Game Theoretic Suppression of Forged Messages in Online Social Networks

Xu Wang, Xuan Zha, Wei Ni, *Senior Member, IEEE*, Ren Ping Liu, *Senior Member, IEEE*, Y. Jay Guo, *Fellow, IEEE*, Xinxin Niu, and Kangfeng Zheng

***Abstract*—Online social networks (OSNs) suffer from forged messages. Current studies have typically been focused on the detection of forged messages and do not provide the analysis of the behaviors of message publishers and network strategies to suppress forged messages. This paper carries out the analysis by taking a game theoretic approach, where infinitely repeated games are constructed to capture the interactions between a publisher and a network administrator and suppress forged messages in OSNs. Critical conditions, under which the publisher is disincentivized to publish any forged messages, are identified in the absence and presence of misclassification on genuine messages. Closed-form expressions are established for the maximum number of forged messages that a malicious publisher could publish. Confirmed by the numerical results, the proposed infinitely repeated games reveal that forged messages can be suppressed by improving the payoffs for genuine messages, increasing the cost of bots, and/or reducing the payoffs for forged messages. The increasing detection probability of forged messages or decreasing misclassification probability of genuine messages also has a strong impact on the suppression of forged messages.**

***Index Terms*—Forged messages, game theory, social network.**

Manuscript received July 18, 2018; revised November 16, 2018 and January 22, 2019; accepted February 5, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0802703, and in part by the University of Technology Sydney, Deputy Vice Chancellor of Research Funding Initiative for Research Strengths. This paper was recommended by Associate Editor T.-M. Choi. *(Corresponding author: Xu Wang.)*

X. Wang is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: xu.wang-3@student.uts.edu.au).

X. Zha is with the Institute of ICT Security Research, China Academy of Information and Communications Technology, Beijing 100191, China, and also with the Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: xuan.zha@student.uts.edu.au).

W. Ni is with Data61, CSIRO, Sydney, NSW 2122, Australia (email: wei.ni@data61.csiro.au).

R. P. Liu and Y. J. Guo are with the Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: renping.liu@uts.edu.au; jay.guo@uts.edu.au).

X. Niu is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Data Security Center, State Key Laboratory of Public Big Data, Guiyang 550025, China (e-mail: xxniu@bupt.edu.cn).

K. Zheng is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: kfzheng@bupt.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2019.2899626

I. Introduction

ONLINE social networks (OSNs), enabling users to interact with one another through the Internet, dramatically reshape the way people are connected and have a strong impact on public decision-making [1]. OSNs, e.g., Facebook and Twitter, have been playing important roles even in politics and commerce [2]. Other OSNs, such as Yelp and TripAdvisor, collect reviews from consumers and can have a direct influence on product sales [3].

With no verification on user-generated content, OSNs are vulnerable to forged messages that intentionally mislead or deceive the recipients. There are a variety of forged messages, such as e-mail spams [4], fake views, e.g., in Amazon [5], [6], and diffusing rumors [7]–[9]. The publishers of forged messages can use social bots [10] or employ water army [11] to propagate forged messages and avoid sanction. Users in OSNs are susceptible to forged messages to different extents [12] and can be sensitive to the content of the messages [13].

Behavioral and linguistic features have been used by individual subscribers to judge the genuineness and forgery of a message [14]–[16]. Many subscribers can feed back their judgments to help correctly classify messages, penalize the publication of forged messages, and inhibit forgery of messages. For instance, Yelp filters reviews and encourages users to report inappropriate reviews [17]. Facebook enables users to provide feedback on any posts which are potentially forged, in addition to independent third-party verification [18]. Once identified, forged posts and/or their publishers can be blocked, which could be too harsh and inadequate in the case where there is a misclassification of genuine messages or unintentional mistakes of publishers [19], [20].

An increasingly popular approach to regulate the publisher's behaviors is to have service providers (or administrators), who send risk alerts to all users on disputable contents without blocking the contents and their publishers, as done in Facebook [18]. However, the effect of risk alerts on the inhibition of forged messages has yet to be properly analyzed and understood. One challenge is that the publisher can interact with the subscribers over an infinite-time horizon and change its strategies over time to potentially mislead the subscribers. The interactions can be sophisticated. Another challenge is that different types of subscribers may coexist in OSNs, including followers, fans, and bots, reacting distinctly to the publisher's behaviors.

The motivation of this paper is to provide a new quantitative understanding on the effect of risk alerts on the inhibition

of forged messages in the presence of multiple types of subscribers and possible miss-detection and false alarm of forged messages. We propose a new game theoretic approach for modeling and analyzing the proliferation of forged messages in OSNs, where there is a network administrator, a message publisher, and message subscribers (including followers, fans, and bots). Infinitely repeated games are constructed to characterize the interactions between the publisher, administrator, and subscribers and to quantify the payoffs of the publisher, first in the absence of misclassification on genuine messages and then in the presence of misclassification. This paper also identifies critical conditions, under which the broadcast of forged messages can be disincentivized and forged messages can be suppressed in OSNs. The contributions of this paper can be summarized as follows.

1) We propose infinitely repeated games to capture the interactions between a publisher and the administrator to suppress forged messages in OSNs.
2) Critical conditions, under which the publisher can be disincentivized to send any forged messages, are identified in the absence and presence of misclassification on genuine messages.
3) Closed-form expressions are established for the maximum number of forged messages of a malicious publisher in the absence and presence of misclassification on genuine messages.

Validated by numerical results, our analysis indicates that forged messages can be suppressed by improving the payoffs for genuine messages, increasing the cost of bots, and/or reducing the payoffs for forged messages. The increasing detection probability of forged messages or decreasing misclassification probability of genuine messages can also benefit the game theoretic suppression of forged messages.

The rest of this paper is organized as follows. In Section II, the related works are surveyed, followed by the proposed social network model and infinitely repeated games of forged messages in Section III. Sufficient conditions, under which the publisher is disincentivized from publishing forged messages, are established, and closed-form expressions for the maximum number of forged messages which can be published before the publisher is disincentivized are derived in Section IV. The proposed model, as well as the policies to suppress forged messages, are numerically validated in Section V, followed by the conclusions in Section VI.

II. Related Work

Interactions among users have been exploited to improve the accuracy of identifying forged messages and misbehaving publishers and penalize the publishers to inhibit forged messages. A subjective trustworthiness model and an objective model were proposed in [21], where trustworthiness was evaluated based on the local knowledge and global information, respectively. The trustworthiness of nodes can be judged by network administrators. For example, a trusted user was defined as the one which has never posted any spam tweet and has posted at least five confident ham tweets [20]. Priority can be given to trusted publishers [21]–[23]. Sirivianos *et al.* [19] proposed a collaborative spam filtering system, where feedback from different reporters was jointly assessed based on the trustworthiness of the reporters. Apart from manual complaints, feedback can be the automated reports, e.g., from honeypots [24].

A popular yet harsh technique of regulating misbehaving publishers is to block any forged messages and their publishers [25]–[27]. The solution would be too harsh and less effective in the case where genuine messages can be misclassified due to biased reviews [19] or unintentional mistakes made by the publishers, such as granting permission to malicious applications [20]. Another increasingly popular yet soft technique of regulating the publishers' behaviors is to have service providers (or administrators), who send risk alerts on forged messages without blocking the contents and their publishers, as done in Facebook [18]. Given the alerts, all subscribers can, by their own decision, distrust and reject the messages, thereby reducing the payoff to the publisher of the messages and disincentivizing the publisher from misbehaving. However, the technique has yet to be appropriately analyzed in the literature, due to sophisticated interactions between the publisher, administrator, and subscribers, as well as potentially different types of coexisting subscribers.

Game theory has been employed to model the interactions in OSNs [28]–[30]. In a game between publishers and administrators [28], the publishers can send more spam messages to gain higher payoff, which increases the successful detection by the administrators of the spam messages. The detected malicious publishers can be added to blacklists and quarantined. In [29], a Stackelberg-type game was set up to capture viral product design and customer satisfaction and optimize product adoption in OSNs. Abbass *et al.* [30] applied game theory to analyze the trustworthiness of N players in an OSN, revealing that the society with no untrustworthy individuals would yield the maximum wealth. However, game theory has not been used to model and analyze a more sophisticated scenario, as the one discussed in this paper, where there can be multiple types of users with different views on (forged) messages, which can even be misclassified by mistake.

To the best of our knowledge, none of the existing game theoretic analyses are able to capture the increasingly popular risk alert technique in OSNs [18]. An understanding of the long-term effect of this technique on the regulation of the publisher's behaviors is important to the design and configuration of the technique though. This paper bridges this gap in the existing literature by developing new game theoretic models to jointly capture the (mis)classification of messages and the interactions between parties in the risk alert systems for OSNs. The new infinitely repeated games are designed to model the sophisticated interactions between the publisher, subscribers (including followers, fans, and bots), and administrator in the presence of intentional and biased misclassification of messages. This paper also provides a new analysis to quantify the impact of different strategies that the administrator can take, such as improving the payoffs for genuine messages and increasing the cost of bots, on the suppression of forged messages. Conditions under which the publisher is disincentivized from sending any forged messages are identified.

