This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130

Effect of Disinformation Propagation on Opinion Dynamics: A Game Theoretic Approach

Zhen Guo, *Student Member, IEEE*, Jaber Valinejad, *Member, IEEE*, and Jin-Hee Cho, *Senior Member, IEEE*

***Abstract*—Disinformation can alter or manipulate our values, opinions, and rational decisions toward any life event because disinformation, such as fake news or rumors, is propagated rapidly and broadly in online social networks (OSNs). Game-theoretic models can help people maximize the benefits from dynamic social interactions. This work presents an opinion framework formulated by repeated, incomplete information games that model OSN users' subjective opinions. The users may update their opinions using various criteria, such as uncertainty, homophily, encounter, herding, or assertion. We demonstrate how Subjective Logic, a belief model explicitly handling opinion uncertainty, can be employed to model attackers' deception strategies, users' opinion update models, and the influences of propagating disinformation through the interactions between users. Through extensive experiments, we investigated how an individual user's information processing type can introduce different impacts on the extent of disinformation propagation. We compared the performance of the five different opinion update models under OSNs characterized by two real OSN datasets. We analyzed their impact on the choices of best strategies, their utilities, and network/opinion polarization. We also examined how the player's choices of best strategies under uncertainty are different from Nash Equilibrium strategies based on correct beliefs towards their opponents' moves.**

***Index Terms*—Disinformation, game theory, subjective opinion, uncertainty, opinion dynamics, polarization.**

I. Introduction

SOCIAL media platforms facilitate everyone to express their opinions and comments to the whole online world. Many people, especially young people, have heavily relied on social media to acquire daily news. They learn about new events from diverse sources. Accordingly, disinformation propagated by malicious users can significantly mislead users, altering their opinions. However, regardless of whether the information is true or not, since information diffuses fast in online social networks (OSNs), it is implausible for OSN users to verify all the information they encounter. Disinformation causes severe privacy violations, ruins reputations, or produces financial losses. Recently, these serious issues caused by false information, fake news, or rumors have demonstrated detrimental effects in the whole society, such as influencing decision-making processes in elections, pandemics, health, or education.

In this work, we aim to investigate how rational users update their opinions in the presence of disinformation propagated in an OSN. We define the rational users as those who seek to maximize their utilities from updating their opinions based on their preferences. In particular, we consider the rational users who have the ability to reason and filter disinformation in terms of the credibility of information based on a source's expertise, the veracity of information, the current opinion they have, and the opinions of their friends. Therefore, this study will examine how these users' rational behaviors in updating their opinions can affect the mitigation or amplification of disinformation propagation, where disinformation propagation can also introduce the network and/or opinion polarization.

OSN users' behaviors towards online information are reflections of their innate propensities or personalities. Some people make decisions based on the competence (or expertise) or certainty of a source. Other people accept other people's opinions based on like-mindedness (i.e., homophily), agreeableness (i.e., conformity), or stubbornness (i.e., not accepting dissimilar opinions) [1, 2, 3]. This can be explained based on confirmation bias [4, 5], which is a cognitive bias representing a tendency a person shows in understanding, interpreting, preferring, or remembering information that confirms one's previous values or beliefs. Another well-known cognitive bias, which explains a subjective belief that can mislead a decision-maker, is the Dunning Kruger effect [6, 7]. This effect also well explains how a subjective belief affected by irrational cognitive bias can leads to a poor decision. The false information diffusion among users in an OSN has been modeled with specific user behaviors by [8, 9, 10]. The theoretical models supported by game-theoretic approaches consider user behaviors and human rationality by cognitive limitations [11, 12, 13, 14, 15]. However, to the best of our knowledge, the existing game theory models rarely help users make rational decisions in updating their opinions.

In this work, we propose a three-player game framework that models online users' behaviors in updating their opinions based on the interactions with other users or attackers. In the framework, we also model attackers' deception tactics to propagate disinformation and the defender (OSN platform system administrator)'s policy to ensure a safe and trustworthy OSN environment. We aim to demonstrate how OSN users' rational information processing behaviors based on opinion update models using various criteria can influence the mitigation of disinformation propagation, which impacts network dynamics and opinion polarization. Via this experiment, we aim to suggest what OSN user behaviors and the OSN platforms can ensure safe and trustworthy cyberspace from disinformation propagation.

This work has the following **key contributions**:

1) We designed and tested a generic game-theoretic opinion

Zhen Guo (*Corresponding author*), Jaber Valinejad, and Jin-Hee Cho are with the Department of Computer Science, Virginia Tech, National Capital Region Campus, Falls Church, VA 22043, USA. Email: {zguo, jabervalinejad, jicho}@vt.edu.

framework in an OSN environment against the spread of disinformation. We demonstrated the flexibility of this framework to accommodate various user interactions and opinion models (OMs), including uncertainty, homophily, encounter, herding, and assertion-based updates. In addition, this framework provided the functionality of analyzing the effects of disinformation processing on users' rational decisions and opinion dynamics and polarization.

