KH and NH. We try different values of k ranging from 50 to 10000. We can observe from Figure 8 that there are much more (around 6 times) NH users in the top k position as compared to the KH users. We got similar results using the number of followers (data not shown).

These results indicate that the NH user group consisted of much more popular users as compared the KH users. Thus the overall popularity of users does not seem to bear any correlation with the spread dynamics of posts on Gab.

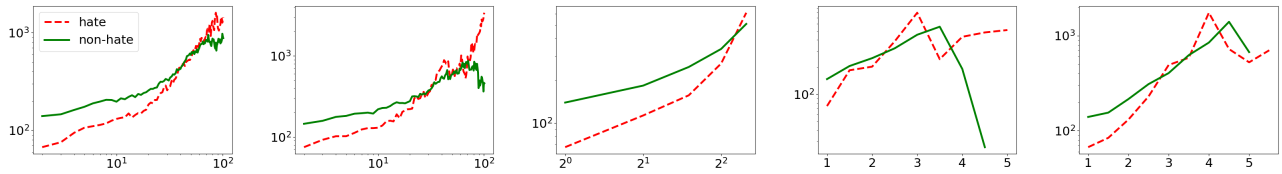
6.2 Domains used

Next, we identify what kind of links were being shared by the posts of the KH and NH users. To this end, we inspect the urls mentioned by the KH and NH users in their posts. We first extract the domains of all the links that are present in the root posts of the cascades for both the KH and NH users. Then, we filter out all the domains which are not present in at least 200 unique posts. Next, we find the fraction of times a domain was used in the post of a KH user to the post of NH user. We report the top domains used by the KH and NH users in Table 6 according to the fraction of usage. We observe that the KH user posts contained domains such as *dailystormer* which is an American neo-Nazi site, white supremacist, and Holocaust denial commentary board. The website advocates for the genocide of Jews and considers itself a part of the alt-right movement. These websites are also responsible for the spread of conspiracy theories. Since we know that fake news tend to spread faster [40], we posit that fake news and hate speech tend to go hand in hand. We would be interested to investigate this relationship between hate speech and fake news in more details in a future work.

6.3 Top 1% viral hateful and non-hateful posts

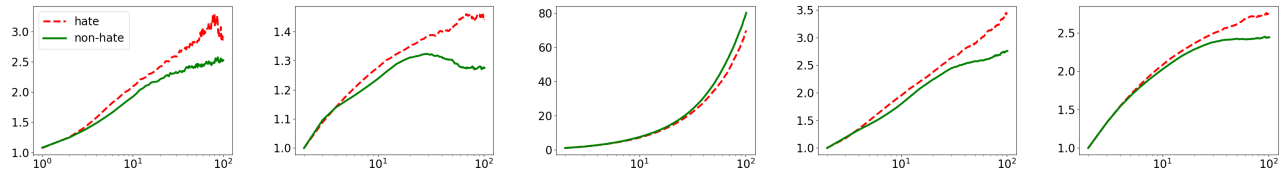
Next, We study the difference of the top 1% of the hate and non-hate posts according to structural virality. We first check the presence of profane words using a lexicon[16]. We found that among the top viral hate posts, 32.91% contained one or more profane words. Only

[16]https://github.com/RobertJGabriel/Google-profanity-words/blob/master/list.txt



(a) size vs time (b) breadth vs time (c) depth vs time (d) avg-depth vs time (e) virality vs time

Figure 6: Temporal profiles of diffusion properties of the cascades generated by the posts of the KH and the NH users. Here, the Y-axis represents time (taken in 10^3 seconds) and X-axis represents the cascade characteristics. The posts of KH users spread farther, wider and deeper more quickly in the initial stages.



(a) breadth vs depth (b) size vs avg. depth (c) size vs breadth (d) size vs depth (e) size vs virality

Figure 7: Diffusion dynamics of different properties of the cascades. For each figure, the first property represents the X-axis and the second property represents the Y-axis.

