

# Modelling the Spread of Toxicity and Exploring its Mitigation on Online Social Networks

Aatman Vaidya<sup>1</sup>, Harsh Bhagat<sup>1</sup>, Seema Nagar<sup>2</sup>, Amit A. Nanavati<sup>1</sup>

<sup>1</sup>School of Engineering and Applied Science, Ahmedabad University  
aatman.v@ahduni.edu.in, harsh.b2@ahduni.edu.in, amit.nanavati@ahduni.edu.in

<sup>2</sup>IBM India Research Lab  
senagar3@in.ibm.com

## Abstract

Hate speech on online platforms has been credibly linked to multiple instances of real world violence. This calls for an urgent need to understand how toxic content spreads and how it might be mitigated on online social networks, and expectedly has been the topic of extensive research in recent times. Prior work has largely modelled hate through epidemic or spread activation based diffusion models, in which the users are often divided into two categories, hateful or not. In this work, users are treated as transformers of toxicity, based on how they respond to incoming toxicity. Compared with the incoming toxicity, users amplify, attenuate, or replicate (effectively, transform) the toxicity and send it forward. We do a temporal analysis of toxicity on Twitter, Koo and Gab and find that (a) toxicity is not conserved in the network; (b) only a subset of users change behaviour over time; and (c) there is no evidence of homophily among behaviour-changing users. In our model, each user transforms incoming toxicity by applying a *shift* to it prior to sending it forward. Based on this, we develop a network model of toxicity spread that incorporates time-varying behaviour of users. We find that the *shift* applied by a user is dependent on the input toxicity and the category. Based on this finding, we propose an intervention strategy for toxicity reduction. This is simulated by deploying peace-bots. Through experiments on both real-world and synthetic networks, we demonstrate that peace-bot interventions can reduce toxicity, though their effectiveness depends on network structure and placement strategy.

## 1 Introduction

Online hate speech has been documented to have produced real-life effects<sup>1</sup>. The research by Müller and Schwarz (2021) demonstrated that anti-refugee sentiment on Facebook directly led to physical attacks against refugees throughout Germany. Hate speech on Facebook led to real world violence in Ethiopia<sup>2</sup>. The Observer Research Foundation identified a direct link between hateful speech online and physical violence in Indian society by tracing cases of mob aggression which started from inflammatory internet

content (Mirchandani 2018). This underscores the urgent need to understand how hate speech propagates through online social networks (OSN), and to develop effective strategies to mitigate its spread. Of late, the study of hate speech and its spread is being examined through multiple perspectives and approaches in order to gain insights into this problem. The expectation from these efforts is to (a) understand the process of spreading of hate and (b) find ways of mitigating it.

Prior research has approached this problem from multiple angles. One line of work classifies messages or users, as hateful vs. non-hateful, in order to study user behaviour changes and the implications of the structure of the network connecting them (Ribeiro et al. 2018; Mathew et al. 2018). Another (often overlapping) line of work attempts to analyse and model the flow of hate through the network. Several variants of Belief propagation (Mathew et al. 2018) and Spreading and Activation (SPA) based models (Nagar et al. 2021b) have been used to model the diffusion of hatred through social networks. Other works have extended epidemic spread models for this purpose as well (Yousefi 2024).

We presuppose that hate is not binary. We use *toxicity* to quantify the degree of hatred in a message (Google Jigsaw 2018). Given an underlying network, each user receives an incoming toxicity in the range [0-1] and outputs an outgoing toxicity in the range [0-1]. Each user is thus a transformer of toxicity. A user responds to a stimulus of the input toxicity they receive by applying a *shift* to it. This leads us to investigate the following questions:

**RQ1:** How does toxicity spread in the network? Does it depend on the structure of the network?

**RQ2:** Are some users more responsible for the spread of toxicity than others? Does user behaviour change with time?

**RQ3:** Are there ways to mitigate the spread of toxicity? Especially, are there soft interventions which do not require the removal of users or connections to achieve this?

To answer these questions, we experimented on three large scale online social networks: Twitter, Koo and Gab, and analysed them for the spread of toxicity over time. Figure 2 describes the flow of this paper. Our contributions in

this work are the following:

- Based on empirical findings, we classify users into three distinct categories based on how they respond to toxicity. *Amplifiers* (who increase it), *Attenuators* (who decrease it), and *Copycats* (who propagate it with little change), see Figure 1. We show that this categorization is observable across all three platforms.
- Through a temporal analysis, we show that standard epidemiological models don’t adequately capture toxicity spread. We find that nearly half of the users remain in a fixed user category over time, while others fluctuate. This is a phenomenon not captured by traditional Susceptible-Infected-Recovered (SIR\*) like frameworks.
- We develop a new model for toxicity spread based on our findings that a user applies a “*shift*” in toxicity depending on their category and the level of toxicity they receive.
- Based on this model, we propose a soft intervention using peace-bots to strategically lower the average toxicity users are exposed to, hence, decreasing the total toxicity observed in the social network.
- We conduct simulations on both real-world and synthetic networks to evaluate our model and peace-bot strategy. Our experiments show that most effective deployment strategy of peace-bots is highly dependent on the underlying network structure, with different strategies proving optimal for different platforms.

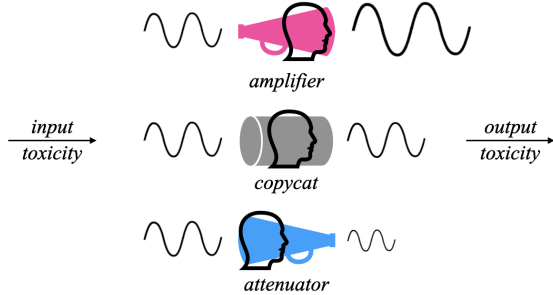


Figure 1: Users viewed as transformers of toxicity: amplifiers, users whose output toxicity is higher than their input toxicity; copycats, whose output toxicity is almost the same as their input toxicity; and attenuators, whose out toxicity is less than their input toxicity.

