Modeling aggression propagation on social media

Chrysoula Terizi [a], Despoina Chatzakou [b], Evaggelia Pitoura [a], Panayiotis Tsaparas [a],
Nicolas Kourtellis [c],*

[a] *University of Ioannina, Greece*
[b] *Centre for Research and Technology Hellas, Greece*
[c] *Telefonica Research, Spain*

ABSTRACT

Cyberaggression has been studied in various contexts and online social platforms, and modeled on different data using state-of-the-art machine and deep learning algorithms to enable automatic detection and blocking of this behavior. Users can be influenced to act aggressively or even bully others because of elevated toxicity and aggression in their own (online) social circle. In effect, this behavior can propagate from one user and neighborhood to another, and therefore, spread in the network. Interestingly, to our knowledge, no work has modeled the network dynamics of aggressive behavior. In this paper, we take a first step towards this direction by studying propagation of aggression on social media using opinion dynamics. We propose ways to model how aggression may propagate from one user to another, depending on how each user is connected to other aggressive or regular users. Through extensive simulations on Twitter data, we study how aggressive behavior could propagate in the network. We validate our models with crawled and annotated ground truth data, reaching up to 80% AUC, and discuss the results and implications of our work.

1. Introduction

Online aggression has spiked in the last few years, with many reports of such behavior across different contexts [1,2]. Indeed, cyberaggression can potentially manifest in any type of platform, regardless of the target audience and utility or purpose envisioned for the platform. In fact, such behavior has been observed in different online social media platforms such as Twitter [3,4], Instagram [5], YouTube [6], Yahoo Finance [7], Yahoo Answers [8], 4chan [9], and across various demographics (e.g., teenagers vs. adults, men vs. women, etc. [1,10, 11]). Interestingly, it is difficult to find a generally accepted definition of cyberaggression across disciplines. As argued in [12], there are different ways to define online aggression, depending on frequency or severity of behavior, power difference of the victim with aggressor, etc.

As shown in [13], users can be influenced to act aggressively and even bully others because of elevated toxicity and aggression in their own social circle. This behavior can manifest in a similar fashion in the online world as well, and aggression can propagate from one user and neighborhood to another, thus, spreading in the network. In fact, some early studies in sociology and psychology already proposed models of computer abuse based on theories of *social learning, social bonds,* and *planned behavior* [14]. To this end, some early work by [15] has studied the pairwise interactions between users and how bullies can influence others, in an online social network. However, to our knowledge, no work has modeled the network dynamics of aggressive behavior using

formal propagation or diffusion models, and study how users' online neighbors and whole social circles affect the propagation of aggression through the network in time. This paper takes the first, but crucial, steps to investigate pending fundamental questions such as: *How can aggressive behavior propagate from one user or neighborhood in the network to another? What propagation model and parameters could best represent the aggression diffusion and its intensity?*

There are several formal models in literature to study propagation of information through a network, such as opinion dynamics [16], Linear Threshold (LT) and Independent Cascades (IC) [17,18]. Given their suitability for behavior modeling, in this paper, we focus on opinion dynamics, widely used to model how opinions propagate through a network, and leave *LT* and *IC* models for future work. We propose new methods for modeling aggression propagation based on classical opinion models, and accommodate how aggression may propagate from one user to another. Indeed, aggression can manifest in a network in various ways: one-time harassment by single (or multiple) contact(s), repetitive bullying behavior by single (or multiple) contact(s), etc. To make the problem tractable, we study aggression propagation in its minimal form of one-time aggressive behavior exhibited from one user to another. If this behavior repeats between the same users, the aggression effect is stronger and cumulative, i.e., bullying [3,4]. Thus, we opt for models that consider important factors such as how much the second user is exposed to aggressive behavior of the first user or its

neighborhood, popularity of interacting users, internal user state, etc. We validate the models' performance on real Twitter data to measure their ability to model this behavior through time. Our contributions are as follows:

- We formally present the problem of aggression propagation in a social network, and the necessary assumptions to study it in an algorithmic fashion on a network of users.
- We propose new network algorithms to model aggression propagation based on opinion dynamics methods informed by properties of aggression found in literature on psychology, sociology and computational social science.
- We implement these methods into a framework that simulates aggression propagation in a network, while controlling for various experimental factors such as: (1) social network used, (2) propagation model applied, (3) selection and ordering of users or edges affected by the propagation. This framework can be applied in different social networks, given appropriate data to bootstrap the models.
- We present extensive experimentation with the simulating framework and Twitter data, and show how model performance depends on the various factors controlled. We find that methods which consider direct interactions between users and users' internal aggression state, better model aggression and how it could evolve on Twitter. We validate the top-performing methods with a case study on Twitter data and measure up to AUC=80% on predicting which users will be aggressive or not in the future. We discuss implications of our findings for curbing cyberaggression on Twitter and other networks with similar structure.
- We open-source our simulation framework and data for reproducibility and extensibility purposes.[1]

2. Modeling aggression propagation

Aggression has been well studied in the past, in online and offline contexts, from sociologists and psychologists [10,12,19–23] and computational social or computer scientists in different forms: cyberbullying and aggression on Twitter and other media [3,4,24–26], hateful speech [27,28], and offensive or abusive language [6,14,29]. Recently, there were also efforts in the detection of such language in the streaming setting, where online models are trained to detect such behavior [30,31]. In the following, we first discuss insights extracted from literature that attempt to model aggression in different ways. Then, we formally propose the *Aggression Propagation* problem and present how the literature insights can be used to inform existing models into modeling aggression propagation.

2.1. Aggression modeling: Literature insights

1. Influence from strong social relationships. Aggressive behavior is reactionary and impulsive, and often results in breaking household rules or the law, and can even be violent and unpredictable [32]. Interestingly, aggressive acts, while reflecting the influence of various mental and physical disorders, in most instances represent learned behaviors from other individuals [33]. In fact, some earlier works proposed that online abusive behavior could be explained using sociology- and psychology-based theories such as *social learning, social bonds* and *planned behavior* [14]. Furthermore, [34] observed that a person's negative mood increases the likelihood of adopting negative behavior, which is easily transmitted from person to person. In addition, [15] identified that bullies may be able to exert influence through their pairwise interactions with others in the network, and thus spread bullying. These works lead us to the first insight on aggressive behavior:

_____
[1] github.com/cterizi/Modeling-Aggression-Propagation-on-Social-Media

due to strong social bonds, users can be influenced by, and learn from others such aggressive behavior.

2. Influence from social groups. Aggressive adolescents may be unpopular in the larger social community of peers and adults, yet they can be accepted by and closely linked to particular subgroups of peers [35]. Furthermore, as it was investigated in [36], the personal responsibility exhibited by individuals or groups can be captured by the General Aggression Model (GAM) [23]. The authors established that when individuals or a group of individuals come to believe either that they are not responsible or that they will not be held accountable by others, the stage is set for the occurrence of violent evil and aggressiveness [36]. In addition, in recent studies on Twitter by Chatzakou et al. [3,4], cyberbully and aggressive users were found to be less embedded in the network, with fewer friends and smaller clustering coefficient. Further, [37] found that exposing a person to negative or positive behaviors of those around him, leads him to experience the same emotions as them. These results lead us to the second insight: *aggressive users may be embedded in small social groups, which can have high impact on their aggression.*

3. Influence due to power difference. Studies have also looked at the emotional and behavioral state of victims of bullying and aggression and how it connects to the aggressor's or victim's network status. [12] observed that a high power difference in the network status of two individuals can be a significant property of bully-victim relationship. Also, [22] noted that the emotional state of a victim depends on the power of the victim's bullies. For example, more negative emotional experiences were observed when more popular cyberbullies conducted the attack. Also, users with higher degree centrality are not only more popular, but potentially more influential on others, since their position in the network offers them greater exposure and thus greater opportunity to have an impact on the behavior of others [15]. These observations lead us to the third insight: *the power difference that a user may have over another (e.g., due to popularity) can be a decisive factor on the exerted aggression to another user.*