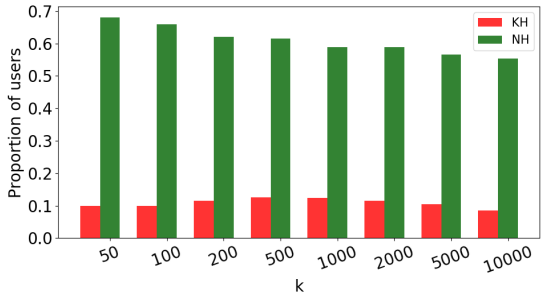


Figure 8: Proportion of KH and NH in the top k PageRank score.

User	Domains
KH	*dailystormer*, imageshack, *radioaryan*, endculturalmarxism, *christophercantwell*, *infostormer*, rationalwiki, skepdic
NH	xxxbios, bring-back-america, yourlawyer, Energy-Ingenuity, petreporters, internetmarketingexperience, strippersforyou

Table 6: Top domain that are used by KH and NH users.

26.23% of the top non-hate posts contained one or more profane words.

We also observed the type of content that is shared by the top 1% of the posts. We found that 16.47% of KH user posts did not contain any form of attachment. The same for NH users was 13.04%. On the other hand, NH users seem to be using urls in their posts a lot; 34.1% of the top NH user posts had at least one url, whereas only 21.24% of the KH user posts had at least one url. We also found that KH users use image and gif in 58.54% of their posts whereas NH users use it in 48.73% of their posts.

7 DISCUSSION

In this section we discuss the implications of this work and note the limitations of the study.

7.1 Real world impact of online hate

The spread of hate speech in the online medium in a grave concern to the society. This is particularly problematic for several unsuspecting victims who might form an unnecessary outgroup prejudice against a particular community [36]. The frequent and repetitive exposure to hate speech leads to desensitization to this form of verbal violence and subsequently to lower evaluations of the victims and greater distancing, thus increasing the outgroup prejudice.

One of the prime examples of this was the Rohingya crisis in Myanmar. Many of the people who helped in disseminating hate on Facebook had not even met a single Rohingya in their life. Their view of the target community was completely manipulated by the rampant spread of hate speech on Facebook. Our results show that the "hate-workers" form cohesive groups is a testimony to why campaigns like the above usually succeed; the spread of the hate content is a well-orchestrated collective effort that helps the content to spread like wildfire as opposed to individual efforts which could never have been so successful.

7.2 Design of online platforms

The online social media platforms facilitates the fast spread of any kind of information. The users with malicious intents normally

make use of such features to disseminate their messages. As we have seen from our analysis, the KH users are most active in the early stages of content spread, our recommendation would be that these social media platforms should curtail the spread of harmful content by *suppressing its initial spread*. This way the harmful posts would appear on the home feed of fewer people and thus cause less damage. The posts that people receive in their home feed reinforces generally their world views. The platforms could thus monitor the spread of hate speech and reduce its effect by showing it in less people's home feed.

Another suitable alternative to fight hate speech without harming freedom of speech would be using counter speech. This strategy is even endorsed by companies like Facebook that has stated in public that it believes counter speech is not only potentially more effective, but also more likely to succeed in the long run [6]. By understanding the spread of hate speech in online social media, the sites could employ appropriate counter speech strategies that could mitigate/neutralize the effects of hate speech.

7.3 Limitations of the current study

In our analysis we have relied on the user account to study the cascade. We assume that the hateful posts of these hateful accounts would generate majority of the reposts. This means that few of the reposts of these hateful accounts might not be hateful in nature. However, while we cannot claim to have captured the full picture, our analysis provided a peek into the cascade dynamics of the hateful posts in Gab.

8 RELATED WORK

Diffusion in online media: To the best of our knowledge there has not been any work that tries to study the diffusion of hate in online social media. However, there are several works that looks into diffusion of fake news [27, 39, 43], Linkedin [3], retweet cascade [8, 9, 18, 39], rumours [11, 17, 20, 22, 47] and Tumblr [1, 2, 7, 44]. Cheng et al. [9] perform a large scale analysis of recurring cascades in Facebook. They observe that content virality is the main driver for recurrence. In Del Vicario et al. [11], the authors perform a large scale analysis of Facebook and observe that selective exposure to content is the primary driver of content diffusion and generates the formation echo chambers. Stuart [37] systematically profiled all Islamist-related terror offenses in the United Kingdom between 1998 to 2015 and found that over a quarter (28%) of Islamist related terror offenses were demonstrably inspired by the rhetoric or propaganda of a proscribed terrorist organisation.

