*Artificial Intelligence Research and Development*
*M. Villaret et al. (Eds.)*


*doi:10.3233/FAIA210122*

# Countering Negative Effects of Hate Speech in a Multi-Agent Society

Arthur MÜLLER[a,b] and Maite LOPEZ-SANCHEZ [a,1]
[a] *Universitat de Barcelona,* [b] *University of the Bundeswehr Munich*

**Abstract.** Hate speech expresses prejudice and discrimination based on personal characteristics such as race or gender. Research has proven that the amount of hateful messages increases on online social media. If not countered properly, the spread of hatred can overwhelm entire societies. This paper proposes a multi-agent model of the spread of hatred. We reuse insights from previous research to construct and validate a baseline model. From this, three countermeasures are modelled and simulated to investigate their effectiveness in containing the spread of hatred: (1) The long-term measure of education is very successful, but it still cannot eliminate hatred completely; (2) Deferring hateful content has a similar positive effect with the advantage of being a short-term countermeasure; (3) Extreme cyber activism against hatred can worsen the situation and even increase the likelihood of high polarisation within societies.

**Keywords.** hate speech, hate spread, countermeasures, social networks, opinion diffusion, education, deferring hate content, cyber activism.

## 1. Introduction

The United Nations define hate speech as the attack or usage of pejorative or discriminatory language with reference to a person or a group based on their religion, ethnicity, nationality, gender or other identity factor[2]. The usage of hateful language has become common on online social media. This is specially the case on platforms with slack content policies—such as Gab—, where the amount of hateful messages has steadily increased over the last years [15]. But also hateful users on platforms such as Twitter have become more extreme [9]. In fact, spread of their messages seems to be inadvertently supported by the algorithms of the social networks [17].

To counter this problem many measures with different temporal horizons have been proposed by researchers and politicians. The long-term effect of education is supposed to introduce positive bias into society by teaching democratic values and tolerance. Awareness campaigns focus on mid-term effects and try to prevent forming negative prejudices against out-groups and minorities. Others propose expensive manual community management or intrinsically motivated counter speech [8]. In contrast, the short-term measure of automatic filtering is criticised as the infringement of the freedom of speech. Exhaus-

[1] Research funded by projects SGR 341, MISMIS (PGC2018-096212-B-C33), Crowd4SDG (H2020-872944), CI-SUSTAIN (PID2019-104156GB-I00), COREDEM (H2020-785907), and nanoMOOC (COMRDI18-1-0010-02).

[2] https://www.un.org/en/genocideprevention/documents/UN Strategy and Plan of Action on Hate Speech 18 June SYNOPSIS.pdf

tive evaluations, however, are still lacking and it is difficult to assess the effectiveness of these initiatives and, much less, to compare them among each other. Therefore, the objective of this paper is to propose an agent-based model as a sandbox for the simulation and comparison of countermeasures against the spread of hatred. Three countermeasures with different temporal effects are selected for this approach: education, deferring hateful content and counter activism.

## 2. Related Work

This section introduces research describing hateful users and behaviours, mathematical models of opinion spread, and existing simulations in the context of hatred.

### *2.1. Characterising Hateful Users in Social Networks*

Hateful users have been found to exhibit a very different profile when compared to other users. From the psychological point of view, they are energetic, talkative, and excitement-seeking [16]. However, other personality traits such as narcissism, lack of empathy, and manipulative are also attributed to them [7]. Haters show high activity on social media and follow more people per day. Although hateful users gain 50% less back-followers for every spawned following relationship per day, they can receive much more followers over the lifetime of their accounts due to their high activity [18]. Surprisingly, the amount of hateful persons is little and does not exceed 1% even on Gab, but they are responsible for an non-proportionally high amount of content. Furthermore, their content can spread faster and diffuse deeper into the network when forwarded by other users [14]. Hateful content seems to be less informative on Twitter, since less URLs and hashtags are added [18], but it is known to be more viral when enriched with images or videos [12]. Finally, hateful users are very densely connected and demonstrate higher reciprocity among themselves compared to normal users [18,14].

