Angry Birds Flock Together: Aggression Propagation on Social Media

Chrysoula Terizi[1], Despoina Chatzakou[2], Evaggelia Pitoura[1], Panayiotis Tsaparas[1], Nicolas Kourtellis[3]
[1]University of Ioannina, Greece, [2]Centre for Research and Technology Hellas, Greece, [3]Telefonica Research, Spain
chterizi@cs.uoi.gr, dchatzakou@iti.gr, pitoura@cs.uoi.gr, tsap@cs.uoi.gr, nicolas.kourtellis@telefonica.com

Abstract

Cyberaggression has been found 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. We look into various opinion dynamics models widely used to model how opinions propagate through a network. We propose ways to enhance these classical models to accommodate 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, and validate our models with ground truth from crawled data and crowdsourced annotations. We 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 (Cyberbullying Research Center 2019; L.S.W. 2018). 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 (Chatzakou et al. 2017; Chatzakou et al. 2019), Instagram (Hosseinmardi et al. 2015), YouTube (Chen et al. 2012), Yahoo Finance (Djuric et al. 2015), Yahoo Answers (Kayes et al. 2015), 4chan (Hine et al. 2017), and across various demographics (e.g., teenagers vs. adults, men vs. women, etc., (Smith et al. 2008; Cyberbullying Research Center 2019; Campbell 1987)). Interestingly, it is difficult to find a generally accepted definition of cyberaggression across disciplines. As argued in Corcoran, Guckin, and Prentice (2015), there are different ways to define online aggression, depending on frequency or severity of the behavior, power difference of the victim with aggressor, etc.

As it has been found in Henneberger, Coffman, and Gest (2017), 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, and thus, spread in the network. In fact, some early works in sociology and psychology already proposed models of computer abuse based on the theories of *social learning*, *social bonds*, and *planned behavior* (Lee and Lee 2002).

However, to our knowledge, no work has modeled the network dynamics of aggressive behavior, and study how online users' connections and interactions may affect the propagation of aggression through an online social network. 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 model and parameters could best represent the aggression propagation and its intensity?* In particular, it studies classical opinion dynamics models widely used to model how opinions propagate through a network, and proposes ways to alter or enhance them to accommodate how aggression may propagate from one user to another. We opt for simple 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 the users, etc. We validate the models' performance on real Twitter data to measure their ability to model this behavior.

The contributions of this work are the following:

- 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 network algorithms to model aggression propagation, based on opinion dynamics methods, and 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 discuss implications of our findings for curbing cyberaggression on Twitter and other networks with similar structure.

2 Modeling Opinion Propagation

In social networks, users may interact with others in their immediate ego-network on a given topic, and consequently may adapt their opinion, or even adopt their friends' personal opinion altogether. Opinion formation is a complex process and many researchers have studied it under different settings. Thus, several mathematical models have been proposed to simulate the propagation of opinion in a network (for a review see Sîrbu et al. (2017)). In this section, we first cover background concepts around opinion propagation, and then basic methods proposed to model opinion dynamics.

2.1 General Problem of Opinion Modeling

Background. Numerous studies have been conducted around the opinion spreading (Sobkowicz 2009). Most published opinion models simulate how an individual's opinion could evolve from the influence he could receive from his immediately environment. In (Newman and Sheth 1985), a simple stochastic model (Voter model) was presented, where an individual is absolutely vulnerable to his neighbor's opinion, which he assimilates if he interacts with. In this class of binary-state dynamics models belongs the Sznajd model (Sznajd-Weron and Sznajd 2000) which states that "if two people share the same opinion, their neighbors will start to agree with them (social validation), and if a block of adjacent persons disagree, their neighbors start to argue with them (discord destroys)". In Deffuant et al. (2000), a model of opinion dynamics showed that agents adjust continuous opinions on the occasion of random binary encounters whenever their difference in opinion is below a given threshold. A classical model of consensus formation was presented in Hegselmann and Krause (2002), the variant of this model due to Friedkin and Johnsen (1990), a time dependent version and a nonlinear version with bounded confidence of the agents. There are also models that combine and compare the aforementioned basic models (Behera and Schweitzer 2003; Bernardes, Stauffer, and Kertsz 2002; Fortunato 2005; Lorenz 2007). Also, a few studies have been published that verify the opinion prediction in real setting (Sobkowicz 2009). For example, real data were used by Behera and Schweitzer (2003), psychological data in Deffuant, Amblard, and Weisbuch (2004), and information from Italian and German elections in Bernardes, Stauffer, and Kertsz; Caruso and Castorina; Fortunato and Castellano (2002; 2005; 2007).