2) Utilizing the rich features in two real-world datasets collected from the Twitter platform, we captured and modeled real OSN users' social activities and propensities, such as information sharing or exchanging.
3) We leveraged a belief model, called *Subjective Logic* (SL), that represents users' subjective, uncertain opinions in real-world, uncertain situations to decide whether to accept other users' opinions. Further, we expanded the SL-based opinion model to consider users' five opinion models and realized attackers' deception strategies [16].
4) We constructed a set of uncertainty-aware payoff equations based on the SL opinion in a game of three agents, i.e., attackers, a defender, and users. Each player can make choices based on the uncertainty, observations, and their inherent preference to maximize utility when interacting with or accepting other users' opinions. Furthermore, we solved each player's preferred strategies by Nash Equilibria (NE), making decisions based on correct beliefs towards the opponents' moves. Since perfect NE strategies may not be realistic in complex real-world scenarios, we compared the performance gap between each player's best strategies chosen under uncertainty and its NE choices.
5) We also examined network community structure and opinion polarization after OSN users interact with each other upon disinformation propagation where network community structure was investigated by three graph partitioning algorithms [17, 18, 19] by calculating network polarization scores using four different methods [17, 20, 21, 22]. From the polarization of opinions under the five opinion models, we investigated how disinformation can influence network topology changes and network power (i.e., social capital).
6) We investigated how different ways of updating opinions can introduce different opinion dynamics and change the social capital of OSN users. We measured a user's bridging social capital by betweenness and bonding social capital by trust, where both of the metrics have been used to represent a user's influence or power in a network [23].

The preliminary results of this work have been included in our prior conference paper [24]. We substantially extended our prior work [24] and made the additional contributions to this work as follows:

1) We extensively conducted simulation experiments using two real social network datasets, including *1KS-10KN* [25, 26] and *Cresci15* [27]. We examined the performance of the five different opinion models in combating disinformation propagation in terms of the distributions of the best strategies and their utilities. We compared the players' best strategies taken under uncertainty with those under certainty derived by Nash Equilibrium (NE).
2) We provided the full details of NE solutions by demonstrating the detailed game trees and payoff matrices for the proposed games of incomplete information. This can offer an in-depth understanding of the proposed games, where what strategies can help combat disinformation propagation, and how users' information processing in each opinion model can influence it. Due to the space constraint, we included this in Appendix A of the supplement document.
3) We substantially extended the analyses to understand the effects of disinformation propagation on opinion polarization using multiple community partitions methods under the five opinion models. We considered three community detection algorithms, including modularity, Kernighan-Lin bipartition, and label propagation under four different methods of calculating polarization scores.
4) We investigated the effect of disinformation propagation on social capital, which measures bridging based on betweenness and bonding based on trust with other users.

The rest of this paper is structured as follows. Section II summarizes research on game-theoretic information diffusion models, opinion dynamics models, and the effect of disinformation propagation on opinion and network dynamics. Section III describes the foundations of SL opinion models, the adaptation of SL to five opinion models, and the interaction models explaining how to share information and make friending decisions. Section IV provides the details of the game-theoretic opinion framework, including the roles, strategies, and payoffs of an attacker, user, and defender. Section V details the experimental setup, including the dataset, metrics, and experiment settings. Section VI presents the simulation results and discusses the overall trends of the results along with their implications. Section VII concludes the paper with the summarized key findings obtained from our study.

II. Related Work

A. Opinion and Information Models

Yang [11] modeled two user strategies, 'cooperative' or 'defective' in prisoner's dilemma game and public goods game for binary opinion diffusion and reached the equilibrium of opinion consensus. Several evolutionary game theory (EGT) work of opinions models solved a stable evolutionary state to model user's strategy transition rate [12, 13]. The goal of EGT model [12, 13, 28, 29, 30] is to consider several factors and the population preferences to influence user decisions. However, EGTs mainly deal with only three behaviors in terms of spreading rumors, not spreading rumors, or spreading anti-rumors without considering the opinions of individual users. Szabó and Tőke [28] studied the likelihood of strategy imitation in the Fermi updating rule, determined by the actual advantages of the fitness of the neighbor. Li et al. [12] simulated rumor diffusion in an OSN with various personal and social attributes, such as users' tie relationship with friends, judgment ability of others, strategy imitation, and the cost of spreading the rumor. Askarizadeh et al. [13] discussed a user's behavior of spreading rumors by the attitude or awareness, community anxiety, and the intensity of rumor and anti-rumor cascades in OSN. Huang et al. [29] developed users' cost-effective defense strategy against rumors where the users'

opinions toward the rumors were updated by a differential game model. Yoshikawa et al. [30] studied a mode where users update their friends' reliability and doubt (or distrust) and then exchange opinions by Bayesian estimation.