In a different yet relevant context, techniques have been proposed to detect forged messages, typically by defining features. The techniques can be used by individual users or network operators to classify messages. Egele *et al.* [14] detected malicious accounts in OSNs based on the finding that malicious accounts exhibit consistent behaviors over time, e.g., time of day, source, text, topic, links in messages, direct user interaction, and proximity. Ruan *et al.* [31] used extroversive behavioral features (i.e., first activity, activity preference, activity sequence, and action latency) and introversive behavioral features (i.e., browsing preference, visit duration, request latency, and browsing sequences) to profile malicious accounts. Chen *et al.* [32] proposed an evolutionary classification algorithm to detect time-varying statistical features of spams, e.g., the number of retweets, by learning detected and manually labeled spams. Yang *et al.* [33] analyzed twitter spammers from the viewpoint of topology and found that spammers have small local clustering coefficients, high betweenness centrality, and small bidirectional link ratios. Neighbor features, such as neighbors' followers, average neighbors' tweets, and followings to median neighbors' followers, were employed to improve the detection of malicious accounts [33]. Review-based features, such as early time frame, rate deviation, the number of first-person pronouns, and the ratio of exclamation sentences, were proposed by Shehnepoor *et al.* [6] in addition to user-based features, to analyze reviews in OSNs (e.g., feedback on a topic or a product). Other review-based features, such as the numbers of views and comments received, ratings, and favorite times, were also used to identify spammers [34]. Linguistic features, such as the structure of sentences and modifiers, were also proposed to analyze reviews [15]. Based on a finding that the majority of collected spams are generated with underlying templates, Tangram [35] divided spams into segments and extracted templates for accurate and fast spam detections. Some of these techniques, such as [15] and [35], can be potentially adopted by the followers in this paper to classify messages. Taking the classification probabilities of the techniques as the input, however, our proposed game theoretic models are general and do not depend on specific classification techniques.

III. Network Model

We consider subscription services of OSNs, as shown in Fig. 1, where a publisher, denoted by $\mathcal{P}$, publishes messages to its subscribers. $\mathcal{P}$ can either publish genuine messages, referred to as the *benign* strategy, or forged messages, referred to as the *malicious* strategy. Genuine messages can benefit the OSNs overall, while forged messages can potentially increase $\mathcal{P}$'s payoff. Typical forged messages include rumors, commercial advertisements, and biased reviews. We consider the case that $\mathcal{P}$ can publish an infinite number of messages. At every round (one message per round), $\mathcal{P}$ takes either the *benign* or *malicious* strategy.

The subscribers consist of N_r followers who feed back objective judgments of $\mathcal{P}$'s messages to an administrator $\mathcal{A}$, and N_n fans who provide subjective (persistently positive) feedback in favor of $\mathcal{P}$. Let p_1 denote the miss-detection probability, at which a follower incorrectly gives positive feedback on a forged message; and p_2 denote the probability of detection, at which a follower correctly gives positive feedback on a genuine message. $(1-p_2)$ is the probability of the false alarm. The followers and fans feed back independently to $\mathcal{A}$. Moreover, $\mathcal{P}$ can also hire N_t bots which always give positive feedback to mislead $\mathcal{A}$ while costing $\mathcal{P}$.

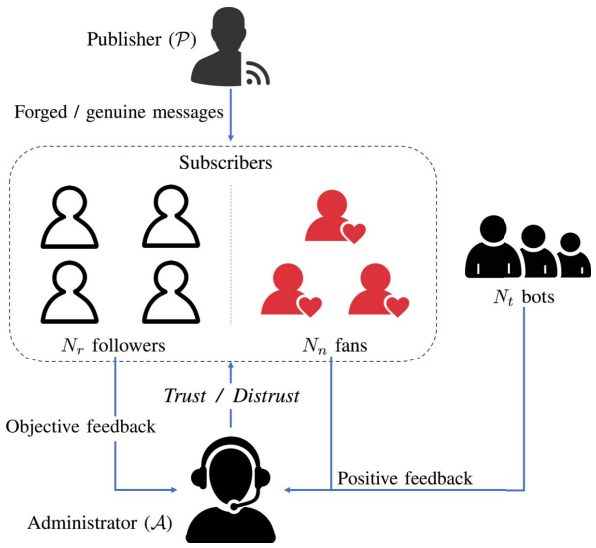


Fig. 1. Illustration of the subscription services in OSNs, where there is a network administrator $\mathcal{A}$, a publisher $\mathcal{P}$ and subscribers (including followers, fans, and bots). $\mathcal{P}$ can publish either genuine or forged messages. The subscribers check the messages and give their feedback to $\mathcal{A}$. The bots are hired by $\mathcal{P}$ to always give positive feedback to $\mathcal{A}$. Based on the feedback from the subscribers, $\mathcal{A}$ can take either a *trust* or *distrust* strategy to deal with $\mathcal{P}$.

Based on the feedback received, the administrator $\mathcal{A}$ can take either a *trust* or *distrust* strategy to encourage $\mathcal{P}$ to publish genuine messages or disincentivize $\mathcal{P}$ from sending forged messages, respectively. When taking the *trust* strategy, $\mathcal{A}$ confirms the genuineness of a message and advises all the subscribers to accept the message. When taking the *distrust* strategy, $\mathcal{A}$ sends risk alerts to notify the subscribers of a potentially forged message. The followers take $\mathcal{A}$'s advice, while the fans always trust $\mathcal{P}$ (at no cost of $\mathcal{P}$, as opposed to the bots). The *distrust* strategy reduces $\mathcal{P}$'s payoff and can penalize and suppress forged messages.

The payoff that $\mathcal{P}$ can receive per message is reasonably assumed to be linear to the number of subscribers who trust the message. The payoff is contributed by N_n fans and/or N_r followers. Let $C_1 > 0$ (or $C_2 > 0$) denote the unit payoff (per message per subscriber) for $\mathcal{P}$ to publish a genuine (or forged) message. If $C_1 \geq C_2$ (i.e., the payoff from a genuine message is higher than that from a forged message), $\mathcal{P}$ has no incentive to publish forged messages. We are particularly interested in the case where $C_1 < C_2$ and $\mathcal{P}$ has an incentive to publish forged messages, as will be analyzed in this paper. Let $C_3 > 0$ denote the cost of a bot to $\mathcal{P}$. The bots increase the cost of $\mathcal{P}$, but may help increase $\mathcal{P}$'s payoff by misleading $\mathcal{A}$ and reducing the probability of $\mathcal{P}$'s misbehaviors being detected.

We take a Facebook group for an example [36]. The group manager plays the role of the administrator $\mathcal{A}$, which can verify every post made by a publisher $\mathcal{P}$ and alert the group

members of disputable posts by using a Pin to Top announcement. As the fans (or friends) of $\mathcal{P}$, some group members choose to ignore $\mathcal{A}$'s advice and always *trust* and accept the posts. As followers of $\mathcal{P}$, other group members accept $\mathcal{A}$'s advice, i.e., *distrust* and reject disputable posts.

By default, $\mathcal{A}$ takes the *trust* strategy and updates the strategy on each of $\mathcal{P}$'s messages based on the detection results. We set the probability that $\mathcal{A}$ classifies a message to be genuine is equal to the ratio of positive feedback. Once a forged message is detected, $\mathcal{A}$ takes the *distrust* strategy against $\mathcal{P}$ and alerts the subscribers for a period of time (or in other words, $\mathcal{P}$ is placed on a "good behavior bond" and can still publish messages during the period). The alerting period can vary under different punishment policies. For example, $\mathcal{P}$ can be declared to be untrusted for a single round, permanently, or with exponentially increasing alert durations [37], [38]. Our analysis focuses on the exponentially increasing durations which double every time, a forged message is detected at $\mathcal{A}$. Thus, the alert duration in response to the kth detected forged message of $\mathcal{P}$ is 2^{k-1} (rounds), during which $\mathcal{P}$ may still choose to publish another forged message at the potential consequence of a further doubled alert duration following the current one [38]. $\mathcal{A}$ takes the *trust* strategy after the duration completes. $\mathcal{P}$ can become aware of the strategy that $\mathcal{A}$ takes and adjust its own behavior accordingly within and after the duration to maximize its payoff. However, our analysis can be readily applied to the other aforementioned punishment policies, i.e., distrust for a single round or permanently.

IV. Games of Forged Messages

In this section, we identify the critical conditions to suppress forged messages, where the interaction between $\mathcal{P}$ and $\mathcal{A}$ is modeled as an infinitely repeated game. We first consider a misclassification-free repeated game where genuine messages can be always correctly deduced. Sufficient conditions are established under which $\mathcal{P}$ only publishes forged messages, or is disincentivized from publishing any forged messages. Then, we extend to more sophisticated infinitely repeated games with possible misclassification on genuine messages. The conditions, under which forged messages are disincentivized, are dependent on the cost of messages as well as the classification results.

