Mitigating Influence of Disinformation Propagation Using Uncertainty-Based Opinion Interactions

Zhen Guo, *Graduate Student Member, IEEE*, Jin-Hee Cho, *Senior Member, IEEE*, and Chang-Tien Lu, *Senior Member, IEEE*

***Abstract*— For decades, the spread of disinformation in online social networks (OSNs) has been a serious social issue. Disinformation via social media can easily mislead people's beliefs toward or against an event that may mislead their behaviors based on the misbeliefs. The game theory approaches have been proposed under dynamic settings to limit the adverse influences of disinformation. It is a challenge to expand the users' game strategies from the spreading decisions to the possible opinion updating choices. This work proposes a game-theoretic opinion framework that can formulate dynamic opinions by a belief model called *Subjective Logic* (SL) and provide opinion updates on five types of users' interactions on OSN platforms. The opinions are updated based on user choices and user types through the game interactions among legitimate users, attackers, and a defender in an OSN. Via the extensive simulation experiments, the effectiveness of the opinion models of five decision-makers (DMs) is analyzed in terms of users believing or disbelieving disinformation in an epidemic model with parameter optimization. Our results show that while homophily-based DMs (H-DMs) introduce the highest opinion polarization, uncertainty-based DMs (U-DMs) can effectively filter untrustworthy users propagating disinformation.**

***Index Terms*— Disinformation, influence, opinion dynamics, opinion/network polarization, subjective opinion, uncertainty.**

NOMENCLATURE

$\omega_i = \{b, d, u, a\}$	SL opinion by belief, disbelief, uncertainty, and base rate.
$P(b_i)$ and $P(d_i)$	User i's projected belief and projected disbelief from ω_i.
ω_F, ω_T, ω_U	False, true, and uncertain initial opinion.
$\oplus, \otimes$	Consensus and trust operator from SL.
$c_i^j (uc_i^j$ or $hc_i^j)$	SL's uncertainty or homophily-based discounting factor.
$\ddot{\omega}_i$	Uncertainty maximized opinion of user i.
U-DM, H-DM, E-DM	Uncertainty, Homophily, and Encounter-based decision makers.
A-DM, HE-DM	Assertion and Herding-based decision makers.
ξ	Threshold of uncertainty maximization.
P_i^f and P_i^p	User i's feeding preference and posting preference.
PD_{ij}	Projected discrepancy between two opinions.
ϕ_i	Threshold to accept or request a friend.
$EP_k^{A_i}$	Expected payoff of an attacker taking strategy k.
EP_ℓ^D	Expected payoff of a defender taking strategy ℓ.
$EP_m^{U_i}$	Expected payoff of a decision maker taking strategy m.
c_ℓ	Defender's cost of strategy ℓ.
$u_{k\ell m}^{ij}$, $u_{\ell k}^D$	Utility value of an element in $EP_k^{A_i}$ and EP_ℓ^D.
$p_{U_i}^{A_j}$	Probability of a decision-maker j as an attacker.
N_R	Number of reports to alert a defender.
ρ	Tolerance to report a malicious user.
η	Learning rate in gradient decent.
I	Number of interactions in simulation.
N	Number of nodes in the OSN.

Manuscript received 9 July 2022; revised 15 October 2022; accepted 25 November 2022. Date of publication 6 December 2022; date of current version 3 April 2023. This work was supported by NSF under Grant III-2107450 and Grant CNS-2141095. *(Corresponding author: Zhen Guo.)*

The authors are with the Department of Computer Science, Virginia Tech, Falls Church, VA 22043 USA (e-mail: zguo@vt.edu; jicho@vt.edu; clu@vt.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCSS.2022.3225375, provided by the authors.

Digital Object Identifier 10.1109/TCSS.2022.3225375

I. INTRODUCTION

THANKS to the popularity of online social networks (OSNs) and their highly advanced features, communications via social media or OSNs become part of our daily life. In various OSN platforms, people exchange their opinions without high confidence or share them without going through any verification process. It is well known that disseminating false information, including unverified rumors, misinformation, or disinformation, can easily destroy individuals' reputations or lives. In this work, we use the terms false information or disinformation interchangeably where it refers to false information propagated with malicious intent [24]. As a result, manipulating public opinions toward sensitive issues can easily happen when disinformation propagates extremely fast. Further, disseminating disinformation can be highly detrimental in affecting critical decision-making processes in our real life at the levels of individuals, communities, and global society [5], [16], such as in elections, pandemics, health, or education.

In an OSN, a person can take advantage of different activities to connect to other users and share opinions. The level of a person's acceptance of a given opinion has been estimated based on various aspects, such as personality traits (e.g., agreeableness, open-mindedness, and stubbornness), a tendency to relying on others' opinions (e.g., herding), homophily (e.g., like-mindedness), competence (e.g., domain expertise), or confidence (e.g., certainty) [7], [8], [20]. There has been a rich volume of approaches modeling and simulating the behaviors of OSN users in updating their opinions and propagating (false) information [41], [42], [43]. Based on OSN user's bounded rationality caused by inherent cognitive bias or incapability of humans [1], [17], [32], [37], [40], most existing game theory diffusion models are grounded by users' decisions of spreading rumors or not. To limit disinformation cascades, the incentives and punishments of spreading unverified rumors were accessed by environment factors, network topology, neighbors' strategy preferences, and individual factors.

However, mitigating disinformation propagation in the existing game models by network users as decision-makers (DMs) meets several main challenges: 1) *Users' decisions of updating uncertain opinions.* More real user behaviors, such as updating opinions from social interactions, can serve as a complement of spreading decision. However, little work has leveraged game theory to justify the significant tendency of users' information processing. 2) *Defense from both individual and network levels.* Individual users' type of subjective opinion updates can mitigate the effect of disinformation in the OSN by the dynamic opinion-based epidemic model. 3) *Network polarization.* It is critical to address the divergent influences of users' opinion updates on network communities as disinformation is often related to the polarization of users [29].

This work aims to demonstrate how OSN users' rational information processing behaviors based on various opinion update criteria and models can influence the mitigation of disinformation propagation and further impact network dynamics and opinion polarization. To model OSN users' social interactions by real individual behavioral features, this work proposes an opinion game framework with three players, including an attacker, defender, and user. An attacker refers to malicious users (i.e., *false informers*) who have the intent of disseminating disinformation for misleading legitimate users to (dis)believe in false (or true) information. A defender may be an OSN system administrator whose policy ensures a safe and trustworthy OSN environment. Users mean other legitimate OSN DMs who interact with their friends and make rational decisions to update opinions in this game model.

The followings are the **key contributions** in this study.

1) We develop a robust belief model to formulate users' subjective and dynamic opinions by *Subjective Logic* (SL) theory. This model characterizes opinion update rules of five types of DMs' disinformation processing.
2) We propose a game-theoretic opinion update strategies framework. It defines the goals, strategies, and payoffs of three networked agent roles. This opinion game investigates how different ways of updating opinions can help DMs combat disinformation propagation.
3) We demonstrate each player role's (i.e., an attacker, defender, or user) best strategy by decision-making under uncertainty in an OSN. The best strategy of each player is compared to the strategy identified by Nash equilibrium (NE). NE unrealistically assumes all players can have correct beliefs about the moves of their opponents.
4) We optimize epidemic model parameters in an effective gradient decent algorithm. Since the opinion dynamics of all the network agents can reflect the transition of states in susceptible-infected-recovered (SIR) model, we optimize their infection (i.e., believing in disinformation) and recovery (i.e., disbelieving in disinformation) rates.
5) We analyze the opinion and network polarization under five opinion models. The dominant opinions of all users in the identified modularity-based communities [9] indicate which opinion model can defend disinformation better.

This work significantly extends the previous work [10] and [11] with the following additional contributions.

1) We distinguish the homophily-type users' game strategies preferences from other DM types in the same network. This extended study elaborates how DMs' choices can combat disinformation.
2) We validate the defense of disinformation in the network level by an uncertain opinion-based SIR model. The SIR model quantifies the rates of disinformation spreading and recovery in the given network where various types of DMs update their opinions.
3) We extend the defense analysis to the simulations of different initial attacker ratios in the network and demonstrate the network polarization caused by a high ratio of attackers.
4) We clarify the defense against disinformation propagation based on the three network views in Fig. 1, including a number of DMs' social interactions and opinion updating decisions, each DMs opinion scale in opinion polarization network, and the overall epidemic status in the SIR network. This enables users' interactions and opinion models to affect disinformation propagation, which also influences the segregation of two parties in believing true or false information and opinion polarization in a given OSN.

This article is structured by the following sections. Section II discusses the research on existing information diffusion game models and the effect of network-level polarization cause by disinformation. Section III introduces the SL belief model and the five opinion models used in our work. We also describes DMs' social interactions and accordingly decisions in sharing information and making friends. Section IV describes the game-theoretic opinion update framework of three agent roles, attackers, users, and defender, and their aims, strategies, and payoff functions. Section V-A describes the experimental setup, metrics, and experiment settings. Section V-B presents the findings and discussions of the simulation results. Section VI conclude our article with the summarized key findings and future research directions.

II. Related Work

In this section, we provide the brief overview of the state-of-the-art research about opinion models, game theoretic information diffusion, the effect of disinformation on polarization.