Problem Definition. The opinion propagation problem can be formally presented as follows (Ising 1925; Glauber 1963; Sznajd-Weron and Sznajd 2000). There is a population of M individuals connected in a network over weighted edges. This weight can signify the intensity of interaction or closeness between two individuals. Each user, at time t has its initial interior opinion o_i on a topic. After an interaction between two users i and j who are neighboring in the network, each of the individuals is led to a new state on the topic.

In general, the interaction between users during the opinion propagation can be assumed to happen in a pairwise or group fashion. In effect, one user may be influenced by another user, or multiple users in his neighborhood. In pairwise models, there are two connected individuals, i and j, and each one can have their personal opinion on a topic. User i's opinion can only be affected by the influence of user j's opinion that he is connected. In group fashion models, there is user i and his neighborhood N_i of users with opinions which can influence i's opinion. Next, we characterize each model based on this distinction and its fundamentals.

Overall, this kind of procedure can be considered as a mechanism of making a decision in a closed community. The general problem of opinion modeling presents a plurality of models that differentiate the final state of the individual and the manner in which it is formed. For easiness, we use the following general notations in the text:

o_i^t	The opinion value for user i at moment t
w_{ij}	The weight of edge (i, j), between users i and j
N_i	The set of friends (neighborhood) of user i

2.2 Classic Opinion Propagation Models

Voter Model. One of the simplest and well-known pairwise models in opinion dynamics is the Voter model (Clifford 1973). In this model, there are only two discrete opinion types: $\{0, +1\}$. At each time step, an edge (i, j) is selected from the network, and user i adopts the opinion that his neighbor j had in the previous time step:

$$o_i^{t+1} = o_j^t \tag{1}$$

For undirected networks, the model reaches consensus to one of the possible initial opinions. For directed networks, the model was modified (Zschaler et al. 2012) to one that fixes the users' out-degree that induce early fragmentation.

Deffuant Model. Another pairwise model suited for interaction within large populations is the Deffuant model (Deffuant et al. 2000) that captures confirmation bias, i.e., people's tendency to accept opinions that agree with their own. In this model, users adjust continuous opinions from their initial binary opinion whenever their difference in opinion is below a given threshold. At each time step, an edge between users (i, j) is selected and user i takes into account the opinion of its neighbor j when the absolute value of their opinions' difference is less than a specific selected value d:

$$\text{If } |o_i^t - o_j^t| < d, \text{then} \tag{2}$$

$o_i^{t+1} = o_i^t + \mu \left(o_j^t - o_i^t\right)$ and $o_j^{t+1} = o_j^t + \mu \left(o_i^t - o_j^t\right)$

where, μ is the convergence parameter, with $\mu \in [0, 0.5]$. High threshold values lead to convergence of opinions whereas low values lead in several opinion clusters.

DeGroot Model. Another model in literature that allows a user to consider all or some of their neighbors' score, is the DeGroot model (Degroot 1974). Here, there is an undirected network and at time t, all users change their opinion by taking the average of their own opinion and the opinion of their neighbors. After a number of iterations (i.e., opinion changes or time steps), the network will reach consensus and each user in the network will have the same opinion.

$$o_i^{t+1} = \frac{w_{ii}o_i^t + \sum_{j \in N_i} w_{ij}o_j^t}{w_{ii} + \sum_{j \in N_i} w_{ij}} \tag{3}$$

FJ Model. A variation of the DeGroot model was proposed by Friedkin and Johnsen (Friedkin and Johnsen 1990). The main difference between the two models is that in the FJ model each user has an intrinsic initial opinion that remains the same, and an expressed opinion that changes over time. User i's new opinion is estimated as:

$$o_i^{t+1} = \frac{w_{ii}o_i^0 + \sum_{j \in N_i} w_{ij}o_j^t}{w_{ii} + \sum_{j \in N_i} w_{ij}} \tag{4}$$

The network consensus is not reached every time, but only in specific cases. Also, the calculation of the opinion's convergence can be modeled as a random walk in the graph, and if an absorbent node is attached in a node, it maintains the node's opinion stable (Gionis, Terzi, and Tsaparas 2013).