A. Misclassification-Free Infinitely Repeated Game

We first consider a misclassification-free game, where genuine messages are always correctly classified, i.e., $p_2 = 1$. $\mathcal{P}$ hires bots only when publishing forged messages. The payoff matrix of every message can be given by Table I, where the payoffs for $\mathcal{P}$ are listed under the specific strategy combinations. Bots are hired with the cost of C_3N_t, only when forged messages are published.

In the case of a forged message, $\mathcal{A}$ can collect $(N_r + N_n + N_t)$ pieces of feedback in total, $p_1N_r + N_n + N_t$ of which are positive. As a result, a forged message can be misjudged to be genuine with probability $q_1 = [(p_1N_r + N_n + N_t)/(N_r + N_n + N_t)]$, and correctly detected to be forged with probability $(1 - q_1)$.

TABLE I
Payoff Matrix to $\mathcal{P}$ Per Round in the Case of the Misclassification-Free Game

		$\mathcal{A}$	
		Trust	*Distrust*
$\mathcal{P}$	*Benign*	$C_1(N_r + N_n)$	C_1N_n
	Malicious	$C_2(N_r + N_n) - C_3N_t$	$C_2N_n - C_3N_t$

Theorem 1: In the misclassification-free infinitely repeated game, under the condition of

$$C_2N_n - C_3N_t - C_1(N_r + N_n) > 0. \tag{1}$$

$\mathcal{P}$ always publishes forged messages for a higher payoff than that when it publishes genuine messages. $\mathcal{A}$ cannot suppress forged messages by taking the distrust punishment. Equation (1) is a sufficient condition of $\mathcal{P}$ only publishing forged messages.

Under the condition of

$$(N_r + N_n)C_2 - ((2 - q_1)N_r + N_n)C_1 - C_3N_t < 0. \tag{2}$$

$\mathcal{P}$ has no incentive to publish forged messages; (2) is a sufficient condition of $\mathcal{P}$ not publishing forged messages.

Proof: Condition (1) can be proved by letting the minimum payoff from a forged message be greater than the maximum payoff from a genuine message. If $\mathcal{P}$ always takes the *malicious* strategy, it can at least get the payoff of $(C_2N_n - C_3N_t)$ per round. On the other hand, if $\mathcal{P}$ always takes the *benign* strategy, it can get at most $C_1(N_r + N_n)$ payoff per round. As a result, if $C_2N_n - C_3N_t - C_1(N_r + N_n) > 0$, the better strategy is *malicious* for $\mathcal{P}$ although it can be punished. In this case, $\mathcal{A}$ cannot suppress forged messages by taking the *distrust* punishment.

Condition (2) can be proved by comparing the cumulative payoff of a malicious publisher and that of a benign punisher over a sufficiently long period of time. Let τ_1 denote the cumulative rounds of punishment on the malicious publisher that publishes f forged messages. τ_1 is given by

$$\tau_1 = \sum_{k=0}^{f}\left(2^k - 1\right)\binom{f}{k}(1 - q_1)^k q_1^{f-k} \tag{3a}$$

$$= \sum_{k=0}^{f}\binom{f}{k}(2 - 2q_1)^k (q_1)^{f-k} - \sum_{k=0}^{f}\binom{f}{k}(1 - q_1)^k q_1^{f-k} \tag{3b}$$

$$= (2 - 2q_1 + q_1)^f - (1 - q_1 + q_1)^f \tag{3c}$$

$$= (2 - q_1)^f - 1 \tag{3d}$$

where (3a) is because k forged messages are detected with probability $\binom{f}{k}(1 - q_1)^k q_1^{f-k}$ in the total f forged messages. $\mathcal{P}$ is to be punished for $\sum_{l=1}^{k} 2^{l-1} = (2^k - 1)$ rounds in total when its forged messages are detected k times. Equation (3c) is obtained based on the Binomial theorem [39].

Let t denote the instant when the publication of f forged messages and the τ_1 corresponding punishments have ended.

Here, $t \geq (f+\tau_1)$. Note that $\mathcal{P}$ can publish either forged or genuine messages during the *distrust* period, resulting in different cumulative payoffs. If $\mathcal{P}$ publishes genuine messages during the period and publishes forged messages only when the punishments are over, the cumulative payoff in t rounds, denoted by π_{a1}, can be given by

$$\begin{aligned}\pi_{a1} =& f(C_2(N_r+N_n)-C_3N_t)+\tau_1 C_1 N_n \\ &+ (t-\tau_1-f)C_1(N_r+N_n) \end{aligned} \tag{4}$$

where the payoff of each round is based on the payoff matrix in Table I. The payoffs from f forged messages are $f(C_2(N_r+N_n)-C_3N_t)$. $\mathcal{P}$ with f forged messages is punished for τ_1 rounds on average, as given in (3), and can gain the payoff of $\tau_1 C_1 N_n$ during the punishment. The payoff of the remaining $(t-f-\tau_1)$ rounds is $(t-f-\tau_1)C_1(N_r+N_n)$, where $\mathcal{P}$ publishes genuine messages and $\mathcal{A}$ uses the default *trust* strategy.

If the malicious $\mathcal{P}$ publishes δ, $\delta > 0$ forged messages during the punishment and $(f-\delta)$ forged messages when it is not punished, the cumulative payoff, denoted by π'_{a1}, can be given by

$$\begin{aligned}\pi'_{a1} =& (f-\delta)(C_2(N_r+N_n)-C_3N_t) \\ &+ \delta(C_2N_n - C_3N_t) + (\tau_1-\delta)C_1N_n \\ &+ (t-\tau_1-f+\delta)C_1(N_r+N_n). \end{aligned} \tag{5}$$

We have $\pi_{a1}-\pi'_{a1} = \delta(C_2-C_1)N_r > 0$. As a result, the best strategy for $\mathcal{P}$ is to publish genuine messages in the punishments and publish forged messages only when the punishments are over. The maximum payoff is given as π_{a1} in (4).

In the case that $\mathcal{P}$ always publishes genuine messages, $\mathcal{A}$ takes *trust* in return. The cumulative payoff of a benign $\mathcal{P}$ in t rounds, denoted by π_{b1}, can be given by

$$\pi_{b1} = tC_1(N_r+N_n) \tag{6}$$

where $C_1(N_r+N_n)$ is the payoff per round in Table I.

Let $g_1(f)$ denote the gap of payoff (referred to as "extra payoff") between a malicious behavior of $\mathcal{P}$ with f forged messages and a benign behavior, i.e.,

$$\begin{aligned} g_1(f) &= \pi_{a1}-\pi_{b1} \\ &= f(C_2(N_r+N_n)-C_3N_t)+\tau_1C_1N_n \\ &\quad -(\tau_1+f)C_1(N_r+N_n) \\ &= f((C_2-C_1)(N_r+N_n)-C_3N_t) \\ &\quad - C_1N_r\Big((2-q_1)^f-1\Big). \end{aligned} \tag{7}$$

Here, t is suppressed. In other words, the payoff gap depends on the number of forged messages and is unaffected by the order of forged and genuine messages.

In the case that $g_1(1) < 0$, $\mathcal{P}$ cannot get a higher payoff even if it publishes a single forged message. This is because $g_1(0)=0$ and $g_1(f)$ first increases and then decreases, as will be proved in Theorem 2 [see (10)]. The sufficient condition of $\mathcal{P}$ not publishing forged messages can be obtained. ■

Theorem 2: $\mathcal{P}$ gains the maximum extra payoff by publishing $\lfloor x_{m_1} \rfloor$ or $\lceil x_{m_1} \rceil$ forged messages. The malicious $\mathcal{P}$ has an incentive to publish less than $\lceil x_{u_1} \rceil$ forged messages. x_{m_1} and x_{u_1} can be given by

$$\begin{aligned} x_{m_1} &= -\log_{2-q_1}(\lambda_1 \ln(2-q_1)) \\ x_{u_1} &= -\frac{W_{-1}\big(-\lambda_1(2-q_1)^{-\lambda_1}\ln(2-q_1)\big)}{\ln(2-q_1)} - \lambda_1 \\ \lambda_1 &= \frac{C_1N_r}{(C_2-C_1)(N_r+N_n)-C_3N_t} \end{aligned} \tag{8}$$

where $\lfloor \cdot \rfloor$ stands for flooring and $\lceil \cdot \rceil$ stands for ceiling.