### *2.2. Models of Opinion Diffusion*

A social network can be mathematically represented as a graph $G = (E, V)$, where users correspond to the set of vertices $V$ and their relationships (i.e., interconnections) to the edges in $E$. Usually, individual opinions about a given topic are represented as numerical values in the interval $[0, 1]$. Both limits of the interval are associated with the extreme stances about the considered topic.

Although opinions are iteratively formed by considering how users influence one another, different opinion diffusion models have been proposed in the literature. For example, the aim of DeGroot model [4] is to come to a consensus by using trust as a means to induce differences in the influence of users. DeGroot model was recently applied to the research on hateful behaviours on Twitter and Gab platforms to adjust the score for hate intensity of users [15,18]. In contrast, bounded confidence models follow the intuition that people usually do not accept opinions too far from their own, which is known as *confirmation bias*. For instance, Friedkin-Johnson [6] adds some kind of stubbornness, distinguishing between an intrinsic initial opinion, which remains the same, and an expressed opinion, which changes over time. Hegselmann-Krause (HK) [10] introduces confidence level—a threshold for opinion difference. Deffuant-Weisbuch (DW) [22] was

the first to use asynchronous random opinion updates of two users considering the confidence level. Finally, Terizi et al. [21] conducted extensive simulations showing that HK and DW outperform other models in describing the spread of hateful content on Twitter.

### *2.3. Multi-Agent Simulations in the Context of Hatred and Polarisation*

To the best of our knowledge, the majority of multi-agent simulations in the context of hatred and polarisation consider a two-dimensional grid as communication topology. Jager & Amblard [11] conducted a general simulation based on the Social Judgement Theory to demonstrate consensus and bi-polarisation. Stefanelli & Seidl [20] used the same theory to model opinion formation on a polarised political topic in Switzerland. The authors used empirical data to set up the simulation and validate their results. Bilewicz & Soral [3] proposed their own model of the spread of hatred. As apposed to this, Schieb & Preuß [19] employed the Elaboration Likelihood Model on a message-blackboard. In contrast to the models presented in Section 2.2, where the underlying psychological models use multiple influence factors to model opinion, these works rely on a simple combination of one-dimensional opinion values. In general, none of mentioned models considered more complex typologies of social networks, neither they studied countermeasures against the spread of hatred, which is the main contribution of this paper.

## 3. Terminology

Next, we introduce some terminology required to describe our models.

*Hate score* represents user's attitude and behaviour in terms of hatred. It is a real number in $[0,1]$, where both extremes correspond to a very non-hateful and hateful opinions respectively. We use the hate score as a user opinion value in our diffusion models. The same concept was also employed by Mathew et al. [15] who showed that hate score distribution on Gab is positively biased towards non-hateful stance. Similarly, we define a user as hateful when $\textit{hate score} \geq 0.75$, else as normal in accordance to and for better comparability with previous work. As stated before, the amount of haters is known to be a minority of ca. 1%. Therefore, we model the hate score using the Gamma distribution $\Gamma(\alpha,\lambda)$ as depicted in Figure 3, so that the area under curve for $x > 0.75$ is ca. 0.01. In rare cases when the Gamma distribution naturally exceeds the value of 1 we artificially set users' hate score to the extreme stance of 1.

*Hate core* is a network component consisting of densely connected hateful users as shown in Figure 1. Such components can be the result of high cohesiveness among hateful users and their higher activity. Although single users within a hate core do not exhibit the same influence as some famous mainstream users, as a compound they can achieve similar effects and attract other users. *Hate strains* can then emerge from a hate core as the result of opinion diffusion under negative influence as shown in Figure 2.

*Swap to a hateful society* takes place when the amount of hateful users exceeds 30% of all users in the social network. We consider such situation as the outcome of an irreversible process, which destabilises the society in a very severe way. Experiments have shown that after having trespassed this limit there is no return to a non-hateful society within the time scope of our simulation.

