Combating Behavioral Deviance via User Behavior Control

Chenxi Qiu, Anna Squicciarini
College of Information Science and Technology, Pennsylvania State University
University Park, PA, USA
{czq3,acs20}@psu.edu

Christopher Griffin
Applied Research Laboratory, Pennsylvania State University
University Park, PA, USA
griffinch@ieee.org

Prasanna Umar
College of Information Science and Technology, Pennsylvania State University
University Park, PA, USA
pxu3@ist.psu.edu

ABSTRACT

Compared to traditional behavioral deviance, online deviant behavior (like cyberbullying) is more likely to spread over online social communities since it is not restricted by time and space, and can occur more frequently and intensely. To control risks associated with the spread of deviant and anti-normative behavior, it is essential to understand online users' reaction when they interact with other users. In this paper, we model online users' behavior interaction as an *evolutionary game* on a graph and analyze users' behavior dynamics under different network conditions. Based on this theoretical framework, we then investigate behavior control strategies that aim to eliminate behavioral deviance. Finally, we use a real world dataset from a social network to verify the accuracy of our model's hypothesis. We also and test the performance of our behavior control strategy through simulations based on both real and synthetically generated data. The experimental results demonstrate that our behavior control methods can effectively eliminate the impact of bullying behavior even when the proportion of bullying messages is higher than 60%.

KEYWORDS

Cyberbullying; evolutionary game theory; online deviant behavior

ACM Reference Format:
Chenxi Qiu, Anna Squicciarini, Christopher Griffin, and Prasanna Umar. 2018. Combating Behavioral Deviance via User Behavior Control. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 9 pages.

1 INTRODUCTION

The fast growth of online social networks (e.g., Facebook, Instagram, and etc.) has led to an increase of abusive incidents and cyberaggression, including *cyberbullying*[1] [31, 34, 36]. Cyberbullying is a new emerging phenomena that has seen a steep rise in the recent years. While there is not a universal definition, a working definition of cyberbullying is given as "using information technology to willfully and repeatedly hurt, insult or harass others"[12].

According to a recent report, 19% of teens engaged in online social networking activities are reported being victims of some form of cyberbullying [12]. Compared with traditional deviant behavior, online behavioral deviance (i.e., referred as any behavior that pollutes negatively online) tends to be more sinister because it is not restricted by time and space and can occur more frequently and intensely, making it more difficult to control [14]. As episodes of online peer-to-peer abuse continue to increase in frequency and severity, several disciplines have actively engaged in research projects surrounding cyberbullying [39]. Several studies have investigated the dynamics of cyberbullying, bullies' motives and interactions [3, 23, 29, 33, 53, 57].

Within the computer science community, a growing body of work has focused on detecting instances of cyberbullying, like labelling offensive content through natural language processing (NLP) [8, 19, 35, 56]. For instance, text features are often used to extract attributes that can be used for supervised approaches (e,g. URLs, part-of-speech, n-grams, Bag of Words as well as lexical features such as sentiment or dictionary) [14, 20, 46].

One peculiar feature of cyberaggression (and even more of cyberbullying) is the role of peer and social pressure. Compared with other proactive deviant behavior (i.e., intentionally attacking others), the health of online communities relies heavily on individual peers' responses to certain triggers, which users may choose to emulate or disengage from. As reported by [1], exposure to aggressive behaviors allows for observational learning of such behavior and hence increases their likelihood to display aggressive behavior, especially for children, adolescents, and teenagers. For example, a recent research has found that adolescents were more likely (e.g., by up to 183% [25]) to carry out some acts of violence if their friends had also committed the same act.

The above observations motivate us to address the problem of online behavioral deviance from a different perspective. That is, we hypothesize that it is possible to reduce negative acts of cyberaggression by controlling exposure and direct influence of bystanders. To achieve this objective, it is essential to understand how users' behavior will evolve when interacting with others on these online social platforms. Here, *evolutionary game theory (EGT)* offers a conveniently adjustable and straightforward model for well-characterized strategic interactions [30]. Particularly, researchers have begun to use EGT on graphs to understand generic network social behaviors [32]. In this paper, we first transform a fundamentally discrete-time model, i.e., influencing behavior in a social network, into a graphical evolutionary game model, which is defined in a continuous time region. We then study users' behavior dynamics based on this continuous time model, which has stronger theoretical properties.

Specifically, we model an online social network of users on a graph, where each user is considered as a specific species. At any

[1] For the purpose of this paper, we use the terms cyberbullying and cyberaggression interchangeably. Yet, recent works tend to differentiate between the two forms of aggression. Our model can in fact apply to any deviant behavior where social influence plays a significant role in adoption.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

time point, each user may take one of two "strategies", i.e. post bullying or non-bullying messages. The basic hypothesis is that each user's intention of posting (non-)bullying will be increased if their neighbors (those who have social connections with this user) post the same type of messages. Based on this hypothesis, we define a payoff matrix for each user. By analyzing users' behavior dynamics, we then derive the relationship between users' payoff matrices and the *evolutionary stable state (ESS)* of (non-)bullying messages. This relationship provides us a theoretic foundation to adjust the parameters in the payoff matrix, to achieve *behavior control*. The adjustments enable convergence of the distribution of two types of behavior to the preferred ESS, where the proportion of non-bullying messages equals to 1. We formulate a new math optimization problem, called *user behavior control* problem, that aims to converge user behavior to the preferred ESS with minimum time, while guaranteeing the total change of payoff matrix does not exceed a constraint. Due to the hardness of the problem, we derive a theoretic lower bound of the optimal solution, and also devise a greedy algorithm, called *fast behavior control*, as a solution.

Finally, we analyzed a dataset from an online social network, where users' messages are labeled as bullying or non-bullying. The dataset verifies our hypothesis in the game theory model, i.e., users are more likely to post (non-)bullying messages if their neighbor post messages of the same type. Based on this seed dataset, we then test the performance of FBC via simulation. Our results demonstrate that FBC can effectively reduce or nearly eliminate the effects of deviant behavior and (over time) change messages in network to non-deviant. Specially, we observe that, when the initial proportion of bullying messages is higher than 60%, which is likely to move to 100% quickly, FBC can alter the moving direction in a short time (i.e., 1 time slot) and eventually converge the bullying message proportion to 0%. Even under the scenarios where the initial proportion of bullying messages will move to 0% without any control, FBC can still be applied to increase the convergence speed by up to 122.2%.

Simply put, our contributions can be summarized as:

1. *User behavior analysis through graphical EGT*: We first transform a discrete-time influencing behavior model into a continuous-time graphical EGT, and then analyze users' behavior dynamics over time. As far as we know, this is the first work to apply graphical EGT for modeling and developing strategies to address the deviant behavior problems.
2. *User behavior control*: Based on the above framework, we then formally formulate a new problem that aims to eliminate deviant behavior via controlling users' interaction. We derive a lower bound of the optimal solution for this problem, and also propose a time efficient algorithm FBC as a solution.
3. *Experiments*: We perform extensive experiments based on both real data and synthetic data. The experimental results demonstrate the efficiency of FBC, in terms of both convergence time and cost.

The remainder of the paper is organized as follows: We introduce the model in Section 2 and propose the behavior control approach in Section 3. In Section 4, we verify the hypothesis of our model and evaluate the performance of our method. Finally, we present related work in Section 5 and conclude in Section 6.

2 MODEL

In this part, we introduce the model, including notations and assumptions, that will be used throughout the paper. Specifically, we first introduce how we model online users' interaction by graphical EGT in Section 2.1 and then analyze the dynamics of users' behavior (including EES) based on the model in Section 2.2.

2.1 Finite Population Replicator Dynamics on a Graph

We consider a set of users $\mathcal{V} = \{1, 2, ..., N\}$ in an online social platform (e.g., Instagram, Facebook), and use a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ to describe the social topology of all users, where each edge $e_{j,i} \in \mathcal{E}$ denotes the social connection between users i and j. We assign a weight $\eta_{j,i}$ to each link $e_{j,i}$ to represent the probability that user i is affected by user j when user j posts a new content. We consider a binary action space: non-bullying (N) or bullying (B) for the content posted by users. Each user creates a collection of such messages. Let $x_{i,l}^{(k)}$ be the number of messages of type $l \in \{N, B\}$ user i has generated at time epoch k. The messages will form the population in an evolutionary game played on the social network graph structure.