Proof: This theorem can be proved by relaxing the discrete function $g_1(f)$ in (7) to be a continuous function, denoted by $\widetilde{g_1}(x)$, as given by

$$\begin{aligned}\widetilde{g_1}(x) =& ((C_2-C_1)(N_r+N_n)-C_3N_t)x \\ &- C_1N_r\big((2-q_1)^x-1\big),\ x \geq 0. \end{aligned} \tag{9}$$

The derivative of $\widetilde{g_1}(x)$ can be given by

$$\begin{aligned}\frac{d\widetilde{g_1}(x)}{dx} =& (C_2-C_1)(N_r+N_n)-C_3N_t \\ &- (C_1N_r\ln(2-q_1))(2-q_1)^x. \end{aligned} \tag{10}$$

As a result, $\widetilde{g_1}(x)$ is a monotonically increasing function when $x < x_{m_1}$, and a monotonically decreasing function when $x > x_{m_1}$. Here, x_{m_1} can be given by

$$\begin{aligned} x_{m_1} &= -\log_{2-q_1}(\lambda_1\ln(2-q_1)) \\ \lambda_1 &= \frac{C_1N_r}{(C_2-C_1)(N_r+N_n)-C_3N_t}. \end{aligned} \tag{11}$$

This is achieved by letting $(d\widetilde{g_1}(x)/dx)=0$. Only an integer number of messages can be published. The malicious $\mathcal{P}$ gains the maximum extra payoff, i.e., $\max(g_1(\lfloor x_{m_1} \rfloor), g_1(\lceil x_{m_1} \rceil))$, by publishing forged messages.

The publisher $\mathcal{P}$ has an incentive to publish less than $\lceil x_{u_1} \rceil$ forged messages, where

$$x_{u_1} = -\frac{W_{-1}\big(-\lambda_1(2-q_1)^{-\lambda_1}\ln(2-q_1)\big)}{\ln(2-q_1)} - \lambda_1. \tag{12}$$

This is achieved by applying the Lambert's W Function [40], i.e., $W(\cdot)$, to $\widetilde{g_1}(x)=0$, where $-(1/e) \leq -\lambda_1(2-q_1)^{-\lambda_1}\ln(2-q_1) \leq 0$ is within the domain of the Lambert W function, as proved in the Appendix. The payoff of a malicious $\mathcal{P}$ is no more than the payoff of a benign $\mathcal{P}$ in the case that $f \geq \lceil x_{u_1} \rceil$ forged messages are published, i.e., $g_1(f)=\widetilde{g_1}(f) \leq 0$ and $f \geq \lceil x_{u_1} \rceil$. ■

B. Infinitely Repeated Game With Message Misclassification

In a more general case that the OSNs subscribers may provide incorrect feedback or comments on a genuine message due to personal opinions, the administrator can be misled and misjudge a genuine message to be a forged one. Therefore, the case of misclassification can be defined as such that $\mathcal{A}$ takes a *distrust* strategy when $\mathcal{P}$ publishes genuine messages. To avoid misclassification of genuine messages and reduce the detection of forged messages, $\mathcal{P}$ hires bots for positive feedback and pays the cost of C_3N_t when publishing genuine and forged messages. As a result, the payoff matrix of a single message can be given by Table II.

TABLE II
PAYOFF MATRIX OF $\mathcal{P}$ PER ROUND IN THE GAME WITH MISCLASSIFICATIONS

		$\mathcal{A}$	
		Trust	*Distrust*
$\mathcal{P}$	*Benign*	$C_1(N_r+N_n)-C_3N_t$	$C_1N_n-C_3N_t$
	Malicious	$C_2(N_r+N_n)-C_3N_t$	$C_2N_n-C_3N_t$

TABLE III
CONFUSION MATRIX FOR DETECTION RESULTS

		Classification result	
		Genuine	Forged
Message	Genuine	q_2	$1-q_2$
	Forged	q_1	$1-q_1$

In the infinitely repeated game, every genuine message is classified as a genuine message with probability $q_2 = [(p_2N_r+N_n+N_t)/(N_r+N_n+N_t)]$. The detection probability of a forged message is $(1-q_1)$ as given in Section IV-A. The confusion matrix, showing the detection and misclassification probabilities, is given by Table III.

We assume that $\mathcal{P}$ can become aware of incorrect punishments after one round, and complain to $\mathcal{A}$. $\mathcal{A}$ rectifies the misclassifications in response to the complaints by stopping the punishment. Meanwhile, $\mathcal{A}$ keeps the current punishment/alert duration for the next punishment instead of doubling it. In this way, the incorrect punishments due to misclassifications remain a single round and do not change the punishment duration of the coming forged messages. As a result, the duration of punishments on detected forged messages only depends on the number of detected forged messages and is independent of the number of genuine messages no matter whether they are misclassified or not. The malicious $\mathcal{P}$ does not complain about the correct punishments for its publication of forged messages.

Theorem 3: Under the condition of

$$\frac{(1-p_2)N_r^2N_t}{(N_r+N_n+N_t)(N_r+N_n)}C_1 - C_3N_t > 0 \tag{13}$$

a benign $\mathcal{P}$ hires bots for a higher payoff than it does not. Equation (13) is a sufficient condition of a benign $\mathcal{P}$ hiring bots.

Under the condition of

$$C_2N_n - C_1(q_2N_r+N_n) > 0. \tag{14}$$

$\mathcal{P}$ always publishes forged messages for a higher payoff than that when it publishes genuine messages. $\mathcal{A}$ cannot suppress forged messages by taking the distrust punishment. Equation (14) is a sufficient condition of $\mathcal{P}$ only publishing forged messages.

Under the condition of

$$C_2(N_r+N_n) - C_1((q_1+q_2-1)N_r+N_n) < 0. \tag{15}$$

$\mathcal{P}$ has no incentive to publish forged messages; (15) is the sufficient condition of $\mathcal{P}$ not publishing forged messages.

Proof: Condition (13) can be proved by comparing the payoffs from genuine messages with and without bots. In the case that a benign $\mathcal{P}$ does not hire bots and always publishes genuine messages, $\mathcal{A}$ can collect (N_r+N_n) pieces of feedback, $p_2N_r+N_n$ of which are positive. Thus, a genuine message is correctly classified with probability $q_3 = [(p_2N_r+N_n)/(N_r+N_n)]$ and misclassified with probability $(1-q_3)$. On the other hand, the punishment due to misclassification can be rectified in one round. As a result, the benign $\mathcal{P}$ is punished with probability $(1-q_3)$ per round. Its payoff per round, denoted by β_{b1}, can be given by

$$\begin{aligned}\beta_{b1} &= q_3C_1(N_r+N_n) + (1-q_3)C_1N_n \\ &= C_1(q_3N_r+N_n).\end{aligned} \tag{16}$$

Likewise, $\mathcal{P}$ can be punished with probability $(1-q_2)$ per round, even if it hires N_t bots and persistently publishes genuine messages, due to the misclassification of genuine messages. Its payoff per round, denoted by β_{b2}, can be given by

$$\begin{aligned}\beta_{b2} &= q_2(C_1(N_r+N_n)-C_3N_t) + (1-q_2)(C_1N_n-C_3N_t) \\ &= C_1(q_2N_r+N_n) - C_3N_t.\end{aligned} \tag{17}$$

By letting $\beta_{b2} > \beta_{b1}$, the condition under which the benign $\mathcal{P}$ hires bots is proved.

Likewise, Condition (14) can be proved by letting $C_2N_n - C_3N_t > \beta_{b2}$, where $C_2N_n - C_3N_t$ is the minimum payoff from a forged message of $\mathcal{P}$, as given in Table II.

Condition (15) can be proved by comparing the cumulative payoff of a malicious publisher and that of a benign punisher over the same period of time. If a malicious $\mathcal{P}$ publishes f forged messages, it is first punished for $\tau_1 = ((2-q_1)^f - 1)$ rounds, as proved by (3). Note that during the punishment, $\mathcal{P}$ can publish τ_1 number of genuine messages and be punished $(1-q_2)\tau_1$ rounds again due to misclassifications. As a result, $\mathcal{P}$ is able to publish τ_2 number of genuine messages during punishment in total, where $\tau_2 = \sum_{i=0}^{\infty}\tau_1(1-q_2)^i = (\tau_1/q_2)$. Thus, the payoff of the malicious $\mathcal{P}$ publishing f forged messages in $f+\tau_2$ rounds, denoted by π_{a2}, can be given by

$$\begin{aligned}\pi_{a2} &= (C_2(N_r+N_n)-C_3N_t)f \\ &\quad + (C_1N_n - C_3N_t)\frac{(2-q_1)^f-1}{q_2}.\end{aligned} \tag{18}$$

In the case that $\mathcal{P}$ always publishes genuine messages, its cumulative payoff of $(f+\tau_2)$ rounds, denoted by π_{b2}, can be given by

$$\pi_{b2} = \beta_{b2}(f+\tau_2). \tag{19}$$

Let $g_2(f)$ denote the gap between the payoffs (extra payoff) of a malicious behavior of $\mathcal{P}$ with f forged messages and a benign behavior, i.e.,

$$\begin{aligned}g_2(f) &= \pi_{a2} - \pi_{b2} \\ &= (C_2(N_r+N_n)-C_3N_t)f\end{aligned}$$