4. Influence due to internal state and external input. GAM is an integrative approach to understanding aggression that incorporates the best aspects of many domain-specific theories of aggression and takes into account a wide range of factors that affect aggression. It is separated into two layers, representing the distal causes and proximate causes. The distal processes express how biological (e.g., hormone imbalances, low serotonin, and testosterone) and persistent environmental (e.g., difficult life conditions, victimization, and diffusion of responsibility) factors work together to influence a user's personality and increase the likelihood of developing an aggressive personality. The proximate processes have three stages: (i) inputs, (ii) routes, and (iii) outcomes, that can affect the person's level of aggression and possible reactions to the input. The reaction that is selected then influences the encounter, which in turn influences the person and situation factors, beginning a new cycle of influence. The findings from this important study (i.e., GAM) lead us to a fourth insight: *users can be influenced by external inputs, but they also try to consolidate them with their internal state of arousal, cognition and affect, before moving to a new state.*

5. Influence can appear in cycles. The overall process outlined by the GAM, and the previously extracted insights (1–4) can be captured in an aggression propagation model. The user (who could be an aggressor or a victim/normal user) is allowed under monitoring to: (1) have an internal aggression state of his own, (2) interact with his neighbors and close friends and receive and/or exert influence of aggression, (3) assess if he will be changing his stance on aggression, i.e. to become more (or less) aggressive after his interactions, (4) act by changing (or not) his stance, (5) repeat these steps in the next cycle (or time step). This insight allows to build on adapted opinion models that work in simulated rounds or cycles to solve the Aggression Propagation problem, as presented next.

2.2. Aggression propagation problem

Online users may be "friends" and connected in an online social network. In our setup, user i, at time t has its own aggression score that represents his internal, continuous state, S_i^t. While he interacts with his followers or followings, he may be influenced to be more or less aggressive, thus changing his internal aggression state at every time instance. The impact that others (i.e., his direct friends or neighborhood) have on his aggression state, can be a function of the strong social relationships with him (w_{ij}), his power score P_i (e.g., degree centrality), the size of the user's neighborhood (N_i), etc. The change of aggression state is continuous, i.e. at every time instance, users are influencing each other's aggression state, partially or totally. Therefore, the problem of aggression propagation is to model how aggression among users will diffuse or propagate in a network for some time window W. Obviously, this problem has clear similarities with the opinion propagation problem, and techniques to model opinion dynamics could be adapted to model how users influence each other to change aggression state.

As this is the first investigation on this problem, we opt to establish a solid baseline of solutions to the problem at hand, and propose simple, parameter-free models that are generalizable and applicable to different social networks. Since aggression can manifest in different ways and intensities, in this work, we study its propagation in a minimal form of one-time aggressive behavior exhibited from one user to another. Repetition, or accumulation of aggressive behavior (via multiple same-pair interactions) can lead to bullying [3,4]. Interestingly, this extreme behavior can also be captured in our models with repetitive increase of aggression state.

The following lists the additional notations used in text:

S_i^t	Aggression score of user i, at time t, $S_i^t \in [0,1]$
w_{ij}	Weight of edge (i,j) of users i and j, defined as Jaccard overlap of neighbor sets: $\frac{N_i \cap N_j}{N_i \cup N_j} \in [0,1]$
P_i	Power score of user i: ratio of in-degree over out-degree: $\frac{inDegree_i}{outDegree_i} \in [0,1]$
A_x	Selector for applying a factor out of the options: $1, w_{ij}, P_i, P_j, w_{ij}P_i$ and $w_{ij}P_j$, for user x

Next, we take the five insights identified and embed them in popular opinion models, to construct our proposals for modeling aggression propagation through the network. Overall, we investigate 26 propagation models: Voter [38], Deffuant [39], HK [40], DeGroot [41], FJ [42], Averaging DeGroot and FJ models, and variants of all these.

Voter & Deffuant models & variants. Firstly, we propose four pairwise models based on the Voter model, i.e. one of the simplest and well-known pairwise models in opinion dynamics. We assume that after an interaction between two users i and j, the aggression score of i changes, because he was influenced (positively or negatively). The formulation of the first set of proposed models is the following:

$$S_i^{t+1} = A_j S_j^t \tag{1}$$

The model names depend on factor A_x: Voter, Voter_W, Voter_P and Voter_WP.

These models consider the strong relationship (1st insight) between user i and j. User i's aggression score does not consider its own state but only the aggression score of the neighbor j. The four versions reflect different variations of the Voter model, where the user i assumes the aggression of his neighbor: (1) all of it (i.e., the neighbor is completely affecting the user), (2) weighted by their edge weight (i.e., the neighbor

has an influence but only depending on the strength of their relationship), (3) weighted by the Power score of the neighbor (i.e., to capture the concept of power difference that aggressors take advantage of), and (4) weighted by the combination of Power and edge weight.

Another pairwise model suited for interaction within large populations is the Deffuant model that captures confirmation bias, i.e. people's tendency to accept opinions that agree with their own. To this end, based on Voter and Deffuant models, we propose a second set of models, where user i at $t+1$ does not only take into account the aggression state of his neighbor j (1st insight), but also includes his personal state before making changes in his aggression (4th insight). Consequently, this set of models can be formalized as follows:

$$S_i^{t+1} = A_i S_i^t + A_j S_j^t \tag{2}$$

The model names depend on factor A_x: Deffuant_W, Deffuant_P and Deffuant_WP. If factor A_x is equal to 1, it is not taken into account. To maintain the limits of aggression score in a closed interval $[0,1]$, we normalize the final aggression score of user i using the maximum aggression score from all the neighbors of i at time $t+1$.

Deffuant & HK models & variants. Another set of pairwise models we propose rely on the combination of the Deffuant and HK models (in HK, opinions take values in continuous interval, and a bounded confidence limits the interaction of user i with his neighbors), as follows:

$$\text{If } |A_i S_i^t - A_j S_j^t| < d, \text{then, } S_i^{t+1} = A_i S_i^t + A_j S_j^t \tag{3}$$

The model names depend on factor A_x: HK_d_W, HK_d_P and HK_d_WP. It does not include the case of factor A_x equal to 1. This set of pairwise models uses the condition about the bounded confidence limits from the HK model (3rd insight), and updates the aggression score accordingly. The proposed model is affected by the strong relationship with its neighbor (1st insight) and internal personal state (4th insight) at the previous moment. We normalize the final aggression score using the maximum aggression score from all of i's neighbors, for those that the treaty is valid at $t+1$.

DeGroot model & variants. The next set of proposed models take into account the neighborhood of user i for deciding what aggression score to give to the user (2nd insight). The aggression of a user can be influenced by all of user's neighbors and its internal behavior (4th insight). As a result, this set of models are variants of DeGroot model (allows a user to consider all or some of their neighbors' score) which considers an average effect across all the neighborhood of the user, and are calculated as follows:

$$S_i^{t+1} = \frac{A_i S_i^t + \sum_{j \in N_i} A_j S_j^t}{A_i + \sum_{j \in N_i} A_j} \tag{4}$$

The model names depend on factor A_x: Degroot, DeGroot_W, DeGroot_P and DeGroot_WP. $A_x=1$ corresponds to the original DeGroot model.

FJ model & variants. We also propose variants of the FJ model (i.e., a variation of the DeGroot model, with the main difference lying in the fact that in the FJ model each user has an intrinsic initial opinion that remains the same, and an expressed opinion that changes over time), integrating the initial aggression state of an individual in the network (4th insight), along with the user's neighborhood (2nd insight):

$$S_i^{t+1} = \frac{A_i S_i^0 + A_i S_i^t + \sum_{j \in N_i} A_j S_j^t}{2A_i + \sum_{j \in N_i} A_j} \tag{5}$$

The model names depend on factor A_x: FJ_W, FJ_P, FJ_WP.