Research on Gab: There is little research done on Gab. Zannettou et al. [45] performed the first study in which the author collected and analyzed a large dataset of Gab and found that the site is predominantly used for discussion of news, world events, and politics. They also found that Gab contains 2.4 times more hate speech as compared to Twitter. Lima et al. [23] also found that Gab is very politically oriented and users who abuse the lack of moderation disseminate hate. Zannettou et al. [46] perform a large scale measurement study of the meme ecosystem by introducing a novel image processing pipeline. Gab has substantially higher number of posts with racist memes. Gab shares hateful and racist memes at a higher rate than mainstream communities. In similar lines, Finkelstein et al. [16] study millions of comments and images from alt-right web communities like 4chan's Politically Incorrect board (/pol/) and the Twitter clone, Gab and quantify the escalation and spread of antisemitism.

Research on hate speech: The majority of the research in hate speech has been done in automatic detection in various social media platforms like Twitter [4, 10, 32, 42], Facebook [12], Yahoo! Finance and News [13, 28, 41] and Whisper [26]. In another online effort, a Canadian NGO, the Sentinel Project[17], launched a site in 2013 called HateBase[18], which invites Internet users to add to a list of slurs and insulting words in many languages. There are some works which have tried to characterize the hateful users [24, 34]. In Ribeiro et al. [34], the authors study the user characteristics of hateful accounts on Twitter and found that the hateful user accounts differ significantly from normal user accounts on the basis of activity, network centrality, and the type of content they produce. In ElSherief et al. [15], the authors perform a comparative study of the hate speech instigators and target users on Twitter. They found that the hate instigators target more popular and high profile Twitter users, which leads to greater online visibility. Mathew et al. [25] studies the effect of counterspeech in hateful YouTube videos and develops machine learning models to automatically detect counterspeech in YouTube comments. In ElSherief et al. [14], the authors focus on studying the target of the hate speech - directed and generalized. They observe that while directed hate speech is more personal, informal and express anger, the generalized hate is more of religious type and uses lethal words such as 'murder', 'exterminate', and 'kill'. Ottoni et al. [30] analyze the hate, violence, and discriminatory bias in a selection of right-wing YouTube channels. The authors found that these channels are more specific in their content, discussing topics such as war and terrorism, and have a higher percentage of negative category words such as aggression and violence. In Olteanu et al. [29], the authors study the effect of external events on hate speech in two social media: Twitter and Reddit. They observe that extremist violence tends to increase hate speech in online medium, especially messages which advocate violence.

9 CONCLUSION AND FUTURE WORK

In this paper, we perform the first study which observes the nuances of the diffusion characteristics of the posts made by hateful and non-hateful users. We used high precision keywords to select hateful users and provide them as input to DeGroot's model to identify the hateful and non-hateful set of users. We then analyse the diffusion characteristics of the posts of these users. We found that the posts made by hateful users tend to spread farther, faster, and wider. These hateful users are densely connected with each other and generate almost 1/4th of the content in Gab despite comprising 0.67% of the users.

Our work also points to several open research avenues. A large fraction of the posts were in the form of images in case of hate users. For future work, we would like to take up the task of building a classification system that can distinguish between images/videos that are hateful in nature. Another interesting direction would be

[17] https://thesentinelproject.org/
[18] https://www.hatebase.org/

to look into the diffusion characteristics of the individual hateful posts instead of the accounts.

REFERENCES

[1] Nora Alrajebah. 2015. Investigating the structural characteristics of cascades on Tumblr. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, 910–917.

[2] Nora Alrajebah, Leslie Carr, Markus Luczak-Roesch, and Thanassis Tiropanis. 2017. Deconstructing diffusion on Tumblr: structural and temporal aspects. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 319–328.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, Jure Leskovec, and Mitul Tiwari. 2015. Global diffusion via cascading invitations: Structure, growth, and homophily. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 66–76.

[4] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets *(WWW)*. 759–760.

[5] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 65–74.