## 4. Baseline Model

Our baseline model is a multi-agent social network where users distribute content. The type of content depends on the user profile, which can be normal or hateful. Subsection 4.1 details how such users are added and connected in the network. Then, Subsection 4.2 reuses insights from the previous research in Section 2 to model the spread of hatred. Our aim is to ensure that the baseline model is close to reproduce findings from previous work. We name this model *baseline* because subsequent sections enrich it with different countermeasures and study their effectiveness in containing the spread of hatred.

### *4.1. Network Construction*

The structure of a social network can be reproduced by the *preferential attachment* iterative method [2]. Briefly, in each round, when joining the network, new users connect to existing users with a probability corresponding to the *node degree* (amount of followers). So that users with many followers are more likely to receive new followers. Since this method does not distinguish between different user profiles, we extend it for hateful users. Firstly, we include hateful users according the Gamma distribution $\Gamma(10,25)$ in Figure 3 (see also Section 3). Secondly, we mimic their behaviour on Gab and Twitter as described in Subsection 2.1:

- When joining the network, a new hateful user will create twice new connections than a normal user. Specifically, we assign the connection variables $c_h = 2$ and $c_n = 1$, where $h$ stands for hater and $n$ for normal.
- A new hateful user will opt-in to connect to the group of hateful users with the probability $p_{h\to h} = 0.9$, else to normal users (the arrow $\to$ indicates connection). After that, the preferential attachment is applied to select a specific user within each group. A hateful followee[3] will answer with the same probability $p_{h\leftarrow h} = 0.9$ and spawn a following connection (the arrow $\leftarrow$ indicates following back).
- Hateful users receive less followers from normal users per time interval. Hence, following back by normal users is modelled with $p_{n\leftarrow n} = 0.8$ and $p_{h\leftarrow n} = 0.4$. Lastly, haters will be less likely to follow back normal users with $p_{n\leftarrow h} = 0.08$ compared to the opposite direction.

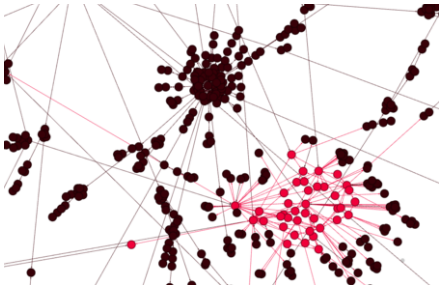

**Figure 1. Hate core**
Densely connected hateful users (red nodes)

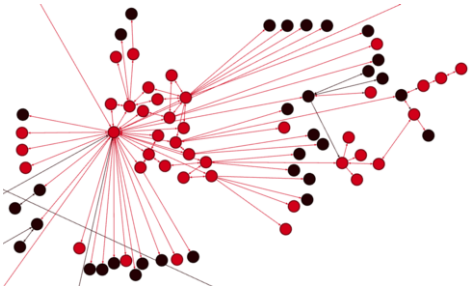

**Figure 2. Hate strains**
Hate core disseminating hatred in strains to nodes with lesser network density

[3] In the terminology of Twitter, a followee is a person who is followed by someone.

### 4.2. Opinion Diffusion by Content

In the bounded confidence models described in Section 2.2, the influence of users is limited to its followers. However, in social networks such as Twitter, the content created by users can be re-posted and, thus, arrive to and influence further audience. Here we reuse the concept of confidence level. Also, the formula of opinion adaption is borrowed from the DW model [22] (see Subsection 2.2), but applied to the author's opinion carried within a post. Thus, a post of user $j$ will influence the opinion of user $i$ by a factor $\mu = 0.05$ at the round $k$, if the difference of both opinions is below a confidence level $\tau_i$:

$$x_{i,k} = x_{i,k-1} + \mu \cdot (x_{j,k-1} - x_{i,k-1}), \quad \text{iff} \quad |x_{i,k-1} - x_{j,k-1}| < \tau_i \tag{1}$$

where $x_{i,k}$ represents the opinion of user $i$ at time $k$ and the $\tau_i$ threshold is modelled as a triangular function on the users' opinion (hate score), so that extreme users will exhibit a rather fixed opinion.