## 2 Related Work

**Modelling Spread of Hateful Content.** Several studies have examined how hateful or toxic content spreads through online social networks. Mathew et al. (2018) used belief propagation to model the diffusion of hateful content, showing that it travels farther, faster, and reaches a wider audience than non-hateful content. In a follow-up study, Mathew et al. (2019) analysed hate speech on Gab using a DeGroot model, finding that hateful users tend to become central quickly, and that newer users adopt hateful behaviour faster. Ribeiro et al. (2017) studied the user characteristics of hateful users

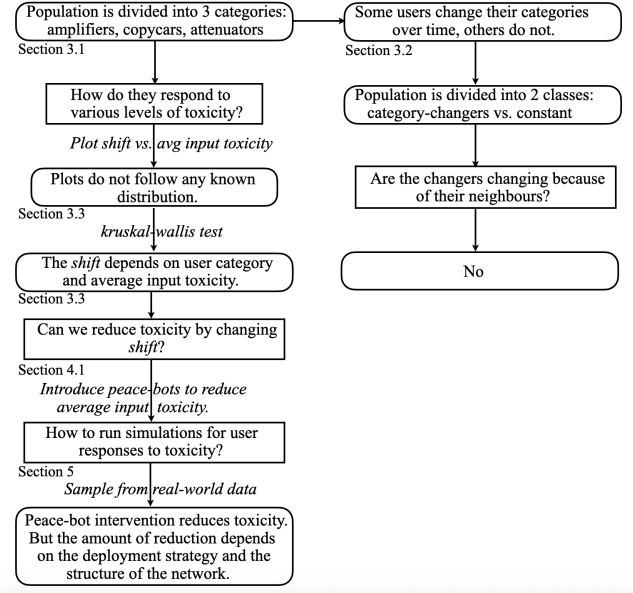


Figure 2: The flow of questions explored in the paper, beginning with the understanding that there are three categories of users.

on Twitter, and highlighted that hateful accounts differ from normal users in activity, network centrality, and content type.

Recent work has explored more nuanced modelling approaches. Vaidya, Nagar, and Nanavati (2024) proposed a model of toxicity spread on Twitter, classifying users into behavioural categories based on their responses to incoming content. Lerman et al. (2024) examined affective polarization, showing that toxicity-driven polarization is not limited to group-based divides but is instead a structural property of social networks. Maarouf, Pröllochs, and Feuerriegel (2022) analysed virality, finding that hateful content from verified users is disproportionately more likely to spread widely.

Topic based modelling approaches have also been proposed. Nagar et al. (2021a) developed a graph autoencoder that integrates user and textual features to track the spread of hate. Goel et al. (2023), consistent with Mathew et al. (2018), found that hateful users play a more crucial role in governing the spread of information compared to singled-out hateful content. They also observe that hatemongers dominate the echo chambers in a network. Masud et al. (2021) introduced a topic-aware diffusion model with attention mechanisms that leverages news data to predict hateful retweets, while Gupta et al. (2021) combined topic modelling and ensemble classification to show that hateful tweets receive more retweets than non-hateful ones.

Another prominent line of work applies epidemiological models to capture the spread of hate and toxicity. Obadimu et al. (2020) introduced the STRS model (Susceptible–Toxic–Recovered–Susceptible) on YouTube, showing that toxicity can escalate into an epidemic if the reproduction number  $R_0 > 1$ , and highlighting the importance of interventions that reduce exposure. Addai, Yousefi,

and Agarwal (2024) adapted the SEIQR model on Twitter, incorporating user history and index of memory can provide a nuanced modelling. Maleki et al. (2022) applied the SEIZ model (Susceptible–Exposed–Infected–Skeptic) to study toxicity propagation on Twitter, demonstrating its effectiveness compared to alternative epidemiological models. Dagtas, Agarwal, and Yousefi (2024) evaluated five epidemiological models on Reddit data, achieving high accuracy with less than 2% fitting error. Building on this, Yousefi, Agarwal, and Addai (2024) extended epidemic models to differentiate between moderate and highly toxic users, yielding improved predictive performance.

**Homophily** also described as users being “birds of a feather flock together” highlights that users with similar traits are more likely to connect (McPherson, Smith-Lovin, and Cook 2001; Kossinets and Watts 2009). Halberstam and Knight (2016) measured homophily among politically engaged users on Twitter, and found that majority group users have more connections, greater information exposure, and faster access to information than minority group users, showing the tendency for like-minded content to circulate within groups.

**Bots on Online Social Networks.** Social bots have played a key role in spreading digital content on social networks (Ferrara et al. 2016; Morgan 2018). Social media bots are defined as automated accounts that engage in content creation, distribution, and collection (Ng and Carley 2025). Empirical studies show that bots can shape information dynamics in important ways. For instance, Mønsted et al. (2017) demonstrated through controlled bot experiments that diffusion on Twitter follows complex contagion dynamics rather than simple contagion. Shao et al. (2018) found that bots amplify articles from low-credibility sources during the early stages of diffusion, helping such content gain traction. Uyheng and Carley (2020) linked bot activity to increased levels of hate, particularly in dense and isolated communities in a network.

### 3 Temporal Analysis

#### 3.1 Datasets and Pre-processing

In this section, we describe the three datasets used in our analysis to understand the temporal dynamics of toxicity spread on social networks. Each dataset contains a directed graph  $G$ , the structure and edge type of the graph vary across datasets. All datasets include a collection of text-only posts gathered over time from multiple users<sup>3</sup>. (See Table 1).

**(1) Twitter:** The dataset is published by (Ribeiro et al. 2017) and presents a Twitter dataset of 100K users along with up to 200 tweets from their timelines with a random walk-based crawler on the retweet graph ( $G$ ), and select a subsample of 4,972 to be manually annotated as hateful or not through crowdsourcing with information about hateful tweets and users. The authors also look at user activity patterns of hateful and normal users.

<sup>3</sup>The Koo and Gab datasets are publicly available under the CC-BY Attribution 4.0 licence.

**(2) Gab:** The dataset is created by (Saha et al. 2023) and studies prevalence of fear and hate speech on the social network. It contains 700K hateful and 400K fear speech posts collected from Gab.com. The dataset contained information about posts reshared by users, and we constructed a directed graph based on these reposts.

**(3) Koo:** The dataset is published by (Mekacher, Falkenberg, and Baronchelli 2024) and presents largest publicly available Koo dataset, spanning from the platform’s founding in early 2020 to September 2023, containing 72M posts, 40M shares and other metadata. Koo was a popular microblogging platform based in India. Similar to Gab, the dataset contained information about posts re-shared by users and we constructed a directed graph based on shares.

For each dataset, we assign a toxicity score using Perspective API to every post (Google Jigsaw 2018). Toxicity is defined as a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.