Averaging DeGroot & FJ models & variants. We propose the following set of models based on DeGroot and FJ, where the aggression score of each user has been inspired by the 2nd, 3rd and 4th insights. The

models are modified by taking the average power score and aggression score from all of user's neighbors, individually:

$$S_i^{t+1} = \left(\frac{A_i + \sum_{j \in N_i} A_j}{1 + \sum_{j \in N_i} 1} \right) \left(\frac{S_i^t + \sum_{j \in N_i} S_j^t}{1 + \sum_{j \in N_i} 1} \right) \quad (6)$$

The model names depend on factor A_x: avg DeGroot_W, avg DeGroot_P, and avg DeGroot_WP.

The final set of proposed models is similar to FJ models, but in these we consider the initial aggressive state of each user i (2nd–4th insights). Thus, the models are as follows:

$$S_i^{t+1} = \left(\frac{A_i + A_i + \sum_{j \in N_i} A_j}{1 + 1 + \sum_{j \in N_i} 1} \right) \left(\frac{S_i^0 + S_i^t + \sum_{j \in N_i} S_j^t}{1 + 1 + \sum_{j \in N_i} 1} \right) \quad (7)$$

The model names depend on factor A_x: avg FJ_W, avg FJ_P, and avg FJ_WP.

Next, we explain how all presented models are implemented into a simulator for exhaustive experimentation with different parameter settings using real Twitter data.

3. Simulation methodology

In this section, we outline the methodology followed to simulate propagation of aggression in a social network, given each one of the models proposed earlier (See Fig. 1 for an overview). First, some users in the network are assumed to be aggressive, and the rest as normal, formalizing the network's initial state. As time passes, users interact with, and may affect, each other to become more or less aggressive, thus changing the overall state of aggression of the network through time. Different models can be used to describe these user interactions and aggression change. To identify which model is better for this task, we compare each model's imposed aggression changes with real (ground truth) data of aggression propagation. Each model performs differently through the simulation, and may match best with the ground truth data at different point in the simulated time. Thus, at regular time intervals during each simulation we capture snapshots of the network's aggression state and compare each model with the validation data.

Next, we address in the simulator design important factors that can impact the exploration of this complex problem:

1. Online social network & ground truth data
2. Aggressive and normal users
3. Users (edges) to perform propagation
4. Ordering of users to perform propagation
5. Propagation models applied to modify users' scores
6. Metrics used to capture (change of) state of aggression
7. Metrics to compare state of aggression in simulated and validation networks

3.1. Online social network & ground truth data

For our experimental exploration, two main datasets are considered, as presented in Table 1. Firstly, a Labeled Twitter (LT) network was constructed using 401 annotated users from [3] as seeds for crawling their ego-network, to record their friends and followers. This set of data constitutes our ground truth, as explained next. We performed two crawls of these 401 users' ego networks: in 06-09/2016 and in 04-05/2019. These 401 users were originally annotated in 2016 as *normal*, *aggressive*, *bullying*, or *spammers* (Step 1, Fig. 1). We removed the spam class and merged all aggressive and bullying users under the *aggressive* label. Thus, the 401 users have two labels for 2016. Their friends and followers do not have an aggression label. However, since during the crawling they were found as active, we make the conservative assumption that they were normal (even though this is a lower bound on aggressive labeling).

When the re-crawl of these users in 2019 was completed (Step 2, Fig. 1), ∼ 650K were found to be active, ∼ 30K were found suspended, ∼ 37K had deleted their account. For ∼ 452K users the status was not

retrievable. We make the assumption that active and deleted users in 2019 can be considered *normal*, and suspended users are *aggressive*, since at some point in the past they violated Twitter rules. We justify this assumption with the following reasoning. The authors in [4] classified users as *aggressive* or *normal* based on their activity, account and network characteristics, with high confidence based on crowdsourced annotations. When they examined the account status of these 401 users after regular time periods (in 2017 and 2018), many of the *aggressive* users were suspended by Twitter, thus justifying their original annotation.[2] Therefore, the characterization of the re-crawled users as *aggressive* or *normal* based on their account status can be considered realistic. To this end, the overall **Crawled Network** LT has $1,171,218$ users (including the 401 seeds) and $1,787,543$ edges connecting them. Indeed, we note this dataset reflects only the ego-networks of 401 users, i.e. it is not a complete social network of 1.1M users. Also, 477,113 edges were dropped, since their endpoints' status was not retrievable. In both crawls of 2016 and 2019 for this network, these users are labeled as *normal* or *aggressive*, following the above process.

Secondly, an **Unlabeled Twitter** (UT) network was considered, with 81,306 users, and $1,768,149$ directed edges between them, acquired from [43]. We specifically focus on the network's Strongly Connected Component (SCC) and the final UT network has 68,413 users and $1,685,163$ edges, with properties of a proper social network, in contrast to the LT ego-networks. Also, and following past work on network analysis [44], we apply weights on the edges based on the Jaccard overlap of social circles between two users i and j, $w_{ij} = \frac{N_i \cap N_j}{N_i \cup N_j} \in [0, 1]$. As explained next, we apply a classification process to characterize these unlabeled users of this dataset as either *aggressive* or *normal*, using the 401 labeled users from LT.

3.2. Which users should be aggressive?

In the UT, we have no labels of aggression, and thus, we do not know which users exhibit aggression and which are normal. An ideal process for labeling users as aggressive or not would be to subject this dataset to annotation by use of crowdsourcing or experts, as done in past studies [3]. However, this process is expensive, slow and non-scalable [25]. It also assumes that there is access to each user's activity (i.e., tweets) to judge if they are behaving aggressively or not. Therefore, and in order to identify users who should be labeled as aggressive, we proceed as follows: First, we use the ego-network data of the 401 users from LT, to train a classifier (CL = Random Forest[3]) on users' network features,[4] and thus to infer the likelihood that a user will be aggressive (Steps 3 and 4, Fig. 1). Overall, by training and testing the CL on the LT's 401 annotated users using repeated 10 cross-fold validation (i.e., train on 90% of users and test on 10%), we achieved 93.24% Accuracy and 93.2% Precision and Recall. Second, we extract the same network properties from the UT network (Step 5, Fig. 1), and apply CL on them, in order to label UT's users as *aggressive* or not, based on some threshold (Step 6, Fig. 1). Thus, we transfer the label "knowledge" or "learning" from the past labeled network (LT) to the unlabeled one (UT), using only network features which are present in both datasets.

By applying CL on the UT network, we got 8.5% or 5,820 users who were labeled as aggressive. We verified that users selected to be aggressive (or normal) in UT had similar distributions for their network properties with the 401 annotated users in LT. Each aggressive (normal) user i was given a score $S_i = 1$ ($S_i = 0$). Interacting users can modify each other's aggression state, leading to users with scores between 0 and 1.

---

[2] Twitter suspends an account (temporarily or even permanently, in some cases) if it has been hijacked/compromised, is considered spam/fake, or if it is abusive.