When a post is created or re-posted by some user, then it can influence all of its followers as readers. Thereby:

- Hateful users will be more active and post at every round (i.e., with probability $p_{h\_pub} = 1$), whereas normal users only with $p_{n\_pub} = 0.2$.
- A post cannot be re-posted twice by the same user. However, it can be re-posted with some low probability even if the opinion does not correspond to re-poster's own opinion. We align here w.r.t. retweet statistics provided by Ribeiro et al. [18] and set re-posting probabilities between normal and hateful users to $r_{n\to n} = 0.15$, $r_{h\to h} = 0.45$, $r_{n\to h} = 0.05$ and $r_{h\to n} = 0.15$ (here, the arrow $\to$ indicates content flow). In this manner, a hater will re-post a normal post with the lowest probability.
- In order to consider different users' activity profiles, we limit the amount of re-posts by the same user per round by setting variables to $m_h = 6$ and $m_n = 2$.

## 5. Modelling Countermeasures

We enrich the baseline model with three alternative countermeasures aimed at containing the spread of hatred. Next subsections detail them.

### 5.1. Educational Bias

As mentioned in the introduction, one of the long-term effects of the education could be a positive bias introduced into the population. This can be modelled by skewing the distribution used for sampling of the *hate score* for society members (see Section 3). The mean value of the whole distribution should then move into the direction of non-hateful persons, hence decrease. However, we assume that despite the educational bias on the majority of the population, the group of very hateful persons will still be present in the population with the same proportion of ca. 1%. So, as depicted in Figure 3, we change the parameters of the Gamma distribution $\Gamma(\alpha, \lambda)$ so that their mean values $\mu$ are decreased. Hence, rather than modelling how this positive bias is actually introduced, we simply apply the positively-skewed distributions during the network construction phase (see Subsection 4.1).

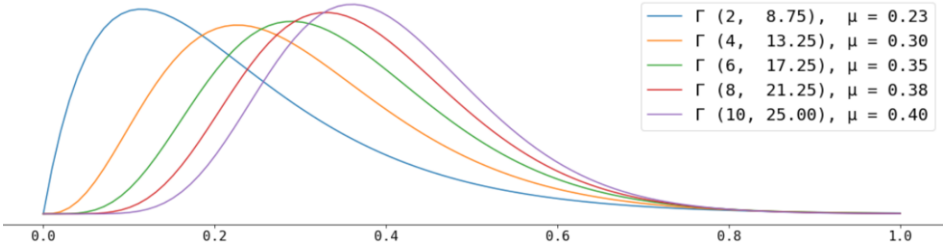


**Figure 3.** Alternative hate score probability distributions modelled with $\Gamma(\alpha, \lambda)$ distribution

### *5.2. Deferring Hateful Content*

In contrast to the educational bias, deferring hateful posts is applied during the opinion dynamics phase described in Subsection 4.2. According to Dharmapala and McAdam [5], the perceived amount of hate speakers play a key role for motivating other users to join and engage in hate speech, making conversations more viral. It is also known that the lifetime of hateful conversations does not exceed a few days and has a culmination within the margins of one day [13]. Hence, the assumption is that deferring hateful content might decrease the willingness of responses. As far as we know, such reactions give more weight to the content and lead to better promotion of it by internal algorithms of social networks [17]. In this work, the response to such content is interpreted as re-posting. We employ a variable probability $p_{defer}$ of deferring a post at some round. Any hateful post can be deferred again, if it is re-posted in further rounds. In addition to parameters in Section 4.2, a cumulative factor $f_{deferred} = 0.5$ is used to decrease the probability of being re-posted.