We wanted to examine how toxicity evolves over time in a network and how users respond to toxicity. To do so, we filtered all three datasets to find a set of users who were present during a time period. Depending on the dataset, the temporal unit of the time period was either weeks or months. Table 1 provides details of the selected timelines and the corresponding user statistics.

| Dataset                 | Timeline        | Time Unit | No. of Posts | Nodes  | Edges |
|-------------------------|-----------------|-----------|--------------|--------|-------|
| <i>Complete Dataset</i> |                 |           |              |        |       |
| Twitter                 | 01/2017–10/2017 | Week      | 17.2M        | 99.8K  | 2.27M |
| Gab                     | 10/2016–06/2018 | Month     | 20.1M        | 62.3K  | 10.4M |
| Koo                     | 01/2020–09/2023 | Month     | 16.9M        | 214.9K | 1.93M |
| <i>Temporal Subset</i>  |                 |           |              |        |       |
| Twitter                 | 07/2017–10/2017 | Week      | 1.34M        | 40.9K  | 737K  |
| Gab                     | 11/2017–06/2018 | Month     | 6.08M        | 53.2K  | 2.19M |
| Koo                     | 08/2022–03/2023 | Month     | 2.96M        | 49.1K  | 1.14M |

Table 1: Datasets Overview and Temporal Subsets.

In the next section, we list our findings from analysing toxicity spread across the three datasets.

#### 3.2 Preliminary Analysis

We begin by examining how toxicity changes over time. First, we analyse the Twitter dataset and then repeat similar experiments on the Gab and Koo datasets. Figure 3 shows the average toxicity changes across time on all three datasets. Nagar et al. (2021b) used spread activation (SPA) models to capture toxicity in networks, however, our analysis in Figure 3 indicates that these models don’t adequately capture its spread.

To better understand how users react to toxicity from a network perspective, we investigated whether a user’s neighbourhood plays a role in influencing their behaviour. To measure the influence of a user  $u$ ’s neighbourhood on  $u$ , we calculate the difference between  $u$ ’s average toxicity and

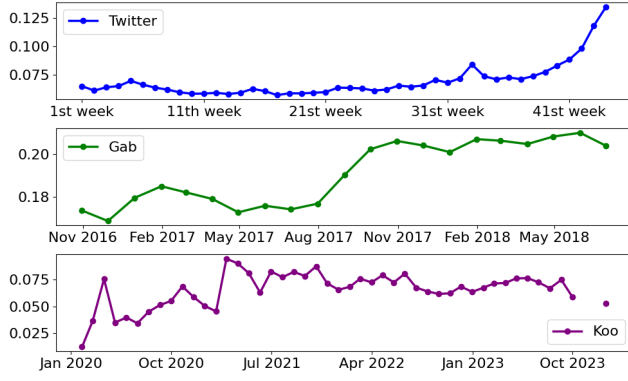


Figure 3: Average Toxicity over time in all the datasets.

the average toxicity of  $u$ 's in-degree neighbours over time. We call this difference **shift**. A shift represents the average change a user applies to the incoming toxicity before transmitting the message further.

Figure 4 shows the distribution of shifts in all the three datasets. We then applied the Shapiro–Wilk test to check for normality of these distribution's, and the results showed that is not a normal distribution (Shapiro and Wilk 1965).

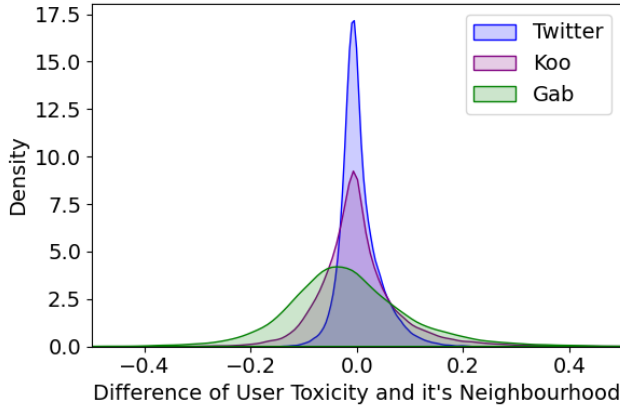


Figure 4: Distribution of the difference between User's Average Toxicity and the Average Toxicity of its in-degree neighbourhood in all three datasets. All the distribution's fail the Shapiro–Wilk (SW) test for normality.

In such a case where the distribution is not normal, we used the Interquartile Range (IQR) measure to categorise users based on their *shift* behaviour. This helps detects outliers in the distribution and separate users. Figure 5 shows the box-and-whisker plot of the distribution in 4. This approach naturally divides the set of users into 3 disjoint subsets and help us categorise user behaviour (Vaidya, Nagar, and Nanavati 2024). These categories reflect how users respond to toxicity. The outlier users on the right in Figure 5 are called *amplifiers*, on the left are called *attenuators* and the remaining (typical users) are called *copycats*. The amplifiers send out more toxicity than they receive, the attenuators send out less toxicity than they

receive and the copycats send out almost the same toxicity as they receive (see Figure 1).

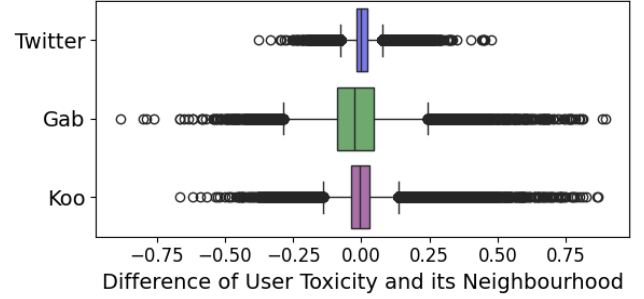


Figure 5: Box and Whisker plot for the distributions in Figure 4. Since the data is not normal, we use the IQR method to detect outliers and separate them from normal users. The outlier users on the right are called *amplifiers*, on the left are called *attenuators* and the remaining are called *copycats*.

Vaidya, Nagar, and Nanavati (2024) employed the same methodology and showed that average toxicity in the network changes over time. They further demonstrated that the placement of amplifiers, attenuators, and copycats influences the extent of toxicity spread. While the authors analysed the dataset in aggregate, this does not preclude the possibility that user behaviour changes over time. If toxicity is viewed as a disease, one may ask whether users exhibit temporary symptoms of increasing/decreasing it (i.e., becoming amplifiers/attenuators), and by default, all users are copycats. Prior work has explored epidemiological models such as SIR and SIER to study the diffusion of hate (Yousefi 2024; Maleki et al. 2022; Dagtas, Agarwal, and Yousefi 2024). A temporal analysis can therefore provide valuable insights into whether such approaches offer an effective framework for modelling hate.