Recently, psychological factors were commonly considered to model the defense strategies in rumor spreading game models [14, 15]. Xiao et al. [14] modeled the spreading behaviors from their psychological factors and investigated the competition by messages supporting rumors or clarifying rumors. Zhang et al. [15] studied the resistance of malicious users based on the reputation dynamics of a user's neighbors. The authors have not explored any game-theoretic opinion framework aiming to mitigate disinformation influence. Further, they did not investigate the OSN network and opinion dynamics when users interact for their opinion exchanges and updates.

One of the pivot questions on opinion models is how to accept new evidence to update the current opinions from pair-wise interactions. Zinoviev and Duong [31] proposed an assertion model to update users' opinions in two aspects, the amount of knowledge and the degree of belief, where the exchange of knowledge level was determined by users' forgetfulness, rate of learning, and trust of one friend. Sonowal et al. [32] proposed a herding-based opinion model which counts the pair-wise social interactions of all existing friends. Cho [2] considered uncertainty-based trust models to update uncertain opinions through dyadic interactions between two agents. The authors investigated the effect of the opinion model on opinion convergence and divergence compared to homophily-based trust model. Zhan et al. [33] computed an uncertainty interval boundary update weighted by the number of close opinion neighbors. They calculated the distance of two opinions based on the range of the uncertainty interval length if an agent has uncertain opinions. However, the above opinion models [2, 31, 32, 33] cannot be comparable because they are applied in vastly different scenarios. In addition, they have not considered any game-theoretic information processing based on players' rationality.

B. Impact of False Information on Network Polarization

Research has shown that false information spreading can polarize users [34], which can facilitate false information circulation [35]. Polarized users tend to gain access to similar content, and they may have a long response lag for their fake news posts [36]. A network consisting of polarized users can be divided into several polarized communities due to the echo chambers effect that users assigned to the same group are highly interconnected [37, 38]. The studies in [34]–[38] modeled the spreading patterns of false information, but not how the opinions can be updated. Hence, polarized users and polarized communities were examined only based on users' binary behaviors, such as whether to spread false information or not. A social network with a polarized topology can cause a non-trivial reduction of the access of social capital (e.g., cognitive and relational capital) [37]. There is a strong correlation of users' opinion homophily and activities towards false information [35] so that clusters can be predicted by the fact that like-minded users with similar polarization scores can gather easily. In the meanwhile, homophile clusters can expedite the speed of false information spreading [35]. In addition, the authors have not considered any game-theoretic decision-making process. In particular, the studies in [35, 37] investigated the influence of polarized users on the rate of false information spreading in the sense inverse to our research.

A number of social science studies [39, 40, 41] have been conducted to draw a conclusion. However, their findings have not been reflected in the simulation models. Homophily increases with the larger community size [39]. In most communities, intra-community information diffuses more quickly and broadly than inter-community information because more people tend to be exposed to information. We usually observe this phenomenon in political campaigns where disinformation can increase conflicts and break strong social ties and social capital between inter-communities [40]. In addition, the long-term effects of political disinformation on social capital have not been well studied [41]. Thus, this research tried to show this effect on social capital in our model.

Unlike the above works [34]–[41], our work pioneers in investigating the effects of disinformation propagation on network and opinion polarization and the distribution of network power using social capital where OSN users interact with other uses based on different types of exchanging their opinions.

III. Uncertain, Subjective Opinion Model

This work leverages a belief model, called *Subjective Logic* [42, 3], in an OSN to quantify users' binomial opinions. Through pair-wise interactions among users, an initial uncertain opinion can be updated by a user.