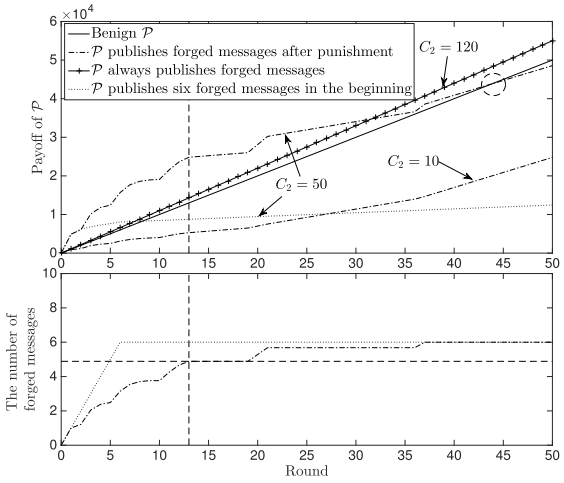


Fig. 2. Cumulative payoff and the number of forged messages of $\mathcal{P}$, where $p_1 = 0.03$, $N_r = 90$, $N_n = 10$, $N_t = 10$, $C_1 = 10$, and $C_3 = 10$. $C_2 = 10$, 50, and 120. The payoff is the average of 5000 independent simulations.

$$
\begin{aligned}
&+ (C_1 N_n - C_3 N_t)\frac{(2-q_1)^f - 1}{q_2} \\
&- (C_1(q_2 N_r + N_n) - C_3 N_t)(f + \tau_2) \\
&= f(C_2(N_r + N_n) - C_1(q_2 N_r + N_n)) \\
&- C_1 N_r (2-q_1)^f + C_1 N_r. \qquad (20)
\end{aligned}
$$

Condition (15) can be proved by letting $g_2(1) < 0$. ■

Theorem 4: The publisher $\mathcal{P}$ gains the maximum extra payoff by publishing $\lfloor x_{m_2} \rfloor$ or $\lceil x_{m_2} \rceil$ forged messages. The malicious $\mathcal{P}$ has an incentive to publish less than $\lceil x_{u_2} \rceil$ forged messages. x_{m_2} and x_{u_2} can be given by

$$
\begin{aligned}
x_{m_2} &= -\log_{2-q_1}(\lambda_2 \ln(2-q_1)) \\
x_{u_2} &= -\frac{W_{-1}\left(-\lambda_2 (2-q_1)^{-\lambda_2} \ln(2-q_1)\right)}{\ln(2-q_1)} - \lambda_2 \\
\lambda_2 &= \frac{C_1 N_r}{C_2(N_r + N_n) - C_1(q_2 N_r + N_n)}. \qquad (21)
\end{aligned}
$$

Proof: This theorem can be proved in the same way as Theorem 2 by replacing $g_1(f)$ with $g_2(f)$. Therefore, the proof is suppressed for brevity. ■

V. Numerical Result

Fig. 2 demonstrates 50 rounds of the proposed infinitely repeated games in the misclassification-free case, where $p_1 = 0.03$, $q_1 = [(p_1 N_r + N_n + N_t)/(N_r + N_n + N_t)] = 0.2$, $C_3 = 10$ and C_2 is set to be 10, 50, and 120. The upper part of the figure plots the cumulative payoffs of $\mathcal{P}$, where the solid line provides the payoff of a benign $\mathcal{P}$ for reference purpose. In the case of $C_2 = 120$, the solid line with marker "+" shows the payoff of $\mathcal{P}$ when $\mathcal{A}$ and $\mathcal{P}$ always take the *distrust* and *malicious* strategies, respectively. We can see that the malicious $\mathcal{P}$ can receive a higher payoff than that a benign one receives, even after being declared publicly to be distrusted. $\mathcal{A}$ cannot suppress forged messages, as stated in Theorem 1. In the case of $C_2 = 10$, $\mathcal{P}$ has no incentive to publish any forged messages at all. In the case of $C_2 = 50$, the payoffs of $\mathcal{P}$ are compared between two cases: 1) *case 1:* $\mathcal{P}$ continues to publish forged messages once an alert duration ends and 2) *case 2:* $\mathcal{P}$ publishes forged messages only in the very beginning. In both cases, six forged messages are published, and the rest 44 messages are genuine. We can see that $\mathcal{P}$ can gain a higher payoff if it sends forged messages when $\mathcal{A}$ takes the default *trust* strategy, as stated in Theorem 1.

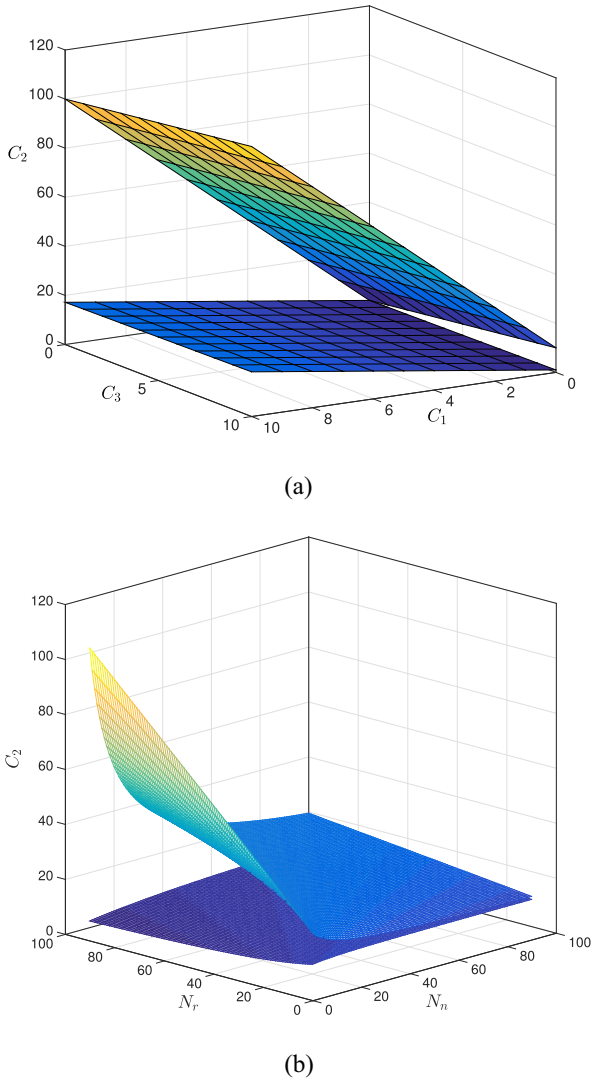


Fig. 3. Visualization of Theorems 1 and 3. (a) C_2 with the growth of C_1 and C_3, where the upper and lower surfaces are obtained with (1) and (2) in Theorem 1, respectively. $N_r = 90$, $N_n = N_t = 10$, and $q_1 = 0.2$. (b) C_2 with the growth of N_n and N_r, where the upper and lower surfaces are obtained with (14) and (15) in Theorem 3, respectively. $N_t = 10$, $p_1 = 0.1$, $p_2 = 0.9$, $C_1 = 10$, and $C_3 = 0$.

The lower part of Fig. 2 shows the corresponding number of published forged messages of $\mathcal{P}$ with the increase of rounds, where the two aforementioned cases of $\mathcal{P}$'s behaviors, i.e., cases 1 and 2, are plotted. We can see that $\mathcal{P}$ can obtain the maximum payoff gap when it publishes the fifth forged message at the 13th round. The number of forged messages is achieved by first identifying the largest payoff gap between the *malicious* and *benign* strategies of $\mathcal{P}$ in the upper part of Fig. 2, then mapping the corresponding round number onto the lower part of the figure (as illustrated by the dashed line), and finally checking the number of forged messages. We also see that the exponentially increasing alert durations are effective to suppress forged messages, and that $\mathcal{P}$ is disincentivized

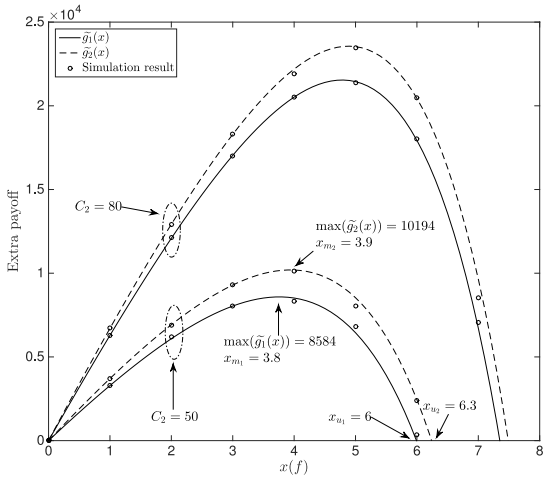


Fig. 4. Auxiliary continuous functions $\widetilde{g}_i(x)$, indicating the extra payoffs from forged messages with the growth of x, where $N_r = 90$, $N_n = 10$, $N_t = 10$, $C_1 = 10$, $C_3 = 5$, $p_1 = 0.1$, and $p_2 = 0.5$. Two values of C_2, i.e., 50 and 80, are considered. Every dot is the average extra payoff of 1000 independent simulations.

from further publishing any forged messages beyond the five forged messages. The payoff from the misbehaviors of $\mathcal{P}$ in case 1 with six forged messages and 38 genuine messages is equal to the payoff of 44 genuine messages, as highlighted by a circle. The payoff from the misbehaviors can be outrun by that from genuine messages when the number of published messages is more than 44.