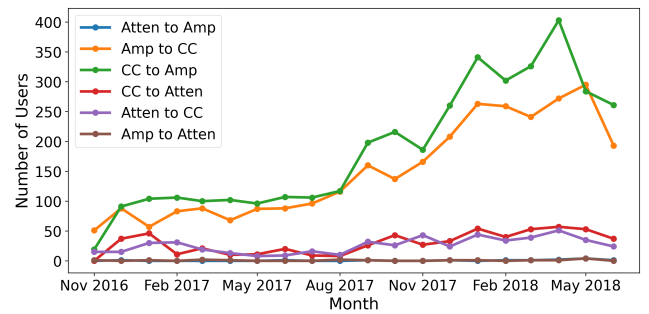


Figure 6: Change in User Category in the Gab Dataset. We observed that a small set of users' change behaviour. Similar Figures for Twitter and Koo have been added in Appendix A.1.

Hence, we looked at how users change behaviour over time. Figure 6 shows that (i) not all users are consistently present in every time interval, (ii) only a set of users change

behaviour across time, (iii) users do not necessarily experience all categories, and their transitions show no fixed patterns. This is unlike diseases, where individuals typically transition between well-defined states (susceptible, infected, recovered). This suggests that epidemiological models such as SIR may have limited effectiveness in fully capturing the spread of toxicity on a network.

An immediate question that follows is, *why do some users change their behaviour?* To check if users change their behaviour because of their neighbours, we calculated the network-level homophily of users (Easley, Kleinberg et al. 2010). We defined users who changed their behaviour over time as *changing* users and those who did not as *non-changing* users. The results, presented in Table 2, show that there is no evidence of homophily within the set of changing users, within the set of non-changing users, or between the two groups.

| User Category    | Actual Edges ( $x$ ) | Probability ( $x^2$ ) |
|------------------|----------------------|-----------------------|
| changing (p)     | 0.012234             | 0.008665              |
| not changing (q) | 0.869270             | 0.822492              |

Table 2: Homophily between Changing and Non Changing users in the Network. Actual edges ( $x$ ) represent the fraction of edges within a given category to the total number of edges in the graph. The probability ( $x^2$ ) is the squared value of actual edges. A user category is said to exhibit network homophily if  $x$  is much larger than  $x^2$ .

We next ask, *how do users in the three categories respond to incoming toxicity?* How does their output toxicity change with incoming toxicity? In order to understand this, we inspect the shift distributions in the next section, of all categories of users across all three datasets.

### 3.3 Shift Analysis

A key aspect to understanding user behaviour is, how do different categories of users respond to different ranges (say low, medium, high) in input toxicity? Is their reaction consistent regardless of the input range? If not, then how does the variation differ across the three categories of users? These questions lead us to inspecting the shift behaviour.

For this analysis, we require a timeline with a consistent set of users. Hence, we filter each dataset to retain such subsets of users. For Twitter we analyse 7,170 users and their neighbourhoods over 13 weeks; for Gab, 4,808 users over 8 months; and for Koo, 5,645 users over 8 months. Further details about these temporal subsets are provided in Table 1, where the reported totals include both users and their neighbourhoods.

| Dataset | Amplifier | Attenuator       | Copypcat |
|---------|-----------|------------------|----------|
| Twitter | [0, 0.50] | <b>[0, 0.80]</b> | [0,0.3]  |
| Gab     | [0, 0.50] | <b>[0, 0.70]</b> | [0,0.6]  |
| Koo     | [0, 0.35] | [0, 0.40]        | [0,0.3]  |

Table 3: Range of Average Incoming Toxicities experienced by each category.

| Dataset | Amplifier   | Attenuator   | Copypcat     |
|---------|-------------|--------------|--------------|
| Twitter | [ 0.0, 0.9] | [-0.7, 0.20] | [-0.20,0.20] |
| Gab     | [-0.2, 0.8] | [-0.7, 0.20] | [-0.40,0.30] |
| Koo     | [-0.1, 0.8] | [-0.3, 0.15] | [-0.15,0.15] |

Table 4: Range of the shifts applied by each category of users.

| Dataset | Category Effect | Predecessor Effect | Interaction Effect |
|---------|-----------------|--------------------|--------------------|
| Twitter | 0.0000          | 0.0363             | 0.0631             |
| Koo     | 0.0000          | 0.0000             | 0.0000             |
| Gab     | 0.0000          | 0.4745             | 0.4595             |

Table 5:  $p$ -values of Kruskal–Wallis test.

Figure 7 shows the shift distributions of users in each category for all datasets in temporal subsets. We summarise some of the information in Tables 3 and 4. We make the following observations:

- (Table 3) The range of average input toxicities experienced by each category of users is different. The attenuators in Twitter and Gab seem to experience higher inputs toxicities compared to the rest of the users.
- (Table 4) The range of the shifts applied by each category of users is different. Expectedly, the amplifiers’ positive shifts are the largest and the attenuators’ negative shifts are the largest.

For each type of user category, we checked if the distribution of shifts by input toxicity follows any known distribution. We tested for normality, power-law, lognormal, exponential, stretched-exponential and truncated-power-law, and none of them passed <sup>4</sup> (Alstott, Bullmore, and Plenz 2014). Given that this data does not follow any of the popular named distributions, we conducted the  $k$ -sample Anderson-Darling test (Scholz and Stephens 1987) to check if the various shifts at least belong to the same (un-named) distribution. We found that no two of them belong to the same distribution either.

Then we conducted the Kruskal-Wallis test (Kruskal and Wallis 1952) to check if the shift applied by a user to average incoming toxicity depends upon (a) the user category and (b) the value of incoming toxicity. We found that for the Twitter and Koo datasets, it depends on both; while for Gab, it depends only on the user category. We surmise that the reason for this is that the Gab dataset was created to specifically study hate speech and thus contains mostly hateful posts (see Table 5). The dependence of the shift on these two factors forms the basis of a model for the spread of toxicity, which we discuss in the next section.