HK Model. In Hegselmann and Krause (2002) model, opinions take values in a continuous interval, where a bounded confidence limits the interaction of user i holding opinion o_i to neighbors with opinions in [o_i-ε, o_i+ε], where $\varepsilon \in [0, 1]$ is the uncertainty. Also, users interact with all of his friends:

$$o_i^{t+1} = \frac{w_{ii}o_i^t + \sum_{j \in N_i : |o_i^t - o_j^t| < \varepsilon} w_{ij}o_j^t}{w_{ii} + \sum_{j \in N_i : |o_i^t - o_j^t| < \varepsilon} w_{ij}} \tag{5}$$

The model has been proven to converge in polynomial time and leads to consensus when $\varepsilon > \varepsilon_c$ and each user or group of users are polarized when $\varepsilon < \varepsilon_c$.

3 Modeling Aggression Propagation

Previously, we outlined some of the most classical models for opinion propagation in a network. Here, we build on these methods to model how aggression could propagate in the network, as an opinion would. Next, we discuss insights extracted from literature that attempt to model aggression in different ways. Building on this background, we formally propose the *Aggression Propagation* problem, in a way that can be aligned with opinion propagation. Finally, we present how the literature insights can be used to inform existing opinion models into modeling aggression propagation.

3.1 Aggression Modeling: Literature Insights

Aggression has been well studied in the past, in online and offline contexts, from sociologists and psychologists (Farver 1996; Bandura, Ross, and Ross 1961; Xie, Cairns, and Cairns 1999; Smith et al. 2008; Pieschl et al. 2013; Allen, Anderson, and Bushman 2018; Corcoran, Guckin, and Prentice 2015) and computational social or computer scientists Chatzakou et al. (2017; 2019) and (Davidson et al. 2017; Founta et al. 2018; Lee and Lee 2002; Waseem and Hovy 2016; Hariani and Riadi 2017; Dinakar, Reichart, and Lieberman 2011; Chen et al. 2012; Nobata et al. 2016).

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 (E. Gabbey and Jewell 2016). Interestingly, aggressive acts, while reflecting the influence of various mental and physical disorders, in most instances represent learned behaviors from other individuals (Gardner and Moffatt 1990). 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* (Lee and Lee 2002). Furthermore, Cheng et al. (2017) observed that a person's negative mood increases the likelihood of adopting negative behavior, which is easily transmitted from person to person. These works lead us to the first insight on aggressive behavior: *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 (Cairns et al. 1988). Furthermore, as it was investigated by Anderson and Carnagey (2004), the personal responsibility exhibited by individuals or groups can be captured by the General Aggression Model (GAM) (Allen, Anderson, and Bushman 2018). 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 (Anderson and Carnagey 2004). In addition, in recent studies on Twitter by Chatzakou et al. (2017; 2019), cyberbully and aggressive users were found to be less embedded in the network, with fewer friends and smaller clustering coefficient. Further, Kramer, Guillory, and Hancock (2014) 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. In (Corcoran, Guckin, and Prentice 2015) it was observed that a high power difference in network status of the two individuals can be a significant property of bully-victim relationship. The authors in (Pieschl et al. 2013) 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. 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.*

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 lead us to a forth 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 also 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 existing, but adapted opinion models that work in simulated rounds or cycles, to solve the Aggression Propagation problem, presented next.

3.2 Aggression Propagation Problem

Online users may be "friends" and connected in an online social network. In this 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 presented earlier 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.

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 the mentioned opinion models, to construct our proposals for modeling aggression propagation through the network.