[6] Jamie Bartlett and Alex Krasodomski-Jones. 2015. Counter-speech examining content that challenges extremism online. *Demos. Available at: http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf* (2015).

[7] Yi Chang, Lei Tang, Yoshiyuki Inagaki, and Yan Liu. 2014. What is tumblr: A statistical overview and comparison. *ACM SIGKDD explorations newsletter* 16, 1 (2014), 21–29.

[8] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can Cascades Be Predicted?. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, New York, NY, USA, 925–936. https://doi.org/10.1145/2566486.2567997

[9] Justin Cheng, Lada A Adamic, Jon M Kleinberg, and Jure Leskovec. 2016. Do cascades recur?. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 671–681.

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. (2017).

[11] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.

[12] Fabio Del Vigna, Andrea Cimino, Felice DellâĂŹOrletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. (2017).

[13] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *WWW '15 Companion*. 29–30.

[14] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Y. Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media *(ICWSM '18)*.

[15] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets. (2018).

[16] Joel Finkelstein, Savvas Zannettou, Barry Bradlyn, and Jeremy Blackburn. 2018. A Quantitative Approach to Understanding Online Antisemitism. *arXiv preprint arXiv:1809.01644* (2018).

[17] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades.. In *ICWSM*.

[18] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2015. The structural virality of online diffusion. *Management Science* 62, 1 (2015), 180–196.

[19] Benjamin Golub and Matthew O Jackson. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics* 2, 1 (2010), 112–49.

[20] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 8.

[21] Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive Embeddings for Hate Speech Detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 26–32.

[22] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2007. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*. SIAM, 551–556.

[23] Lucas Riguete Pereira De Lima, Julio C. S. Reis, Philipe F. Melo, Fabricio Murai, Leandro Araújo Silva, Pantelis Vikatos, and Fabrício Benevenuto. 2018. Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System. *ASONAM* (2018).

[24] Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv preprint arXiv:1812.02712* (2018).

[25] Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherje. 2018. Thou shalt not hate: Countering Online Hate Speech. *arXiv preprint arXiv:1808.04409* (2018).

[26] Mainack Mondal, Leandro Araujo Silva, and Fabricio Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *HT*.

[27] Eni Mustafaraj and Panagiotis Takis Metaxas. 2017. The fake news spreading plague: was it preventable?. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 235–239.

[28] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *WWW '16*. 145–153.

[29] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The Effect of Extremist Violence on Hateful Speech Online. (2018).

[30] Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernardina, Wagner Meira Jr, and Virgílio Almeida. 2018. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 323–332.

[31] Arie Perliger and Ami Pedahzur. 2011. Social Network Analysis in the Study of Terrorism and Political Violence. *Political Science and Politics* 44, 2 (2011).

[32] Jing Qian, Mai ElSherief, Elizabeth M. Belding-Royer, and William Yang Wang. 2018. Hierarchical CVAE for Fine-Grained Hate Speech Classification. *EMNLP* abs/1809.00088 (2018).

[33] Manoel Ribeiro, Pedro Calais, Yuri Santos, VirgÃŋlio Almeida, and Wagner Meira Jr. 2018. Characterizing and Detecting Hateful Users on Twitter *(ICWSM '18)*.

[34] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2017. " Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. In *WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. 8.

[35] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: Detecting Hate speech against Women. *arXiv preprint arXiv:1812.06700* (2018).

[36] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior* 44, 2 (2018), 136–146.

[37] Hannah Stuart. 2017. *Islamist Terrorism: Analysis of Offence and Attacks in the UK (1998-2015)*. Henry Jackson Society.

[38] Io Taxidou and Peter M. Fischer. 2014. Online Analysis of Information Diffusion in Twitter. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. ACM, 1313–1318.

[39] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[40] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[41] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media (LSM '12)*. 19–26.

[42] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[43] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 637–645.

[44] Jiejun Xu, Ryan Compton, Tsai-Ching Lu, and David Allen. 2014. Rolling through tumblr: characterizing behavioral patterns of the microblogging platform. In *Proceedings of the 2014 ACM conference on Web science*. ACM, 13–22.

[45] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. 1007–1014.

[46] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. *arXiv preprint arXiv:1805.12512* (2018).

[47] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1395–1405.