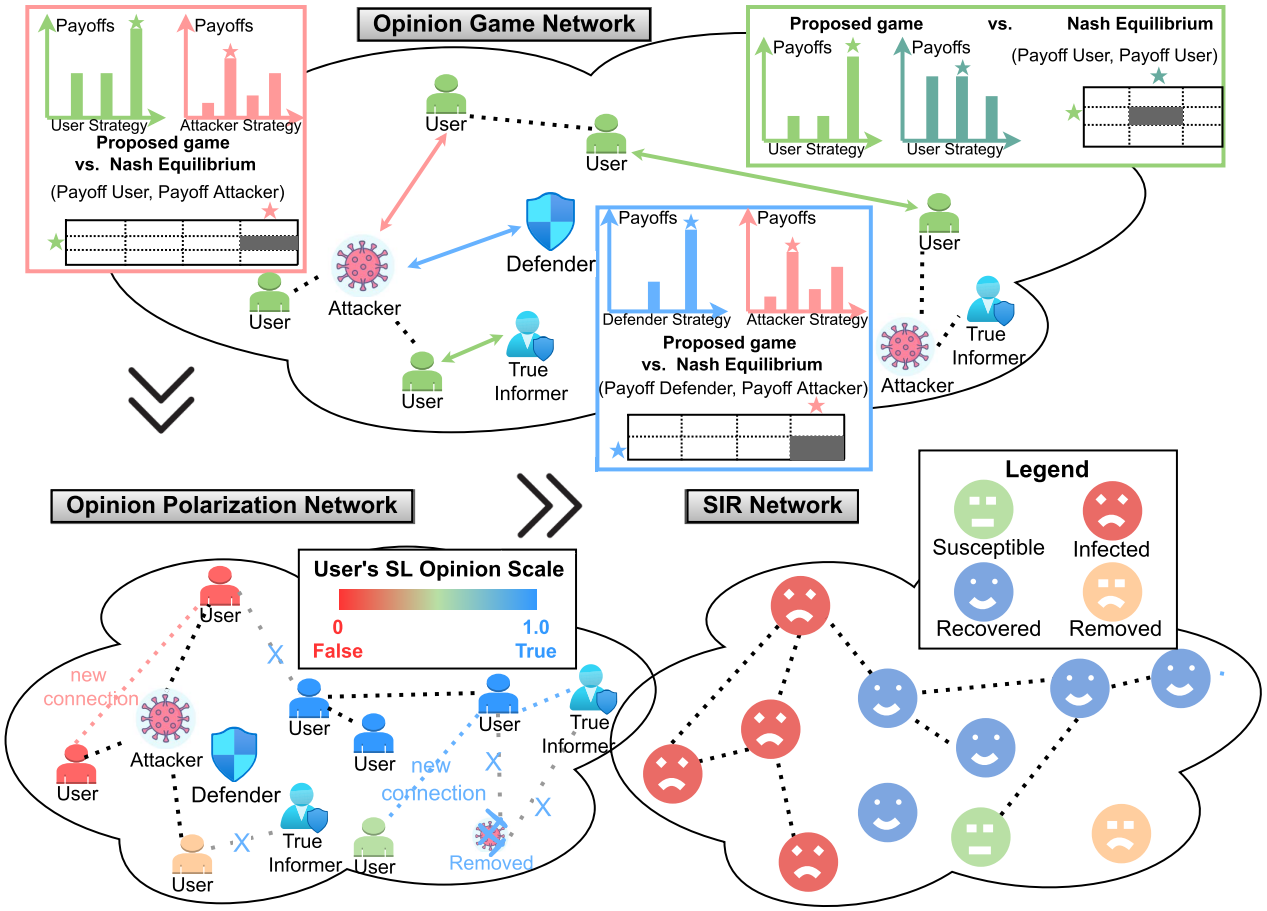

Fig. 1. Overview of the proposed opinion update game in an OSN. The top figure illustrates the three game examples in one interaction, compared to their related NEs. The middle figure represents the opinion polarization in the last interaction. The bottom figure depicts the network status of the middle figure by the epidemic model.

A. Opinion Models

Opinion models show how OSN users update their opinions when encountering another OSN user with a different opinion. An assertion model provides two criteria for users to update an opinion [44]: 1) *the amount of knowledge* and 2) *the degree of belief*. The users can determine the level of knowledge exchange regarding their forgetfulness, learning capability, and trust in another user. A herding-based opinion model is also developed by considering the amount of pair-wise social interactions with all other friends [23]. Uncertainty-based opinion model was also discussed where a user could update an opinion when an encountered user's opinion has higher certainty, such as high expertise [7]. Similarly, an uncertainty range interval-based opinion model is used for a user to update an opinion [39]. This model calculated the distance between two opinions based on the range of the uncertainty interval length if an agent has uncertain opinions. However, the works above [7], [23], [44], and [39] did not consider a user's rational behaviors in updating opinions. Unlike those works, we leveraged game theory in diverse opinion models to investigate the impact of the user's rational behavior on the spread of disinformation.

B. Game Theoretic Information Diffusion

Yang [37] formulated an opinion diffusion game where players with binary opinions can take the benefits of a "cooperative" or "defective" strategy to reach an opinion consensus. Li et al. [17] studied the payoffs of rumor diffusion decisions with a punishment cost in an OSN. Askarizadeh et al. [1] examined the dynamics of rumor diffusion and control strategies (e.g., spreading either rumor, anti-rumor, or neutral messages) by the evolutionary game theory (EGT). A few factors, such as an attitude toward rumor, the anxiety of society, or strength of rumor and anti-rumor, can influence the evolutionary stable state by the ordinary differential equation (ODE). Xiao et al. [32] modeled the competition of rumors and anti-rumor messages for the same recipient and investigated the user psychology factors by EGT. Zhang et al. [40] modeled the evolution of neighbors' reputation, which influenced the payoffs of a user's spreading decision to reduce the negative effects of malicious users.

Szabó and Tőke [25] investigated the Fermi updating rule to understand the likelihood of strategy imitation, estimated based on the benefit of neighbors' fitness. Li et al. [17] studied rumor propagation in OSNs by considering diverse social and individual characteristics, including friend relationships, cognitive judgment capability, strategy imitation, and rumor propagation cost. Askarizadeh et al. [1] also examined rumor propagation in OSNs by considering individuals' attitudes or awareness toward rumors, a level of community anxiety, and the spread intensity of rumors and anti-rumor cascades. Huang et al. [14] used diverse game models to investigate cost-effective defense strategies against rumor propagation.

Yoshikawa et al. [38] estimated users' opinions using Bayesian networks based on trust and reliability in users.

Unlike the above works [1], [14], [17], [25], [32], [37], [38], [40], our work used game theory to model users' opinion propagation behaviors and strategies to deal with disinformation. In addition, we investigated how the game-theoretic opinion models and rational user behaviors affect opinion dynamics and the patterns of disinformation spreading.

Xiao et al. [33] proposed a Diffusion2pixel algorithm to transform the relationship network of online users with topic diffusion into an image pixel matrix. This work showed the performance of their proposed approach in effectively predicting the group diffusion trends of rumor and considered the competitive relationship between rumor and anti-rumor. Li et al. [19] developed a tripartite cognitive model that disseminates information based on the symbiosis and antagonism of multiple types of messages and the polymorphism of users' cognitive process under the influence of multimessages. This work proved the non-coexistence relationship of multiple messages while those messages can form a game situation. Xiao et al. [34] explored the adversarial game relationship between rumor and anti-rumor in the propagation process and proposed a rumor propagation model. The proposed model is designed to enhance homomorphism data in the sample space while using the evolutionary game to devise a mutual influence model of rumor and anti-rumor, and predict the group behavior with rumor topics. However, the above works [19], [33], and [34] did not consider uncertain opinion updates in dealing with disinformation where rational agents may have different utility functions to maximize their objectives as our work does.

C. Effect of Disinformation on Polarization

False information spreading can increase the polarization of users [29] while polarization can expedite the diffusion of false information [3]. Polarized users exposed to similar content have a divergent and longer response time when posting fake news [30]. The network can be segmented into several polarized groups because of the echo chambers and users being interconnected within the same group [22]. Polarized social network structures can also significantly reduce access to social capital in terms of relational and cognitive social capital [22]. A user's activity on false information diffusion correlates with homophily [3] in opinions. Polarization can predict homophile clusters because users connect with a similar polarization. In the homophile clusters, false information can be easily propagated [3]. Unlike the above works [3], [22], [29], [30], we first investigate how certain types of user interactions can reduce opinion polarization caused by disinformation in an OSN.

The relationships between homophily and disinformation propagation have been examined to study their effect on opinion or network polarization. A large community tends to have high homophily [12]. Intracommunity enables information diffusion faster than intercommunity as it exposes people to interact with other communities. This phenomenon can be easily observed in political campaigns where disinformation is exploited to trigger and increase conflicts and break social ties and social capital between different parties [2]. However, how disinformation propagation influences social capital has not been deeply studied [27]. Unlike the works above [2], [12], [27], our work studies how users' rational interaction models can influence the segregation of opinion communities where two extreme parties believe in true or false information.

III. Uncertainty-Based Opinion Model

This section uses *SL* [8], [15] to formulate each user's subjective opinion. We describe the key dimensions and initialization of SL, users' different preferences on exchanging and updating opinions, and user interactions in an OSN. All the variables in this model are summarized in Nomenclature.

A. Opinion Formation

For a given proposition, such as a piece of news, an opinion in SL, $\omega = \{b, d, u, a\}$, is defined by four dimensions as

$$b, d, u, a \in [0, 1]^4, \quad b + d + u = 1 \tag{1}$$

where the degree of belief (i.e., *pro*, agree, b) means that an agent believes the given proposition is true without knowing the real truth. The degree of disbelief (i.e., *con*, disagree, d) describes that the agent opposes the proposition, thus disbelieving it. The u refers to uncertainty, often called *vacuity*, representing uncertainty due to a lack of evidence. The *base rate* a reflects an agent's prior knowledge which can represent expertise or bias [15]. In SL, an agent can update his opinion including base rate a by interacting with other agents. Each agent's initial opinion is formulated by observed evidence following the following mapping rule:

$$b = \frac{r}{r + s + W}, \quad d = \frac{s}{r + s + W}, \; u = \frac{W}{r + s + W} \tag{2}$$

where r is the amount of positive evidence and s is the negative evidence for a certain proposition. The W is the amount of uncertain evidence as inherent errors from a system. Based on agent i's binomial opinion $\omega_i = \{b_i, d_i, u_i, a_i\}$, the expectations of projected belief $P(b_i)$ and disbelief $P(d_i)$ are

$$P(b_i) = b_i + a_i u_i, \quad \text{and } P(d_i) = d_i + (1 - a_i)u_i \tag{3}$$

where $P(b_i) + P(d_i) = 1$. The a_i is critical in interpreting uncertainty u_i when b_i and d_i are almost the same.