When a message interacts with another, i.e., when users view their neighbor's messages, new messages are produced. We can encode the birth process of these messages in the following chemical diagrams:

$$N + N \to \kappa_N N \qquad N + B \to (\kappa_N - \alpha)N \tag{1}$$

$$B + B \to \kappa_B B \qquad B + N \to (\kappa_B - \beta)B \tag{2}$$

These expressions can be read as: when user i, having produced a non-bullying message that interacts with a neighbor having produced a non-bullying message, then κ_N non-bullying messages are produced as a result of comfort with the non-bullying messages. On the other hand, when user i's message interacts with a neighbor having produced a bullying message, then $\kappa_N - \alpha$ non-bullying messages are produced through a peer-pressure effect. Here, *comfort* is the users preferred behavioral mode on the social network. Similar explanation can be applied to the case when user i has produced a bullying message (i.e., the 2nd line in the chemical diagrams).

We assume that $\alpha, \beta, \kappa_N, \kappa_B \geq 0$. We can think of the message multipliers as entries in a payoff matrix:

$$\mathbf{A} = \begin{bmatrix} \kappa_N & \kappa_N - \alpha \\ \kappa_B - \beta & \kappa_B \end{bmatrix}. \tag{3}$$

Let $\mathbf{x}_i^{(k)} = [x_{i,N}, x_{i,B}]^\top \in \mathbb{Z}^2$ be a vector giving the number of messages of each type (N and B) generated by user i at time epoch k. Let

$$s^{(k)} = \sum_{l=N,B} \sum_{i \in N} x_{i,l}^{(k)} \tag{4}$$

be the total number of messages generated by all users at time epoch k, respectively. Then, we use $\mathbf{p}_i^{(k)} = \mathbf{x}_i^{(k)} / s^{(k)}$ to represent the proportion of messages in each type (N and B) generated by user i at time epoch k.

For the moment, we model time incrementally. Suppose at each time epoch, each user interacts randomly with a neighbor by comparing her messages to a random message of the neighbor and new messages are generated using the update rule. On the other hand, considering that the more users' messages are posted in the platform, the more messages are possibly omitted (e.g., users usually only pay attention to the most recent or popular messages posts in front of the message list), we assume each message has a probability ρ to "die". Then we can write:

$$x_{i,l}^{(k+1)} = x_{i,l}^{(k)} + \gamma x_{i,l}^{(k)} \frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_l^\top \mathbf{A}\mathbf{x}_j^{(k)} \eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{x}_j^{(k)} \eta_{j,i}} - \rho x_{i,l}^{(k)} \tag{5}$$

$$= x_{i,l}^{(k)} + \gamma x_{i,l}^{(k)} \frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_l^\top \tilde{\mathbf{A}}\mathbf{p}_j^{(k)} \eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_j^{(k)} \eta_{j,i}} \tag{6}$$

where

$$\tilde{\mathbf{A}} = \begin{bmatrix} \kappa_N - \frac{\rho}{\gamma} & \kappa_N - \alpha - \frac{\rho}{\gamma} \\ \kappa_B - \beta - \frac{\rho}{\gamma} & \kappa_B - \frac{\rho}{\gamma} \end{bmatrix}. \tag{7}$$

Here, $\mathbf{1}$ is a vector of 1's and $\gamma \in [0, 1]$ is a probability that any message in type l is responded (or response ratio) and $\mathcal{N}_i$ is the neighborhood of user i. Formally, each message of type l generated by user i interacts with a random message from user j. This message distribution is given by $\mathbf{p}_j^{(k)}$.

Then the expected number of new messages in type l per interaction is:

$$\frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_l^\top \tilde{\mathbf{A}}\mathbf{p}_j^{(k)} \eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_j^{(k)} \eta_{j,i}}.$$

In words, the user i considers all messages of type l she has ever sent in round $k - 1$ and inspects the messages of user j. Then generates new messages of type l based on these rules .

We can now pass to the mean-field approximation by observing the approximation:

$$x_{i,l}(t + \Delta t) \approx x_{i,l}(t) + \frac{\mathrm{d}x_{i,l}(t)}{\mathrm{d}t}\Delta t + O(\Delta t^2) \tag{8}$$

To facilitate the approximation, we must assume that $\gamma \sim \hat{\gamma}\Delta t$. That is, as the time interval between epochs shrinks, the probability of interaction varies linearly in Δt. Substituting Equation (8) into Equation (6) and simplifying yields:

$$\Delta t \frac{\mathrm{d}x_{i,l}(t)}{\mathrm{d}t} = \Delta t \gamma x_{i,l}(t) \frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_l^\top \tilde{\mathbf{A}}\mathbf{p}_j^{(k)} \eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_i^{(k)} \eta_{j,i}} \tag{9}$$

Thus, we conclude:

$$\dot{x}_{i,l}(t) = \gamma x_{i,l}(t) \frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_l^\top \tilde{\mathbf{A}}\mathbf{p}_j^{(k)} \eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_i^{(k)} \eta_{j,i}} \tag{10}$$

This equation gives the mean-field approximation on the number of messages of type l produced by user i. However, it is more instructive to know the proportions of different message types produced by user i. This can be computed by applying the quotient rule to compute the derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{x_{i,l}}{s}\right) = \frac{\dot{x}_{i,l}}{s} - x_{i,l}\frac{\dot{s}}{s^2} = \frac{\dot{x}_{i,l}}{s} - p_{i,l}\frac{\dot{s}}{s} \tag{11}$$

Here $p_{i,l}$ is the l^{th} element of $\mathbf{p}_i$. Note:

$$\frac{\dot{x}_{i,l}(t)}{s(t)} = \gamma p_{i,l}(t) \frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_l^\top \tilde{\mathbf{A}}\mathbf{p}_j^{(k)} \eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_i^{(k)} \eta_{j,i}}, \tag{12}$$

since $x_{i,l}(t)/s(t) = p_{i,l}(t)$. Also:

$$\dot{s}(t) = \sum_i \sum_l \dot{x}_{i,l}(t) = \gamma \sum_i \frac{\sum_{j\in\mathcal{N}_i} \mathbf{x}_i(t)^\top \tilde{\mathbf{A}}\mathbf{p}_j(t)\eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_j \eta_{j,i}} \tag{13}$$

which implies that

$$\frac{\dot{s}(t)}{s(t)} = \gamma \sum_i \frac{\sum_{j\in\mathcal{N}_i} \mathbf{p}_i(t)^\top \tilde{\mathbf{A}}\mathbf{p}_j(t)\eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_j \eta_{j,i}}. \tag{14}$$

Combining terms yields the cyber-bullying network equation:

$$\dot{p}_{i,l} = \gamma p_{i,l}\left(U_{i,l}(\mathbf{p}) - \overline{U}(\mathbf{p})\right) \tag{15}$$

where

$$U_{i,l}(\mathbf{p}) = \frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_l^\top \tilde{\mathbf{A}}\mathbf{p}_j \eta_{j,i}}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_j \eta_{j,i}} \tag{16}$$

$$\overline{U}(\mathbf{p}) = \sum_r \sum_l p_{r,l} U_{i,l}(\mathbf{p}) \tag{17}$$

By fitting $\mathbf{A}$, this estimates the proportion of bullying and non-bullying messages sent by user i at any time. This is a special form of multi-species/subspecies games studied in [51].

2.2 Evolutionary stable state

After modeling users' behavior by graphic EGT, the next step is to analyze the behavior dynamics of users, especially to derive the stable equilibrium of their behaviors after a period of strategic interactions, namely the *evolutionarily stable state (ESS)*.