TABLE I
Notations of Design Parameters and Their Meanings

Notation	Description
ω, ω_i	User i's SL-based binary opinion
(b, d, u, a)	Belief, disbelief, uncertainty, and base rate
$P(b_i)$	User i's projected belief from ω_i
$P_{true}, P_{false}, P_{uc}$	Proportion of true informers, false informers, and normal users
$\oplus$	Consensus operator from SL
$\otimes$	Trust operator from SL
c_i^j, uc_i^j, hc_i^j	Discounting factors from SL, uncertainty-based and homophily-based
$\ddot{\omega}_i$	Uncertainty maximized opinion of user i
ξ	Threshold of uncertainty maximization
P_i^f, P_i^p	User i's feeding probability and posting probability
PD_{ij}	Projected discrepancy between two opinions
ϕ_i^1	Threshold to accept or request a friend
a_k^A	Attacker's strategy k in $\{DG, C, DN, S\}$
a_ℓ^U	Normal user's strategy ℓ in $\{SU, U, NU\}$
a_m^D	Defender's strategy m in $\{T, M\}$
ω_F	False opinion $(0, 1, 0, 0)$
ω_T	True opinion $(1, 0, 0, 1)$
$EP_k^{A_i}$	Expected payoff of attacker's strategy k
$u_{k\ell m}^{ij}$	Utility of an element in $EP_k^{A_i}$
EP_ℓ^D	Expected payoff of defender's strategy ℓ
$u_{\ell k}^D$	Utility of an element in EP_ℓ^D
c_ℓ	Defender's cost of strategy ℓ
U, H, A, HE, E	Uncertainty, Homophily, Assertion, Herding, and Encounter-based user types
$EP_m^{U_i}$	Expected payoff of user's strategy m
$p_{U_i}^{A_j}$	Probability of user j as an attacker
stc_i	User i's structural social capital
T_i	User i's trust by other friends
N_R	Number of reports to alert a defender
ρ	Tolerance to report a malicious user
I	Number of interactions in simulation
N	Number of nodes in the OSN

A. Opinion Formation

A binomial *opinion*, $\omega = (b, d, u, a)$, is represented by *belief* (b), *disbelief* (d), *uncertainty* (u), and *base rate* (a) in SL [42]. Those four dimensions are simply formed as:

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

where belief (b) means the degree of *pro*, agree, or true information held by an agent to believe a proposition even if the real truth is not available. Disbelief (d) means the degree of *con*, disagree, or false information for an agent to oppose to a proposition or disbelieve it. Uncertainty (u) means the level of *vacuity* normally due to an insufficient amount of evidence. *Base rate* (a) is the prior belief, expertise, or bias [42] for an agent's prior knowledge in a given domain. An agent updates its opinion in the four dimensions where each dimension is updated by considering an interacted agent's opinion. Based on the obtained evidence of each view, a user's binomial opinion can be represented by the observed or available evidence by the following mapping rule:

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

where r is the amount of positive evidence and s is the amount of negative evidence for a certain proposition. W is the amount of uncertain evidence as inherent errors from a system or an environment, such as unavailable source of evidence or limited and partial ability of observation. In a binomial opinion, W is commonly set to 2, representing the number of beliefs (i.e., two belief masses including belief and disbelief) [42].

For user i, the binomial opinion is $\omega_i = (b_i, d_i, u_i, a_i)$. The expected belief $P(b_i)$ and expected disbelief $P(d_i)$ are the factors for user i to make a decision by:

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

where $P(b_i) + P(d_i) = 1$ as $b_i + d_i + u_i = 1$. If user i needs a decision but b_i or d_i is very similar, a_i can be a critical decision factor because a_i is used to interpret uncertainty u_i.

B. Initialization of Opinions

There exist *true informers* and *false informers*, who can be zealots [43] to support extreme opinions in our network. The zealots not only propagate their true or false opinions but also refuse to change their own opinions. All other users who have initial uncertain opinions are willing to learn new opinions via sharing or feeding behaviors towards other users. That is, we set the fractions of users with $P_{true} + P_{false} + P_{uc} = 1$ where P_{true}, P_{false}, and P_{uc} refer to the fractions of true informers, false informers, and normal users, respectively. When the false informers (i.e., attackers) take the action *Subversion* (S), they propagate the false opinion (see Section IV). We initialize a true opinion, false opinion (i.e., disinformation), and uncertain opinion, held by false informers (i.e., attackers), true informers, and legitimate users, respectively, by:

- *True opinion,* ω_T, is initialized with $(b \to 1, d \to 0, u \to 0, a = 1)$, implying true opinion's belief (i.e., believing true information) is close to 1 (highly true) while disbelief (i.e., disbelieving true information) is close to 0.
- *False opinion (i.e., disinformation),* ω_F, is initialized with $(b \to 0, d \to 1, u \to 0, a = 0)$. This means that belief of false opinion is close to 0 while disbelief is close to 1.
- *Uncertain opinion* is formulated to represent the opinion of the rest of users, except true and false informers, initialized as $(b, d, u, a) = (b \to 0, d \to 0, u \to 1, a = 0.5)$, without showing strong preference. Their opinions can be updated depending on an interacted user's opinion and the way to update his/her opinion (i.e., opinion models).

C. Opinion Update

False and true informers are zealots and do not change their opinions while influencing other users' opinions. Legitimate users with a lack of confidence will refresh their opinions with new information from the pairwise interactions when they interact with their friends. Since uncertainty-based OM, homophily-based OM, and encounter-based OM are all grounded by SL's consensus update mechanism, we describe how SL framework updates user i's opinion ω_i as below.