[3] Number of trees generated: 10; Maximum depth set: unlimited

[4] User's friends and followers, their ratio, user's clustering coefficient score, hub score, authority score, and eigenvector score.
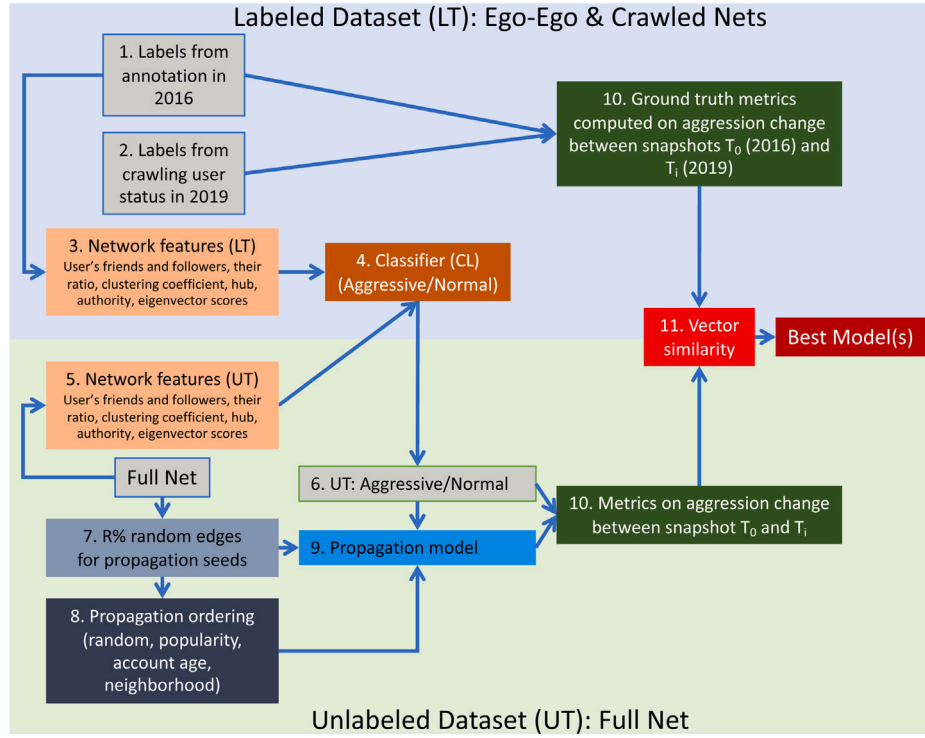
Fig. 1. Overview of the simulation methodology.

Table 1
Overview of datasets used with our framework.

	Ground truth labeled data (LT)		Unlabeled data (UT)	
	Crawled Net	Ego-Ego Net (SCC)	Full Net	SCC
# nodes	1,171,218	314 (57)	81,306	68,413
# edges	1,787,543	540 (395)	1,768,149	1,685,163
labels	Aggressive Normal	Bullying → Aggressive Aggressive Normal Normal	Aggressive Normal	

3.3. Users to perform propagation

We select a set $\mathbb{R}$ of random users for executing the propagation model (Step 7, Fig. 1). A large $\mathbb{R}$ can cover a larger portion of the network, but can be extremely costly to simulate. We opt for 10% of random edges to make the problem computationally tractable, covering 65.3% of total users of the UT network. For completeness, we also tried for limited number of experimental configurations the full network (i.e., 100% of edges) and the results were similar.

3.4. Propagation changes applied

Users in $\mathbb{R}$ that were selected for propagation may interact with each other in different ways (Step 8, Fig. 1):

1. Randomly, i.e. the selected users are randomly shuffled before their aggression is propagated.
2. Based on the most popular (or least popular) user (e.g., using their degree centrality).
3. Based on the neighborhood involved (i.e., group users based on neighborhood and propagate between them).
4. Based on their network id (a proxy of their account age, a trend confirmed in our data).

We measure how each method impacts the aggression change of each model during simulation.

3.5. Propagation models used

A propagation model can be used to simulate how information or influence or type of behavior can propagate from one user to another, and consequently from one part of the network to another. The model used may have parameters to be tuned, or can also be parameter-less. In total, we consider all models and variants proposed in Section 2 (Step 9, Fig. 1). All are parameter-less, making them simple and generalizable on different networks and setups.

3.6. Metrics used to measure aggression change

We measure the state of aggression of users and network, and how it changes through simulated time using 26 different metrics, as explained in Table 2. These elements capture the state of the network with respect to users and edges and their label at time $t = i$, and how they changed through the simulation between time t_i and t_j (declared as →). Each of these metrics can be computed at regular snapshots ($T_0, T_1, T_2, \dots, T_N$), by comparing the network state at a given snapshot (T_N) with initial state T_0.

3.7. Measuring ground truth metrics

In the previous paragraphs, we described the list of metrics we measure to quantify the changes in state of aggression in the network. We compute these metrics in the LT network using the previously

Table 2
Metrics used to measure aggression change.

Metric of state at time $t = i$	Explanation
$\{n\}_i$ || $\{a\}_i$	Portion of normal (|| aggressive) users in the network
$\{N\text{-}N\}_i$ || $\{A\text{-}A\}_i$	Portion of edges that both users i and j are normal (|| both users are aggressive)
$\{N\text{-}A\}_i$ || $\{A\text{-}N\}_i$	Portion of edges that user i is normal & j aggressive (|| user i is aggressive & j is normal)

State change between $t = i$ and $t = j$	Explanation
$\{n\}_i \rightarrow \{n$ || $a\}_j$	Portion of normal users in the initial state who remain normal or become aggressive, respectively
$\{a\}_i \rightarrow \{n$ || $a\}_j$	Portion of aggressive users in the initial state who become normal or remain aggressive, respectively
$\{N\text{-}N\}_i \rightarrow \{N\text{-}N$ || $N\text{-}A$ || $A\text{-}N$ || $A\text{-}A\}_j$	Portion of edges that users i, j were normal, and remain normal, or one / both become aggressive, respectively
$\{N\text{-}A\}_i \rightarrow \{N\text{-}N$ || $N\text{-}A$ || $A\text{-}N$ || $A\text{-}A\}_j$	Same as above for edges where j is aggressive at the initial state
$\{A\text{-}N\}_i \rightarrow \{N\text{-}N$ || $N\text{-}A$ || $A\text{-}N$ || $A\text{-}A\}_j$	Same as above for edges where i is aggressive at the initial state
$\{A\text{-}A\}_i \rightarrow \{N\text{-}N$ || $N\text{-}A$ || $A\text{-}N$ || $A\text{-}A\}_j$	Same as above for edges where both i and j are aggressive at the initial state

mentioned time crawls (2016 and 2019). Thus, we construct a ground truth or validation vector for the above mentioned 26 metrics (Table 2), which capture the change of aggression of users and type of edges (A-A, A-N, etc.) (Step 10, Fig. 1). However, we do not take into account the following metrics: n, n → n, N-N, and N-N → N-N, as they are dominant and overtake the rest of (evidently smaller) changes.

3.8. Comparing simulation & ground truth data

The above set of metrics is computed for all models and for 10 time snapshots per simulation (Step 10, Fig. 1). Using a pre-selected threshold T_A for each user's aggression score, we binarize their final state and thus, compute overall aggression change in nodes and edges. Then, we compare with the validation vector from the ground truth data (Step 11, Fig. 1). This comparison is executed using standard similarity metrics such as Cosine similarity, Pearson correlation, Spearman rank correlation, and Euclidean distance. This comparison establishes how close a model changes the state of aggression of the network (in both nodes and edges) to match the ground truth.

4. Analysis of simulation results

In this section, we show the results from the extensive (>5k) simulations performed, under different experimental settings used: 26 propagation models, 10 thresholds for $T_A = 0.05, \ldots, 0.9$, comparisons with 10 time snapshots, 4 metrics for comparing ground truth with model performance in each snapshot, 5 types of orderings of users to propagate aggression, and 10% random edges and their users. Note: We also tested the full network (i.e., 100% of edges) for a limited number of experimental configurations, and the results were similar to using 10% of random edges. Therefore, and to reduce computational overhead (each experimental setup on the full network takes 2–3 days to complete) we performed more in-depth experiments using the 10% of random edges.

4.1. Which models are stable and perform best?

Different aggression thresholds T_A: The first step towards analyzing the simulation results is to compare the proposed models with respect to their performance.