It is intractable to derive the closed form of ESS in arbitrary graph. In fact, many community detection (e.g., graph clustering) have been proposed to scale down and simplify the structure of dynamic and complex networks [6, 45]. In this part, we make a simplifying assumption that the graph $\mathcal{G}$ is a homogeneous clique of identical users, i.e., $\mathcal{N}_i = \mathcal{V}\backslash i$ for each user i and $\eta_{j,i} = \eta$ for each pair of users i and j, which means we target on analyzing the dynamics of a single online community.

As we set our goal as minimizing the influence of bullying messages within the network, in the following, we analyze the dynamics of the total number (or the proportion) of two types of messages from all users:

EES of the total number of two types of messages from all users. Let $x_l(t) = \sum_i x_{i,l}(t)$. We can derive that

$$\dot{x}_N = \sum_i \dot{x}_{i,N} = \gamma \sum_i x_{i,N} \sum_{j\in\mathcal{V}\backslash i} \mathbf{e}_N^\top \tilde{\mathbf{A}}\mathbf{p}_j, \tag{18}$$

from which we obtain

$$\dot{x}_N = \gamma\left(a_N^1\left(\frac{x_N^2}{x_N + x_B} - \frac{\sum_i x_{i,N}^2}{x_N + x_B}\right) + a_N^2\left(\frac{x_B x_N}{x_N + x_B} - \frac{\sum_i x_{i,B} x_{i,N}}{x_N + x_B}\right)\right), \tag{19}$$

where $a_N^1 = (\kappa_N - \frac{\rho}{\gamma})$ and $a_N^2 = (\kappa_N - \alpha - \frac{\rho}{\gamma})$. Considering that $\sum_i x_{i,N}^2 \ll x_N^2$ and $\sum_i x_{i,B} x_{i,N} \ll x_B x_N$, we can approximate $\dot{x}_N(t)$ by

$$\dot{x}_N \approx g_N(x_N, x_B) = \gamma x_N \left(a_N^1 - \frac{\alpha x_B}{x_N + x_B}\right). \tag{20}$$

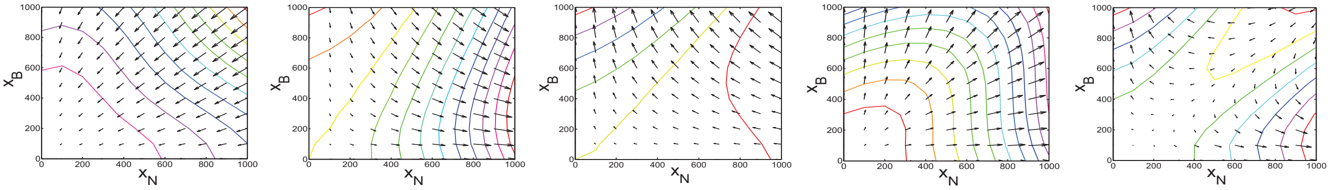
Similarly, we can obtain

$$\dot{x}_B \approx g_B(x_N, x_B) = \gamma x_B \left(a_B^1 - \frac{\beta x_N}{x_N + x_B}\right), \tag{21}$$

where $a_B^1 = (\kappa_B - \frac{\rho}{\gamma})$ and $a_B^2 = (\kappa_B - \beta - \frac{\rho}{\gamma})$.

Figure 1 gives several examples of x_N and x_B's direction fields with different payoff matrix parameters. The figure illustrates that the differences in the payoff matrix structure yield bifurcations in the flow.

EES of the proportion of two types of messages from all users. Let $p_l(t) = \sum_i p_{i,l}(t)$ and hence $\dot{p}_l = \sum_i \dot{p}_{i,l}$ ($l \in \{N, B\}$). We first



(a) $\max\{\kappa_N, \kappa_B\} \le \frac{\rho}{\gamma}$ (b) $\kappa_B \le \frac{\rho}{\gamma} \le \kappa_N - \alpha$ (c) $\kappa_N \le \frac{\rho}{\gamma} \le \kappa_B - \beta$ (d) $\frac{\rho}{\gamma} \le \min\{\kappa_N - \alpha, \kappa_B - \beta\}$ (e) $\kappa' \le \frac{\rho}{\gamma} \le \min\{\kappa_N, \kappa_B\}$

Figure 1: Variations of the dynamics of X_N and X_B in the direction field with different payoff matrices. The arrows indicate the motion direction as t increases. $^*\kappa' = \max\{\kappa_N - \alpha, \kappa_B - \beta\}$.

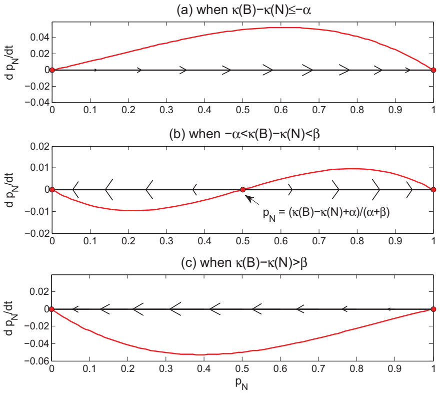


Figure 2: The dynamics of p_N in the three different cases.

derive $U_{i,N}(\mathbf{p})$, $U_{i,B}(\mathbf{p})$, and $\overline{U}(\mathbf{p})$ and embed them to Equations (15)-(17).

$$U_{i,N}(\mathbf{p}) = \frac{\sum_{j\in\mathcal{N}_i} \mathbf{e}_N^\top \tilde{\mathbf{A}}\mathbf{p}_j}{\sum_{j\in\mathcal{N}_i} \mathbf{1}^\top \mathbf{p}_j} = \sum_{j\in\mathcal{N}_i} \mathbf{e}_N^\top \tilde{\mathbf{A}}\mathbf{p}_j$$

$$= \left(\left(\kappa_N - \frac{\rho}{\gamma}\right) \sum_{j\in\mathcal{V}\setminus i} p_{j,N} + \left(\kappa_N - \alpha - \frac{\rho}{\gamma}\right) \sum_{j\in\mathcal{V}\setminus i} p_{j,B}\right)$$

Considering that $p_{j,l} \ll p_l$, we approximate $\sum_{j\in\mathcal{V}\setminus i} p_{j,l}$ by p_l, and then we obtain

$$U_{i,N}(\mathbf{p}) \approx \left(\kappa_N - \frac{\rho}{\gamma}\right) - \alpha p_B. \tag{22}$$

Similarly, we can derive

$$U_{i,B}(\mathbf{p}) \approx \left(\kappa_B - \frac{\rho}{\gamma}\right) - \beta p_N. \tag{23}$$

Finally, by embedding Equations (22) and (23) into Equation (15), we obtain

$$\overline{U}(\mathbf{p}) = \sum_i \sum_l p_{i,l} U_{i,l}(\mathbf{p}) = p_N\kappa_N + p_B\kappa_B - (\alpha+\beta)p_N p_B - \frac{\rho}{\gamma}. \tag{24}$$

Consequently, we can derive $\dot{p}_N$ ($\dot{p}_B = -\dot{p}_N$):

$$\dot{p}_N = \gamma p_N(1-p_N)((\alpha+\beta)p_N + \kappa_N - \kappa_B - \alpha) \tag{25}$$

According to the above equation, we discuss the dynamics of p_N in the following three cases:

I When $\kappa_B - \kappa_N \le -\alpha$ (Figure 2(a)): Two rest points $p_N^{*0} = 0$ and $p_N^{*1} = 1$ are *repulsive* and *attractive*, respectively. As $\dot{p}_N$ is always positive in the interval $(0, 1)$, p_N will always move to 1 under this case.

II When $-\alpha < \kappa_B - \kappa_N \le \beta$ (Figure 2(b)): Three rest points $p_N^{*0} = 0$, $p_N^{*} = \frac{\kappa_B-\kappa_N+\alpha}{\alpha+\beta}$, and $p_N^{*1} = 1$ are *attractive*, *repulsive*, and *attractive*, respectively. Since $\dot{p}_N$ is negative in the interval $\left(0, \frac{\kappa_B-\kappa_N+\alpha}{\alpha+\beta}\right)$ and is positive in the interval $\left(\frac{\kappa_B-\kappa_N+\alpha}{(\alpha+\beta}, 1\right)$, p_N will move to 0 in $\left(0, \frac{\kappa_B-\kappa_N+\alpha}{\alpha+\beta}\right)$ and to 1 in $\left(\frac{\kappa_B-\kappa_N+\alpha}{\alpha+\beta}, 1\right)$.