Fig. 3(a) and (b) demonstrates Theorems 1 and 3, respectively. In the case that C_2 is between the two surfaces in each of the figures, $\mathcal{P}$ has an incentive to publish forged messages for an extra payoff but can be disincentivized by the *distrust* strategy from $\mathcal{A}$. Above the upper surface, $\mathcal{P}$ is expected to always publish forged messages for high payoffs, as stated in Theorems 1 and 3. Below the lower surface, $\mathcal{P}$ has no incentive to publish any forged messages at all, as stated in Theorems 1 and 3. From both figures, we can see that C_1 has a stronger impact on $\mathcal{P}$'s selection of its strategies than C_3, in the case of $N_t \ll N_r + N_n$. We also see there is a peak on the upper surface where $N_r = 100$ and $N_n = 0$. In other words, $\mathcal{P}$ would not infinitely publish forged messages when it does not have fans.

Fig. 4 validates Theorems 2 and 4 by plotting the auxiliary continuous functions, i.e., $\widetilde{g}_1(x)$ and $\widetilde{g}_2(x)$. We find that the extra payoff of $\mathcal{P}$, which is the payoff gap at the end of an alert duration between *malicious* and *benign* strategies that $\mathcal{P}$ takes (as defined in Section IV), first grows and then decreases rapidly. The extra payoff is upper-bounded, e.g., $\max(\widetilde{g}_1(x)) = 8584$ when $x_{m_1} = 3.8$, as shown in the figure. As revealed in Theorem 2, $\mathcal{P}$ only has an incentive to broadcast a limited number of forged messages, e.g., $x_{u_1} = 6$, for the extra payoff. Theorem 4 is validated when $x_{u_2} = 6.3$ and $x_{m_2} = 3.9$, which leads to $\max(\widetilde{g}_2(x)) = 10\,194$, in the case of message misclassification. We can also see that the maximum extra payoff and the maximum forged messages of $\mathcal{P}$ increase with C_2. Moreover, $\mathcal{P}$ can gain higher payoffs in the case where the genuine messages can be misclassified, as shown by the dashed curves.

We proceed to evaluate the impact of different parameters on the suppression of forged messages. Particularly, we will show

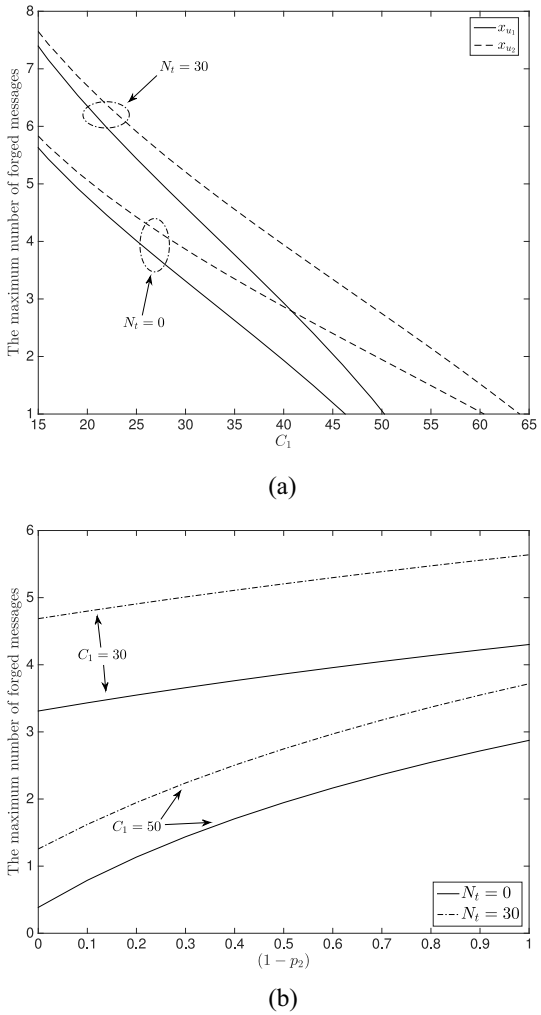


Fig. 5. Maximum number of forged messages of a malicious $\mathcal{P}$ with the growth of C_1 and $(1 - p_2)$, where $N_r = 90$, $N_n = 10$, $C_3 = 5$, $p_1 = 0.1$, and two values of N_t, i.e., 0 and 30, are considered. (a) Maximum number of forged messages of a malicious $\mathcal{P}$ with the growth of C_1, where $p_2 = 0.5$. (b) Maximum number of forged messages of a malicious $\mathcal{P}$ with the growth of $(1 - p_2)$. Two values of C_1, i.e., 30 and 50, are considered.

that both increasing the reward for genuine messages, i.e., C_1, and reducing the misclassification of genuine messages, i.e., decreasing $(1 - p_2)$, are helpful to suppress forged messages and to incentivize the publication of genuine messages, as will be shown in Fig. 5. Moreover, $\mathcal{A}$ can encourage $\mathcal{P}$ to publish genuine messages by improving the detection rate of forged messages, i.e., $(1 - p_1)$, as will be shown in Fig. 6.

Fig. 5(a) shows the maximum number of forged messages with the growth of C_1. Two values of N_t, i.e., 0 and 30, are considered. From the figure, we can see that the maximum number of forged messages decreases with the increasing value of C_1. The maximum number of forged messages can be less than 1, indicating that $\mathcal{P}$ has no incentive to publish forged messages. In other words, the increasing value of C_1 can effectively suppress forged messages. We also see that the misclassification of genuine messages gives malicious $\mathcal{P}$ chances to publish more forged messages, since the dashed curves are above the solid curves, especially for large C_1. For a small reward of genuine messages, e.g., $C_1 = 15$ in Fig. 5(a), hiring bots helps $\mathcal{P}$ publish more forged messages without being detected. This

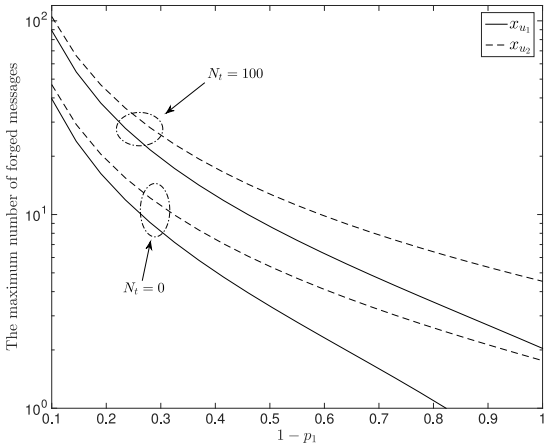


Fig. 6. Maximum number of forged messages of $\mathcal{P}$, i.e., x_{u_1} and x_{u_2}, by publishing forged messages, where $N_r = 90$, $N_n = 10$, $C_1 = 30$, $C_2 = 50$, $C_3 = 5$, and $p_2 = 0.5$. $N_t = 0$ and 100, are considered.

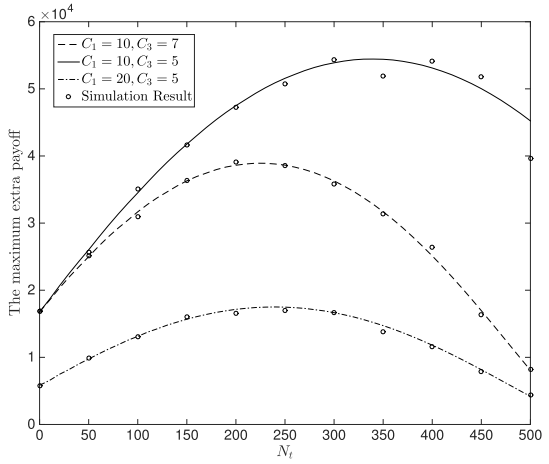


Fig. 7. Maximum extra payoff of $\mathcal{P}$ in the misclassification-free game with the growth of N_t, where $N_r = 90$, $N_n = 10$, $C_2 = 50$, and $p_1 = 0.5$. Every dot is an average result of 2000 independent runs.

is also confirmed in Fig. 5(b), where the maximum number of forged messages grows with the increasing misclassification probability of genuine messages, i.e., $(1 - p_2)$. Reducing the misclassification of genuine messages can help suppress forged messages.