In SL, the first step to update agent i's opinion ω_i is to consider how much agent j's opinion can be accepted by agent i. Agent i discounts agent j's opinion by discounting operator c_i^j, which implies agent i's trust in agent j. Hence, agent j's opinion is considered by agent i based on $\omega_{i\otimes j} = (b_{i\otimes j}, d_{i\otimes j}, u_{i\otimes j}, a_{i\otimes j})$. Each opinion element is given by:

$$\begin{aligned} b_{i\otimes j} &= c_i^j b_j,\ d_{i\otimes j} = c_i^j d_j, \\ u_{i\otimes j} &= 1 - c_i^j(1 - u_j),\ a_{i\otimes j} = a_j, \end{aligned} \tag{4}$$

where the quantity of $u_{i\otimes j}$ above is the same as $u_{i\otimes j} = 1 - b_{i\otimes j} - d_{i\otimes j}$ since $b_i + d_i + u_i = 1$. These trust opinion calculations from Eq. (4) are all at time t.

The second step to update SL opinion ω_i is to integrate the discounted opinion $\omega_{i\otimes j}$ by the *consensus* operator [42]. Using the consensus operator, agent i's new opinion at time $t+1$ after meeting agent j is formulated as $\omega_i \oplus \omega_{i\otimes j} = (b_i \oplus b_{i\otimes j}, d_i \oplus d_{i\otimes j}, u_i \oplus u_{i\otimes j}, a_i \oplus a_{i\otimes j})$. Each element is given by:

$$\begin{aligned} b_i \oplus b_{i\otimes j} &= \frac{b_i(1 - c_i^j(1 - u_j)) + c_i^j b_j u_i}{\beta}\ , \\ d_i \oplus d_{i\otimes j} &= \frac{d_i(1 - c_i^j(1 - u_j)) + c_i^j d_j u_i}{\beta}\ , \\ u_i \oplus u_{i\otimes j} &= \frac{u_i(1 - c_i^j(1 - u_j))}{\beta}\ , \\ a_i \oplus a_{i\otimes j} &= \frac{(a_i - (a_i + a_j)u_i)(1 - c_i^j(1 - u_j)) + a_j u_i}{\beta - u_i(1 - c_i^j(1 - u_j))}\ , \\ \beta &= 1 - c_i^j(1 - u_i)(1 - u_j) \neq 0, \end{aligned} \tag{5}$$

where the uncertainty $u_i \oplus u_{i\otimes j}$ is identical to $1 - b_i \oplus b_{i\otimes j} - d_i \oplus d_{i\otimes j}$. The right side is for the time step t while the left side shows $\omega_i(t+1)$. We omitted the time for simplicity.

Based on the above SL operations and opinion structure, we describe the five OMs for performance analysis in our work:

- **Uncertainty-based OM**: The uncertainty-based discounting operator, uc_i^j, as a specific c_i^j in Eq.(5), is derived by judging two users' uncertainties as:

$$uc_i^j = (1 - u_i)(1 - u_j). \tag{6}$$

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130

Uncertainty (or lack of confidence) from one's opinions has been investigated as a deciding factor to reflect the information from the updates of users' opinions [2]. Although uncertainty comes from many sources, we refer to uncertainty from two reasons: insufficient evidence and conflicting evidence. To represent the uncertainty property from both vacuity and conflict, the *uncertainty (or vacuity) maximization* technique [42] is also applied to prevent the uncertainty from going down to zero. This is because if an agent collects enough evidence for both belief and disbelief, it will reach zero uncertainty and then stop accepting new evidence. The uncertainty maximization technique [42] can prevent this from not being updated and transfer evidence supporting the belief and disbelief masses to the uncertainty mass. This means moving conflicting evidence to vacuity. We use a threshold, ξ, to determine whether to use this uncertainty maximization, i.e., apply only when $u_i < \xi$ (i.e., only when uncertainty is sufficiently low).

The vacuity-maximized opinion for user i is defined 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 computed by:

$$\ddot{u}_i = \min\Big[\frac{P(b_i)}{a_i}, \frac{P(d_i)}{1-a_i}\Big], \tag{7}$$
$$\ddot{b}_i = P(b_i) - a_i \cdot \ddot{u}_i, \;\; \ddot{d}_i = P(d_i) - (1-a_i)\cdot \ddot{u}_i,$$

where the projected belief and disbelief $P(b_i)$ and $P(d_i)$ are from Eq. (3). The uncertainty maximization also plays a critical role in uc_i^j, such that if $u_i < \xi$, the vacuity-maximized $\ddot{u}_i$ would replace the u_i in Eq. (6).