III When $\kappa_B - \kappa_N > \beta$ (Figure 2(c)): two rest points $p_N^{*0} = 0$ and $p_N^{*1} = 1$ are *attractive* and *repulsive.* respectively. As $\dot{p}_N$ is always negative in the interval $(0, 1)$, p_N will always move to 0 under this case.

3 USER BEHAVIOR CONTROL

According to the analysis in Section 2.2, whether p_N will converge to 0 or 1 depends on the locations of the rest points, which are determined by the payoff matrix. Hence, an intuitive idea for behavior control is to move the rest points by adjusting the parameters in the payoff matrix (Table 1 lists several examples of social network actions enabling a change of κ_N, κ_B, α, and β) such that p_N is always in the area that flows to 1. The details on how to adjust these parameters will be introduced in Section 3.

Table 1

Control category	Algorithm action
Social link control κ_N, κ_B	Delay/Block messages Block friend requests
Social capital control α, β	Alter like count Alter follower counts

In this section, we introduce our behavior control method. We first formulate the problem in Section 3.1 and propose a greedy algorithm as a solution in Section 3.2.

Before introducing our control strategy, we note that when $\kappa_B - \kappa_N \le -\alpha$ and $\kappa_B - \kappa_N > \beta$, temporary control won't affect the EES of p_N since p_N will eventually converge to 1 and 0 (respectively) once the control stops[2]. In fact, the first case indicates a healthy network that will automatically eliminate the influence of bullying messages without any control. The second case seldom happens since β is usually much higher than the difference between κ_N and κ_B in practical (which is also illustrated in our dataset described in Section 4.1). Accordingly, we only discuss the case when $-\alpha < \kappa_B - \kappa_N \le \beta$ in what follows.

3.1 Problem formulation

For simplicity, we first let $z = p_N^* = \frac{\kappa_B-\kappa_N+\alpha}{\alpha+\beta}$. Then, Equ. (25) can be rewritten as

$$\dot{p}_N = f(z, p_N) = \gamma(\alpha+\beta)p_N(1-p_N)(p_N - z) \tag{26}$$

[2] We assume that the behavior control is temporary, which means once the control actions stop, the payoff matrix will be recovered to its original value.

As analyzed in Section 2.2, to converge p_N is essentially to relocate the rest point z (i.e., to be smaller than p_N) such that p_N is in the area flowing to 1, as Figure 3(a) shows. Moreover, p_N should converge to 1 as fast as possible, which means we need to find the location of z that maximizes the value of $\dot{p}_N = f(z, p_N)$.

On the other hand, changing z too quickly and too drastically may be detrimental to users, due to blatant modifications to usersâĂŹ social network experience (e.g., users may be unhappy with social network experience when many of their messages are delayed by the system). Particularly, if p_N is pushed to the convergence area such that it will flow to 1 quickly, then to keep controlling z can only make tiny improvement of p_N's convergence speed at expense of unnecessary cost. In this case, z should be recovered to its original value to reduce the cost, as Figure 3(b) shows. Next, by considering both p_N's convergence speed and the limit of z's modification, we formulate the problem as an optimization problem.

We assume that actions to achieve behavior control can only be taken in a series of discrete time points $\delta, 2\delta, ..., K\delta$, where $K\delta$ is the maximum acceptable time for convergence. We use $p_N^{(k)}$ to represent $p_N(k\delta)$ and call the time interval $[(k-1)\delta, k\delta)$ time slot k. We assume that δ is small enough such that

$$p_N^{(k+1)} - p_N^{(k)} = \int_{(k-1)\delta}^{k\delta} f(z, p_N)\mathrm{d}t \approx f\left(z^{(k)}, p_N^{(k)}\right)\delta. \tag{27}$$

We also use $z^{(k)}$ to represent z value in time slot k. For each positive integer k, we have

$$p_N^{(k)} = \sum_{l=1}^{k-1} f\left(z^{(l)}, p_N^{(l)}\right) + p_N^{(1)}. \tag{28}$$

Our objective is to find the minimal integer y_{m} such that $p_N^{(y_{\mathrm{m}})} = 1$. In addition, we need to set a constraint Λ to limit the total change of z in the whole process: $\sum_{k=1}^{y_{\mathrm{m}}} |z^{(0)} - z^{(k)}| \le \Lambda$. Consequently, we formulate the *user behavior control* problem as:

$$\begin{aligned}
\min \quad & y && (29)\\
\text{s.t.} \quad & p_N^{(k)} = \sum_{l=1}^{k-1} f\left(z^{(l)}, p_N^{(l)}\right) + p_N^{(1)},\ \forall k = 1, 2, \ldots, K && (30)\\
& \min\left\{k \,\middle|\, p_N^{(k)} = 1, 1 \le k \le K\right\} \le y, && (31)\\
& \sum_{k=1}^{K} \left|z^{(0)} - z^{(k)}\right| \le \Lambda, && (32)\\
& z^{(1)}, \ldots, z^{(K)} \in [0, 1] \text{ and } y \in \mathbb{Z}^+ && (33)
\end{aligned}$$

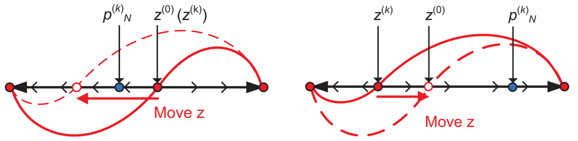
where $z^{(1)}, ..., z^{(K)}$ and y are decision variables, $z^{(0)}$ is the original value of z, $K\delta$ is p_N's maximum convergence time allowed by the platform, and $\mathbb{Z}^+$ represents positive integers.

The above problem is a mixed integer problem (MIP), which is NP-hard in general. Even if we relax y to a continuous region, the relaxed problem is still non-convex. Considering the computational tractability, in the following we propose a greedy algorithm, namely the *fast behavior control (FBC)* algorithm, that can effectively converge p_N to 1 with low time complexity. Before introducing the algorithm, we first derive a lower bound of y_{m} in Proposition 3.1. By comparing this lower bound with the solution of FBC, we can find how close that FBC can achieve to the optimal.

Proposition 3.1. *The minimum element in the set Ω, defined by*

$$\Omega = \left\{ k \in \mathbb{Z}^+ \,\middle|\, \begin{array}{l} \left(1 + \frac{\gamma\delta(\alpha+\beta)}{4}\right)^{k-1} p_N^{(1)} \\ -\frac{\gamma\delta(\alpha+\beta)}{4}\left((k-1)z^{(0)} - \Lambda\right) \ge 1 \end{array} \right\} \tag{34}$$

must be a lower bound of y_{m}.



(a) When $p_N < z$, move z to be smaller than p_N to push p_N to 1.

(b) When $p_N > z$, move z back to its original value to reduce cost.

Figure 3: Two cases in behavior control.

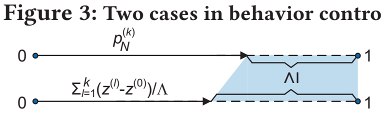


Figure 4: The constraint of FBC's z change in time slot k.