Fig. 6 plots the maximum number of forged messages, i.e., x_{u_1} and x_{u_2}, with the growth of the detection rate of forged messages, i.e., $1 - p_1$. $N_t = 0$ and 100 are considered. The y-axis is in a logarithmic scale. We can see that the improvement of the detection rate can effectively suppress forged messages when the detection rate is low, e.g., $1 - p_1 = 0.1$. We also see that the maximum number of forged messages can be more than one even in the case where $1 - p_1 = 1$, especially where many bots are hired, i.e., $N_t = 100$. As a result, the improvement of the detection rate cannot stop $\mathcal{P}$ from publishing forged messages. Based on Figs. 5(a) and 6, $\mathcal{A}$ can suppress forged messages by improving the detection rate in the beginning, and by increasing the reward for genuine messages.

Fig. 7 plots the maximum extra payoff of $\mathcal{P}$ with the growing number of bots in misclassification-free games. We can see that bots can increase the maximum extra payoff from the start, but decrease the maximum extra payoff later. Thus, there

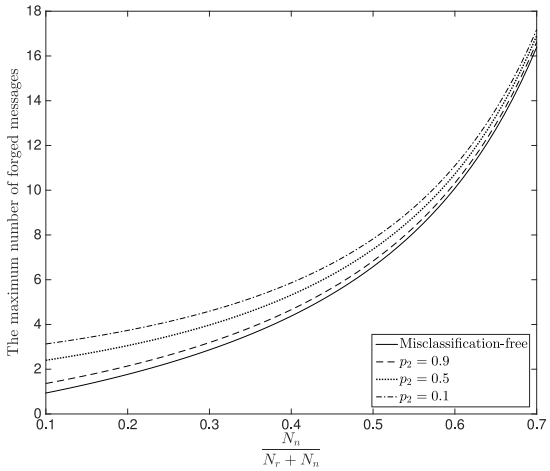


Fig. 8. Maximum number of forged messages with the growth of the ratio of fans, i.e., $[N_n/(N_r + N_n)]$, where $N_r + N_n = 100$, $N_t = 10$, $C_1 = 30$, $C_3 = 5$, $C_2 = 50$, and $p_1 = 0.1$.

exists the best number of bots for the highest extra payoff. This is because the bots benefit the extra payoff by reducing the detection rate of forged messages. The detection rate is non-negative. As a result, the benefit of bots is upper bounded. However, the cost of bots increases linearly with the number of bots. In extreme cases, the benefit from the decreased detection rate cannot counteract the cost of bots, e.g., $N_t = 500$ in the case of $C_1 = 10$ and $C_3 = 7$. It can be found that the bots are more effective in the case where genuine messages provide lower payoffs, as can be seen by comparing the curves of $C_1 = 10$ and $C_1 = 20$ under the same unit payoff of forged messages (i.e., $C_2 = 50$). We also see that low-cost bots, e.g., $C_3 = 5$ versus $C_3 = 7$, can bring a higher extra payoff. As a result, the malicious $\mathcal{P}$ has an incentive to hire bots if the bots are low-cost and forged messages bring high payoffs.

Fig. 8 shows the maximum number of forged messages with the growing ratio of fans, i.e., $[N_n/(N_r + N_n)]$. We can see that $\mathcal{P}$ can publish more forged messages with more fans, because fans persistently trust the messages from $\mathcal{P}$. We note that the increasing probability of misclassification, i.e., from $p_2 = 0.9$ to $p_2 = 0.1$, also offers chances for $\mathcal{P}$ to publish more forged messages, especially in the case of a low ratio of fans. This is because the misclassification enlarges the payoff gap between forged and genuine messages by reducing rewards for genuine messages.

VI. Conclusion

In this paper, infinitely repeated games were developed to evaluate the effect of risk alerts on the inhibition of forged messages in the presence of multiple types of subscribers and possible miss-detection and false alarm of forged messages. The proposed games captured the interactions between a message publisher and the network administrator in OSNs. In the absence and presence of misclassification on genuine messages, sufficient conditions under which the publisher is disincentivized from publishing forged messages were identified. Closed-form expressions were derived for the maximum number of forged messages of a malicious publisher. Confirmed by simulations, our analysis indicates that forged messages

can be suppressed by improving the payoffs for genuine messages, increasing the cost of bots, and/or reducing the payoffs for forged messages. The increasing detection probability of forged messages or decreasing misclassification probability of genuine messages can also have a strong impact on the suppression of forged messages.

APPENDIX

By letting $h(\lambda_1) = \lambda_1(2-q_1)^{-\lambda_1}\ln(2-q_1)$, $\lambda_1 > 0$, we have

$$0 \le h(\lambda_1) \le h\left(\frac{1}{\ln(2-q_1)}\right) = \frac{1}{e}. \tag{22}$$

Here, $h(\lambda_1) \ge 0$ because every item is no less than 0. The maximum value of $h(\lambda_1)$ is achieved when $\lambda_1 = (1/[\ln(2-q_1)])$. This can be obtained by letting $([d\ln(h(\lambda_1))]/d\lambda_1) = 0$. $\ln(h(\lambda_1))$ is given by

$$\begin{aligned}\ln(h(\lambda_1)) &= \ln\left(\lambda_1(2-q_1)^{-\lambda_1}\ln(2-q_1)\right)\\ &= \ln(\lambda_1) - \lambda_1\ln(2-q_1) + \ln(\ln(2-q_1)).\end{aligned} \tag{23}$$

Therefore, its derivative is given by

$$\frac{d\ln(h(\lambda_1))}{d\lambda_1} = \frac{1}{\lambda_1} - \ln(2-q_1). \tag{24}$$

REFERENCES

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *J. Comput. Mediated Commun.*, vol. 13, no. 1, pp. 210–230, 2007.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Manag. Sci.*, vol. 62, no. 12, pp. 3412–3427, 2016.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *Proc. 7th Annu. Collaboration Electron. Messaging Anti Abuse Spam Conf. (CEAS)*, 2010, pp. 1–10.

[5] J. Malbon, "Taking fake online consumer reviews seriously," *J. Consum. Policy*, vol. 36, no. 2, pp. 139–157, Jun. 2013.

[6] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: A network-based spam detection framework for reviews in online social media," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 7, pp. 1585–1595, Jul. 2017.

[7] Y. Wang, A. V. Vasilakos, J. Ma, and N. Xiong, "On studying the impact of uncertainty on behavior diffusion in social networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 2, pp. 185–197, Feb. 2015.

[8] X. Wang, W. Ni, K. Zheng, R. P. Liu, and X. Niu, "Virus propagation modeling and convergence analysis in large-scale networks," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2241–2254, Oct. 2016.

[9] X. Wang *et al.*, "Group-based susceptible-infectious-susceptible model in large-scale directed networks," *Security Commun. Netw.*, vol. 2019, p. 9, Jan. 2019. [Online]. Available: https://doi.org/10.1155/2019/1657164

[10] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016.

[11] C. Chen, K. Wu, V. Srinivasan, and X. Zhang, "Battling the Internet water army: Detection of hidden paid posters," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min. (ASONAM)*, Aug. 2013, pp. 116–120.

[12] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Predicting susceptibility to social bots on Twitter," in *Proc. IEEE 14th Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2013, pp. 6–13.

[13] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders, "Social networks and context-aware spam," in *Proc. ACM Conf. Comput. Support. Cooper. Work (CSCW)*, New York, NY, USA, 2008, pp. 403–412.

[14] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards detecting compromised accounts on social networks," *IEEE Trans. Depend. Secure Comput.*, vol. 14, no. 4, pp. 447–460, Jul. 2017.

[15] V. L. Rubin, "Deception detection and rumor debunking for social media," in *The SAGE Handbook of Social Media Research Methods*, Los Angeles, CA, USA: SAGE, 2017, p. 342.

[16] X. Zha *et al.*, "The impact of link duration on the integrity of distributed mobile networks," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2240–2255, Sep. 2018.

[17] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?" in *Proc. Int. AAAI Conf. Weblogs Soc. Media (ICWSM)*, 2013, pp. 409–418.

[18] (2018). *How Is Facebook Addressing False News Through Third-Party Fact-Checkers?* [Online]. Available: https://www.facebook.com/help/1952307158131536?locale=en_US

[19] M. Sirivianos, K. Kim, and X. Yang, "SocialFilter: Introducing social trust to collaborative spam mitigation," in *Proc. INFOCOM*, Apr. 2011, pp. 2300–2308.

[20] S. Sedhai and A. Sun, "Semi-supervised spam detection in Twitter stream," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 1, pp. 169–175, Mar. 2018.

[21] M. Nitti, R. Girau, and L. Atzori, "Trustworthiness management in the social Internet of Things," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1253–1266, May 2014.

[22] X. Wang *et al.*, "Survey on blockchain for Internet of Things," *Comput. Commun.*, vol. 136, pp. 10–29, Feb. 2019.

[23] L. Liu and H. Jia, "Trust evaluation via large-scale complex service-oriented online social networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 11, pp. 1402–1412, Nov. 2015.

[24] J. D. Falk and M. S. Kucherawy, "Battling spam: The evolution of mail feedback loops," *IEEE Internet Comput.*, vol. 14, no. 6, pp. 68–71, Nov. 2010.