B. Initialization of Opinions

In the beginning, the zealots [28] of *true informers* or *false informers*, who support two extreme opinions as *true opinion*, $\omega_T = \{b \to 1, d \to 0, u \to 0, a = 1\}$, and *false opinion* (i.e., disinformation), $\omega_F = \{b \to 0, d \to 1, u \to 0, a = 0\}$. The initial opinions of the rest of users are defined as *Uncertain opinion*, by $\omega_U = \{b \to 0, d \to 0, u \to 1, a = 0.5\}$, without showing strong preference.

C. Opinion Update

When two users i and j interact, with opinions ω_i and ω_j, they may exchange and update their opinions based on their preferences. Agent i's trust opinion $\omega_{i\otimes j}$ in j's opinion is defined by a discounting operator [15], $c_i^j \in [0, 1]$, as

$$\omega_{i\otimes j} = \Big\{ b_{i\otimes j} = c_i^j b_j, \; d_{i\otimes j} = c_i^j d_j \times \; u_{i\otimes j} = 1 - c_i^j \big(1 - u_j\big), \; a_{i\otimes j} = a_j \Big\}. \tag{4}$$

The c_i^j can be specified based on user i's type of decision-making. We may consider a user i as three types: 1) uncertainty-based DM (U-DM); 2) homophily-based DM (H-DM); and 3) encounter-based DM (E-DM). The following discounting operators are used accordingly.

1) *Uncertainty-Based Discounting (uc_i^j):* Uncertainty (or lack of confidence) on given information has been used as a basis to decide whether to reflect the information in a user's opinion update [7]. In SL, the uncertainty-based discounting operator, uc_i^j is the combination of two uncertainties as [7]

$$uc_i^j = (1-u_i)\big(1-u_j\big). \tag{5}$$

We mainly consider uncertainty u derived from a lack of evidence and conflicting evidence. To consider both, we leverage the so-called *uncertainty (or vacuity) maximization* technique [15]. In SL, an opinion update stops when uncertainty is zero, which will prevent new information from being effectively applied in the latest opinion. The uncertainty maximization technique [15] enables the amount of conflicting evidence to be transformed to the vacuity (i.e., u_i) of an opinion. Given user i's opinion ω_i, $P(b_i)$ and $P(d_i)$, the corresponding vacuity-maximized opinion is estimated by $\ddot{\omega}_i = (\ddot{b}_i, \ddot{d}_i, \ddot{u}_i, a_i)$ where $\ddot{u}_i$, $\ddot{b}_i$ and $\ddot{d}_i$ are

$$\begin{aligned} \ddot{u}_i &= \min\left[\frac{P(b_i)}{a_i}, \frac{P(d_i)}{1-a_i}\right] \\ \ddot{b}_i &= P(b_i) - a_i\ddot{u}_i, \quad \ddot{d}_i = P(d_i) - (1-a_i)\ddot{u}_i. \end{aligned} \tag{6}$$

Only when u_i is sufficiently low by a threshold, ξ, the above $\ddot{u}_i$ can replace uc_i^j in (5).

2) *Homophily-Based Discounting (hc_i^j):* Homophily (or like-mindedness) is an important influence factor of an opinion update [18]. We use the *cosine similarity* to measure the extent of homophily of two users' belief and disbelief masses in the range of [0,1], as the hc_i^j by

$$hc_i^j = \frac{b_i b_j + d_i d_j}{\sqrt{b_i^2 + d_i^2}\sqrt{b_j^2 + d_j^2}}. \tag{7}$$

3) *Encounter-Based Discounting ($c_i^j = 1$):* The agent i encounters and trusts j's full opinion by $\omega_{i\otimes j} = \omega_j$ in (4) when $c_i^j = 1$. Otherwise, c_i^j can be either uc_i^j or hc_i^j in SL. When interacting with new information from the trust opinion of j by the *consensus* operator [15], agent i can update the opinion as $\omega_i \oplus \omega_{i\otimes j}$. $= \{b_i \oplus b_{i\otimes j}, d_i \oplus d_{i\otimes j}, u_i \oplus u_{i\otimes j}, a_i \oplus a_{i\otimes j}\}$. The details of each element are given by [15]

$$\begin{aligned} b_i \oplus b_{i\otimes j} &= \Big[b_i\Big(1-c_i^j\big(1-u_j\big)\Big) + c_i^j b_j u_i\Big]/\delta; \\ d_i \oplus d_{i\otimes j} &= \Big[d_i\Big(1-c_i^j\big(1-u_j\big)\Big) + c_i^j d_j u_i\Big]/\delta; \\ u_i \oplus u_{i\otimes j} &= \Big[u_i\Big(1-c_i^j\big(1-u_j\big)\Big)\Big]/\delta; \\ a_i \oplus a_{i\otimes j} &= \frac{\big(a_i-(a_i+a_j)u_i\big)\Big(1-c_i^j\big(1-u_j\big)\Big)+a_j u_i}{\delta-u_i\Big(1-c_i^j\big(1-u_j\big)\Big)} \end{aligned} \tag{8}$$

where $\delta = u_i + 1 - c_i^j(1-u_j) - u_i(1-c_i^j(1-u_j)) = 1 - c_i^j(1-u_i)(1-u_j)$ and $\delta \neq 0$ is assumed. The level of uncertainty $u_i \oplus u_{i\otimes j}$ is the same as $1-(b_i \oplus b_{i\otimes j} + d_i \oplus d_{i\otimes j})$.

In addition to above three types of DMs, models of other DMs can also take advantage of users' preferences by existing opinion update rules [23], [44]. We consider the assertion-based (A-DM) and herding-based DMs (HE-DM) as comparing counterparts for extensive experiments of investigating their effectiveness in combating the spread of disinformation. These are designed based on the following assertion and herding opinion update rules as follows.

a) Assertion opinion update model: Comparable to the proposition in SL, the amount of knowledge and degree of subjective prior belief can form the assertion opinion by $A_i = \{k_i, \text{spb}_i\}$ [44]. When treating k_i and spb_i as the belief, disbelief, and base rate in ω_i, and matching the ranges of spb_i from $[-1, 1]$ to base rate a_i in [0, 1], the assertion-based opinion update rule for $\omega_{i\oplus j}$ is by

$$\begin{aligned} k_{i\oplus j} &= k_i + k_j(1-k_i) \quad \text{for} \quad k \in [b, d, u] \\ a_{i\oplus j} &= a_i + b_j a_j(1-a_i). \end{aligned} \tag{9}$$

Since the assertion model can obtain more knowledge of k_i and collect more evidence of b, d, u, the total of b, d, u can exceed 1. To fit the sum of $b_{i\oplus j}$, $d_{i\oplus j}$, $u_{i\oplus j}$ to 1, the next step is to redistribute the SL masses of three dimensions by $k_{i\oplus j}/(\sum_{k\in\{b,d,u\}} k_{i\oplus j})$.

b) Herding opinion update model: Similar to the imitation of all the neighbors' behaviors, a user can update the opinion by evaluating all the neighbor's opinion similarities by the *convincing power* [23]. Fitting to the SL opinion, for $x \in [b, d, a]$, the herding update rule for $\omega_{i\oplus j}$ is

$$x_{i\oplus F_i} = \min\left[1, x_i + \frac{u_i}{|F_i|}\sum_{j\in F_i}\big(1-u_j\big)\big(x_j - x_i\big)\right] \tag{10}$$

where $u_i = 1-(b_i + d_i)$. This herding influence implies that user i will rely more on the neighbors j's opinions whose u_j is low particularly when user i is not confident of his/her opinion.

D. Interaction Model for Opinion Update

Users can update their opinions when they are exposed to other users' opinions from user interactions in OSN. The interaction model covers two real-world user interactions: 1) the activities of sharing information and 2) connecting or disconnecting with other users. The details of those activities as follows.

1) Sharing Opinion: A user can have a chance to update his/her opinion through sharing his/her opinion as follows.

a) Pair-wise interaction: Leaving comments or receiving messages or providing feedback (e.g., leaving sentiments, such as likes) with one friend, which is called *feeding behavior*. This feeding probability of user i, denoted by P_i^f, is collected from real datasets.

b) Posting: Sharing user opinion with all friends by posting messages. We also extract this posting probability of user i

from real datasets, denoted by P_i^p, for the user to interact with other users.

In this interaction model, users may update opinions with one neighbor by feeding or with all neighbors by posting. User i interacts with neighbor j, who has two types of sharing, P_j^f and P_j^p, and a relative degree of activities P_{ij}, by

$$P_{ij} = \frac{P_j^f + P_j^p}{\sum_{k \in F_i} \left(P_k^f + P_k^p \right)} \tag{11}$$

where F_i is the set of user i's friends. We assume that a user will interact with other users who are more active, which is reasonable in real OSNs. If users i and j interact, the feeding probabilities of both (i.e., P_i^f and P_j^f) will increase. If user i takes the "updating and sharing" (SU) strategy (see Section IV-B), then the posting P_i^p will increase.