Proof. First, according to the constraint Equation (32), we have

$$\sum_{k=1}^{y_{\mathrm{m}}-1}\left(z^{(0)} - z^{(k)}\right) \le \sum_{k=1}^{y_{\mathrm{m}}-1}\left|z^{(0)} - z^{(k)}\right| \le \Lambda, \tag{35}$$

and hence $\sum_{k=1}^{y_{\mathrm{m}}-1} z^{(k)} \ge (y_{\mathrm{m}} - 1)z^{(0)} - \Lambda$. Then based on Equation (26) and (27), and the inequality $\left(1 - p_N^{(k)}\right)p_N^{(k)} \le \frac{1}{4}$, we have

$$p_N^{(k+1)} - p_N^{(k)} \le \frac{\gamma\delta(\alpha+\beta)}{4}\left(p_N^{(k)} - z^{(k)}\right) \text{ for each } k = 1, \ldots, y_{\mathrm{m}} - 1 \tag{36}$$

from which we can further derive that (to save space we let $\xi = \frac{\gamma\delta(\alpha+\beta)}{4}$)

$$\begin{aligned}
p_N^{(y_{\mathrm{m}})} & \le (1+\xi)^{y_{\mathrm{m}}-1} p_N^{(1)} - \xi \sum_{k=1}^{y_{\mathrm{m}}-1}\left(z^{(k)}(1+\xi)^{y_{\mathrm{m}}-k-1}\right)\\
& \le (1+\xi)^{y_{\mathrm{m}}-1} p_N^{(1)} - \xi \sum_{k=1}^{y_{\mathrm{m}}-1} z^{(k)} \quad (\text{since } \xi > 0)\\
& \le (1+\xi)^{y_{\mathrm{m}}-1} p_N^{(1)} - \xi\left((y_{\mathrm{m}} - 1)z^{(0)} - \Lambda\right). \qquad (37)
\end{aligned}$$

By embedding $p_N^{(y_{\mathrm{m}})} = 1$ into Equation (37), we obtain the constraint defined in Equation (34)

$$\left(1 + \frac{\gamma\delta(\alpha+\beta)}{4}\right)^{y_{\mathrm{m}}-1} p_N^{(1)} - \frac{\gamma\delta(\alpha+\beta)}{4}\left((y_{\mathrm{m}} - 1)z^{(0)} - \Lambda\right) \ge 1. \tag{38}$$

Hence y_{m} must be in Ω, indicating the minimum value in Ω is a lower bound of y_{m}. □

To find the minimum element of Ω, we simply check the integers from 1 to K one by one by increasing order, and pick up the first one satisfying Ω's inequality, where the time complexity is $O(K)$. The comparison of this lower bound and the solution of FBC will be shown in Section 4 (Figure 8).

3.2 The FBC algorithm

The basic idea of FBC is to maximize the *convergence efficiency* of p_N, $v^{(k)}$, in each time slot k, where $v^{(k)}$ is a relative measure of p_N's convergence speed to the change of $z^{(k)}$:

$$v^{(k)} = \left(\dot{p}_N^{(k)}\right)^r \Big/ \left|z^{(0)} - z^{(k)}\right|. \tag{39}$$

Here, r is FBC's *speed parameter* that affects the convergence speed and cost of the algorithm. The higher r is set, the faster p_N will converge to 1, and the more z will be modified.

On the other hand, we need to limit the modification of z in each step. As Fig. 4 shows, in each step, FBC is essentially to move p_N to 1 and simultaneously to prevent the sum of $(z - z^{(0)})/\Lambda$ from reaching 1. Accordingly, after adjusting $z^{(k)}$ in each time slot k, if we can make sure that the relative remaining space allowed to modify z, i.e., $1 - \frac{\sum_{l=1}^{k}\left(z^{(0)}-z^{(l)}\right)}{\Lambda}$, is larger than $1 - p_N$, then we can guarantee the total change of z can never exceed Λ before p_N converging to 1. Consequently, we formulate the following problem for FBC in each time slot k:

$$\max\ v^{(k)} \text{ s.t. } 1 - \frac{\sum_{l=1}^{k}\left|z^{(0)}-z^{(l)}\right|}{\Lambda} \geq 1 - p_N \qquad (40)$$

The above problem is a constrained single variable minimization problem with decision variable $z^{(k)}$. It is essentially to find $z^{(k)}$ that maximizes the value of non-liner function $\frac{\gamma(\alpha+\beta)p_N^{(k)}\left(1-p_N^{(k)}\right)\left(p_N^{(k)}-z^{(k)}\right)}{\left|z^{(0)}-z^{(k)}\right|}$ in interval $\left[\sum_{l=1}^{k-1}\left|z^{(0)}-z^{(l)}\right| - p_N\Lambda + z^{(0)}, \Lambda - \sum_{l=1}^{k-1}\left|z^{(0)}-z^{(l)}\right| + z^{(0)}\right]$, and we can directly apply the existing well-developed solutions to address this problem [28].

4 EMPIRICAL VALIDATION

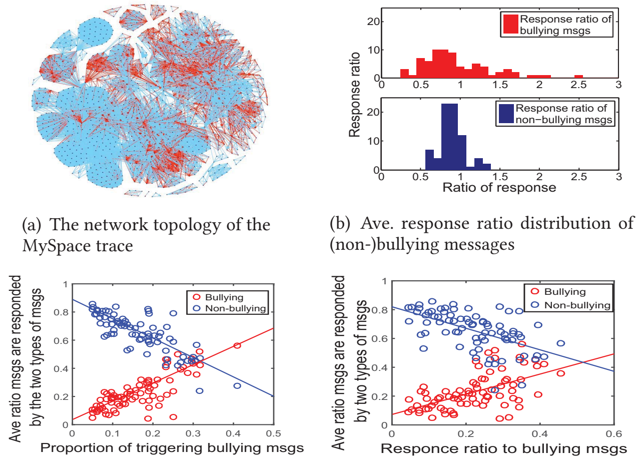
In this section, we turn our attention to practical applications of the proposed behavior control mechanism. We first validate the main hypotheses of our model (in Section 4.1). We test our FBC algorithm through simulations, based on both real-world data (in Section 4.2) and synthetically generated data (in Section 4.3). The main metrics we will measure include:
1) *Proportion of bullying triggering messages (denoted by $p_{\rm B}^{\rm tr}$)*;
2) *Average ratio that messages are responded by bullying (non-bullying) messages (denoted by $\kappa_{\rm B}^{\rm rsp}$ ($\kappa_{\rm N}^{\rm rsp}$))*;
3) *Convergence time*, the num. of slots to converge p_N to 1;
4) *Total change cost*, the total change of the rest point z over time.

4.1 Datasets

To verify the feasibility of our model and behavior control approach, we first used a real-world dataset from a real social network. Further, we created our own synthetic dataset to further validate and assess our model and related algorithms in more controlled settings.

MySpace dataset. The MySpace dataset in [35] contains 3032 posts on the MySpace social network generated by 1129 distinct users in 118 distinct threads and over the course of approximately four years. In each thread, a set of *triggering messages* are first posted, and then users can respond to either the triggering comments or to other users' comments. When posting messages, users may read the messages left by other users in the same thread. All the messages have been labeled by either "bullying" or "non-bullying". The graph depicted in Figure 5(a) records how users (represented by nodes) sent response messages (represented by edges) to others, where the bullying and non-bullying messages are marked by red and blue color, respectively (each edge may have multiple messages and we marke it by red color if it has at least one bullying message). As small number of bullying comments can hardly generate any influence, in this part, we only focused on the 76 threads that have at least 3 triggering bullying comments from the trace.

We first calculated the *average response ratio* to (non-)bullying messages in each thread, and compared their distribution over 76 threads in Figure 5(b). The average response ratios to bullying and non-bullying messages are 1.033 and 0.896, respectively, which suggests users are more likely to respond bullying messages than



(a) The network topology of the MySpace trace

(b) Ave. response ratio distribution of (non-)bullying messages

(c) Correlation between the proportion of triggering bullying and its peer pressure to the two types of responses

(d) Correlation between the response ratio to bullying messages and its pressure to two types of responses

Figure 5: Analysis of the MySpace dataset

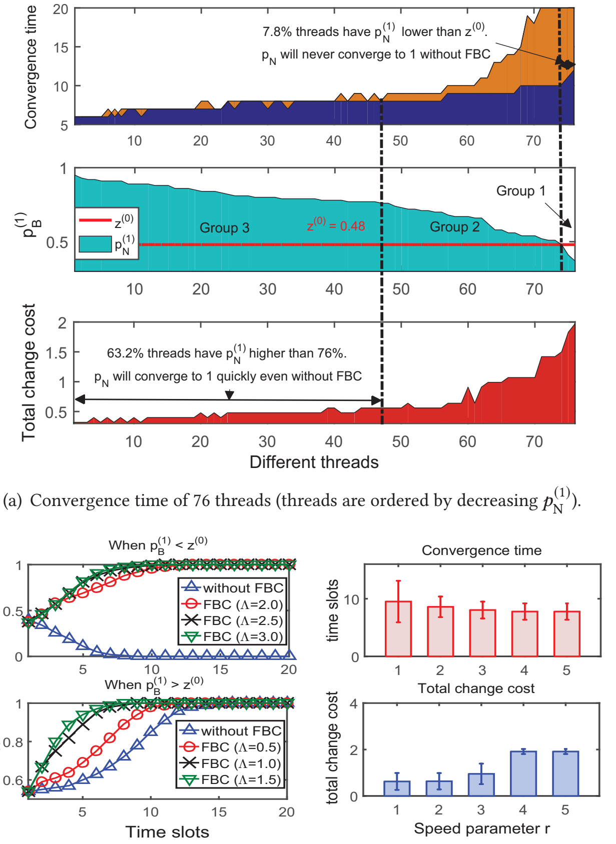
non-bullying ones. We also obtain the average response ratio of all the messages $\hat{\gamma} = 0.906$.