Cosine similarity is used for the comparison, with the threshold above which a user is characterized as aggressive or not, to be varying between $T_A = 0.05, \ldots, 0.9$. Fig. 2 plots the cosine similarity for all considered models in relation to the validation vector of real data, for four threshold values as examples ($T_A = 0.3, 0.5, 0.7$ and 0.9). The results for the rest of the thresholds are omitted for brevity. Overall, we observe models showing stable (low or high) performance, independent of T_A selected, or models highly depended on T_A. Specifically, *Deffuant_P*, *Voter*, and *HK_*_P* achieve high performance (similarity > 0.8), regardless of the T_A selected. On the contrary, *Voter_W*, *Voter_WP*, and *Averaging_*_** show lower performance (similarity < 0.7), irrespective of T_A selected. Alternatively, there are models whose performance highly

depends on the change of T_A. *Deffuant_W** and *HK_*_W** fluctuate from average to high performance ($0.7 <$ similarity < 0.85), and *Degroot** and *FJ_** show highly varying performance ($0.45 <$ similarity < 0.85). Based on the comparison across all thresholds, we conclude that $T_A = 0.5$ is a reasonable threshold and we use it for the rest of our experiments.

Performance within models: Fixing $T_A = 0.5$, we observe that with *Deffuant_P*, we achieve best performance. Also, *Voter*, *Deffuant_W*, and *HK_*_W* are among the top models. We note that when the edge weight (W) is considered, the performance in some cases is adversely affected. For example, the *Voter_W* model reaches similarity with ground truth less than 0.5. Mixed results are observed when the Power score (P) is used. For example, in *Deffuant* and *DeGroot* models, the performance increases, indicating that the influence of user j to its neighbor i is more important when it is not constrained to the given edge (i, j), but instead, when the neighborhood of i, N_i, is considered. *Degroot* and *FJ* models perform similarly, indicating that the state of neighborhood of the user has no significant influence to the overall performance. Finally, averaging *DeGroot* and *FJ* models perform the worst, regardless of whether the edge weight or power score are considered, separately or in combination.

Takeaways. From the top performing models, i.e., *Deffuant_P*, *Voter*, and *HK_*_P*, two main observations can be made: (i) a user's internal aggression state is highly dependent on their mate's aggression state (i.e., with whom there is a direct interaction/relationship), and (ii) the internal aggression state of a user constitutes an important factor in aggression propagation. This observation aligns well with the 4th insight in Section 2.1. In situations with various options for reaction, the inner state (in our case, the aggression state) of individuals, as well as those with whom there is a direct connection with, are key factors in the subsequent state of the individuals themselves [45]. This is also reflected by the top models, as, on the one hand, they are all pairwise models, while on the other, apart from *Voter*, they consider a user's internal aggression state before making aggression changes. Overall, based on the best models, we observe that online aggression (especially when taking place on Twitter) is propagating from one user to another; users are not so influenced by their neighborhood. Aggressive users have been shown to be less popular (i.e., smaller number of followers and friends) than normal users [4] which could explain the fact that aggressive users are more affected by direct relationships rather than their neighborhood aggression state.

4.2. How do models perform over time?

Next, we examine how model performance is affected when different time snapshots of the network's aggression state are considered. This analysis is done to: (i) compare with ground truth each model's state at progressing simulation times, and (ii) detect which of the snapshots was better fitting the real data. We remind the reader that there is no point in simulating propagation until the models converge to some steady state, since the time taken for this may not match the
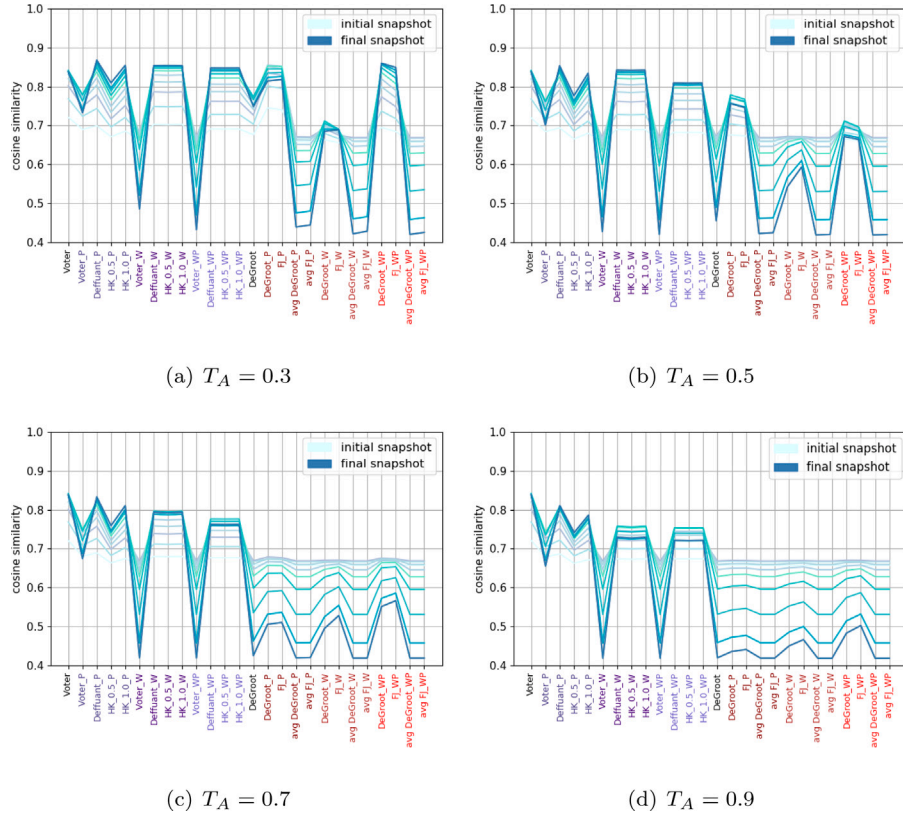
Fig. 2. Cosine similarity of all proposed models with the validation vector, for 10% of selected edges and different thresholds T_A. The different time snapshots of the simulation are colored from light to dark blue. We group the models in sets based on their commonality. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
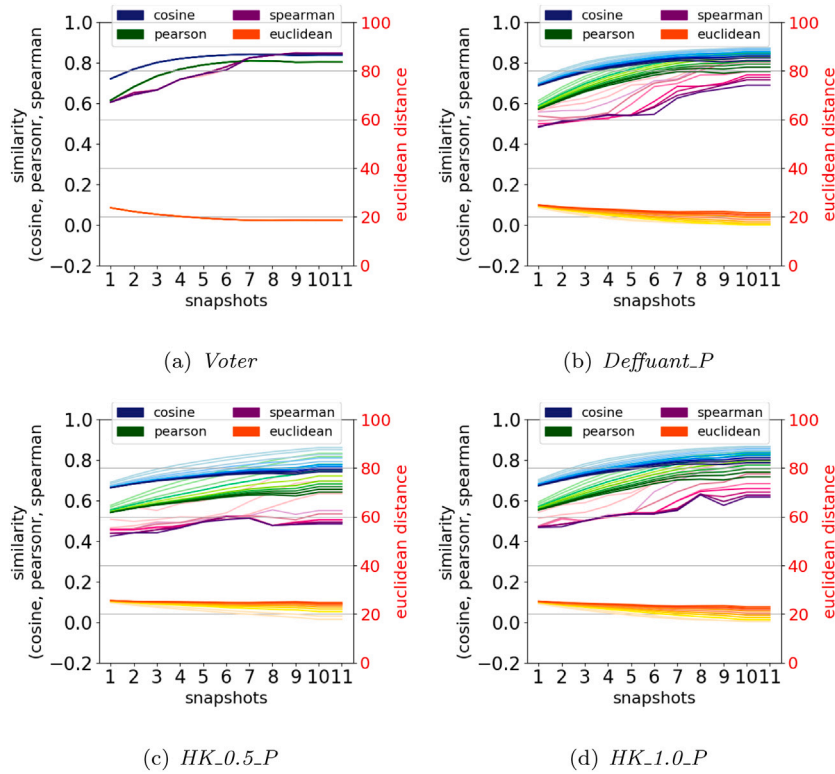
Fig. 3. Similarity of the top 4 performing models with ground truth, through 10 time snapshots: (a) *Voter*, (b) *Deffuant_P*, (c) *HK_0.5_P*, and (d) *HK_1.0_P*. We show 4 similarity metrics: cosine similarity, Pearson correlation, spearman correlation, and euclidean distance. Variation of colors for a metric illustrate the performance of the given model and metric for different thresholds $T_A = 0.05, \ldots, 0.90$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

timing the ground truth data were captured. For our analysis, we focus on the top four performing (regardless of T_A) models, i.e., *Deffuant_P*, *Voter*, *HK_0.5_P*, and *HK_1.0_P*.