- **Homophily-based OM**: Homophily (or like-mindedness) between two opinions is a critical factor for opinion updates [44]. Like [3] and [44], we also use *cosine similarity* [45] within the range of [0,1], as the homophily-based discounting operator, hc_i^j. We chose this cosine similarity because it can properly capture the measure of the distance between beliefs, including belief and disbelief masses, where we formulated an opinion based on Subjective Logic. This hc_i^j can replace the c_i^j in Eq. (5) by the following definition [3]:

$$hc_i^j \;\; = \;\; \frac{b_i b_j + d_i d_j}{\sqrt{b_i^2 + d_i^2}\sqrt{b_j^2 + d_j^2}}. \tag{8}$$

We ignore uncertainty for two opinions' dissimilarity above because of the assumption of $b + d + u = 1$, where belief and disbelief can indirectly reflect the uncertainty.

- **Encounter-based OM**: We use this model as a baseline model in which the opinion is simply updated by the existing consensus method (i.e., $w_i \oplus w_j$) in SL [42] without applying any filters such as uncertainty, homophily, or assertion. This OM is implemented as $c_i^j = 1$ in Eqs. (4) and (5).

The following two opinion models were proposed from other existing research [31, 32]. For the fair comparison of the five OMs in this work, we extend the subjective opinion to fit those two models in [31, 32] as baseline counterparts to the OMs supported by the consensus operator as follows:

- **Assertion-based OM**: We use this opinion model as an existing counterpart. This model uses the so-called *assertion* [31], by $A_i = \{k_i, spb_i\}$, which is formulated based on the knowledge and a subjective prior belief. The original update rules for the two values are $k_{i\oplus j} = k_i + k_j(1-k_i)$ and $spb_{i\oplus j} = spb_i + k_j spb_j(1 \pm spb_i)$, where '+' is for negative spb_i and '-' is for positive spb_i. We convert this model to our SL's opinion with (b_i, d_i, u_i, a_i) where k_i quantifies the evidence of b_i and d_i and spb_i is the base rate a_i. By converting spb_i's range in $[-1, 1]$ to $[0, 1]$ for a_i and maintaining k_i's range in $[0, 1]$ for b_i and d_i, the opinion update rule for this opinion model is formulated by:

$$b_{i\oplus j} = b_i + b_j(1-b_i), \; d_{i\oplus j} = d_i + d_j(1-d_i), \tag{9}$$
$$u_{i\oplus j} = 1 - b_{i\oplus j} - d_{i\oplus j}, \; a_{i\otimes j} = a_i + b_j a_j(1-a_i).$$

- **Herding-based OM**: We adopt this model to update an opinion considering the bias towards a user's neighbor (i.e., leaning more towards his/her neighbors' opinions) [32] to emphasize the *convincing power* of the neighbors' opinions. As this model mainly relies on the neighbors' opinions, we consider this *herding-based* opinion update. The term 'herding' has been previously used in the network science domain when *herding behavior* is used to indicate one's behavior following his/her friends or neighbors [46]. Since we use SL-based opinion format, for a fair comparison, we consider the following opinion update operator when each user updates his/her opinion upon the interaction:

$$x_i = \min[1, x_i + \frac{u_i}{|F_i|}\sum_{j\in F_i}(1-u_j)(x_j - x_i)], \; x \in \{b, d, a\},$$
$$u_i = 1 - (b_i + d_i). \tag{10}$$

Eq. (10) above implies that user i will consider his/her neighbors j's opinions when being unsure of the opinion with high uncertainty (u_i). Neighbor j's opinion with higher certainty (i.e., $(1-u_j)$) will be more considered when updating user i's opinion. This implies that neighbor j's opinion with lower uncertainty (u_j) has a more convincing power to user i. In SL, an opinion's uncertainty represents how much confidence the owner of the opinion has on the opinion. The rationale of the uncertainty-based opinion model is well aligned with the prior research that expert sources can influence persuasion because they can motivate recipients to more seriously consider the information provided by them compared to the information provided by non-experts [47, 48].

D. Interaction Model for Opinion Update

Users are assumed to share opinions with other friends and update opinions during their interactions. A user's high posting frequencies tend to attract more interactions with other users. Hence, if a user tends to be more exposed to the information, the user will have higher chances to interact with the users posting more. In addition, a user can interact with another user based on the probability that the two users interact directly. In this work we consider the following user activities:

- *Sharing*: This behavior is the precondition of an opinion update. A user can share his/her opinion by:
  - *Pair-wise interaction*: The pair-wise sharing includes receiving tweets or leaving comments or feedback, such as

likes or other sentiments. We use P_i^f as i's feeding rate to model the *feeding behavior* between two users.

– *Posting*: The posting behavior is sharing posts or messages with all the friends. We use P_i^p as i's posting probability to share with all the friends.