Voter & Deffuant models & variants. Firstly, we propose four pairwise models based on the Voter model. 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:

$$S_i^{t+1} = A_j S_j^t \tag{6}$$

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

These models take into account 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.

Based on Voter and Deffuant models, we propose a 2nd set of models, in which user i, at time $t+1$ does not only take into account the aggression state of his neighbor j (1st insight), but also includes his personal state before making any changes in his aggression (4th insight). Consequently, this set of models can be formalized as follows:

$$S_i^{t+1} = A_i S_i^t + A_j S_j^t \tag{7}$$

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 Deffuant and HK models, as follows:

$$\text{If } |A_i S_i^t - A_j S_j^t| < d, \text{then, } S_i^{t+1} = A_i S_i^t + A_j S_j^t \tag{8}$$

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 user 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 which considers an average effect across all the neighborhood of the user, and are calculated as follows:

$$S_i^{t+1} = \frac{A_i S_i^t + \sum_{j \in N_i} A_j S_j^t}{A_i + \sum_{j \in N_i} A_j} \tag{9}$$

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, integrating the initial aggression state of an individual in the network (4th insight), along with the user's neighborhood (2nd insight):

$$S_i^{t+1} = \frac{A_i S_i^0 + A_i S_i^t + \sum_{j \in N_i} A_j S_j^t}{2A_i + \sum_{j \in N_i} A_j} \tag{10}$$

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:

$$S_i^{t+1} = \left(\frac{A_i + \sum_{j \in N_i} A_j}{1 + \sum_{j \in N_i} 1}\right)\left(\frac{S_i^t + \sum_{j \in N_i} S_j^t}{1 + \sum_{j \in N_i} 1}\right) \tag{11}$$

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:

$$S_i^{t+1} = \left(\frac{A_i + A_i + \sum_{j \in N_i} A_j}{1 + 1 + \sum_{j \in N_i} 1}\right)\left(\frac{S_i^0 + S_i^t + \sum_{j \in N_i} S_j^t}{1 + 1 + \sum_{j \in N_i} 1}\right) \tag{12}$$

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.

4 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. 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. Therefore, 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
2. Aggressive and normal users
3. Users (edges) to perform propagation
4. Ordering of users to perform propagation
5. Propagation model 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

4.1 Online social network

We use an unlabeled, directional Twitter network (McAuley and Leskovec 2012) (UT). This dataset has 81306 users or nodes, and 1768149 directed edges between them. We focus on the network's strongly connected component, that has 68413 nodes and 1685163 directed edges. Following past work on network analysis (Zuo et al. 2016), we apply weights on edges based on the Jaccard overlap of social circles between two users i and j, $w_{ij} = \frac{N_i \cap N_j}{N_i \cup N_j} \in [0, 1]$.

4.2 Which users should be aggressive?

In UT, we have no labels of aggression: we do not know which users exhibit aggression and which are normal. To identify users who should be labeled as aggressive in UT:

1. We use a small, Twitter network LT (Chatzakou et al. 2017), with users labeled as aggressive or not, to train a classifier (CL) on users' network features,[1] to infer the likelihood that a user will be aggressive. CL was trained with 93.24% accuracy and 93.2% precision and recall.
2. We extract the same network properties from UT, and apply CL on UT to label its users as aggressive or not, based on some threshold.

By applying CL on UT, we got 8.5% or 5820 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 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.

[1] User's followers and followees and their ratio, user's clustering coefficient score, hub score, authority score, and eigenvector score.

4.3 Users to perform propagation

We select a set $\mathbb{R}$ of random users for executing the propagation model. 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, covering 65.3% of total users.

4.4 Propagation changes applied

Users in $\mathbb{R}$ selected for propagation may interact with each other in different ways:

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).

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

4.5 Propagation models used

We test all models explained earlier. They are parameterless, making them simpler and more generalizable on different networks and setups.