We then checked how $p_{\rm B}^{\rm tr}$ is correlated to users' response in different threads, say $\kappa_{\rm B}^{\rm rsp}$ and $\kappa_{\rm N}^{\rm rsp}$. As shown in Figure 5(c), the correlation between $p_{\rm B}^{\rm tr}$ and $\kappa_{\rm B}^{\rm rsp}$ equals 0.80, and the correlation between $p_{\rm B}^{\rm tr}$ and $\kappa_{\rm N}^{\rm rsp}$ equals -0.81. Hence, there is a strong correlation between users' different types of response and $p_{\rm B}^{\rm tr}$, supporting the hypothesis introduced in our chemical diagrams (Equation (1)-(2)) which assumes that higher proportion of bullying comments will increase (decrease) users' intention of generating bullying (non-bullying) messages. In addition, Figure 5 shows that the response ratio to bullying messages is positively and negatively correlated to $\kappa_{\rm B}^{\rm rsp}$ and $\kappa_{\rm N}^{\rm rsp}$, respectively, where their correlations equal to 0.5628 and -0.5522. Intuitively, this indicates that reducing the response ratio of bullying messages (e.g., via blocking messages) may help decrease the influence of bullying messages.

Payoff matrix parameter learning. To simulate users' behavior, we first need to estimate the payoff matrix parameters α, β, κ_N, and κ_B. As opposed to the metrics analyzed in Figure 5, α, β, κ_N, and κ_B cannot be directly measured. However, these parameters can be inferred from other observable metrics, like $p_{\rm N}^{\rm tr}$, $p_{\rm B}^{\rm tr}$, $\kappa_{\rm N}^{\rm rsp}$, and $\kappa_{\rm B}^{\rm rsp}$. Specifically, we have $\left[p_{\rm N}^{\rm tr} \quad p_{\rm B}^{\rm tr}\right]\mathbf{A} = \left[\kappa_{\rm N}^{\rm rsp} \quad \kappa_{\rm B}^{\rm rsp}\right]$, and by linear regression we can infer that $\hat{\beta} = 1.66$, $\hat{\kappa}_N = 0.87$, $\hat{\kappa}_B = 1.23$, and $\hat{\alpha} = 0.84$. We will use these estimated parameters in both real-trace and synthetic simulation.

4.2 Real-Trace Driven based Simulation

In this part, we carry out a simulation based on the 76 threads in MySpace dataset. We consider the triggering messages and their responses in each thread as the messages in time slot 1 and 2, respectively. We then simulate users' behaviors (posting bullying or non-bullying messages) in the later 18 slots based on the estimated payoff matrix. For the settings of FBC, we select to control the rest point z via adjusting κ_B, with the total change cost no larger than 3, and we set the speed parameter $r = 3$ by default.



(a) Convergence time of 76 threads (threads are ordered by decreasing $p_N^{(1)}$).

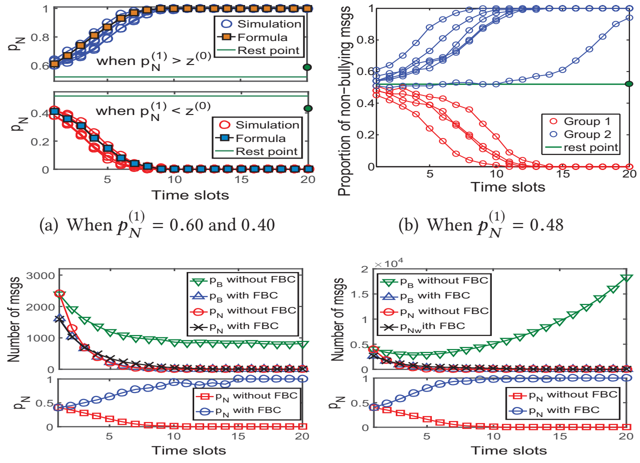
(b) Examples when $p_N^{(1)}$ is lower and higher than the original rest point $z^{(0)}$.

(c) The performance of the FBC algorithm with different r.

Figure 6: Simulation based on real-world data.

We first compare the convergence time with and without FBC over 76 threads in Figure 6(a), in which we also depict $p_N^{(1)}$ and the total change cost of different threads. In the figure, we order the threads by decreasing $p_N^{(1)}$, and divide all the threads into three groups, which have $p_N^{(1)} \in [0, 0.48)$, $p_N^{(1)} \in [0.48, 0.76]$, and $p_N^{(1)} \in [0.76, 1]$, respectively. Here, 0.48 is the value of initial rest point $z^{(0)}$ and 0.76 is the value of p_N^{peak} that maximizes the convergence speed $\dot{p}_N$ (defined in Equation (26)). Both 0.48 and 0.76 are calculated based on the estimated payoff matrix.

Not surprisingly, we find that the threads in group 1 (7.8% of all) cannot move their p_N to 1 without FBC, as the initial $p_N^{(1)}$ is in the area converging to 0. However, FBC can alter the moving direction of p_N and eventually converge p_N to 1. In contrast to group 1, the threads in group 3 (63.2% of all) have their p_N moving to 1 quickly even without any control. Further, there is no significant difference between the convergence time with and without FBC in group 3, as when $p_N \geq p_N^{peak}$, moving z will either reduce the convergence speed (when moving to 0) or only make a tiny improvement in speed (when moving to 1). Therefore, FBC needs to take fewer actions in group 3 and the total change cost is at most 0.5. Finally, from group 2, we observe that FBC increases the convergence speed of p_N significantly (by up to 122.2%), especially when $p_N^{(1)} \leq 0.7$, though most threads can move p_N to 1 automatically.



(a) When $p_N^{(1)} = 0.60$ and 0.40

(b) When $p_N^{(1)} = 0.48$

(c) $\max\{\kappa_N, \kappa_B\} \leq \frac{\rho}{\gamma}$

(d) $\kappa' \leq \frac{\rho}{\gamma} \leq \min\{\kappa_N, \kappa_B\}$

Figure 7: Simulation based on synthetically generated data.

In the next experiment, we check how FBC helps p_N converge to 1 with different constraint Λ. We randomly pick one thread in group 1 and 2, respectively, and depict the convergence process of p_N over time without and with FBC (using different Λ) in Figure 6(b). Specifically, we adjust Λ from 2 to 3 in group 1 and from 0.5 to 1.5 in group 2. From the two figures, we observe that 1) when $p_N^{(1)} < z^{(0)}$, FBC can alter the motion direction of p_N in a short time (1 time slot) and eventually move p_N to 1; 2) when $p_N^{(1)} > z^{(0)}$, FBC can increase the convergence speed of p_N; 3) FBC has higher convergence speed when Λ is higher.

Finally, we test how the change of FBC's speed parameter r will affect FBC's performance. We adjust r from 1 to 5 and depict the median, 5th and 95th percentile of the convergence time and the total change cost of FBC over all 76 threads in Figure 6(c). The two figures demonstrates that, with the increase of r, the convergence time of p_N is reduced and the total change of z enhances, indicating higher speed requires higher cost. Also, we observe that compared with $r = 3$, the convergence time of $r = 4, 5$ is reduced slightly (by at most 3.3%), but the cost is increased considerably (by at least 101.2%), which suggests that r should be set lower than 4.