Fig. 3 shows that for all models, performance is lower within the first snapshots and gradually increases, to stabilize in the last snapshots. As for the similarity metrics, they follow a similar pattern across models, indicating that either of them can be used to do the performance analysis, and that the comparison results are stable against analysis that considers ranks or absolute numbers in the vectors compared.

Takeaways. These models successfully capture how a network's aggression status changes across time. The notion of snapshots constitutes a valid process in representing the aggression propagation, since it captures the way aggression is (or will be) expected to be in real time in a network. By focusing on user interactions (*Voter* model), or user's internal aggression state (*Deffuant_P*, *HK_0.5_P*, and *HK_1.0_P*), the aforementioned models can be used to track how aggression propagates in networks with similar properties.

4.3. Is the order of changes important?

Fig. 4 shows the aggression evolution of three sets of users (i.e., normal, aggressive, all users) in relation to how they could interact (i.e., randomly, based on popularity (most/least), involved neighborhood, and network id), for the top four models. Specifically, the top and bottom parts of the plots depict how the aggression changes for the aggressive and normal users, respectively, while the middle includes all users. If aggressive users were to interact randomly, it would lead to faster decline of the aggression compared to the rest of the ordering methods. In contrast, aggressive users show greater resistance in reducing their aggression if they were to interact based on users' popularity (from highest to lower). For instance, authors in [12] observed that more powerful (i.e., popular) users have a stronger impact on others' feelings. Interestingly, assuming the least popular users were to interact (act on their aggression) first, it would lead to slower propagation. If aggression propagated from one neighborhood to the next, it would also lead to a slow rate of propagation, at times approaching the least popular user ordering. For normal users, aggression status could not be significantly affected by the way they interact; the difference between initial and final aggression scores is subtle. As we see in Fig. 4, the difference between the initial and final aggression scores of normal users is subtle, indicating that normal users are particularly resistant in the aggression that may exist within their network. Instead, the effect of normal users is greater on aggressors than the inverse. For all considered models (apart from *Voter*) after a number of interactions (i.e., $175K$), the aggression score converges, indicating that within a network, even with faster or slower changes, the network state stabilizes.

Takeaways. Overall, the way users could interact and "exchange" or propagate aggression impacts the overall network state. Popularity of users can be a great predictor of how aggression will move in the network. This could be attributed to the fact that more popular users can have stronger impact within a network if they are aggressive (or not) due to their high degree centrality, as they can affect many users at the same time, leading to high rate of aggression propagation. This is also aligned with phenomena already evident in the wild (3rd insight: *Influence due to power difference*, Section 2.1). At the same time, normal users are more resistant to aggression, due to their expected higher power status and larger neighborhoods with non-aggressive state.

4.4. What is users' final aggression state?

Fig. 5 shows how users' aggression has settled at the end of the simulation (on 10th snapshot) across the top four performing models; we consider the more realistic random ordering of changes. Fig. 5(a) shows that regardless of model, at least 60% of normal users in the end

Table 3
Model performance at predicting which users will remain (or change into) normal or aggressive. (Showing only the results for AUC ≥ 0.70, on the *Ego-Ego Net* SCC *LT*).

Model	Order	T_A	#edges	AUC	Precision	Recall
Deffuant_P	sorted	0.10	395	0.71	0.59	0.71
Deffuant_P	sorted	0.20	395	0.74	0.60	0.74
HK_1.0_P	sorted	0.10	395	0.72	0.60	0.72
HK_1.0_P	sorted	0.20	395	0.75	0.60	0.75
HK_1.0_P	random	[0.05, 0.30]	50	0.76	0.76	0.76
Deffuant_P	random	[0.05, 0.30]	50	0.79	0.70	0.79

remain unaffected (i.e., their aggression score is zero), while at most 40% of users gain some aggression (at different levels, varying from 0 to 1). For instance, based on *Deffuant_P*, almost 10% of users end up with maximum aggression score, with the rest of users varying between the lower and upper limits. In *Voter*, because of its formulation, the aggression score is either 0 or 1, with about 20% of normal users affected. Finally, from Fig. 5(b), about 20%–30% of aggressive users maintain their high aggression score (depending on the model), with only about 30% of aggressive users turning normal (*Deffuant_P* and *HK_*_P*). Contrary to normal users, when *Voter* is used, a high portion of $\sim 80\%$ aggressive users is positively affected by turning into normal. Overall, in consistency with Fig. 4, normal users are more resistant to adopting aggression, as opposed to aggressive users.

4.5. Validation of top models in prediction

We conclude this investigation by evaluating the performance of the top four models in predicting the aggression status of users, by testing the models on the ground truth dataset *Crawled Net LT* (Table 1). However, this network is comprised of the 401 users' ego-networks, i.e., their friends and followers, but not their 2nd-hop and beyond (we could not collect this information due to Twitter limitations). To perform fair application of the models, we filtered this graph and kept a subgraph that includes all ego users (out of the 401) that have edges connecting them. This way, all users participating in the experiment have their ego-graphs crawled (i.e., i and j in the models). As shown in Table 1, this filtering resulted in a subgraph of 314 users and 540 edges (**Ego-Ego Net** *LT*). Then, we kept the graph's SCC and concluded with 57 nodes and 395 edges.

Due to its small size, we select all the edges of this graph for the interaction (and not only a random 10% of them). This simulation tests all the different types of users' ordering (Section 3.4) and the four best models: *Voter*, *Deffuant_P*, *HK_0.50_P*, and *HK_1.0_P*. Different aggression thresholds $T_A = 0.05, \ldots, 0.9$ are used to characterize a user as aggressive if their final state is greater than the threshold, and normal otherwise. The goal of this experiment is to let the models simulate propagation of aggression, thus altering users' state. Then, we measure their performance in predicting correctly the final state of users (aggressive or normal, i.e. true positives or negatives, respectively), as found in **Ego-Ego Net** *LT*, using standard ML metrics such as Precision, Recall and AUC (using different operating T_As).