4.6 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 next:

- **n**: portion of normal users in the network
- **a**: portion of aggressive users in the network
- **N-N**: portion of edges that both users i and j are normal
- **N-A**: portion of edges that user i is normal & j aggressive
- **A-N**: portion of edges that user i is aggressive & j normal
- **A-A**: portion of edges that i and j are aggressive users
- **n** $\rightarrow$ {**n** || **a**}: portion of normal users in the initial state who remain normal or become aggressive, respectively
- **a** $\rightarrow$ {**n** || **a**}: portion of aggressive users in the initial state who become normal or remain aggressive, respectively
- **N-N** $\rightarrow$ {**N-N** || **N-A** || **A-N** || **A-A**}: portion of edges that users i and j at initial state were normal and remain normal, or one, or both users become aggressive, respectively
- **N-A** $\rightarrow$ {**N-N** || **N-A** || **A-N** || **A-A**}: same as above for edges where j is aggressive at the initial state
- **A-N** $\rightarrow$ {**N-N** || **N-A** || **A-N** || **A-A**}: same as above for edges where i is aggressive at the initial state
- **A-A** $\rightarrow$ {**N-N** || **N-A** || **A-N** || **A-A**}: same as above for edges where both i and j are aggressive at the initial state

These elements capture the state of the network with respect to users and edges and their label at time t, and how these are changing through the simulated time between time t_i and t_{i+1} (declared as $\rightarrow$). Each of these metrics can be computed at regular snapshots ($T_0, T_1, T_2, \ldots, T_N$), by comparing the network state at a given snapshot (T_N) with initial state T_0.

4.7 Measuring ground truth metrics

We compute these metrics in LT as follows. First, we used the snapshot from 2016, with 401 users labeled as normal, aggressive, bully, or spammers. We removed the spam class and merged all aggressors and bullies under the aggressive class. Note: in this set of labeled users, their friends or followers are not labeled, so they are considered normal. Then, during 04-05/2019, we re-crawled these users' ego-networks. This crawl involved $\sim 717k$ users (401 egos - unique users). When this crawl was completed, $\sim 650k$ users were found to be active, $\sim 30k$ users were found suspended, and $\sim 37k$ users were found to have deleted their account. We make the assumption that active users in 2019 can be considered "normal", and suspended users are "aggressive", since at some point in the past they violated Twitter rules. We ignore users who deleted their profiles, as it is a user-decided action. Using the above two time crawls (2016 and 2019), we computed the ground truth or validation vector for the above mentioned 26 metrics, which capture the change of aggression of users and type of edges (A-A, A-N, etc.).

4.8 Comparing simulation and ground truth data

The above set of metrics is computed for all models and for 10 time snapshots per simulation. 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. 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.

5 Analysis of Simulation Results

In this section, we show the results from the extensive 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).

5.1 Which models are stable and perform best?

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, set to $T_A = 0.5$. Figure 1 plots the cosine similarity for all considered models in relation to the validation vector of real data. 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; e.g., in *Deffuant* and *DeGroot* models the

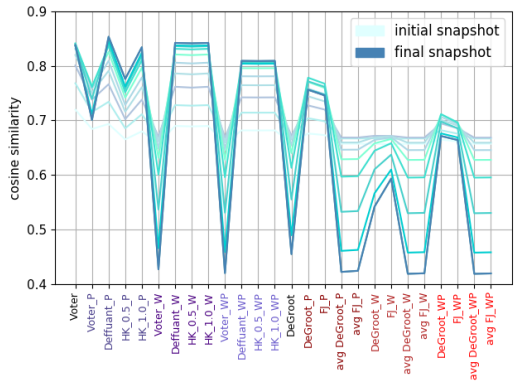


Figure 1: Cosine similarity of all proposed models with the validation vector, for 10% of selected edges and $T_A = 0.5$. The different time snapshots of the simulation are colored from light to dark blue. We group the models in sets based on their commonality.

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.

Different Aggression Thresholds T_A. We evaluate further the proposed models by investigating a wider range of thresholds $T_A = 0.05, \ldots, 0.9$ (results omitted due to space limits). 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). On the contrary, *Voter_W*, *Voter_WP*, and *Averaging_*_* show lower performance (similarity < 0.7) regardless 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).

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 Sec. 3.1. 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 (Hatfield, Cacioppo, and Rapson 1993). 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 (Chatzakou et al. 2019) which could explain the fact that aggressive users are more affected by direct relationships rather than their neighborhood aggression state.