4.3 Synthetic Data based Simulation

For these experiments, we again use the payoff matrix estimated via MySpace data, and manually change other parameters, like $p_N^{(1)}$, the death ratio ρ, and the response ratio γ to test the model and FBC under different scenarios. The simulation is run for 20 time slots, and over 4,000 messages are generated at the first time slot. The goal of these experiments was to further validate our algorithm under various configurations, and specifically to assess evolution of the game over multiple time slots, as well as its theoretical boundaries.

First, we test how initial status of p_N affects the convergence of p_N. We set $p_N^{(1)}$ by 0.60, 0.40, and 0.48, and ran the simulation with each setting for 100 times. We then randomly pick up 5 results for both $p_N^{(1)} = 0.60, 0.40$ and 10 results for $p_N^{(1)} = 0.48$, where p_N's moving process over time of all the picked results are depicted in

Figure 7(a)(b). We also compare these simulation results with the curves derived from formulas (Equation (20) and (21)) in the two figures. From Figure 7(a), we observe that all p_Ns move to their stable points (1 when $p_{\text{N}}^{(1)} = 0.60$ and 0 when $p_{\text{N}}^{(1)} = 0.40$) at the beginning, and p_N's motions in different trials are all consistent with their corresponding formula curves. However, when p_N is located at an unstable rest points, like in Figure 7(b), the moving direction is uncertain at the beginning and is possibly oppose in different trials. But once p_N leaves the unstable rest point 0.48, the moving direction seldom changes unless it reaches a stable rest point.

In addition to learning the proportion of (non)-bullying messages, we are also interested in observing how the number of the two types of messages evolve over multiple time slots. Here, we take two cases as examples:
Case I: when $\max\{\kappa_N, \kappa_B\} \le \frac{\rho}{\gamma}$ (we set $\rho = 0.6$ and $\gamma = 0.4$)
Case II: when $\max\{\kappa_N - \alpha, \kappa_B - \beta\} \le \frac{\rho}{\gamma} \le \min\{\kappa_N, \kappa_B\}$ (we set $\rho = 0.6$ and $\gamma = 0.6$), where the direction fields of the two cases are shown in Figure 1(a) and (e). In Case II, we also set $p_N^{(1)}$ to be lower than 0.45 (Note that, in this case, the number bullying messages and non-bullying messages increases and decrease in Figure 1(e)). In Figure 7(c) and (d), we depict the number&proportion of (non)-bullying messages over time with and without FBC in the two cases. We observe that without any control, 1) the number of both types of messages decrease in Case I, which is consistent with Figure 1(a), 2) the number of bullying and non-bullying messages increases and decreases, respectively in Case II, which is consistent with Figure 1(e), and 3) p_N in both cases moves to 0. But FBC can still push p_N to 1 eventually in both cases.

Finally, in Figure 8- for theoretical interest- we also compare the solution of FBC with the lower bound derived in Proposition 3.1. In this experiment, we change $p_N^{(1)}$ from 0.3 to 1. We observe that the ratio of the lower bound to FBC ranges from 0.6 to 1, which is no smaller than the approximation ratio, as the optimal solution must be in the gap between the lower bound and FBC. The figure also shows that with the decrease of p_N, the gap between FBC and the lower bound increases. This is partially due to the relaxation of $\dot{p}_N$'s upper bound when deriving the inequality constraint in Proposition 3.1. The relaxation takes place in each time slot, and hence the error is likely to accumulate over time.

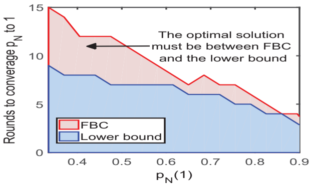


Figure 8: Convergence time of FBC and the lower bound

5 RELATED WORK

There are many social, biological, and physical systems in which a number of discrete individuals adjust an internal variable based on mutual interactions, leading the group to converge towards some sort of *consensus* (see [44] and its references). These models also depend sensibly on the connectivity of the individuals [7, 16, 38], so many of these models have been considered in the context of networks, like social networks [2, 18, 26]. For example, KrauseâĂŹs original consensus model [38] has been studied in networks, in particular by Olfati-Saber Murray [49] and Blondel et al. [4, 5] . Algorithms for consensus are proposed in [15, 17, 22]. Recent work by Proskurnikov et al. [52] studies the opinion consensus problem with hostile camps. Distributed consensus in a stochastic setting is studied in [55] and in second-order multi-agent systems in [43].

A distinct class of models consider the spread of a behavior on a social network as a contagion and applied mathematical epidemiological models. Epidemic models are also graph-based, and focus on individual accounts and connections between them. Individuals in the graph are exposed to a contagion through interaction with a neighbor in the network. Nodes are influenced by or altered by this exposure with some probability [27]. These models have been successfully used to explain the dynamics of diverse processes from emotional contagion on Facebook [13, 37] to health-related behaviors such as smoking, alcohol consumption and obesity [9–11]. Negative online behavior spread, like anti-normative commenting can also be modeled by a variation on an epidemic model (see [40, 41] and its references).

Among the models for interactions in dynamic systems, evolutionary game theory offers a conveniently adjustable and straightforward model for well-characterized strategic interactions, which include aspects such as sub-optimal stable equilibria and multiple equilibria [30]. Work in theoretical biology has begun to use evolutionary games on graphs in similar ways to understand network topologies for which evolutionary stability can be expected [47, 48] and develop variations on the replicator dynamic (see e.g., [21, 54]). Hussein [32] investigated a similar problem for generic network social behaviors while Pantoja and Quijano [50] investigate a distributed optimization problem on a network with the replicator. We note that recent work by Madeo and Mocenni [42] has developed a general replicator dynamic on graph structures, extending previous results [47, 48]. A result most closely related to this paper is found in [24], which studies convergence of best-response strategies on graphs. As far as we know, our paper is the first work to apply graphical EGT for modeling and developing strategies to address the deviant behavior problems.

6 CONCLUSIONS

In this paper, we used graphical evolutionary game to model the interaction among online users' behavior, and analyzed the dynamics of (non-)deviant behaviors of all users given the payoff matrix. Based on this framework, we then proposed a behavior control strategy, namely FBC, to converge the proportion of deviant behavior to 0 by adjusting the payoff matrix with limited change. The simulation results based on both real world data and synthetic data demonstrate the efficiency of FBC.

In the future, we plan to investigate the problem in a heterogenous setting, in which users may have different social influence to their neighbors and hence the payoff matrices among users might be different. Further, we plan to carry an extensive real user study online to further test and revise the model, and also develop a behavior control prototype in an actual online social platform.

7 ACKNOWLEDGEMENTS

This work was supported in part by AFOSR grant FA9550-15-1-0149 and PSU CSRE grant, and CG was supported by the National Science Foundation under grant number CMMI-1463482.

REFERENCES

[1] Albert Bandura, Dorothea Ross, and Sheila A. Ross. 1961. Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology* (1961).

[2] N. Bellomo, M. Herrero, and A. Tosin. 2013. On the dynamics of social conflicts: Looking for the black swan. *Kinetic Related Models* 6 (2013), 459–479.

[3] Ilene R Berson, Michael J Berson, and Michael J Berson. 2002. Emerging risks of violence in the digital age: Lessons for educators from an online study of adolescent girls in the United States. *Journal of School Violence* 1, 2 (2002), 51–71.

[4] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis. 2005. Convergence in multiagent coordination, consensus, and flocking. In *Proceedings of the 44th Conference on Decision and Control*. IEEE, 2996–3000.

[5] V. D. Blondel, J. M. Hendrickx, and J. N. Tsitsiklis. 2009. On krauseâĂŹs multi-agent consensus model with state-dependent connectivity. *IEEE Trans. Automatic Control* 54, 11 (2009), 2586–2597.

[6] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. 2012. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* 20, 2 (2012), 172–188.