2) Maintaining a Friend Network: A user will make friending or unfriending decisions based on uncertainty or projected difference (PD) between his/her opinion and friends' or other users' opinions, depending on the user type. The PD between two user opinions is [7]

$$\mathrm{PD}_{ij} = \frac{|b_i - b_j| + |d_i - d_j|}{2}. \tag{12}$$

Each user's friending and unfriending activities will follows.

a) Friending: Each user will invite a friend based on his/her tendency to connect. The probability of a new edge connecting to any node with degree k is by the price model [21] as $((k+1)p_k/m + 1)$, where m is the mean out-degree and p_k is the fraction of nodes with degree k. User j's threshold ϕ_j is a random real number in $[0, 1]$ following the Gaussian distribution. U-DM j accepts friending requests from i only when $u_i < \phi_j$; Other DM j accepts the request if $\mathrm{PD}_{ji} < \phi_j$. Otherwise, j will reject it.

b) Unfriending: A user can unfriend a current friend when he/she finds his/her opinion is way different from the friend user's. Given uncertainty is less than 0.5 (i.e., $u_j < 0.5$) for j to ensure sufficient interactions to update opinion, U-DM user i will unfriend j based on the threshold ϕ_i when $\phi_i < u_j < 0.5$; while other DM user i will unfriend j if the PD in (12) satisfies $PD_{ji} > \phi_i$.

IV. Game Theoretic Agent Model

In Fig. 1, we demonstrate a whole game considered by three types of players (i.e., an attacker, defender, and user). This game is represented by an undirected graph, denoted by $G(V, E)$, where V is a set of vertices representing game players and E is a set of edges representing the friend relationship (i.e., $e_{ij} = 1$ when players i and j are friends; otherwise $e_{ij} = 0$). Each player i (i.e., v_i) can form his/her opinion based on an amount of information received from nearby friends (i.e., adjacent players in a given G).

Since the social network mediates the interactions among players, a single game can only occur when two players are friends (i.e., directly connected in a topology). Accordingly, a player can start a single game with another friend at a different interaction time. Although one player can participate single games repeatedly from all interactions' times, this section will define a single game by three players' roles, attackers, a defender, and legitimate users. We define and formulate each player's objectives, strategies, and its payoff function in a single game. Supposing two players select each of their strategies, they can decide how to update their opinions based on their innate preferences, roles, or adopted opinion models. Then they can judge each utility from the results of opinion updates. A player's payoff function by taking a particular strategy is a weighted sum of the utilities, considering all the possible strategies of an interacting friend. Finally, in a single game, each player can find a preferred strategy associated with the highest payoff value defined by (13), (16), and (19). We summarize the input (i.e., strategies) of each player's payoff function and the output (i.e., expected payoff) of the payoff function in Table I.

The proposed game framework uses a repeated game consisting of multiple single games. A repeated game is a form of an extensive game where the same game (called a stage game) is played repeatedly, and the stage game influences decisions in future games. We use a repeated game to describe continuous interactions between attackers, users, and a defender (a social network platform administrator). In the repeated game, each player can take an action by predicting other players' actions based on their actions observed in the previous stage games. Each stage game belongs to a game of incomplete and imperfect information, such that each player does not know the type of an opponent. For example, the attacker may not know a user's type. A defender may not know exactly whether a given user is a legitimate user or an attacker. The user may not know whether an encountered user is an attacker or another user. We also assume that each player knows his/her opponent's move based on limited observability. This is modeled based on the probability distribution of the opponent's actions with 90% accuracy, representing each player's belief about his/her opponent. All above assumptions considering incomplete and imperfect information of the game is well aligned with real-world, uncertain situations.

A. Attacker Agent Model

An attacker aims to maximize the influence of disinformation by directly disseminating disinformation or disrupting users from obtaining true information by taking various deception strategies [16].

1) Strategies: The four attack deception strategies as follows.

a) Degradation (DG; a_1^A): This strategy is to confuse legitimate users by injecting noise into true information. This is realized by forwarding a highly uncertain opinion, represented by $\{b, d, u, a\} = \{1/n, 1/n, (n-2)/n, 0.5\}$, where n is a large number of evidence.

b) Corruption (C; a_2^A): This strategy produces false beliefs by injecting disinformation or replacing true information with disinformation. We realize this by propagating an opposite opinion of exchanging b and d to the friends.

c) Denial (DN; a_3^A): This strategy is to prevent users from accessing true information by not forwarding true information to their friends. It can lead to a lack of information, creating uncertainty and making users' judgment difficult in discerning the truthfulness of information.

TABLE I
INPUT AND OUTPUT OF THE GAME MODEL WITH THREE PLAYERS

	Attackers	Users	Defender
Input (Strategies)	$a_1^A, a_2^A, a_3^A, a_4^A$	a_1^U, a_2^U, a_3^U	a_1^D, a_2^D
Output (Expected Payoff)	$EP_k^{A_i}(a^U, a^D) = \sum_{\ell \in D} \sum_{m \in U} p_\ell^D \cdot p_m^U \cdot u_{k\ell m}^{ij}$	$EP_m^{U_i}(a^{U_j}) = p_{U_i}^{A_j} \cdot u_m^{U_i A_j} + (1 - p_{U_i}^{A_j}) \cdot u_m^{U_i U_j}$	$EP_\ell^D(a^A) = \sum_{k \in A} p_k^A \cdot u_{\ell k}^D$

d) Subversion (S; a_4^A): This strategy is to mislead a user's decision by altering his/her information process, such as considering noncredible information more or credible information less. For S, the attacker will feed false opinions, ω_F, as described in Section III.

Note that an attacker taking actions of C, or DN will use opinions received from other users and perform the attack on the deception opinions as described above.

2) Payoffs: We assume that the attacker performs the above strategies by sharing messages rather than performing pair-wise interactions for efficiency. Attacker i can estimate its **expected payoff** of taking strategy k by

$$\mathrm{EP}_k^{A_i}\left(a^U, a^D\right) = \sum_{\ell \in D} \sum_{m \in U} p_\ell^D \cdot p_m^U \cdot u_{k\ell m}^{ij} \tag{13}$$

where D is the set of the defender's strategies, U is the set of the user's strategies, ℓ is an element strategy in D, and m is an element strategy in U. The p_ℓ^D refers to the probability for a defender to take strategy ℓ (i.e., either detect/terminate the attacker's account or continue monitoring, a_1^D or a_2^D). The p_m^U is the probability of a user to take strategy m (i.e., $a_1^U - a_3^U$). An attacker obtains p_ℓ^D and p_m^U based on the historical record. The $u_{k\ell m}^{ij}$ is the utility when attacker i takes k strategy when the defender takes ℓ strategy and user j takes m strategy. The $u_{k\ell m}^{ij}$ is given by

$$u_{k\ell m}^{ij} = \mathrm{ds}\left(k, m, \omega_i, \omega_j\right) - g_\ell \tag{14}$$

where the utility is estimated by the gain minus loss. The gain is estimated by $\mathrm{ds}(k, m, \omega_i, \omega_j)$ which represents the improvement made for the mean similarity between the false opinion, ω_F, and users j's opinion by taking attack strategy k and not taking it. The $\mathrm{ds}(k, m, \omega_i, \omega_j)$ is given by

$$\mathrm{ds}\left(k, m, \omega_i, \omega_j\right) = s\left(k, m, \omega_F, \omega_j\right) - s\left(\neg k, m, \omega_F, \omega_j\right) \tag{15}$$

where $s(k, m, \omega_F, \omega_j)$ refers to the cosine similarity in (7) of j's opinion and false opinion ω_F when the attacker takes k strategy when interacting with user type m. The $s(\neg k, m, \omega_F, \omega_j)$ refers to the similarity score for the attacker not taking k strategy in the same context. The g_ℓ is the attacker's loss when the defender takes ℓ strategy, estimated based on the average similarity in (7) between the true opinion and each of all users' opinions.

B. User Agent Model

A user aims to judge information correctly based on his/her propensity. A user can be either of the five DMs types described in Section III-C: 1) U-DM; 2) H-DM; 3) E-DM; 4) A-EM; and 5) HE-DM (i.e., U: uncertainty, H: homophily, E: encounter, A: assertion, and HE: herding).

1) Strategies: A user's three strategies as follow.

a) Updating and sharing (SU; a_1^U): A user updates his/her opinion and shares the updated opinion with other friends.

b) Updating (U; a_2^U): A user updates his/her opinion only but does not share the updated opinion with other friends.

c) No updating (NU; a_3^U): A user discards a received opinion.

2) Payoffs: User i plays with other user or an attacker, not a defender. However, user i does not know whether that player is an attacker or a legitimate user. Hence, user i estimates his/her **expected payoff** of taking strategy m by

$$\mathrm{EP}_m^{U_i}\left(a^{U_j}\right) = p_{U_i}^{A_j} \cdot u_m^{U_i A_j} + \left(1 - p_{U_i}^{A_j}\right) \cdot u_m^{U_i U_j} \tag{16}$$

where $p_{U_i}^{A_j}$ is the probability that user j is an attacker. The $u_m^{U_i A_j}$ or $u_m^{U_i U_j}$ is the utility user i can obtain by taking strategy m when an encountered user is an attacker or a legitimate user (can be either one of the five DM types), respectively. The $u_m^{U_i A_j}$ is given by

$$u_m^{U_i A_j} = \begin{cases} \sum_{k \in A} p_k^{A_j} \cdot -s\left(m, \omega_F, \omega_i, \omega_j\right), & \text{if } m = a_1^U \text{ or } a_2^U \\ 0, & \text{if } m = a_3^U \end{cases} \tag{17}$$

where k is a strategy taken by attacker j and $p_k^{A_j}$ is user i's belief of attacker j performing strategy k. User i can estimate $p_k^{A_j}$ based on its historical experience with attackers in the past. The $s(m, \omega_F, \omega_i, \omega_j)$ represents the mean similarity between false opinion, ω_F, and users i's expected opinion after interacting with attacker j via m strategy. The $s(m, \omega_F, \omega_i, \omega_j)$ is estimated by the cosine similarity.