5.2 How do models perform over time?

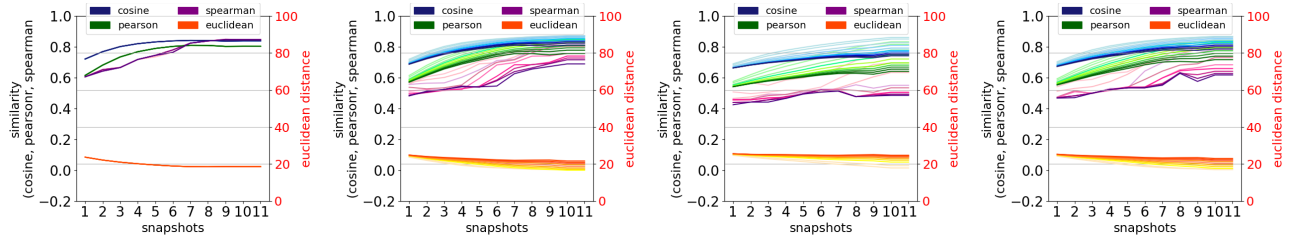
Next, we examine how model performance is affected when different time snapshots of the network's aggression state are considered. This analysis is done to: (i) compare with ground truth each model's state at progressing simulation times, and (ii) detect which of the snapshots was better fitting the 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 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*.

Figure 2 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.

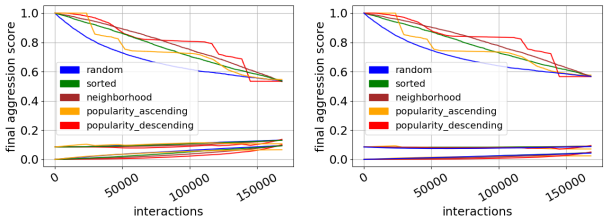
5.3 Is the order of changes important?

Figure 3 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 two models (similar results for all top models, omitted due to space). 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). 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 the initial and final aggression scores is subtle. Instead, the effect of normal users is greater on the aggressors than the inverse. For all considered models (apart from



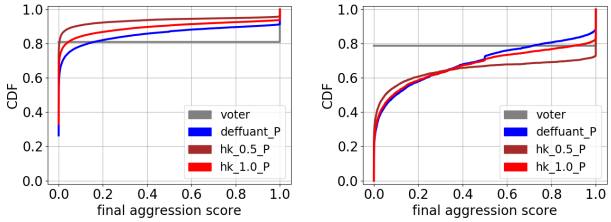
(a) *Voter* (b) *Deffuant_P* (c) *HK_0.5_P* (d) *HK_1.0_P*

Figure 2: 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 $T_A = 0.05, \ldots, 0.90$.



(a) Deffuant_P (b) HK_0.5_P

Figure 3: 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.



(a) Normal users (b) Aggressive users

Figure 4: CDFs of the final aggression scores of (a) normal and (b) aggressive users who participated in the aggression propagation.

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, since 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*, Sec. 3.1). 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.

5.4 How is users' final aggression state?

Figure 4 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. Figure 4(a) shows that regardless of model, at least 60% of normal users in the end 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 Figure 4(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 Figure 3, normal users are more resistant to adopting aggression, as opposed to aggressive users.

6 Discussion

Despite the consequences that abusive behavior has on individuals (e.g., embarrassment, depression, isolation from other community members), there are still important cases of aggression that stay under the radar of social networks, e.g., (O'Sullivan 2018). In fact, how such behavior propagates within networks has not been studied extensively. 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 fac-

tors to be considered in aggression propagation modeling, (4) influence by users' neighborhood is of less importance.

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). An interesting extension of this work would be to attempt aggression propagation modeling on a dynamic network, in which links are added or removed through time, since evolving networks are the most realistic but notoriously difficult to model. Also, it would be worthwhile to investigate the effectiveness of the proposed models to predict aggression on other platforms.

Acknowledgements

This research has been partially funded by the European Union's Horizon 2020 Research and Innovation program under the Marie Skodowska-Curie ENCASE project (Grant Agreement No. 691025) and CONCORDIA project ((Grant Agreement No. 830927).