A user updates his/her opinion by interacting with one of the neighbors j's. Each neighbor user j is characterized by P_j^f and P_j^p, the probabilities of leaving feedback (e.g., sentiments such as likes or comments) and posting, respectively. Each user i judges the relative level of the neighbor's sharing behavior to find user j to interact. We assume that users like to interact with more active users than less active users in a given OSN. We quantify user i's likelihood, P_{ij}, to select j for possible interaction, assuming F_i is i's friends:

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

where P_j^f and P_j^p are initialized by the features in the datasets (see Section V). Users will interact with other users and accordingly take actions, including updating and sharing (i.e., a_1^A). Hence, when they share their opinions, their feeding and sharing probabilities are also dynamically updated accordingly. For example, if users i and j interact, it will increase the feeding probabilities, P_i^f and P_j^f, for both of them. If user i takes sharing strategy SU (see Section IV-D), user i's posting probability (i.e., P_i^p) will increase.

- *Maintaining a friend network*: A user can add new friends or make unfriending decisions based on the corresponding opinion differences. The projected difference PD_{ij} between two opinions held by users i and j is obtained by [2]:

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

This PD is symmetric such that $PD_{ij} = PD_{ji}$. All of the elements are at time t. A user will make the friending or unfriending decisions based on the PD by:

– *Friending*: Users can invite a friend if they have a tendency to make friends. This tendency is considered as the probability of inviting a friend derived from the current number of friends. The probability of a new edge connecting to any node with degree k is from the Price Model [49] as $p_k(k+1)/(m+1)$, where m is the mean out-degree and p_k is the fraction of nodes with degree k. For any user i, ϕ_i^1 is a threshold to accept a friend and is scaled in the range of $[0, 1]$ at random following the Gaussian distribution. In an uncertainty-based OM, user j will accept a friending request only by $u_i < \phi_j^1$; user j in other OM types will accept it when $PD_{ji} < \phi_j^1$. Otherwise, user j will always ignore a new friend request.

– *Unfriending*: It is the opposite process to friending. A user can dismiss a current friend if the user finds an opinion discrepancy from the friend user. To ensure a sufficient amount of updating j's opinion, we bound uncertainty to $u_j < \phi_i^2$. Then user i using uncertainty-based OM will unfriend j if $\phi_i^1 < u_j < \phi_i^2$; while i using other OMs (i.e., by assertion, herding, homophily, and encounter) will unfriend j when $PD_{ji} > \phi_i^1$.

These two operations will change a network topology and accordingly affect a user's influence in a network (e.g., centrality or social capital) as discussed in Section II.

IV. Game Theoretic Agent Model

The social network in this work is denoted by an undirected graph structure, $G(V, E, \Omega)$, that holds users as V and holds all the friendship connections as E ($e_{ij} = 1$ only when i and j are friends). Ω is the collective of the subjective opinion ω_i for each user v_i. We describe how each agent is characterized by a set of features. In this game model, the agents have three roles: attackers, users, and a defender (i.e., a service provider). They all take calibrated actions in response to disinformation in this OSN.

Our proposed game-theoretic opinion models belong to the category of a networked game in social network (i.e., game on networks), which has several properties [50]: (1) A large number of participating players exist in a social network; (2) The given social network mediates the interactions between players and payoffs where the players only interact with their friends and each player's payoff is estimated based on their opinion status updated based on their interaction with other friends and the way they interact (i.e., an opinion model); (3) It is not feasible to derive an exhaustive table to specify payoffs because there is randomness for online users (players) to interact; and (4) The proposed game considers dynamic, gradual interactions among players. Nash Equilibrium (NE) can predict the strategy distributions if all players know the moves of the other players [51, 52]. However, the actual choices of the human players (i.e., online users) may be different. According to behavioral game theory [53], this finding is likely due to the bounded rationality (e.g., limited memory or a lack of perfect observability) and simultaneous moves in each interaction; and (5) The collective user behaviors may vary over time due to different choices of actions taken in each single-shot interaction.

Each user's opinion model reflects one's preference in updating an opinion based on one's behavioral propensity, so that the formulation of the utility in a game follows this preference. That is, users will aim to maximize their utility based on their preferences. This may look irrational to other types of users who use different types of opinion models. However, this also reflects real-world situations.

Now we explain the goals, strategies, and payoffs of the key three players in detail below.