For $u_m^{U_i U_j}$, depending on user j's type, it should be estimated as user j can choose one of his/her strategies by

$$u_m^{U_i U_j} = \begin{cases} \sum_{m' \in \mathcal{U}_j} p_{m'}^{U_j} \cdot \mathrm{uc}_{\mathrm{im}'}^j, & \text{if } j \text{ is } U - \text{DM type} \\ \sum_{m' \in \mathcal{U}_j} p_{m'}^{U_j} \cdot \mathrm{hc}_{\mathrm{im}'}^j, & \text{if } j \text{ is other DM types} \end{cases} \tag{18}$$

where $\mathrm{uc}_{\mathrm{im}'}^j$ and $\mathrm{hc}_{\mathrm{im}'}^j$ are calculated by (5) and (7), respectively, with the opinions of users i and j updated by strategy m' where $\mathcal{U}_j$ is a set of actions by user j.

C. Defender Agent Model

A defender (e.g., a given OSN system administrator) plays a game against an attacker if a malicious user receives N_R misconduct reports provided by other users.

1) Strategies: The defense strategies as follows.

a) Terminating a malicious user (T; a_1^D): The defender will suspend the account of a reported user based on its evaluation using the expected payoff function in (19).

b) Monitoring a suspect user (M; a_2^D): A defender finds a suspect user but regards this user as nonmalicious.

2) Payoffs: The defender can estimate its **expected payoff** of taking strategy ℓ by only considering the attacker's strategy as follows:

$$\mathrm{EP}_{\ell}^{D}(a^{A}) = \sum_{k \in A} p_{k}^{A} \cdot u_{\ell k}^{D} \tag{19}$$

where A is the set of the attacker's strategies and k is an element strategy in A. The p_k^A refers to the probability for an attacker to take strategy k. If the reported attacker is not a real attacker, but a legitimate user, the defender will assume that the reported user is an attacker and estimates its utility based on (19). The $u_{\ell k}^{D}$ is the utility when the defender takes ℓ strategy when the attacker takes the k strategy. Unlike the attacker's utility, the defender will consider the utility of each strategy based on the overall impact introduced to the system. Hence, $u_{\ell k}^{D}$ is estimated where the overall strategies of the attacker and users are considered by

$$u_{\ell k}^{D} = \mathrm{ds}(\ell, k, \omega_T, \omega') - c_{\ell} \tag{20}$$

where $\mathrm{ds}(\ell, k, \omega_T, \omega')$ refers to the improvement made by taking strategy ℓ for the mean similarity between the truthful opinion, ω_T, and the expected opinion ω' of all users (including all active users known as legitimate) given an attacker taking k strategy, compared to not taking defense strategy ℓ. The $\mathrm{ds}(\ell, k, \omega_T, \omega')$ is given by

$$\mathrm{ds}(\ell, k, \omega_T, \omega') = s(\ell, k, \omega_T, \omega') - s(\neg\ell, k, \omega_T, \omega') \tag{21}$$

where $s(\ell, k, \omega_T, \omega')$ and $s(\neg\ell, k, \omega_T, \omega')$ are estimated as the mean value of the cosine similarity in (7) between the truthful opinion and a user's opinion and the maximum possible $\mathrm{ds}(\ell, k, \omega_T, \omega')$ is 1. The c_ℓ is a constant cost incurred for the defender to take strategy ℓ where the cost of T (i.e., a_1^D) is 0.1 and the cost of M (i.e., a_2^D) is 0.

D. Opinion-Based SIR Epidemic Model

The propagation of disinformation can alter the opinions of users. In the population view, each user can be assigned a status based on the projected belief $P(b)$ and disbelief $P(d)$. By the SIR model, the Susceptible S users have $P(b) \leq 0.5$ and $P(d) \leq 0.5$. A user of $P(d) > 0.5$ belongs to status Infected (I) and a user of $P(b) > 0.5$ stays in recovered state (R). The SIR model quantifies the dynamics of transition from S to I and I to R by the infection rate β_t and recovery rate γ_t. We use time-dependent β and γ because our opinion propagation is influenced by the decision-making process in the game model, but not naturally transmitted by contact like the spread of disease. The ODEs to solve this SIR model at any time t are

$$\frac{\mathrm{dS}}{\mathrm{dt}} = -\beta_t S_t I_t, \quad \frac{\mathrm{dI}}{\mathrm{dt}} = \beta_t S_t I_t - \gamma_t I_t, \quad \frac{\mathrm{dR}}{\mathrm{dt}} = \gamma_t I_t \tag{22}$$

where $S_t + I_t + R_t = N$ with N nodes in a given network. This SIR status and infection and recovery rates can represent the effect of disinformation propagation under the different opinion update models. However, the parameters $\theta = \{\beta_1, \ldots, \beta_T, \gamma_1, \ldots, \gamma_T\}$ are only available from the game model simulation results. We investigate how each opinion model generates β_t and γ_t which determines the extent of disinformation propagation in the network.

1) Parameter Optimization and Gradient Decent: Given the simulation of users' opinions, we can calculate the S_t, I_t, and R_t for $t \in [1, T]$. Then we need to fit those values to the ODEs in (22). We use gradient descent [4] to optimize the parameters of interest, i.e., $\theta = \{\beta_1, \ldots, \beta_T, \gamma_1, \ldots, \gamma_T\}$. The objective function considering all T interactions are defined as below

$$\mathcal{J}(\theta) = \sum_{t=1}^{T} \left(\tilde{I}_{\theta,t} - I_t\right)^2 \tag{23}$$

where $\tilde{I}_\theta(t)$ is the estimated number of infectious people at time t via the SIR model with parameters θ. However, the gradients are intractable since they are parameters of the SIR model, which is an ODE. We use the small difference (1%) between the objective function divided by the difference between the parameter to approximate a parameter's gradient. For approximation, the β_t's gradient is derived by

$$\nabla_{\beta_t} \mathcal{J}(\theta) = \frac{\mathcal{J}_t(\theta) - \mathcal{J}_t(1.01 \times \theta)}{-0.01 \times \beta_t} \tag{24}$$

where $\mathcal{J}_t(\theta) = (\tilde{I}_{\theta,t} - I_t)^2$. Note that the small difference (i.e., 1%) is applied by adding 1.01 in the numerator and -0.01 in the denominator in the above equation. After we get the gradients of parameters, i.e., $\nabla_\beta(\theta_k)$ and $\nabla_\gamma(\theta_k)$ for the kth iteration, we use the gradient descent's update rule to update the parameters to obtain θ_{k+1} by

$$\beta_{k+1} = \beta_k - \eta \nabla_\beta(\theta_k), \quad \text{and} \quad \gamma_{k+1} = \gamma_k - \eta \nabla_\gamma(\theta_k) \tag{25}$$

where η is the learning rate.

2) SIR Status Prediction: When the lists of β_t and γ_t under each time points are available from the data-fitting step, we can estimate the future count of each S_t, I_t, and R_t users from the ODE model in (22). Given the initial S_0, I_0, and R_0, we can predict the next-step values of each role, i.e., S_1, I_1, and R_1, from (22), using β_1 and γ_1. This step can be iterated to predict the numbers of S, I, and R at any future interaction time.

V. Experiment Results and Analysis

A. Experiment Setup

1) Dataset: The *1KS-10KN* dataset [35], [36] includes Twitter accounts of 10 000 normal users and 1000 human attackers. The legitimate users and spammers have an average number of friends of 7744 and 2520, respectively. The dataset has a full spectrum of profiles, social activities, and tweets texts for each user. From the user behaviors data, the initial P_j^f was calculated as the sum of the frequencies of reply and favorite tweets and initial P_j^p was the sum of the tweeting and retweeting frequencies.

2) Metrics: We use the following performance metrics.

a) Agent's opinion ($\omega_i = \{b_i, d_i, u_i, a_i\}$): It consists of the four masses of an agent i's opinion, which can show how convergent or divergent users' opinions are after T number of interactions.

b) Ratios of S, I, and R: This counts the user status based on their SL-based opinions at each interaction time.

c) Infection rate β and recovery rate γ : These rates under each opinion model can show a different effect on the DMs.

TABLE II
KEY PARAMETERS AND THEIR DEFAULT VALUES

Param.	Default value	Param.	Default value
T	200	P_i^f	0.142 (mean)
N	1,000	P_i^p	0.186 (mean)
P_{true}	0.1	c_ℓ	T: 0.1, M: 0
P_{false}	0.1	ρ	$\mathcal{N}(0.5, 0.05)$
ϕ	$\mathcal{N}(0.1, 0.1)$	N_R	3
ξ	0.05	$p_{U_i}^{A_j}$	0.1
ω_T	$\{1, 0, 0, 1\}$	ω_F	$\{0, 1, 0, 0\}$
ω_U	$\{0, 0, 1, 0.5\}$	η	$1e10 - 6$

d) Network community topology: It visualizes the quality of two communities generated from the greedy modularity algorithm [9] before and after disinformation propagation.

e) Probability distribution of taking strategies: The probability distributions of each player role taking their strategies help to understand how those choices of strategies have introduced different ways to process false information and corresponding outcomes.

f) NE analysis: In a noncooperative game, NE solution can be analyzed where each player is assumed to know the other players' strategies and their preferences [26], leading to the best payoff estimation based on the best actions.