References

[Allen, Anderson, and Bushman 2018] Allen, J. J.; Anderson, C. A.; and Bushman, B. J. 2018. The general aggression model. *Current Opinion in Psychology* 19:75 – 80. Aggression and violence.

[Anderson and Carnagey 2004] Anderson, C. A., and Carnagey, N. L. 2004. Violent evil and the general aggression model. *The Social Psychology of Good and Evil* 168 – 192.

[Bandura, Ross, and Ross 1961] Bandura, A.; Ross, D.; and Ross, S. A. 1961. Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology* 63(3):575–582.

[Behera and Schweitzer 2003] Behera, L., and Schweitzer, F. 2003. On spatial consensus formation: Is the sznajd model different from a voter model? *International Journal of Modern Physics C* 14:1331–1354.

[Bernardes, Stauffer, and Kertsz 2002] Bernardes, A.; Stauffer, D.; and Kertsz, J. 2002. Election results and the sznajd model on barabasi network. *Physics of Condensed Matter* 25:123–127.

[Cairns et al. 1988] Cairns, R. B.; Cairns, B. D.; Neckerman, H. J.; Gest, S. D.; and Gariepy, J.-L. 1988. Social networks and aggressive behavior: Peer support or peer rejection? *Developmental Psychology* 24(6):815–823.

[Campbell 1987] Campbell, A. & Muncer, S. 1987. Models of anger and aggression in the social talk of women and men. *Journal for Theory of Social Behaviour* 17(4).

[Caruso and Castorina 2005] Caruso, F., and Castorina, P. 2005. Opinion dynamics and decision of vote in bipolar political systems. *International Journal of Modern Physics C* 16(09):14731487.

[Chatzakou et al. 2017] Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017. Mean birds: Detecting aggression and bullying on twitter. In *WebSci*, 13–22. New York, NY, USA: ACM.

[Chatzakou et al. 2019] Chatzakou, D.; Leontiadis, I.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; Vakali, A.; and Kourtellis, N. 2019. Detecting cyberbullying and cyberaggression in social media. *Transactions on the Web*.

[Chen et al. 2012] Chen, Y.; Zhou, Y.; Zhu, S.; and Xu, H. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *PASSAT and SocialCom*.

[Cheng et al. 2017] Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW 17*.

[Clifford 1973] Clifford, Peter & Sudbury, A. 1973. A model for spatial conflict. *Biometrika* 60(3):581–588.

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

[Cyberbullying Research Center 2019] Cyberbullying Research Center. 2019. https://cyberbullying.org/facts.

[Davidson et al. 2017] Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

[Deffuant, Amblard, and Weisbuch 2004] Deffuant, G.; Amblard, F.; and Weisbuch, G. 2004. Modelling group opinion shift to extreme : the smooth bounded confidence model.

[Deffuant et al. 2000] Deffuant, G.; Neau, D.; Amblard, F.; and Weisbuch, G. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems* 03(01n04):87–98.

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

[Dinakar, Reichart, and Lieberman 2011] Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web* 11(02).

[Djuric et al. 2015] Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; and Bhamidipati, N. 2015. Hate Speech Detection with Comment Embeddings. In *WWW*.

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

[Farver 1996] Farver, J. 1996. Aggressive behavior in preschoolers' social networks: Do birds of a feather flock together? *Early Childhood Research Quarterly* 11(3):333–350.

[Fortunato and Castellano 2007] Fortunato, S., and Castellano, C. 2007. Scaling and universality in proportional elections. *Physical review letters* 99:138701.

[Fortunato 2005] Fortunato, S. 2005. Monte carlo simulations of opinion dynamics. *Complexity, Metastability and Nonextensivity*.

[Founta et al. 2018] Founta, A.-M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*.

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

[Gardner and Moffatt 1990] Gardner, W. I., and Moffatt, C. W. 1990. Aggressive behaviour: Definition, assessment, treatment. *International Review of Psychiatry* 2(1):91–100.

[Gionis, Terzi, and Tsaparas 2013] Gionis, A.; Terzi, E.; and Tsaparas, P. 2013. Opinion maximization in social networks. In *SDM*, 387–395. SIAM.