[7] D. Centola and A. Baronchelli. 2015. Flocks, herds, and schools: A quantitative theory of flocking. In *Proceedings of the National Academy of Sciences*. 1989–1994.

[8] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of International Conference on Privacy, Security, Risk and Trust (PASSAT)*. IEEE, 71–80.

[9] N. A. Christakis and J. H. Fowler. 2007. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* 357, 4 (2007), 370–379.

[10] N. A. Christakis and J. H. Fowler. 2013. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine* 32, 3 (2013), 556–577.

[11] E. Cohen-Cole and J. M. Fletcher. 2008. Is obesity contagious? social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics* 27, 5 (2008), 1382–1387.

[12] Lucie Corcoran, Conor Mc Guckin, and Garry Prentice. 2015. Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression. *Societies* 5, 2 (2015), 245–255.

[13] L. Coviello, Y. Sohn, A. D. I. Kramer, M. Franceschetti C. Marlow, N. A. Christakis, and J. H. Fowler. 2014. Detecting emotional contagion in massive social networks. *PLoS ONE* 9, 3 (2014).

[14] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In *Advances in Artificial Intelligence - 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings*. 275–281. https://doi.org/10.1007/978-3-319-06483-3_25

[15] G. de Campos and A. Seuret. 2011. Improved consensus algorithms using memory effects. In *Proceedings of Conference on Decision and Control*. IEEE.

[16] M. H. DeGroot. 1974. Reaching a consensus. *J. American Stat. Association* 69 (1974), 118–121.

[17] P. DeLellis, M. diBernardo, and F. Garofalo. 2009. Novel decentralized adaptive strategies for the synchronization of complex networks. *Automatica* 45 (2009), 1312–1318.

[18] P. DeLellis, M. diBernardo, F. Garofalo, and D. Liuzza. 2010. Analysis and stability of consensus in networked control systems. *Appl. Math. Comput.* 5 (2010), 988–1000.

[19] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying.. In *The Social Mobile Web*.

[20] Nemanja Djuric, Jing Zhou, and Morris et al. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 29–30.

[21] G. B. Ermentrout, C. Griffin, and A. Belmonte. 2016. Transition matrix model for evolutionary game dynamics. *Physical Review E* 93 (2016).

[22] F. Fagnani and S. Zampieri. 2008. Randomized consensus algorithms over large scale networks. *IEEE J. Selected Areas Commun.* 26 (2008), 634–649.

[23] Ruth Festl and Thorsten Quandt. 2013. Social Relations and Cyberbullying: The Influence of Individual and Structural Attributes on Victimization and Perpetration via the Internet. *Human Communication Research* 39, 1 (2013), 101–126. https://doi.org/10.1111/j.1468-2958.2012.01442.x

[24] B. Gharesifard, B. Touri, T. BaÅ§ar, and J. Shamma. 2016. On the convergence of piecewise linear strategic interaction dynamics on networks. *IEEE Trans. Automat. Control* 61, 6 (2016), 1682–1687.

[25] Jeff Grabmeier. 2016. Violence spreads like a disease among adolescents, study finds Contagion moves from friends to friends of friends and beyond. *Ohio State News* (2016).

[26] D. Helbing. 2010. *Quantitative Sociodynamics: Stochastic Methods and Models of Social Interaction Processes*. Springer-Verlag.

[27] H. W. Hethcote. 2000. The mathematics of infectious diseases. *SIAM Rev.* 42 (2000), 599–653.

[28] Frederick S. Hillier. 2008. *Linear and Nonlinear Programming*. Stanford University.

[29] Sameer Hinduja and Justin W Patchin. 2013. Social influences on cyberbullying behaviors among middle and high school students. *Journal of youth and adolescence* 42, 5 (2013), 711–722.

[30] J. Hofbauer and K. Sigmund. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University.

[31] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *Proceedings of International Conference on Social Informatics*. Springer Link.

[32] I. I. Hussein. 2009. An individual-based evolutionary dynamics model for networked social behaviors. In *Proc. of American Control Conference*. 5789–5796.

[33] Jaana Juvonen and Sandra Graham. 2014. Bullying in schools: The power of bullies and the plight of victims. *Annual review of psychology* 65 (2014), 159–185.

[34] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. 2013. Detecting cyberbullying: Query terms and techniques. In *Proceedings of Web Science Conference*. ACM.

[35] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual ACM Web science conference*. ACM, 195–204.

[36] RM Kowalski, GW Giumetti, AN Schroeder, and MR Lattanner. 2014. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. In *Proceedings of Web Science Conference*. APA.

[37] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. In *Proceedings of the National Academy of Sciences*. IEEE, 8788–8790.

[38] U. Krause. 2000. A discrete nonlinear and non-autonomous model of consensus formation. *In Communications in Difference Equations* (2000), 227– 236.

[39] Colette Langos. 2012. Cyberbullying: The challenge to define. *Cyberpsychology, Behavior, and Social Networking* 15, 6 (2012), 285–289.

[40] C. Liao, A. Squicciarini, C. Griffin, and S. Rajtmajer. 2015. A hybrid epidemic model for the spread of abusive content in online social sites. In *Proc. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[41] C. Liao, A. Squicciarini, C. Griffin, and S. Rajtmajer. 2016. A hybrid epidemic model for deindividuation and antinormative behavior in online social networks. *Social Network Analysis and Modeling* 6, 1 (2016), 1–11.

[42] D. Madeo and C. Mocenni. 2015. Game interactions and dynamics on networked populations. *IEEE Trans. Automat. Control* 60, 7 (2015), 1801–1810.

[43] J. Mei, W. Ren, and J. Chen. 2016. Distributed consensus of second-order multi-agent systems with heterogeneous unknown inertias and control gains under a directed graph. *IEEE Trans. Automat. Control* 61, 8 (2016), 2019–2034.

[44] S. Motsch and E. Tadmor. 2014. Heterophilious Dynamics Enhances Consensus. *SIAM Rev.* 56 (2014), 577–621.

[45] M. E. J. Newman. 2012. Fast algorithm for detecting community structure in networks. *IEEE Transactions on Knowledge and Data Engineering* 69 (2012).

[46] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 145–153. https://doi.org/10.1145/2872427.2883062

[47] H. Ohtsuki and M. A. Nowak. 2006. The replicator equation on graphs. *Journal of Theoretical Biology* 243, 1 (2006), 86–97.

[48] H. Ohtsuki and M. A. Nowak. 2008. Evolutionary stability on graphs. *Journal of Theoretical Biology* 251, 4 (2008), 698 – 707.

[49] R. Olfati-Saber and R. M. Murray. 2004. Consensus problems in network of agents with switching topology and time delays. *IEEE Trans. Automatic Control* 49, 9 (2004), 1520–1533.

[50] A. Pantoja and N. Quijano. 2012. Distributed optimization using population dynamics with a local replicator equation. In *Proc. of EEE Conference on Decision and Control (CDC)*. 3790–3795.

[51] Elisabeth Paulson and Christopher Griffin. 2016. Cooperation can emerge in prisoner's dilemma from a multi-species predator prey replicator dynamic. *Mathematical biosciences* 278 (2016), 56–62.

[52] A. V. Proskurnikov, A. S. Matveev, and M. Cao. 2016. Opinion dynamics in social networks with hostile camps: Consensus vs. polarization. *IEEE Trans. Automat. Control* 61, 6 (2016), 1524–1536.

[53] Philip C Rodkin, Thomas W Farmer, Ruth Pearl, and Richard Van Acker. 2006. They're cool: Social status and peer group supports for aggressive boys and girls. *Social Development* 15, 2 (2006), 175–204.

[54] J. W. Weibull. 1997. *Evolutionary Game Theory*. MIT Press.

[55] X. Wu, Y. Tang, J. Cao, and W. Zhang. 2016. Distributed consensus of stochastic delayed multi-agent systems under asynchronous switching. *IEEE Transactions on Cybernetics* 46, 8 (2016), 1817–1827.

[56] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2 (2009), 1–7.

[57] Carlos P Zalaquett and SeriaShia J Chatters. 2014. Cyberbullying in college. *Sage Open* 4, 1 (2014), 2158244014526721.