3) Experiment Environment: A sample of $N = 1,000$ nodes sub-network contained $P_{\text{false}} = 10\%$ of total users as attackers and another $P_{\text{true}} = 10\%$ of total users as true informers and the rest of them (i.e., 80%) as users. As a result, the initial state of SIR model was $S_0 = 800$, $I_0 = 100$, and $R_0 = 100$. The attackers and true informers were selected from nodes with the top 20% in their degree centrality measures. All the notations used are summarized in Table II. The values of parameters describing user behaviors, such as feeding and posting, are derived from real datasets described in Section III-D. For other user behaviors (e.g., a tolerance level to report a malicious user, ρ, and a threshold to accept a friend request, ξ), we assumed it follows Gaussian distribution with a given mean and standard deviation (i.e., $\mathcal{N}(\mu, \sigma)$) to describe the users' different characteristics. It is possible to investigate the sensitivity of the results when a different set of mean and standard deviation is used to describe a different set of user populations. Other system parameters, such as a threshold for uncertainty maximization and learning rate (e.g., ϕ and η), are selected at their optimal settings. The users were either the same types, or from various ratios of H-DM and other users for in-depth sensitivity analysis. Since the majority of the followers were not in this collected network, the friending network was simulated by allowing each user to select the users with the highest pairwise topic similarity [31], which was a cosine similarity of the top 20 tweets topics generated by latent Dirichlet allocation (LDA) algorithm from their total tweets. An attacker was assumed to connect to normal users or true informers only. To reduce the stochastic errors, one experiment setting was repeated 100 runs where one simulation run covered $T = 200$ number of total interactions. Each player's probability distribution of actions was updated by Dirichlet distribution [13] where an outcome from each game is counted as one piece of evidence.

4) Simulation Procedures: In the 1st interaction, each node interacts with at most one neighbor following the steps.

1) **1.A**: Each attacker starts playing a game by choosing one neighbor to propagate false opinions by strategy *subversion*.
2) **1.B**: Each user i picks one of the friends j based on P_{ij} and chooses an random action. If a_1^U or a_2^U is chosen, user i updates an opinion (see Section III-C) based on i's type.
3) **1.C**: If attacker j interacts with user i who accepted j's opinion to update his/her opinion, attacker j will share the received opinion ω_i with deception to other friends in the next interaction.
4) **1.D**: Each user follows the procedures of the friending and unfriending decisions in Section III-D.

From the 2nd to Tth interaction, one interaction time allowed each player to play a game with at most one neighbor. All the players were considered in a random order. If all friends of a player had participated in games with others at the current interaction round, this player would neither play a game nor update opinion. The steps in one interaction were as follows.

1) **2.A**: Each attacker chose one neighbor i and decided one attacker strategy by evaluating the payoffs of each attacker strategy in (13). Then, the attacker propagated the opinion received from 1.C with the taken deception strategy to i and repeated 1.C.
2) **2.B**: Each user kept performing 1.B but chose the best user's strategy of the highest expected payoff in (16).
3) **2.C**: User i should report another user or attacker j interacting with (i.e., a friend in user i's social network), based on the extent of discrepancy between ω_i and ω_j using (12). The tolerance ρ for user i to report to the defender, an OSN service provider, was modeled as the Gaussian distribution of mean μ and standard deviation σ.
4) **2.D**: The defender decided whether to suspend a detected, malicious user or not by a higher payoff strategy from (19), if a malicious user was reported at least N_R times by other users.
5) **2.E**: Each user maintained the friending network by following step 1.D. If a legitimate user reported a malicious user in step 2.C, their friend relationship was also removed.

For reproducibility of this experiment, we made all extracted user activity features from the original dataset and the source code for the simulation implemented available at `https://anonymous.4open.science/r/opinions-game-22/README.md`.

B. Numerical Results and Analysis

By default the percentages of true and false informers were 10% each and the remaining 80% of users were single type users of U-DM, H-DM, E-DM, A-DM, or HE-DM. The sensitivity analysis of various false informers ratios P_{false} was conducted. In the Appendices, we plotted the distribution of strategies in our opinion game players when there were mixtures of H-DMs and other DMs.

1) Analysis of Uncertain Opinions: Fig. 2 compared users' opinions dynamics from five types of DMs in 200 game interactions in terms of b (belief), d (disbelief), u (uncertainty), and

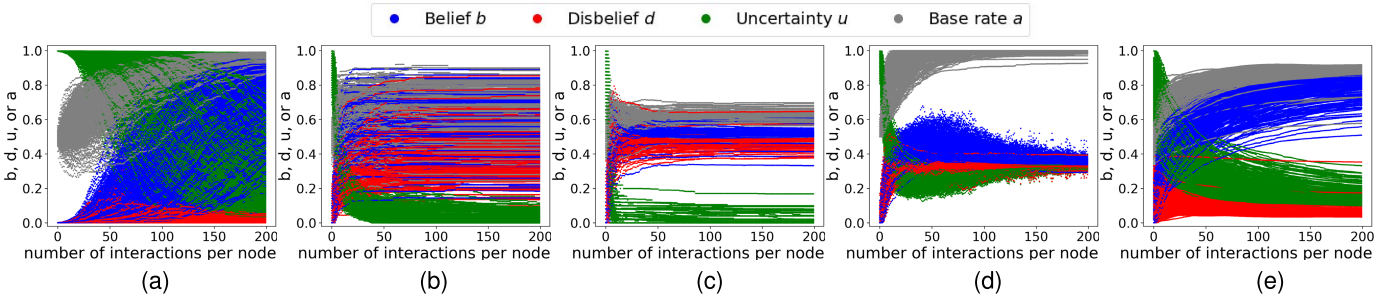

Fig. 2. Evolution of SL-based opinions of all legitimate users over 200 interactions in belief (b, blue), disbelief (d, red), uncertainty (u, green), and base rate (a, gray) under the dataset 1KS-10KN [35], [36]. (a) Uncertainty-DM. (b) Homophily-DM. (c) Encounter-DM. (d) Assertion-DM. (e) Herding-DM.

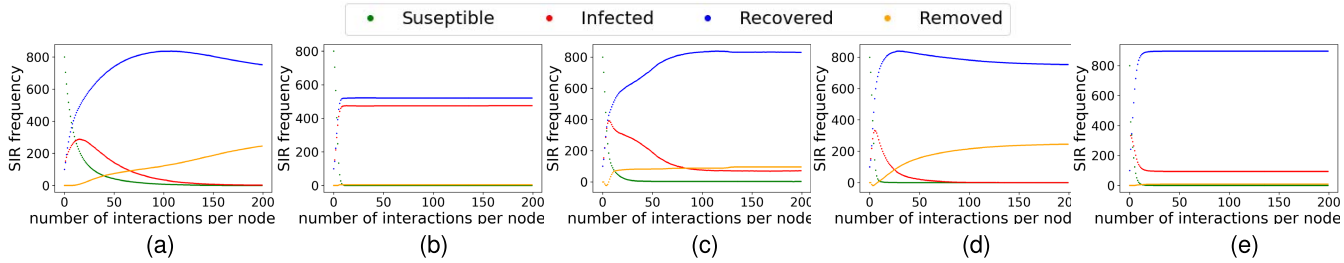

Fig. 3. SIR curves of five opinion models in our proposed game, including R_m as the removed nodes by the defender. (a) Uncertainty-DM. (b) Homophily-DM. (c) Encounter-DM. (d) Assertion-DM. (e) Herding-DM.

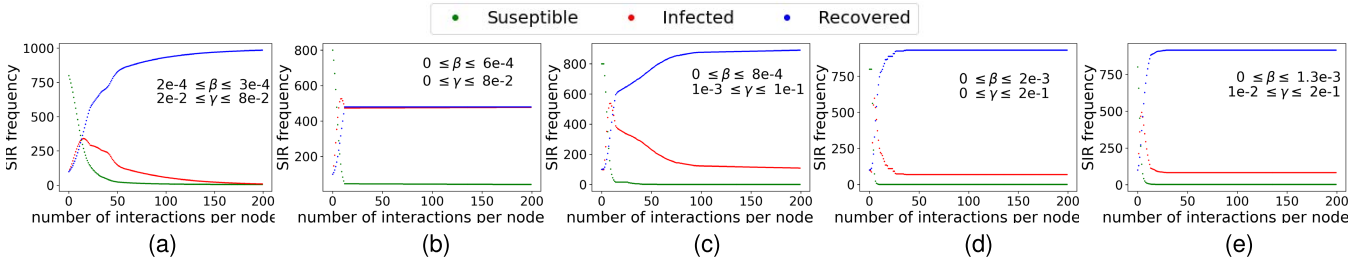

Fig. 4. Simulated SIR curves of five opinion models from the ODE parameter optimization. The β and γ values fluctuate at each interaction time. The common ranges from 200 interactions of each β_t and γ_t are shown inside the figures. (a) Uncertainty-DM. (b) Homophily-DM. (c) Encounter-DM. (d) Assertion-DM. (e) Herding-DM.

a (base rate). Since different types of users had distinctively opinion dynamics, we could observe H-DM type users in Fig. 2(b) introduce much higher polarization to the network by spreading false opinions because some users believe true information while other users believed in false information. In contrast, in Fig. 2(a) of U-DM type users, more users believed true information when they had more interactions with other users, as shown by beliefs greater than 0.5 and low-level disbeliefs. This is because the U-DM type users interact and update their opinions based on uncertainty of an encountered user's opinion, and highly uncertain opinions propagated by the attackers aiming to deter propagation of true information were more likely to be dropped or less considered. The HE-DM type users in Fig. 2(e) can also form high beliefs and low disbeliefs during the interaction. In Fig. 2(c) and (d), all users had opinion consensus after 200 interactions.

When U-DM and HE-DM users updated opinions in Fig. 2(a) and (e), their belief gradually increased while uncertainty gradually decreases. The base rate increased from the initial 0.5 to the final interaction of nearly 1.0. In contrast, in H-DM users' opinion update, the uncertainty dropped rapidly below 0.1 and belief, disbelief and base rate increased to less than 0.9 in Fig. 2(b). Even after interaction 50, the opinions distribution became stable. The belief and disbelief values ranged across the whole range which showed a higher diversity as in other types of users. When users interacted with other users and updated their opinions based on the similarity of opinions, they were more likely to believe disinformation because H-DM type users can accept noisy, uncertain opinions as long as the difference between their own opinions and the received opinion was still below ϕ.