## 4 The Model

From the analysis, we found that a user’s response to average incoming toxicity is not static. The shift applied by a user to average incoming toxicity depends upon (a) the user

<sup>4</sup><https://github.com/jeffalstott/powerlaw>

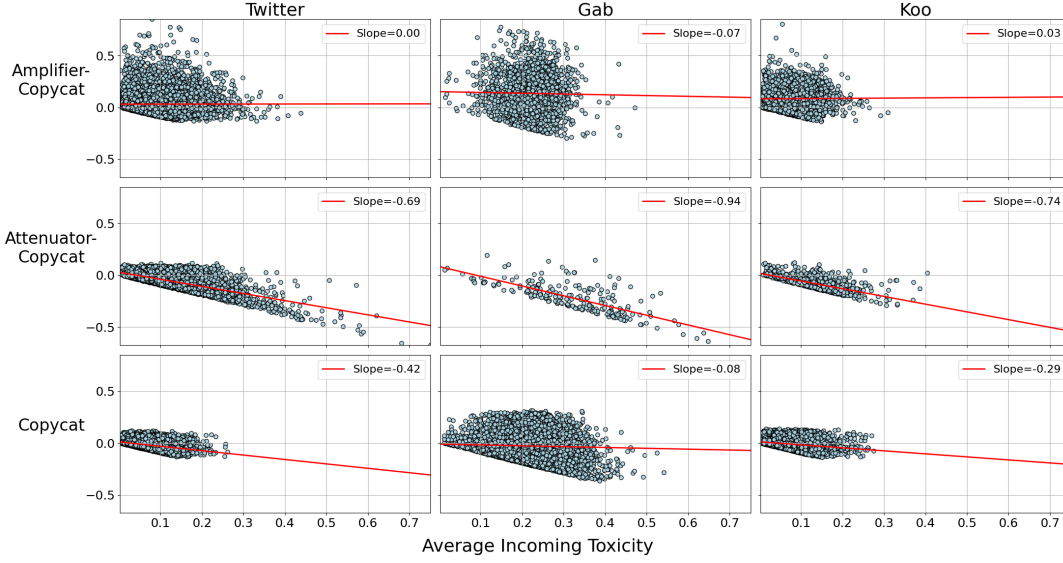


Figure 7: Distribution of shifts as a function of Average Incoming Toxicity across all datasets. Average incoming toxicity for a user  $u$  is the average toxicity value of all in-degree neighbours of  $u$ . Given that most users exist in only one of copypcat $\leftrightarrow$ amplifier, copypcat $\leftrightarrow$ attenuator, and copypcat categories, we plotted the distribution so that each user’s behaviour is reflected in one of the plots only.

category and (b) the value of incoming toxicity. At each time step  $t$ , the output toxicity of a user  $u$  is governed by:

$$O(u, t) = I_{avg}(u, t) + s(c(u, t), I_{avg}(u)) \quad (1)$$

where  $O(u, t)$  is the output toxicity,  $I_{avg}$  is the average of the input toxicity,  $c(u)$  is the category of the user and  $s$  is the shift applied by the user. The general model above can be suitably tailored for datasets such as Gab where the dependence on incoming toxicity is not observed.

From the analysis, we found that the distributions of the shifts as a function of input toxicity do not fit any well known distributions. Due to this, we sample shifts from the Twitter distribution (see Figure 7) for all the experiments in this paper.

We now formulate a model for the spread of toxicity, detailed in Algorithm 1. We specify the following assumptions and model parameters:

- The model process is initialized with one or more starting nodes, each containing a message with initial toxicity  $> 0$ .
- Time is represented as a sequence of forward passes (hops), where a hop denotes the action of a user forwarding a message to its successors (out-degree neighbours).
- A fraction of users change category with time.
- Shifts are sampled from real-world dataset distributions.

We detail the model in Algorithm 1: *tox\_spread*. For each user  $u$  with one or more incoming edges from nodes carrying non-zero toxicities, we compute the average input toxicity. Based on this value and the category of  $u$  at that time step, we sample from the real-world distribution to determine the shift applied by  $u$ .

#### 4.1 Intervention

Given a model that captures the spread of toxicity, a natural next question is how it might be mitigated? Prior work has explored approaches such as removing hateful nodes or links from the network to limit the spread of hate and toxicity (Alorainy et al. 2022; Arttime et al. 2020). Instead, we focus on a gentler approach. Since the shift applied by a user depends on the average incoming toxicity, can we alter the *average* incoming toxicity such that the output toxicity reduces?

Intuitively, since most users are copycats, *reducing* the average incoming toxicity ought to lead to an overall reduction in toxicity. Suppose we deploy “peace-bots”, i.e., bots which send out non-toxic messages (with toxicity = 0). This would lead to a reduction in the average incoming toxicity of the outgoing neighbours of the peace-bots. Several questions arise:

- **Q1:** How can we be assured that the toxicity will reduce? How does this depend on the shape of the shift distributions  $D_{amp}$ ,  $D_{atn}$ ,  $D_{cc}$ ?
- **Q2:** How many peace-bots do we need?
- **Q3:** How to decide where to deploy them in the network?

In order to answer Q1, let us consider the distributions and Equation 1. Rewriting the equation (the parameters have been dropped from the equations for brevity):

$$\begin{aligned} O &= I_{avg} + s(c, I_{avg}) \\ O' &= I'_{avg} + s'(c, I'_{avg}) \end{aligned}$$

The latter equation with  $I'_{avg}$ ,  $s'$ ,  $O'$  denotes the values after the intervention (i.e. after the deployment of peace-bots). So,



---

**Algorithm 1:** *tox\_spread*

---

**Require:** The network of users  $G = (V, \vec{E})$ , shift distributions  $D_{amp}, D_{atn}, D_{cc}$  for each user category.

**Require:** The number of iterations, *kiter*.

**Require:** The initial tweet(s) with their toxicities.

**Ensure:** The final values of toxicity for each user.

```
1: for each timestamp  $t$  in  $[1, \dots, kiter]$  do
2:   for node  $u \in V, tox(u) \neq 0$  do
3:     for  $v \in out(u)$  //  $v$  has an incoming edge from  $u$ 
       do
4:        $curr_{cat} = catg(v)$  //  $catg$  is one of amp, atn, cc
5:        $avg_{tox}(v) = \frac{1}{|in(v)|} \sum_{u \in in(v)} tox(u)$ 
6:       sample  $shift$  from  $D_{curr_{cat}}$  given  $avg_{tox}(v)$ 
7:        $tox(v) = avg_{tox}(v) + shift$ 
8:       if  $tox(v) > 1$  then
9:          $tox(v) \leftarrow 1$ 
10:      else if  $tox(v) < 0$  then
11:         $tox(v) \leftarrow 0$ 
12:      end if
13:      update the category of  $v$  // may remain unchanged
14:    end for
15:  end for
16: end for
```

---

$I'_{avg} < I_{avg}$ . For the intervention to be beneficial,  $O' < O$ , implying that,  $I'_{avg} + s' < I_{avg} + s$ .

Therefore, in the shift distribution  $D$ , if the expected value of the output toxicity is smaller for smaller average input toxicities, then the peace-bot intervention can be expected to work.

To answer Q2, Q3 and to check how effective peace-bot intervention actually is, we turn to experiments.

## 5 Experiments and Results

We evaluate the proposed model and intervention strategy on both randomly generated and real-world networks. We conduct extensive experiments, varying both the number of peace-bots and their positions within the network.