### *5.3. Counter Activism*

The group of counter activists is aimed to be the pole of '*the good*'. They build a counter movement by convincing people to promote anti-hate slogans and to spread positively influencing messages. Such actors exhibit activity behaviours very similar to hateful users, but transport the opinions from the lower hate score interval $[0, 0.25)$. This countermeasure starts during the opinion dynamics phase, where activists are sampled from the group of non-hateful persons with a probability $p_{convince}$. By default, their opinion is not fixed and can change due to opinion diffusion. When it exceeds $hate\ score \geq 0.25$ they change to normal activity, but keep previously created connections. Furthermore:

- On becoming activist (denoted as $a$), a person spawns additional following connections $c_a$ to the group of all activists, which are answered with the probability $p_{a \leftarrow a} = 0.9$.
- Activists publish posts with the probability $p_{a\_pub} = 1$ at every round.
- They never re-post any content of haters and vice versa, but promote non-hateful content frequently. For the rest, the rules of normal users are used. Therefore, the re-posting probabilities are $r_{a \to h} = r_{h \to a} = 0$, $r_{a \to a} = 0.45$ and $r_{a \to n} = r_{n \to a} = 0.15$.
- The maximal amount of re-posts per round is set to $m_a = 6$.

## 6. Simulation Results

The simulation is conducted in two phases. First, the network construction from Section 4.1 is run until $t_1 \in [0, 500, 1000, 2000, 5000]$ rounds, which creates a network with the same size of user nodes. Then, opinion dynamics from Section 4.2 starts so that the network is grown for further 1000 rounds. Each simulation is conducted 100 times for building the following average metrics:

- Fractions of normal or hateful users and posts.
- Mean and standard deviation of hate score distribution within the society.
- Ratio of network densities of hateful to normal users, which shows how much hateful users are more cohesive than normal users.
- Reciprocity of following within normal or hateful users.
- Mean amounts of followers and followees as well as follower-followee ratio.
- Mean path length of re-posts through the network.
- Fraction of swaps to a hateful society (see Section 3). Runs which end with a swap are not taken into account for none of the above metrics due to the instability they introduce. Instead, they are tracked separately through this specific metric.

### *6.1. Validating the Baseline Model*

Validation of the baseline model is an important step for this work, since it normalises the simulation with real statistics on hateful users. In the first phase of simulation—network construction—multiple metrics could be satisfactorily reproduced in accordance to the state-of-the-art. However, runs resulted in extremely high network density ratios of hateful users over normal users: ca. 11 times more than reported by [14]. Also the amount of followers as well as the ratio between followers and followees of haters were to high compared to normal users. During the second phase—opinion dynamics—these metrics decreased very close to reported values for higher network sizes. Only the reciprocity among hateful users were too low compared to normal users. Although this might be repaired by introducing additional rewiring rules for users who switch from normal to hateful state, we advocate for the simplicity of the model and leave this for future work. Overall, it can be stated that our simulation represents hateful behaviours in a convenient way. Further, an interesting fact is that the switch from network growth to opinion diffusion demarcates a structural change in the sub-network of hateful users. It allows hate cores to disseminate hatred in strains to normal users with lesser network densities, showing that true hate cores might be even more densely connected than reported by statistics about real social networks.

### *6.2. Countermeasures simulation results*

#### *6.2.1. Educational Bias*

We perform simulations of using education as a countermeasure by decreasing the $\alpha$ parameter of the Gamma distribution to induce stronger positive bias (see Figure 3 and Section 3). Overall, this countermeasure can be summarised as being very successful. On the one hand, it can substantially decrease the amount of hateful persons, even if not remove them completely from the society as can be seen on the left of Figure 4. Even with

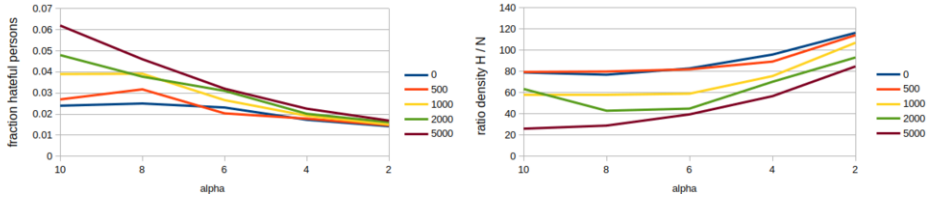


**Figure 4.** Effects of the education as countermeasure for different network sizes referred by the value of $t_1$: (left) fraction of hateful users; (right) ratio of network densities of hateful to normal users depending on the parameter $\alpha$ of the Gamma distribution