2) Analysis of Opinion Dynamics Using the SIR Model: In terms of the opinion status analyzed by the SIR epidemic model, the dynamics of all false informers, true informers, and legitimate users in each of S, I, R were demonstrated in Fig. 3. Another user state *removed* was included because the defender could terminate the malicious accounts if it chose strategy $a_1^D = T$. The curves of the SIR status showed the different effects of disinformation propagation in the five DMs. The *infected* users in all non-H-DMs increased and then decreased. The decrease of R in U-DM and A-DM correlated with the increase of *removed* nodes but H-DM and HE-DM did not remove any users or attackers.

The gradient decent was applied in Section IV-D to optimize the infecting rate β and recovery rate γ in the ODEs of users. The fitting curves of Fig. 3 were presented in Fig. 4 with the

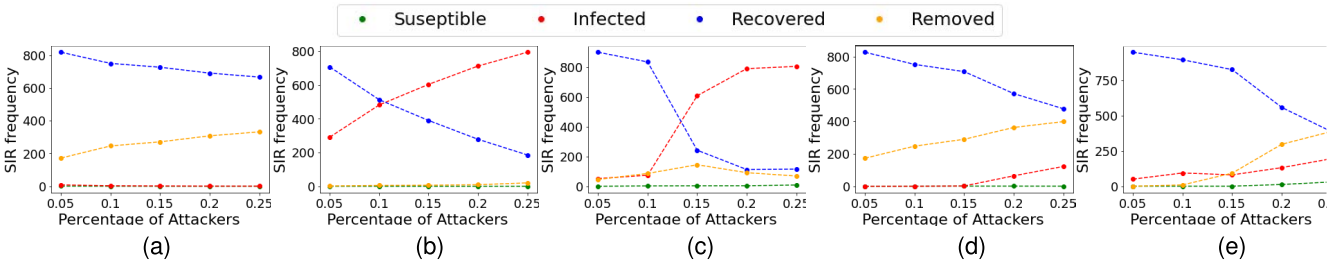

Fig. 5. Ratio of S, I, and R of five initial percentage of attacker, P_{false}, after 200 interaction times in our proposed game. (a) Uncertainty-DM. (b) Homophily-DM. (c) Encounter-DM. (d) Assertion-DM. (e) Herding-DM.

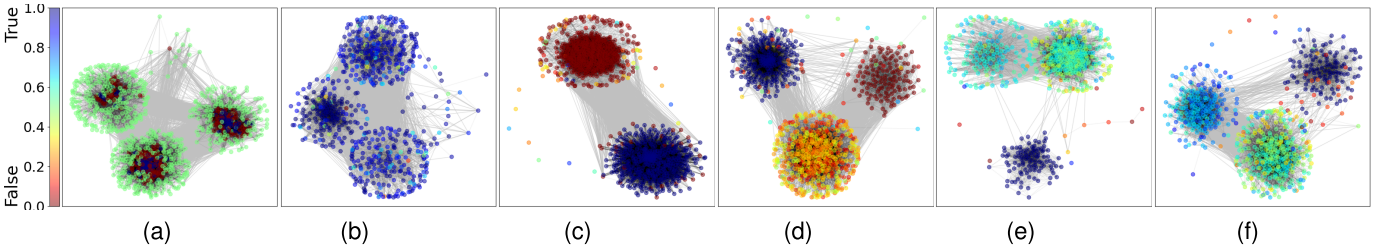

Fig. 6. Community view of all the users in the network with their projected belief $P(b_i)$ and communities, which shows belief in true information by blue color and belief in disinformation by red color. (a) Plots the initial uncertain belief by green color, i.e., $P(b_i) = 0.5$. There are $P_{\text{false}} = 20\%$ attackers in the beginning of each network. (b) Uncertainty-DM. (c) Homophily-DM. (d) Encounter-DM. (e) Assertion-DM. (f) Herding-DM.

ranges of β_t and γ_t. There was no *removed* curve in Fig. 4 because it was not modeled in the SIR model and this caused the higher numbers of R in Fig. 4(a), (c), and (d). The curves of S, I, R in both figures were well matched with each other. Both A-DM and H-DM in Fig. 4(d) and (e) had large infection and recovery rates because the SIR curves stayed stable after the first a few interactions. The H-DM in Fig. 4(b) maintained a large number of I throughout the interactions so that β and γ were close to 0 when the users were polarized to two extreme groups and they hardly changed their opinions. In the first a few interactions, the H-DMs had a higher β and lower γ, compared to the U-DMs in Fig. 4(a).

3) Analysis of Varying the Ratio of Attackers: Further experiments were performed to investigate the influence of initial attackers, i.e., P_{false} in each type of the network dynamics. Fig. 5 plotted the dynamics of S, I, R, and *removed* nodes after 200 interaction by increasing the P_{false} from 5% to 25%. The results showed a clear trend that U-DMs in Fig. 5(a) were the most resistant to the number of attackers because the I was close to 0 for all the ratios of attackers. The I increased greatly for H-DM users and in 25% initial attackers, most of the users believe disinformation. The other E-DM, A-DM, and HE-DM type users showed similar patterns that the I had a low level when there were less attackers but the I increased along with the increase of P_{false}.

4) Analysis of Opinion Dynamics and Polarization: Fig. 6 reflected the topology and dynamics of the network before and after opinion updates by each types of users when there were 20% of initial attackers. We applied the community detection algorithm called *greedy modularity maximization* [9]. The network after U-DMs updating opinions in Fig. 6(b) indicated the lowest polarization and the homogeniety of U-DMs' opinions, which implied that almost all the U-DM nodes had high beliefs in true information, as shown with blue dots, after disinformation propagation in the U-DM network. Both the H-DM and E-DM networks shown in Fig. 6(c) and (d) had highly polarized blue and red communities with distinct boundaries, but H-DMs could cause the most polarized network topology. Fig. 6 supported that the opinion update model in U-DMs helps users unite and share uniform opinions. On the other hand, H-DMs could facilitate more polarized opinion groups, which was aligned with the empirical phenomena observed in the OSN platforms. This agreement of simulation and empirical analysis proved the effectiveness of our game model.

Due to the space constraint, we show the experimental results for additional sensitivity analysis in the submitted supplement document. In the supplement document, Appendix A discusses what strategy selection is made by different players under varying the ratio of H-DM and other users. Appendix B analyze the strategies chosen by our approach and the NE which assumes players to hold accurate beliefs toward the moves of the opponents.

C. Experimental Limitations

In this work, we found the following limitations due to the inherent difficulties in modeling and simulation research.

1) Some user behaviors, such as a threshold to accept or request a friend or a tolerance level to report a malicious user, are set intuitively based on social science research findings. Since they are not derived from real datasets, some inaccuracies may be introduced due to inherent limitations.
2) Game theory assumes that each agent (an OSN user) is a rational entity to achieve its selfish goal and behaves to maximize its utility. It may be arguable whether a human is a rational entity and behaves to maximize its utility. In future work, to describe an OSN user's behaviors, we will consider behavioral game theory [6], which is well known to predict human behavior and decisions.
3) Although the number of followers is available, network topologies representing the relationships between users are unavailable in most datasets. Hence, we generated a

friend network simulated based on the topic similarity, which is reasonable in OSNs [31]. However, since the friend network is not based on real network datasets, there is an inherent limitation about whether the friend network reflects a realistic network topology.

4) To address agents' limited observability, we considered 90% accuracy of observations toward opponents' moves. However, there is an inherent uncertainty that may not be accurately estimated in nature.

VI. Conclusion and Future Work

This work proposed a game-theoretic opinion framework allowing users to make rational decisions in updating their opinions. We developed two opinion models with a user updating an opinion based on perceived uncertainty or homophily of an encountered user's opinion. This model can demonstrate how these two types of users can defend against false information. This study obtained the following **key findings**.

1) Uncertainty-based decision makers (U-DM) had the most effective opinion model in combating the spread of disinformation. This is because U-DM can better inform the defender in reporting malicious users. In addition, the defender can terminate the malicious users, leading to mitigating false information propagation.
2) Homophily-based decision makers (H-DM) showed the least performance in combating the spread of disinformation. H-DM tends to make users rely on opinions similar to their own opinions. It also makes users' opinions easily stuck in the opinion updates made at the beginning because they do not tend to change their opinions.
3) The spread of disinformation often introduces opinion polarization. H-DM users are likely to produce higher opinion polarization than users using other opinion models. In the friending and unfriending process, U-DM users can remove adjacent users with high uncertainty while connecting with users with low uncertainty, leading to less polarization.
4) We observed some discrepancies between NE's and the players' strategies. This is because each player does not have correct beliefs about the opponent's move due to inherent uncertainty, while NE-based strategy selection is based on the true probability distributions of each player's move, which is not true in reality.

We plan to conduct the following **future research directions**: 1) examine the influences of the different numbers of true informers and attackers when they are selected based on different centrality metrics; and 2) conduct additional sensitivity analyses, such as varying ϕ to adjust users' friending and unfriending decisions and ρ to change a user's reporting behavior for malicious users.

Acknowledgment

The authors acknowledge Dr. You Lu for his partial contribution to the implementation of the SIR model.

References

[1] M. Askarizadeh, B. T. Ladani, and M. H. Manshaei, "An evolutionary game model for analysis of rumor propagation and control in social networks," *Phys. A, Stat. Mech. Appl.*, vol. 523, pp. 21–39, Jun. 2019.