From Table 3, we observe that the best performance arises when the interaction of users happens based on the increasing ranking of their user IDs, indicating that account age (which is proxied by the user ID) may have important influence in the propagation process. We notice that *Deffuant_P* and *HK_1.0_P* achieve good performance when low T_A is used, such as 0.10–0.20, and the AUC score ranges from 0.71 to 0.75. In addition, we experimented with executing the models for a portion of the edges available. Using random users' ordering, the *HK_1.0_P* model achieves AUC score, Precision and Recall equal to 0.76, for $T_A = [0.05, 0.30]$, after 50 interactions. Also, satisfactory validity was achieved for *Deffuant_P* method, when 50, 100, 200 and 250 interactions take place, in random order. We observe that after 50
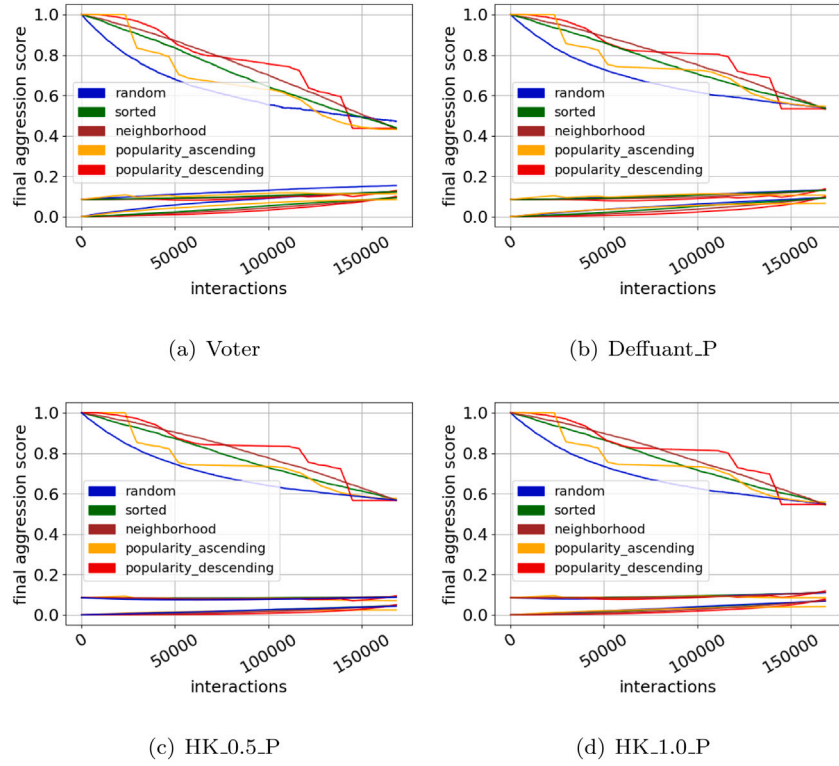
(a) Voter   (b) Deffuant_P

(c) HK_0.5_P   (d) HK_1.0_P

Fig. 4. Final average aggression score for aggressive (top part), normal (bottom part), and all (middle part) users, based on 5 different types of users' ordering, through the simulation time, for the top four models.
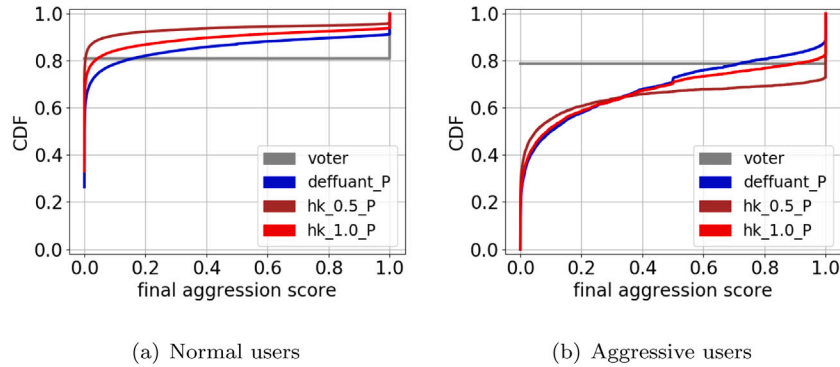

(a) Normal users   (b) Aggressive users

Fig. 5. CDFs of the final aggression scores of (a) normal and (b) aggressive users who participated in the aggression propagation.

interactions, and while T_A ranges from 0.05 to 0.30, the AUC score, Precision and Recall are equal to 0.79, 0.70 and 0.79, respectively.

Takeaways. Overall, with these validation experiments, we show that our top performing models are capable of emulating reliably the propagation of aggression, as validated on this small, annotated Twitter dataset. In addition to the previous results, we find that older accounts propagating first their aggression to newer accounts matches better the ground truth data at hand. This hints that account age may have important influence in the aggression propagation process and renders some extended future work.

5. Conclusion & future work

Despite the consequences that aggressive behavior has on individuals (e.g., embarrassment, depression, isolation from other community members, etc.), there are still important cases of aggression that stay under the radar of social networks, e.g., as in [46]. In fact, how such behavior propagates within networks and which network or other

factors affect this propagation, have not been extensively studied. To address this gap, here, we are the first to propose a pipeline to evaluate various aggression dynamics models, and to conclude in those best emulating aggression propagation in social networks. To simulate how such behavior could spread in the network, we built on top of popular opinion dynamics models, and test and validate our models' performance on real Twitter data. We found that our proposed models based on the *Deffuant* and *Hegselmann & Krause* opinion models, perform best in modeling aggression propagation in a network such as Twitter, regardless of parameters or thresholds used. Important insights embedded in these models are: (1) online aggression tends to propagate from one user to another, (2) power score of a user (e.g., degree centrality) and (3) users' internal aggression state, both constitute top factors to be considered in aggression propagation modeling, (4) influence by users' neighborhood is of less importance than other factors.

Overall, we believe this work makes a significant first step towards understanding and modeling the dynamics of aggression. The outcomes of our work highlight the suitability of the top performing models in simulating propagation of aggression in a network such as Twitter, and

how a campaign to monitor and even stop aggression on Twitter could work. That is, if aggressive users are monitored in their interactions with others (e.g., posting of aggressive messages), and simultaneously, normal users are shielded from this aggression by dropping such communication, the overall aggression in the network will significantly drop. In fact, if the campaign targets highly popular aggressive users, who are encouraged to reduce their aggression via educational tutorials and other interventions, the overall aggression in the network can drop faster than selecting users with different criteria (e.g., random). Also, we found evidence that the age of an account may be influencing the way aggression propagates. However, more in-depth investigation on ground truth data with aggression propagation information is needed to reveal clearer patterns of this aspect.

As first proposed by [15], and moving towards a better, more global or unified model of aggression propagation, in the future, we can consider aspects of this behavior in relation to social media activity, such as the actual content of a post, the time it was posted, which user posted it, which users commented on the post and how long after the original post was done, the reaction of the neighborhood of the original poster and commenters, etc. Moreover, Linear Threshold and Independent Cascades are two alternative sets of models that could be examined and compared with the opinion dynamics based models for modeling aggression propagation. Early work on these models has recently showed promising results [47], but more in-depth analysis of such models and on different datasets is needed. Another interesting extension of this work would be to attempt aggression propagation modeling on a dynamic network. Evolving networks, where links are added or removed through time, are the most realistic but notoriously difficult to model, and render significant future attention. Finally, it would be worthwhile to investigate the effectiveness of the proposed models to predict aggression on other online social network platforms such as Facebook, Instagram, etc.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

References

[1] Cyberbullying research center, 2019, https://cyberbullying.org/facts.

[2] L.S.W. Signe Whitson, Confronting passive aggressive behavior on social media, 2018, http://bit.ly/2QP1yi9.

[3] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, Mean birds: Detecting aggression and bullying on Twitter, in: WebSci, ACM, New York, NY, USA, 2017, pp. 13–22.

[4] D. Chatzakou, I. Leontiadis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, N. Kourtellis, Detecting cyberbullying and cyberaggression in social media, Trans. Web (2019).

[5] H. Hosseinmardi, S.A. Mattson, R.I. Rafiq, R. Han, Q. Lv, S. Mishra, Analyzing labeled cyberbullying incidents on the instagram social network, in: SocInfo, 2015.

[6] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: PASSAT and SocialCom, 2012.

[7] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate speech detection with comment embeddings, in: WWW, 2015.

[8] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, F. Bonchi, The social world of content abusers in community question answering, in: WWW, 2015.

[9] G.E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, J. Blackburn, Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web, in: ICWSM, 2017.

[10] P. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippett, Cyberbullying: Its nature and impact in secondary school pupils, in: Child Psychology and Psychiatry, 2008.