### 5.1 Experimental Setup

We considered random graph models like Barabási–Albert (BA), Erdős–Rényi (ER) graph and Watts–Strogatz (WS) (Barabási and Albert 1999; Erdős and Rényi 1961; Watts and Strogatz 1998). For our experiments, we select ER graphs because BA and WS graphs are less representative of real-world social media networks (Chang et al. 2025). For instance, WS graphs fail to reproduce realistic degree distributions, while BA graphs capture power-law degree distributions but lack community structure and clustering. To address these limitations, we also evaluate our model on real-world graphs of Twitter, Koo, and Gab, as summarized in Table 1. We aim to address the following questions through our experiments:

- **Q1:** Does Intervention strategy of peace-bots help reduce toxicity in the network?

- **Q2:** How does the number of peace-bots matter?
- **Q3:** Does the placement of peace-bots matter?
- **Q4:** Does network structure matter?

We use the following parameters for our experiments:

- **Graph Sizes:** We generate directed ER Graphs ( $G_{n,p}$ ) of sizes  $25K, 50K, 75K$  and  $100K$  nodes with a probability  $p$  of 0.05% for edge creation. We also test the model on real-world graphs with sizes shown in Table 1.
- **Timeline:** To simulate time in the model, we consider 3–4 hops as a week and run the model for 8 weeks. (24–32 hops) (see definition of time and hops in the Algorithm 1)
- **User Distribution:** We keep the proportion of Amplifiers 5.3%, Attenuators 1.4% and Copycats 93.3%. We randomly assign all the nodes in the graph to these user proportions. Along with this, 47% users change their category with some probability detailed in Table 8 in Appendix. The numerical values chosen are observed in the Twitter dataset.
- **Shift Distribution:** To sample shift values, we draw from the distributions observed in the real-world datasets (see Figure 7). For a given value of input toxicity, we consider the corresponding range of shift values and apply a density-based probability sampling approach to select the shift. For our experiments, we chose the shift distribution from the Twitter dataset.

For every week in the simulation, we record total toxicity and average toxicity per user in the network. For each graph size, we run the experiment five times and take average of the total toxicity to find the results. We ran all the experiments on a Mac Studio with 96GB of unified memory.

### 5.2 Results

To understand the placement of peace bots and its effect on total toxicity we devise the following scenarios to run the experiments:

1. Case 1 – No peace-bots assigned.
2. Case 2 –  $N$  peace-bots assigned as indegree neighbours to randomly selected nodes. We call this the Random Placement (RP) strategy.
3. Case 3 –  $N$  peace-bots assigned as indegree neighbours to nodes with the lowest indegree. We call this the Lowest Indegree (LI) strategy.

To measure the effect of peace-bots in reducing the total toxicity, we tabulate the *percentage reduction* in the toxicities based on the RP and LI strategies, with the Case 1 as the baseline.

Table 6 appears to have a clear winner in the LI strategy. This is intuitively easy to explain. All else being “equal”, consider two vertices  $u$  and  $v$ , such that  $indeg(u) < indeg(v)$  and  $avg_{tox}(u) = avg_{tox}(v)$ . The smaller the number of incoming neighbours, the larger the contribution of the peace-bot to the average incoming toxicity. A peace-bot adds zero toxicity, so the updated values of average incoming toxicity for  $u$  and  $v$  after the addition of the peace-bot are updated as follows:

| Nodes | Edges | No. of Peace Bots | RP    | LI     |
|-------|-------|-------------------|-------|--------|
| 25K   | 312K  | 280               | 2.84% | 3.26%  |
|       |       | 560               | 3.58% | 3.78%  |
|       |       | 1120              | 4.89% | 5.94%  |
|       |       | 1250              | 6.61% | 6.42%  |
|       |       | 1400              | 6.18% | 7.80%  |
|       |       | 2800              | 9.08% | 11.65% |
| 50K   | 1.25M | 300               | 1.57% | 1.43%  |
|       |       | 600               | 1.91% | 2.11%  |
|       |       | 1200              | 2.39% | 2.83%  |
|       |       | 1500              | 3.74% | 3.90%  |
|       |       | 2500              | 4.43% | 5.23%  |
|       |       | 3000              | 4.37% | 5.04%  |
| 75K   | 2.81M | 320               | 1.46% | 1.43%  |
|       |       | 640               | 2.41% | 1.96%  |
|       |       | 1280              | 1.89% | 2.23%  |
|       |       | 1600              | 2.05% | 1.87%  |
|       |       | 3200              | 3.12% | 3.60%  |
|       |       | 3750              | 3.44% | 4.34%  |
| 100K  | 5M    | 320               | 1.23% | 1.30%  |
|       |       | 640               | 1.59% | 1.16%  |
|       |       | 1280              | 2.01% | 2.15%  |
|       |       | 1600              | 1.93% | 1.61%  |
|       |       | 3200              | 2.54% | 2.76%  |
|       |       | 5000              | 3.44% | 3.65%  |

Table 6: **ER Graphs Results.** The RP (Random Placement) column shows the percentage reduction in toxicity when the peace-bots are placed randomly in the network, while the LI (Lowest Indegree) column shows the percentage reduction when they are placed as incoming neighbours of lowest indegree vertices. The latter almost always outperforms the former.

$$avgintox(u) = \frac{avgintox(u) \cdot indeg(u) + 0}{indeg(u) + 1} - avgintox(u) \quad (2)$$

$$avgintox(v) = \frac{avgintox(v) \cdot indeg(v) + 0}{indeg(v) + 1} - avgintox(v) \quad (3)$$

Since  $indeg(u) < indeg(v)$ , we expect  $avgintox(u) > avgintox(v)$  after peace-bot intervention.

However, the experiments on the real-life datasets tell a different story. For Twitter, the RP strategy performs better than the LI strategy, with increasing number of peace-bots; For Koo, the RP strategy performs better for higher number of peace-bots; For Gab, they both are very close.