[25] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Fighting spam on social Web sites: A survey of approaches and future challenges," *IEEE Internet Comput.*, vol. 11, no. 6, pp. 36–45, Nov. 2007.

[26] X. Zha, W. Ni, K. Zheng, R. P. Liu, and X. Niu, "Collaborative authentication in decentralized dense mobile networks with key predistribution," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 10, pp. 2261–2275, Oct. 2017.

[27] K. Zhang, X. Liang, R. Lu, and X. Shen, "PIF: A personalized fine-grained spam filtering scheme with privacy preservation in mobile social networks," *IEEE Trans. Comput. Soc. Syst.*, vol. 2, no. 3, pp. 41–52, Sep. 2015.

[28] M. Parameswaran, H. Rui, and S. Sayin, "A game theoretic model and empirical analysis of spammer strategies," in *Proc. Collaboration Electron. Messaging AntiAbuse Spam Conf. (CEAS)*, vol. 7, 2010, pp. 1–7.

[29] F. Zhou, R. J. Jiao, and B. Lei, "Bilevel game-theoretic optimization for product adoption maximization incorporating social network effects," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 8, pp. 1047–1060, Aug. 2016.

[30] H. Abbass, G. Greenwood, and E. Petraki, "The N-player trust game and its replicator dynamics," *IEEE Trans. Evol. Comput.*, vol. 20, no. 3, pp. 470–474, Jun. 2016.

[31] X. Ruan, Z. Wu, H. Wang, and S. Jajodia, "Profiling online social behaviors for compromised account detection," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 1, pp. 176–187, Jan. 2016.

[32] C. Chen *et al.*, "Statistical features-based real-time detection of drifted Twitter spam," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 4, pp. 914–925, Apr. 2017.

[33] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving Twitter spammers," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 8, pp. 1280–1293, Aug. 2013.

[34] F. Benevenuto *et al.*, "Practical detection of spammers and content promoters in online video sharing systems," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 688–701, Jun. 2012.

[35] T. Zhu *et al.*, "Beating the artificial chaos: Fighting OSN spam using its own templates," *IEEE/ACM Trans. New.*, vol. 24, no. 6, pp. 3856–3869, Dec. 2016.

[36] *Groups*. (2018). [Online]. Available: https://www.facebook.com/help/1629740080681586

[37] E. Kalai and E. Lehrer, "Rational learning leads to nash equilibrium," *Econometrica*, vol. 61, no. 5, pp. 1019–1045, 1993. [Online]. Available: http://www.jstor.org/stable/2951492

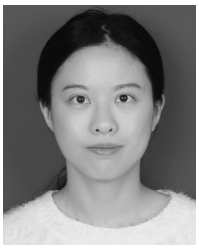
[38] T. G. Papaioannou and G. D. Stamoulis, "An incentives' mechanism promoting truthful feedback in peer-to-peer systems," in *Proc. CCGrid*, vol. 1, May 2005, pp. 275–283.

[39] J. L. Coolidge, "The story of the binomial theorem," *Amer. Math. Monthly*, vol. 56, no. 3, pp. 147–157, 1949.

[40] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996. [Online]. Available: https://doi.org/10.1007/BF02124750

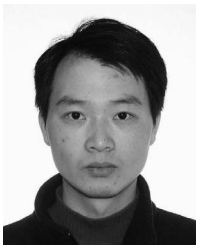
Xu Wang received the B.S. degree in computer science from Beijing Information Science and Technology University, Beijing, China, in 2010. He is currently pursuing the Ph.D. degree in information security with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, and the Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW, Australia.

He visited CSIRO, Sydney, NSW, Australia, in 2014. His current research interests include graph theory, Markov theory, intrusion detection, cyber security, and blockchain.



Xuan Zha received the B.S. degree in mathematics and applied mathematics from Anhui University, Hefei, China, in 2010, and the Ph.D. degree in information security from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree in information security with the Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW, Australia.

She is currently with the Institute of ICT Security Research, China Academy of Information and Communications Technology, Beijing. Her current research interests include wireless network security, blockchain, Markov theory, and blockchain and vehicular ad hoc networks



Wei Ni (M'09–SM'15) received the B.E. and Ph.D. degrees in electronic engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively.

He is currently a Team Leader with CSIRO, Sydney, NSW, Australia, and an Adjunct Professor with the University of Technology Sydney, Ultimo, NSW, Australia. He was a Post-Doctoral Research Fellow with Shanghai Jiao Tong University, Shanghai, from 2005 to 2008, a Deputy Project Manager with the Bell Labs R&I Center, Alcatel/Alcatel-Lucent, Boulogne-Billancourt, France, from 2005 to 2008, and a Senior Researcher with the Devices Research and Development, Nokia, Espoo, Finland, from 2008 to 2009. He also holds honorary positions with the University of New South Wales, Sydney, and Macquarie University, Sydney. His current research interests include stochastic optimization, game theory, graph theory, as well as their applications to network and security.

Dr. Ni has been serving as the Vice Chair for IEEE NSW VTS Chapter and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018, a Secretary of IEEE NSW VTS Chapter from 2015 to 2018, the Track Chair for VTC-Spring 2017, the Track Co-Chair for IEEE VTC-Spring 2016, and the Publication Chair for BodyNet 2015. He also served as the Student Travel Grant Chair for WPMC 2014, a Program Committee Member of CHINACOM 2014, a TPC Member of IEEE ICC'14, ICCC'15, EICE'14, and WCNC'10.

Ren Ping Liu (M'09–SM'14) received the B.E. degree in telecommunication engineering and the M.E. degree in computer engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1985 and 1988, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Newcastle, Callaghan, NSW, Australia, in 1996.

He is currently a Professor and the Head of Discipline of Network and Cybersecurity with the University of Technology Sydney, Ultimo, NSW, Australia. He was a Principal Scientist and the Research Leader with CSIRO, Sydney, NSW, Australia, where he led wireless networking research activities. He is also the Co-Founder and CTO of Ultimo Digital Technologies Pty, Ltd., Sydney, developing IoT and blockchain. He specializes in system design and modeling and has delivered networking solutions to a number of government agencies and industry customers. He has over 150 research publications, and has supervised over 30 Ph.D. students. His current research interests include wireless networking, cybersecurity, and blockchain.

Prof. Liu was a recipient of the Australian Engineering Innovation Award and the CSIRO Chairman Medal. He was the Founding Chair of IEEE NSW VTS Chapter. He served as the Technical Program Committee Chairs and the Organizing Committee Chairs in a number of IEEE conferences.

Y. Jay Guo (F'14) received the B.E. degree (First Class Hons.) in electronic and electrical engineering and the M.E. degree in microwave engineering from Xidian University, Xian, China, in 1982 and 1984, respectively, and the Ph.D. degree in antennas and propagation from Xian Jiaotong University, Xian, in 1987.

He held various senior technology leadership positions in Fujitsu, London, U.K., Siemens, Cambridge, U.K., and NEC, Ruislip, U.K. He served as the Director of CSIRO for over nine years, where he directed a number of ICT research portfolios. In 2014, he joined the University of Technology Sydney, Ultimo, NSW, Australia, where he is a Distinguished Professor and the Founding Director of Global Big Data Technologies Centre. He has published over 400 research papers and holds 24 patents in antennas and wireless systems. His current research interests include antennas, mm-wave and THz communications and sensing systems as well as big data technologies.

Prof. Guo was a recipient of the most prestigious Australian National Awards, and was named one of the Most Influential Engineers in Australia, in 2014 and 2015. He has chaired numerous international conferences. He is the Chair Elect of International Steering Committee, International Symposium on Antennas and Propagation (ISAP). He was the International Advisory Committee Chair of IEEE VTC2017, the General Chair of ISAP2015, iWAT2014, and WPMC2014, and the TPC Chair of 2010 IEEE WCNC, and 2012 and 2007 IEEE ISCIT. He served as a Guest Editor for special issues on Antennas for Satellite Communications and Antennas and Propagation Aspects of 60–90 GHz Wireless Communications, both in IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, Special Issue on Communications Challenges and Dynamics for Unmanned Autonomous Vehicles, in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and Special Issue on 5G for Mission Critical Machine Communications, in *IEEE Network Magazine*. He is a fellow of the Australian Academy of Engineering and Technology and IET, and a member of the College of Experts of Australian Research Council.

Xinxin Niu was born in 1963. She received the Ph.D. degree in electronic engineering from the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, in 1997.

She is currently a Professor with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China, and also with the Data Security Center, State Key Laboratory of Public Big Data, Guiyang, China. She has published over 80 papers and 8 books. Her current research interests include information security, data hiding, digital watermarking, and digital signal processing.

Kangfeng Zheng received the Ph.D. degree in information and signal processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 2006.

He is currently an Associate Professor with the School of Cyberspace Security, Beijing University of Posts and Telecommunications. His current research interests include network security and network data analysis.