[11] A. Campbell, S. Muncer, Models of anger and aggression in the social talk of women and men, J. Theory Soc. Behav. 17 (4) (1987).

[12] L. Corcoran, C. Guckin, G. Prentice, Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression, Societies 5 (2) (2015) 245–255.

[13] A.K. Henneberger, D.L. Coffman, S.D. Gest, The effect of having aggressive friends on aggressive behavior in childhood: Using propensity scores to strengthen causal inference, Social Development 26 (2) (2017) 295–309, http://dx.doi.org/10.1111/sode.12186, URL https://www.ncbi.nlm.nih.gov/pubmed/28553013.

[14] J. Lee, Y. Lee, A holistic model of computer abuse within organizations, Inf. Manag. Comput. Secur. 10 (2) (2002) 57–63.

[15] A. Squicciarini, S. Rajtmajer, Y. Liu, C. Griffin, Identification and characterization of cyberbullying dynamics in an online social network, in: IEEE/ACM ASONAM, 2015.

[16] A. Sîrbu, V. Loreto, V.D.P. Servedio, F. Tria, Opinion dynamics: Models, extensions and external effects, in: Participatory Sensing, Opinions and Collective Awareness, Springer International Publishing, Cham, 2017, pp. 363–401, http://dx.doi.org/10.1007/978-3-319-25658-0_17.

[17] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: ACM SIGKDD, 2003, pp. 137–146.

[18] A. Srivastava, C. Chelmis, V.K. Prasanna, Social influence computation and maximization in signed networks with competing cascades, in: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 41–48.

[19] J. Farver, Aggressive behavior in preschoolers' social networks: Do birds of a feather flock together? Early Child. Res. Q. 11 (3) (1996) 333–350, http://dx.doi.org/10.1016/S0885-2006(96)90011-3, URL http://www.sciencedirect.com/science/article/pii/S0885200696900113.

[20] A. Bandura, D. Ross, S.A. Ross, Transmission of aggression through imitation of aggressive models, J. Abnorm. Soc. Psychol. 63 (3) (1961) 575–582.

[21] H. Xie, R.B. Cairns, B.D. Cairns, Social networks and configurations in inner-city schools: Aggression, popularity, and implications for students with EBD, J. Emot. Behav. Disord. 7 (3) (1999) 147–155, http://dx.doi.org/10.1177/106342669900700303, arXiv:https://doi.org/10.1177/106342669900700303.

[22] S. Pieschl, T. Porsch, T. Kahl, R. Klockenbusch, Relevant dimensions of cyberbullying - Results from two experimental studies, J. Appl. Dev. Psychol. 34 (5) (2013).

[23] J.J. Allen, C.A. Anderson, B.J. Bushman, The general aggression model, Curr. Opin. Psychol. 19 (2018) 75–80, http://dx.doi.org/10.1016/j.copsyc.2017.03.034, Aggression and violence, URL http://www.sciencedirect.com/science/article/pii/S2352250X17300830.

[24] K. Dinakar, R. Reichart, H. Lieberman, Modeling the detection of textual cyberbullying, Soc. Mob. Web 11 (02) (2011).

[25] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of Twitter abusive behavior, in: ICWSM, 2018.

[26] Hariani, I. Riadi, Detection of cyberbullying on social media using data mining techniques, Int. J. Comput. Sci. Inf. Secur. 15 (3) (2017) 244.

[27] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93.

[28] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: ICWSM, 2017.

[29] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: 25th ACM WWW Companion, 2016.

[30] H. Herodotou, D. Chatzakou, N. Kourtellis, Catching them red-handed: Real-time aggression detection on social media, in: 2021 IEEE International Conference on Data Engineering, ICDE, 2021.

[31] H. Herodotou, D. Chatzakou, N. Kourtellis, A streaming machine learning framework for online aggression detection on Twitter, in: 2020 IEEE International Conference on Big Data, Big Data, 2020, pp. 5056–5067, http://dx.doi.org/10.1109/BigData50022.2020.9377980.

[32] A. E. Gabbey, T. Jewell, Aggressive behavior, 2016, http://bit.ly/2QNuFSR.

[33] W.I. Gardner, C.W. Moffatt, Aggressive behaviour: Definition, assessment, treatment, Int. Rev. Psychiatry 2 (1) (1990) 91–100, http://dx.doi.org/10.3109/09540269009028275, arXiv:https://doi.org/10.3109/09540269009028275.

[34] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, J. Leskovec, Anyone can become a troll, in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, ACM Press, 2017, http://dx.doi.org/10.1145/2998181.2998213.

[35] R.B. Cairns, B.D. Cairns, H.J. Neckerman, S.D. Gest, J.-L. Gariepy, Social networks and aggressive behavior: Peer support or peer rejection? Dev. Psychol. 24 (6) (1988) 815–823, URL https://www.researchgate.net/profile/Jean_Louis_Gariepy/publication/232534518_Social_Networks_and_Aggressive_Behavior_Peer_Support_or_Peer_Rejection/links/544ec0520cf2bca5ce90bd04/Social-Networks-and-Aggressive-Behavior-Peer-Support-or-Peer-Rejection.pdf.

[36] C.A. Anderson, N.L. Carnagey, Violent evil and the general aggression model, Soc. Psychol. Good Evil (2004) 168–192.

[37] A.D.I. Kramer, J.E. Guillory, J.T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, Proc. Natl. Acad. Sci. 111 (24) (2014) 8788–8790, http://dx.doi.org/10.1073/pnas.1320040111, arXiv:https://www.pnas.org/content/111/24/8788.full.pdf, URL https://www.pnas.org/content/111/24/8788.

[38] P. Clifford, A. Sudbury, A model for spatial conflict, Biometrika 60 (3) (1973) 581–588, http://dx.doi.org/10.1093/biomet/60.3.581, arXiv:http://oup.prod.sis.lan/biomet/article-pdf/60/3/581/576759/60-3-581.pdf.

[39] G. Deffuant, D. Neau, F. Amblard, G. Weisbuch, Mixing beliefs among interacting agents, Adv. Complex Syst. 03 (01n04) (2000) 87–98, http://dx.doi.org/10.1142/S0219525900000078.

[40] R. Hegselmann, U. Krause, Opinion dynamics and bounded confidence models, analysis and simulation, J. Artif. Soc. Soc. Simul. 5 (2002).

[41] M.H. Degroot, Reaching a consensus, J. Amer. Statist. Assoc. 69 (345) (1974) 118–121, http://dx.doi.org/10.1080/01621459.1974.10480137.

[42] N.E. Friedkin, E.C. Johnsen, Social influence and opinions, J. Math. Sociol. 15 (3–4) (1990) 193–206, http://dx.doi.org/10.1080/0022250X.1990.9990069, arXiv:https://doi.org/10.1080/0022250X.1990.9990069.

[43] J. McAuley, J. Leskovec, Learning to discover social circles in ego networks, in: NIPS - Volume 1, Curran Associates Inc., USA, 2012, pp. 539–547, URL http://dl.acm.org/citation.cfm?id=2999134.2999195.

[44] X. Zuo, J. Blackburn, N. Kourtellis, J. Skvoretz, A. Iamnitchi, The power of indirect ties, Comput. Commun. 73 (2016) 188–199.

[45] E. Hatfield, J.T. Cacioppo, R.L. Rapson, Emotional contagion, Curr. Dir. Psychol. Sci. 2 (3) (1993) 96–100.

[46] D. O'Sullivan, Bomb suspect threatened people on Twitter, and Twitter didn't act, 2018, https://cnn.it/2t8F1na.

[47] M. Poiitis, A. Vakali, N. Kourtellis, On the aggression diffusion modeling and minimization in online social networks, ACM Trans. Web (2021) arXiv:https://arxiv.org/pdf/2005.10646.pdf.