This suggests that the outcome depends considerably on the structure of the network, and we cannot estimate the impact without conducting experiments. So, when deploying peace-bots on any network in practice, one would have to conduct such experiments, until we get a deeper under-

| Real World Graph | Nodes | Edges | No. of Peace Bots | RP     | LI     |
|------------------|-------|-------|-------------------|--------|--------|
| Twitter          | 100K  | 2.28M | 320               | 1.26%  | 0.89%  |
|                  |       |       | 640               | 2.08%  | 0.96%  |
|                  |       |       | 1280              | 3.52%  | 0.68%  |
|                  |       |       | 1600              | 3.71%  | 0.73%  |
|                  |       |       | 3200              | 6.10%  | 1.24%  |
|                  |       |       | 5019              | 8.88%  | 2.49%  |
| Koo              | 275K  | 2.33M | 360               | 4.66%  | 6.31%  |
|                  |       |       | 720               | 4.16%  | 5.57%  |
|                  |       |       | 1440              | 11.55% | 9.34%  |
|                  |       |       | 1800              | 11.16% | 7.57%  |
|                  |       |       | 3600              | 10.15% | 9.24%  |
|                  |       |       | 13751             | 18.33% | 16.44% |
| Gab              | 72.8K | 2.31M | 320               | 1.89%  | 1.65%  |
|                  |       |       | 640               | 2.51%  | 2.78%  |
|                  |       |       | 1280              | 3.90%  | 3.54%  |
|                  |       |       | 1600              | 4.61%  | 4.16%  |
|                  |       |       | 3200              | 6.92%  | 7.32%  |
|                  |       |       | 3644              | 7.78%  | 7.71%  |

Table 7: **Real-world Graphs Results:** The RP (Random Placement) column shows the percentage reduction in toxicity when the peace-bots are placed randomly in the network, while the LI (Lowest Indegree) column shows the percentage reduction when they are placed as incoming neighbours of lowest indegree vertices. For Twitter, the RP strategy performs better than the LI strategy, with increasing number of peace-bots; For Koo, the RP strategy performs better for higher number of peace-bots; For Gab, they both are very close.

standing of the underlying processes which govern user behaviour, as proposed in Section 6.

## 6 Discussion

We now discuss the limitations of our approach and future work.

- *Assumption of peace-bot connectivity:* In our experiments, peace-bots were deployed and users were assumed to follow them. In reality, this would require recommending peace-bot accounts to users (Elmas 2025), and the effectiveness would depend on how many users accept such recommendations.
- *Dependence on shift distributions:* The success of peace-bots in reducing average incoming toxicity relies on assumptions about the shift distributions  $D_{amp}$ ,  $D_{atn}$ ,  $D_{cc}$ . If shifts increase relative to the decrease in incoming toxicity, the intervention may fail. Identifying the properties these distributions must satisfy for the intervention to succeed is essential, though not sufficient. Network structure and the placement of user categories also strongly influence the outcome.



- *Why do some users change their behaviour?* While there was no evidence of homophily, the question as to why some users change their behaviour and others don't remains. Is this driven by intrinsic user traits, or by their past exposure to toxicity? Insights from social science could help us better understand how individuals respond to toxic content, leading to a more complete picture of the psychological+social individual.
- *Finer models:* In this paper, we classified users into three categories based on their overall response across all toxicity values. However, users may behave differently across ranges of input toxicity. For example, a user may act as an amplifier for input toxicity in the range [0.23 - 0.55], but acts as an attenuator for the rest [0.56 - 1]. This has not been addressed in this work and deserves to be explored.
- *Network properties:* Instead of relying on extensive experiments, can we identify properties (if such exist) that might be better predictors of which strategy (random vs. lowest-incoming, etc.) is likely to be more successful for a given network?
- *Holistic models:* Past research has taken the approach of categorising users as being hateful vs. non-hateful. Whereas this work classifies them as amplifiers, attenuators, or copycats (with possible temporal changes). Combining these two perspectives could yield six user categories, enabling a more comprehensive model and new insights.

## 7 Conclusion

We propose a model to capture the spread of toxicity on online social networks. Based on their behaviour as transformers of toxicity, users may be divided into three categories – amplifiers, attenuators, and copycats. In the case of Twitter, we found that nearly half of users remain fixed in their respective categories. Our analysis of Twitter, Gab and Koo datasets show that neither spreading activation nor epidemiological models capture the spread of toxicity effectively. Each user applies a shift to the incoming toxicity and the shift is dependent on the user's category and the value of the incoming toxicity.

Based on these observations, we propose an intervention strategy to mitigate spread of toxicity over time. Our experiments yielded a critical insight: there is no universally optimal strategy for deploying these interventions. While targeting users with the lowest in-degree proved most effective on random networks, this did not hold true for the complex, real-world structures of Twitter, Koo and Gab. On these platforms, a random placement strategy was often more, or equally, effective.

These experiments suggest that under certain conditions, it is possible to mitigate toxicity in a social network. If such conditions exist, then the moderators/providers of a social network application can use these techniques to mitigate toxicity. The insights from this work can also inform moderation policy for online platforms.

## References

- Addai, E.; Yousefi, N.; and Agarwal, N. 2024. SEIQR: an epidemiological model to contain the spread of toxicity using memory-index. In *Fifth International Workshop on Cyber Social Threats, International Conference on Web and Social Media*.
- Alorainy, W.; Burnap, P.; Liu, H.; Williams, M. L.; and Giommoni, L. 2022. Disrupting networks of hate: characterising hateful networks and removing critical nodes. *Social Network Analysis and Mining*, 12.
- Alstott, J.; Bullmore, E.; and Plenz, D. 2014. powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one*, 9(1): e85777.
- Arttime, O.; d'Andrea, V.; Gallotti, R.; Sacco, P.; and Domenico, M. D. 2020. Effectiveness of dismantling strategies on moderated vs. unmoderated online social platforms. *Scientific Reports*, 10.
- Barabási, A.-L.; and Albert, R. 1999. Emergence of scaling in random networks. *science*, 286(5439): 509–512.
- Chang, S.; Chaszczewicz, A.; Wang, E.; Josifovska, M.; Pierson, E.; and Leskovec, J. 2025. LLMs generate structurally realistic social networks but overestimate political homophily. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 341–371.
- Dagtas, S.; Agarwal, N.; and Yousefi, N. 2024. Modeling Toxicity Propagation on Reddit Using Epidemiology. In *International Conference on Complex Networks and Their Applications*, 113–124. Springer.
- Easley, D.; Kleinberg, J.; et al. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*, volume 1. Cambridge university press Cambridge.
- Elmas, T. 2025. Personal communication. July 2025.
- Erdős, P.; and Rényi, A. 1961. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1): 261–267.
- Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM*, 59(7): 96–104.
- Goel, V.; Sahnan, D.; Dutta, S.; Bandhakavi, A.; and Chakraborty, T. 2023. Hatemongers ride on echo chambers to escalate hate speech diffusion. *PNAS nexus*, 2(3): pgad041.
- Google Jigsaw. 2018. Perspective API. <https://www.perspectiveapi.com/>.
- Gupta, S.; Nagar, S.; Nanavati, A. A.; Dey, K.; Barbhuiya, F. A.; and Mukherjee, S. 2021. Consumption of Hate Speech on Twitter: A Topical Approach to Capture Networks of Hateful Users. In *ROMCIR@ ECIR*, 25–34.
- Halberstam, Y.; and Knight, B. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics*, 143: 73–88.
- Kossinets, G.; and Watts, D. J. 2009. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2): 405–450.