the strongest educational bias using $\alpha = 2$ the amount of haters does not fall below of 1%. Also, the amount of hateful posts decreases similarly. The risk of swaps to a hateful society drops from 25% below 5% for the value of $\alpha = 6$. The final mean hate score approaches even lower values than originally introduced by the Gamma distribution. This can be explained by the structure of the network produced using preferential attachment, where some nodes have unproportionally higher influence. Hence, applying a skewed distribution upon it can skew the final distribution even more after opinion diffusion.

On the other hand, the density among hateful users increases as depicted on the right of Figure 4. The same happens for the reciprocity and mean follower-followee ratio. Education impedes the emergence of hate strains, so that hateful persons stay among like-minded within highly densely connected hate cores. Surprisingly, the mean path length of hateful posts increases linearly. This is so because, although hate posts have much less room to unfold by re-posting within hate strains (see Figure 2), hateful posts can still make very long paths by circulating posts between persons within a hate core (see Figure 1).

### 6.2.2. *Deferring Hateful Content*

We conduct simulations by varying the deferring probability of hateful content $p_{defer}$ and a value of 0.7 deems realistic considering the state-of-the-art accuracy in recognition of hate speech [1]. Compared to the education, this countermeasure is less successful in decreasing the fraction of hateful persons as can be seen on the left of Figure 5. There is even some kind of reluctance and increase for $p_{defer} < 0.5$. Something similar happens with the fraction of hateful posts and mean hate score. The reason for such reluctance, which can bear hidden risks especially in case of bad hate recognition accuracy, is not answered in this work and kept for future research. However, this countermeasure has an obvious effect in decreasing the mean path length of hateful content. More outstanding is its property in protection against swaps to a hateful society as shown on the right of Figure 4. This is very similar to the education, but provides a short-term effect. More importantly, it has the advantage of being much aligned with the freedom of speech value than the short-term countermeasure of automatic filtering.

### 6.2.3. *Counter Activism*

In the case of counter activists, we used different simulation setups with the four parameters in Table 1 with the aim of increasing the strength of the counter movement. Surprisingly, none of those simulations lead to a clear decrease of hateful users as depicted on the left of Figure 6. A decrease could be only recorded for settings with bigger networks

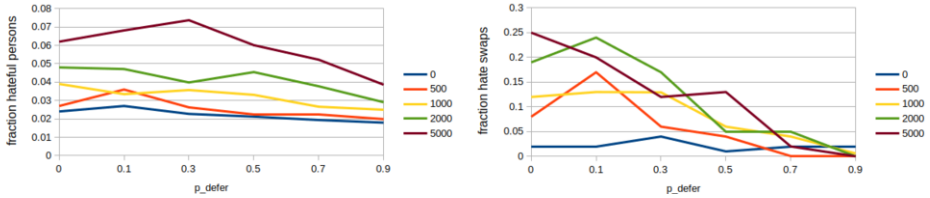


**Figure 5.** Effects of deferring hateful content as countermeasure for different network sizes $t_1$: (left) fraction of hateful users; (right) fraction of swaps to a hateful society depending on the deferring probability $p_{defer}$

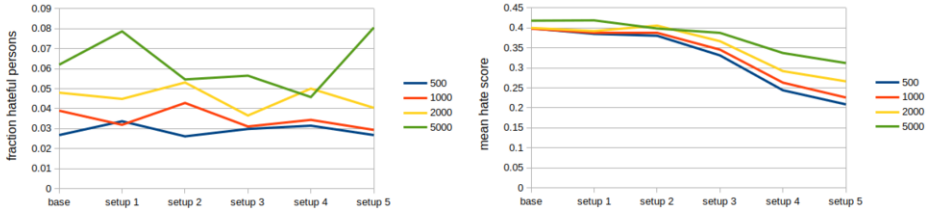


**Figure 6.** Effects of counter activists as countermeasure for different network sizes $t_1$: (left) fraction of hateful users; (right) mean hate score depending on the increasing strength of activist in subsequent setups