[Glauber 1963] Glauber, R. J. 1963. Timedependent statistics of the ising model. *Journal of Mathematical Physics* 4(2):294–307.

[Hariani and Riadi 2017] Hariani, and Riadi, I. 2017. Detection of cyberbullying on social media using data mining techniques. *IJCSIS* 15(3):244.

[Hatfield, Cacioppo, and Rapson 1993] Hatfield, E.; Cacioppo, J. T.; and Rapson, R. L. 1993. Emotional contagion. *Current directions in psychological science* 2(3):96–100.

[Hegselmann and Krause 2002] Hegselmann, R., and Krause, U. 2002. Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of Artificial Societies and Social Simulation* 5.

[Henneberger, Coffman, and Gest 2017] Henneberger, A. K.; Coffman, D. L.; and Gest, S. D. 2017. The effect of having aggressive friends on aggressive behavior in childhood: Using propensity scores to strengthen causal inference. *Social development* 26(2):295–309.

[Hine et al. 2017] Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*.

[Hosseinmardi et al. 2015] Hosseinmardi, H.; Mattson, S. A.; Rafiq, R. I.; Han, R.; Lv, Q.; and Mishra, S. 2015. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *SocInfo*.

[Ising 1925] Ising, E. 1925. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei* 31(1):253–258.

[Kayes et al. 2015] Kayes, I.; Kourtellis, N.; Quercia, D.; and Iamnitchi, A. & Bonchi, F. 2015. The Social World of Content Abusers in Community Question Answering. In *WWW*.

[Kramer, Guillory, and Hancock 2014] Kramer, A. D. I.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24):8788–8790.

[Lee and Lee 2002] Lee, J., and Lee, Y. 2002. A holistic model of computer abuse within organizations. *Information management & computer security* 10(2):57–63.

[Lorenz 2007] Lorenz, J. 2007. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C* 18(12):18191838.

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

[McAuley and Leskovec 2012] McAuley, J., and Leskovec, J. 2012. Learning to discover social circles in ego networks. In *NIPS - Volume 1*, 539–547. USA: Curran Associates Inc.

[Newman and Sheth 1985] Newman, B. I., and Sheth, J. N. 1985. A Model of Primary Voter Behavior. *Journal of Consumer Research* 12(2):178–187.

[Nobata et al. 2016] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In *25th ACM WWW Companion*.

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

[Pieschl et al. 2013] Pieschl, S.; Porsch, T.; Kahl, T.; and Klockenbusch, R. 2013. Relevant dimensions of cyberbullying - Results from two experimental studies . *Journal of Applied Developmental Psychology* 34(5).

[Sîrbu et al. 2017] Sîrbu, A.; Loreto, V.; Servedio, V. D. P.; and Tria, F. 2017. *Opinion Dynamics: Models, Extensions and External Effects*. Cham: Springer International Publishing. 363–401.

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

[Sobkowicz 2009] Sobkowicz, P. 2009. Modeling opinion formation with physics tools: Call for closer link with reality. *JASSS* 12(1):11.

[Sznajd-Weron and Sznajd 2000] Sznajd-Weron, K., and Sznajd, J. 2000. Opinion evolution in closed community. *International Journal of Modern Physics C* 11(06):11571165.

[Waseem and Hovy 2016] Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@ HLT-NAACL*.

[Xie, Cairns, and Cairns 1999] Xie, H.; Cairns, R. B.; and Cairns, B. D. 1999. Social networks and configurations in inner-city schools: Aggression, popularity, and implications for students with ebd. *JEBD* 7(3):147–155.

[Zschaler et al. 2012] Zschaler, G.; Böhme, G. A.; Seißinger, M.; Huepe, C.; and Gross, T. 2012. Early fragmentation in the adaptive voter model on directed networks. *Phys. Rev. E* 85:046107.

[Zuo et al. 2016] Zuo, X.; Blackburn, J.; Kourtellis, N.; Skvoretz, J.; and Iamnitchi, A. 2016. The power of indirect ties. *Computer Communications* 73:188–199.