- Kruskal, W. H.; and Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260): 583–621.
- Lerman, K.; Feldman, D.; He, Z.; and Rao, A. 2024. Affective polarization and dynamics of information spread in online networks. *npj Complexity*, 1(1): 8.
- Maarouf, A.; Pröllochs, N.; and Feuerriegel, S. 2022. The Virality of Hate Speech on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 8: 1 – 22.
- Maleki, M.; Arani, M.; Mead, E.; Kready, J.; and Agarwal, N. 2022. Applying an Epidemiological Model to Evaluate the Propagation of Toxicity related to COVID-19 on Twitter.
- Masud, S.; Dutta, S.; Makkar, S.; Jain, C.; Goyal, V.; Das, A.; and Chakraborty, T. 2021. Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 504–515. IEEE.
- Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2018. Spread of Hate Speech in Online Social Media. *Proceedings of the 10th ACM Conference on Web Science*.
- Mathew, B.; Illendula, A.; Saha, P.; Sarkar, S.; Goyal, P.; and Mukherjee, A. 2019. Hate begets Hate. *Proceedings of the ACM on Human-Computer Interaction*, 4: 1 – 24.
- McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1): 415–444.
- Mekacher, A.; Falkenberg, M.; and Baronchelli, A. 2024. The koo dataset: An indian microblogging platform with global ambitions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1991–2002.
- Mirchandani, M. 2018. *Digital hatred, real violence: Majoritarian radicalisation and social media in India*. Observer Research Foundation.
- Mønsted, B.; Sapiezzyński, P.; Ferrara, E.; and Lehmann, S. 2017. Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PloS one*, 12(9): e0184148.
- Morgan, S. 2018. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of cyber policy*, 3(1): 39–43.
- Müller, K.; and Schwarz, C. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4): 2131–2167.
- Nagar, S.; Gupta, S.; Bahushruth, C.; Barbhuiya, F. A.; and Dey, K. 2021a. Hate speech detection on social media using graph convolutional networks. In *International Conference on Complex Networks and Their Applications*, 3–14. Springer.
- Nagar, S.; Gupta, S.; Barbhuiya, F. A.; and Dey, K. 2021b. Capturing the spread of hate on twitter using spreading activation models. In *International Conference on Complex Networks and Their Applications*, 15–27. Springer.
- Ng, L. H. X.; and Carley, K. M. 2025. What is a Social Media Bot? A Global Comparison of Bot and Human Characteristics. *arXiv preprint arXiv:2501.00855*.
- Obadimu, A.; Mead, E.; Maleki, M.; and Agarwal, N. 2020. Developing an epidemiological model to study spread of toxicity on YouTube. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 266–276. Springer.
- Ribeiro, M.; Calais, P.; Santos, Y.; Almeida, V.; and Meira Jr, W. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 12.
- Ribeiro, M. H.; Calais, P. H.; Santos, Y. A.; Almeida, V. A.; and Meira Jr, W. 2017. ” Like Sheep Among Wolves”: Characterizing Hateful Users on Twitter. *arXiv preprint arXiv:1801.00317*.
- Saha, P.; Garimella, K.; Kalyan, N. K.; Pandey, S. K.; Meher, P. M.; Mathew, B.; and Mukherjee, A. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11): e2212270120.
- Scholz, F. W.; and Stephens, M. A. 1987. K-sample Anderson–Darling tests. *Journal of the American Statistical Association*, 82(399): 918–924.
- Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1): 4787.
- Shapiro, S. S.; and Wilk, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4): 591–611.
- Uyheng, J.; and Carley, K. M. 2020. Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. *Journal of Computational Social Science*, 3: 445 – 468.
- Vaidya, A.; Nagar, S.; and Nanavati, A. A. 2024. Analysing the Spread of Toxicity on Twitter. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, 118–126.
- Watts, D. J.; and Strogatz, S. H. 1998. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684): 440–442.
- Yousefi, N. 2024. *A Comparative Study of Toxicity Propagation Using Epidemiological Models*. Ph.D. thesis, University of Arkansas at Little Rock.
- Yousefi, N.; Agarwal, N.; and Addai, E. 2024. Developing epidemiological models with differentiated infected intensity. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 58–68. Springer.

## A Appendix

### A.1 Preliminary Analysis

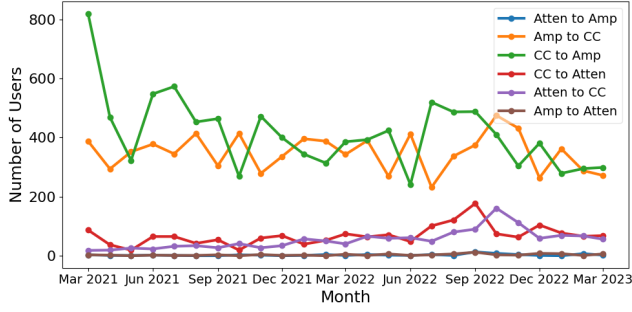


Figure 8: User Category change over time in the Koo Dataset

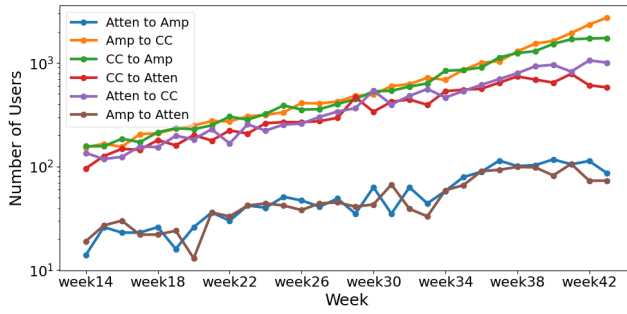


Figure 9: User Category change over time in the Twitter Dataset

| User Category     | CopyCat | Attenuator | Amplifier |
|-------------------|---------|------------|-----------|
| <b>CopyCat</b>    | –       | 0.1737     | 0.2826    |
| <b>Attenuator</b> | 0.1874  | –          | 0.0347    |
| <b>Amplifier</b>  | 0.2903  | 0.0314     | –         |

Table 8: User Category Change Probabilities