| | **setup 1** | **setup 2** | **setup 3** | **setup 4** | **setup 5** |
|---|---|---|---|---|---|
| Convincing probability $p_{convince}$ | 0.01 | 0.01 | 0.04 | 0.01 | 0.01 |
| Additional connections to other activists $c_a$ | 1 | 2 | 1 | 2 | 2 |
| Fixed opinion (stubbornness) | false | false | false | true | true |
| Select activists by their influence | false | false | false | false | true |

**Table 1.** Experiment settings for activists' countermeasure

over 5000 users in setups 2–4. The same happens to the fraction of hateful posts and, even more alarming, the fraction of swaps to a hateful society. Even so, a drop of the mean hate score was recorded—especially for the settings with stubborn activists—as seen on the right of Figure 6. Thus, activists seem to create higher polarisation within the society by dragging some persons into the positive direction without affecting hateful persons. This depletes representatives of the median opinion, so that people with higher hate scores are rather attracted by very hateful users. It shows that activism needs to be carried out in a very sensible way, which is beyond the scope of this paper.

## 7. Conclusions and Future Work

We propose a multi-agent model of the spread of hatred. We reuse insights from previous research to construct and validate a baseline model. Then, we enrich it by adding three countermeasures to study their effectiveness in containing the spread of hatred. As a result, we conclude that: i) The long-term measure of education is very successful, but it still cannot eliminate hatred completely; ii) Deferring hateful content has a similar positive effect with the advantage of being a short-term; iii) Extreme positive cyber activism can increase the society polarisation. Beyond that, we indicate that true hate cores, which are responsible for hate spread, might be even more densely connected than reported by statistics about real social networks. As future work, we plan to further refine the model (as mentioned along the paper) as well as to incorporate additional countermeasures.

## References


[1] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep learning models for multilingual hate speech detection, 2020.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[3] Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020.

[4] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[5] Dhammika Dharmapala and Richard H McAdams. Words that kill? an economic model of the influence of speech on behavior (with particular reference to hate speech). *The Journal of Legal Studies*, 34(1):93–136, 2005.

[6] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[7] Lena Frischlich, Tim Schatto-Eckrodt, Svenja Boberg, and Florian Wintterlin. Roots of incivility: How personality, media use, and online experiences shape uncivil participation. *Media and Communication*, 9(1):195–208, 2021.

[8] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. Unesco Publishing, 2015.

[9] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Countering hate on social media: Large scale classification of hate and counter speech. *arXiv preprint arXiv:2006.01974*, 2020.

[10] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

[11] Wander Jager and Frédéric Amblard. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, 10(4):295–303, 2005.

[12] Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Dissecting the meme magic: Understanding indicators of virality in image memes. *arXiv preprint arXiv:2101.06535*, 2021.

[13] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[14] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182, 2019.

[15] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

[16] Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. Interaction dynamics between hate and counter users on twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 116–124. 2020.

[17] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, 33(4):459–478, 2015.

[18] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[19] Schieb, Carla and Preuss, Mike. Considering the elaboration likelihood model for simulating hate and counter speech on facebook. *SCM Studies in Communication and Media*, 7(4):580–606, 2018.

[20] Annalisa Stefanelli and Roman Seidl. Opinions on contested energy infrastructures: An empirically based simulation approach. *Journal of environmental psychology*, 52:204–217, 2017.

[21] Chrysoula Terizi, Despoina Chatzakou, Evaggelia Pitoura, Panayiotis Tsaparas, and Nicolas Kourtellis. Angry birds flock together: Aggression propagation on social media. *arXiv preprint arXiv:2002.10131*, 2020.

[22] Gerard Weisbuch. Bounded confidence and social networks. *The European Physical Journal B*, 38(2):339–343, 2004.