[2] G. Asmolov. (2018). *The Disconnective Power of Disinformation Campaigns*. [Online]. Available: https://jia.sipa.columbia.edu/disconnective-power-disinformation-campaigns

[3] A. Bessi et al., "Homophily and polarization in the age of misinformation," *Euro. Phys. J. Spec. Topics*, vol. 225, no. 10, pp. 2047–2059, 2016.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[5] L. Brumley, C. Kopp, and K. B. Korb, "Cutting through the tangled web: An information-theoretic perspective on information warfare," *Air Power Aust. Analyses*, vol. 9, pp. 1–25, Oct. 2012.

[6] C. F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ, USA: Princeton Univ. Press, 2011.

[7] J.-H. Cho, "Dynamics of uncertain and conflicting opinions in social networks," *IEEE Trans. Computat. Social Syst.*, vol. 5, no. 2, pp. 518–531, Jun. 2018.

[8] J.-H. Cho, S. Rager, J. O'Donovan, S. Adali, and B. D. Horne, "Uncertainty-based false information propagation in social networks," *ACM Trans. Social Comput.*, vol. 2, no. 2, pp. 1–34, 2019.

[9] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, Dec. 2004, Art. no. 066111.

[10] Z. Guo and J.-H. Cho, "Game theoretic opinion models and their application in processing disinformation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–7.

[11] Z. Guo, J. Valinejad, and J.-H. Cho, "Effect of disinformation propagation on opinion dynamics: A game theoretic approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 5, pp. 3775–3790, Sep. 2022.

[12] Y. Halberstam and B. Knight, "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter," *J. Public Econ.*, vol. 143, pp. 73–88, Nov. 2016.

[13] R. K. S. Hankin, "A generalization of the Dirichlet distribution," *J. Stat. Softw.*, vol. 33, no. 11, pp. 1–18, 2010.

[14] D.-W. Huang, L.-X. Yang, P. Li, X. Yang, and Y. Y. Tang, "Developing cost-effective rumor-refuting strategy through game-theoretic approach," *IEEE Syst. J.*, vol. 15, no. 4, pp. 5034–5045, Dec. 2020.

[15] A. Jøsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Cham, Switzerland: Springer, 2016.

[16] C. Kopp, K. B. Korb, and B. I. Mills, "Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to fake news," *PLoS ONE*, vol. 13, no. 11, pp. 1–35, 2018.

[17] D. Li, J. Ma, Z. Tian, and H. Zhu, "An evolutionary game for the diffusion of rumor in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 433, pp. 51–58, Sep. 2015.

[18] L. Li, A. Scaglione, A. Swami, and Q. Zhao, "Consensus, polarization and clustering of opinions in social networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 6, pp. 1072–1083, Jun. 2013.

[19] Q. Li, T. Xiang, T. Dai, and Y. Xiao, "An information dissemination model based on the rumor & anti-rumor & stimulate-rumor and tripartite cognitive game," *IEEE Trans. Cognit. Develop. Syst.*, early access, Jul. 25, 2022, doi: 10.1109/TCDS.2022.3193576.

[20] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford Univ. Press, 2010.

[21] D. De Solla Price, "A general theory of bibliometric and other cumulative advantage processes," *J. Amer. Soc. Inf. Sci.*, vol. 27, no. 5, pp. 292–306, Sep. 1976.

[22] R. Recuero, G. Zago, and F. Soares, "Using social network analysis and social capital to identify user roles on polarized political conversations on Twitter," *Social Media Soc.*, vol. 5, no. 2, Apr. 2019, Art. no. 205630511984874.

[23] A. Sonowal, A. Idupulapati, D. Booravilli, P. Parimi, and R. R. Rout, "An improved model for dynamic opinion updates in online social networks," in *Proc. IEEE 4th Conf. Inf. Commun. Technol. (CICT)*, Dec. 2020, pp. 1–6.

[24] B. C. Stahl, "On the difference or equality of information, misinformation, and disinformation: A critical research perspective," *Informing Sci.*, vol. 9, no. 15, pp. 83–96, 2006.

[25] G. Szabó and C. Toke, "Evolutionary prisoner's dilemma game on a square lattice," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 58, no. 1, p. 69, Jul. 1998.

[26] S. Tadelis, *Game Theory: An Introduction*. Princeton, NJ, USA: Princeton Univ. Press, 2013.

[27] J. A. Tucker et al., "Social media, political polarization, and political disinformation: A review of the scientific literature," *Political Polarization, Political Disinformation, Rev. Sci. Literature*, pp. 1–95, Mar. 2018.

[28] G. Verma, A. Swami, and K. Chan, "The impact of competing zealots on opinion dynamics," *Phys. A, Statist. Mech. Appl.*, vol. 395, pp. 310–331, Feb. 2014.
[29] M. D. Vicario et al., "The spreading of misinformation online," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 3, pp. 554–559, 2016.
[30] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Trans. Web*, vol. 13, no. 2, pp. 1–22, May 2019.
[31] Z. Wang, J. Liao, Q. Cao, H. Qi, and Z. Wang, "Friendbook: A semantic-based friend recommendation system for social networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 3, pp. 538–551, Mar. 2015.
[32] Y. Xiao, D. Chen, S. Wei, Q. Li, H. Wang, and M. Xu, "Rumor propagation dynamic model based on evolutionary game and anti-rumor," *Nonlinear Dyn.*, vol. 95, no. 1, pp. 523–539, 2019.
[33] Y. Xiao, Z. Huang, Q. Li, X. Lu, and T. Li, "Diffusion pixelation: A game diffusion model of rumor & anti-rumor inspired by image restoration," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 21, 2022, doi: 10.1109/TKDE.2022.3144310.
[34] Y. Xiao, W. Li, S. Qiang, Q. Li, H. Xiao, and Y. Liu, "A rumor & anti-rumor propagation model based on data enhancement and evolutionary game," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 2, pp. 690–703, Jun. 2022.
[35] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers," in *Recent Advances in Intrusion Detection*. Berlin, Germany: Springer, 2011, pp. 318–337.
[36] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter," in *Proc. 21st Int. Conf. WWW*, vol. 2012, pp. 71–80.
[37] H.-X. Yang, "A consensus opinion model based on the evolutionary game," *Europhys. Lett.*, vol. 115, no. 4, p. 40007, Aug. 2016.
[38] K. Yoshikawa, T. Awa, R. Kusano, H. Sato, M. Ichino, and H. Yoshiura, "A fake news dissemination model based on updating reliability and doubt among individuals," in *Proc. 11th Int. Conf. Awareness Sci. Technol. (iCAST)*, Dec. 2020, pp. 1–8.
[39] M. Zhan, H. Liang, G. Kou, Y. Dong, and S. Yu, "Impact of social network structures on uncertain opinion formation," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 4, pp. 670–679, Aug. 2019.
[40] H. Zhang, Y. Li, Y. Chen, and H. V. Zhao, "Smart evolution for information diffusion over social networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1203–1217, 2021.
[41] L. Zhao, Q. Wang, J. Cheng, Y. Chen, J. Wang, and W. Huang, "Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal," *Phys. A, Stat. Mech. Appl.*, vol. 390, no. 13, pp. 2619–2625, Jul. 2011.
[42] L. Zhao, J. Wang, Y. Chen, Q. Wang, J. Cheng, and H. Cui, "SIHR rumor spreading model in social networks," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 7, pp. 2444–2453, 2012.
[43] L. Zhao, X. Qiu, X. Wang, and J. Wang, "Rumor spreading model considering forgetting and remembering mechanisms in inhomogeneous networks," *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 4, pp. 987–994, Feb. 2013.
[44] D. Zinoviev and V. Duong, "A game theoretical approach to broadcast information diffusion in social networks," in *Proc. 44th Annu. Simul. Symp.*, 2011, pp. 47–52.

Zhen Guo (Graduate Student Member, IEEE) received the M.S. degree in biological sciences and the M.S. degree in computer science from Fordham University, New York City, NY, USA, in 2013 and in 2016, respectively. He is currently pursuing the Ph.D. degree in computer sciences with Virginia Polytechnic Institute and State University, Falls Church, VA, USA.

His recent research interests include understanding and combating online social deception through various concepts derived from social and behavioral theories, game theory, and machine/deep learning.

Jin-Hee Cho (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Virginia Tech, Falls Church, VA, USA, in 2004 and 2008, respectively.

She has been an Associate Professor with the Department of Computer Science, Virginia Tech, since 2018. Before joining Virginia Tech, she has been a Computer Scientist at the U.S. Army Research Laboratory (USARL), Adelphi, MD, USA, since 2009. She has published technical papers in leading journals and conferences in trust management, cybersecurity, metrics and measurements, network performance analysis, resource allocation, agent-based modeling, uncertainty reasoning and analysis, information fusion/credibility, and social network analysis.

Dr. Cho is a member of ACM. She received the Best Paper Awards from IEEE TrustCom'2009, BRIMS'2013, IEEE GLOBECOM'2017, 2017 ARL's Publication Award, and IEEE CogSima 2018. She is a Winner of the 2015 IEEE Communications Society William R. Bennett Prize in the field of communications networking. In 2016, she was selected for the 2013 Presidential Early Career Award for Scientists and Engineers (PECASE). She also received the 2022 Faculty Fellow Award from the College of Engineering, Virginia Tech.

Chang-Tien Lu (Senior Member, IEEE) received the Ph.D. degree from the University of Minnesota at Twin Cities, Minneapolis, MN, USA, in 2001.

He is currently a Professor of computer science and the Associate Director of the Sanghani Center for AI and Data Analytics, Virginia Tech, Falls Church, VA, USA. He has published over 170 articles in top rated journals and conference proceedings. His research interests include spatial databases, data mining, urban computing, and intelligent transportation systems.

Dr. Lu is an ACM Distinguished Member. He served as the General Chair for the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems in 2009, 2020, and 2021, and the International Symposium on Spatial and Temporal Databases in 2017. He also served as the Secretary (2008–2011) and the Vice Chair (2011–2014) for the ACM Special Interest Group on Spatial Information (ACM SIGSPATIAL).