A. Agents' Features

We characterize each user i's key attributes by:

$$F(v_i) = [\omega_i, P_i^f, P_i^p, \phi_i^1, \phi_i^2, \rho_i, \mathrm{bw}_i], \tag{13}$$

where ω_i is the subjective opinion as shown in Section III-A by SL. P_i^f and P_i^p are the probabilities of i's feeding and posting activities, ϕ_i^1 and ϕ_i^2 are thresholds for i to identify the opinion difference between i and i's friend as described in Section III-D, and ρ_i is a tolerance threshold of i to judge the current friend to be a suspect attacker upon exchanging opinions (discussed in Section V-C). The bw_i refers to user

i's betweenness which is commonly used to capture user i's structural social capital [54] (i.e., higher betweenness refers to higher structural social capital). In this work, we use bw_i to examine the impact of disinformation on betweenness where individual users' structural social capital (or bridging) is analyzed using the betweenness. We chose 'betweenness' to measure an online user's structural social capital (i.e., bridging). The reason is because the betweenness centrality has been commonly used as the indicator of a person's bridging capability, which is the core aspect of social capital [55, 56, 57]. Based on our investigation in [23], we found that 'betweenness' represents one of the highly comparable centrality metrics (e.g., much more powerful than k-shell [58], collective influence [59], or redundancy [54]).

B. Attacker Model

The malicious users, as attackers, deploy information deception tactics [16] to disseminate disinformation, fabricate or block true information to mislead the beliefs of legitimate users. SL opinion framework featuring uncertainty can materialize the theoretical **deception strategies** [16] where the whole set of strategies is denoted by $A = \{a_1^A, a_2^A, a_3^A, a_4^A\}$:

- *Degradation* (DG; a_1^A) is to confuse legitimate users by injecting noises into true information. In an SL-based opinion, DG is modeled by sending out a highly uncertain opinion, as $(b, d, u, a) = (0, 0, 1, 0.5)$.
- *Corruption* (C; a_2^A) generates false beliefs by injecting disinformation or replacing true information with false information. We model this by replacing a received opinion with a completely opposite opinion and then sharing it with a friend. For example, if an attacker receives an opinion $(b, d, u, a) = (0.7, 0.2, 0.1, 0.3)$ from a friend, the attacker forwards $(b, d, u, a) = (0.2, 0.7, 0.1, 0.3)$ to other friends.
- *Denial* (DN; a_3^A) is to prevent users from accessing true information by the way of inhibiting true information flow. It can cause the vacuity of information sources which increases uncertainties and difficulties to users in judging the truthfulness of information.
- *Subversion* (S; a_4^A) refers to an attacker's deception by changing the user's processing of perceived inputs. The aim of this attack is to make targeted user trust credible information less while using more of non-credible information. To launch an S attack, the attacker will always forward false opinions, as discussed by ω_F in Section III-B, to consistently increase the volume of disinformation.

As detailed above, an attacker's strategy, DG, C, or DN, will modify the opinions received from its friends and share them for efficient propagation of the uncertain, noisy, conflicting opinions, instead of pair-wise interactions. Each attacker chooses a strategy with the highest expected payoff value in our game model. We calculate an attacker i's **expected payoff** of a strategy k by considering the weighted sum of utility $u_{k\ell m}^{ij}$, caused by a specific condition of defender's strategy ℓ and user's strategy m. The expected payoff of strategy k is:

$$EP_k^{A_i}(a^D, a^U) = \sum_{\ell \in D} \sum_{m \in U} p_\ell^D \cdot p_m^U \cdot u_{k\ell m}^{ij}, \tag{14}$$

where D and U are the collections of all strategies for the defender and users. p_ℓ^D is a defender's expectation of strategy ℓ (i.e., either a_1^D or a_2^D by terminating a suspect immediately or monitoring with caution). p_m^U is a user's expectation of strategy m (i.e., one of $a_1^U - a_3^U$). The attacker can obtain the probability distribution of taking each strategy by the defender and user based on the historical observations of p_ℓ^D and p_m^U.

The utility of a specific condition based on all assumptions, $u_{k\ell m}^{ij}$, is defined by benefit over an incurring loss as:

$$u_{k\ell m}^{ij} = \mathrm{ds}(k, m, \omega_i, \omega_j) - g_\ell, \tag{15}$$

where the attacker's benefit $\mathrm{ds}(k, m, \omega_i, \omega_j)$ refers to how much attack strategy k contributes to making user j's opinion ω_j closer to false opinion ω_F in Section III-B, by the cosine similarity of ω_j and ω_F, when k is taken and not taken ($\neg k$):

$$\mathrm{ds}(k, m, \omega_i, \omega_j) = \mathrm{s}(k, m, \omega_F, \omega_j) - \mathrm{s}(\neg k, m, \omega_F, \omega_j), \tag{16}$$

where the first cosine similarity $\mathrm{s}(k, m, \omega_F, \omega_j)$ from Eq. (8) considers the updated user's ω_j when attacker i has strategy k and user j has strategy m. The second cosine similarity $\mathrm{s}(\neg k, m, \omega_F, \omega_j)$ refers to j's opinion when attacker i shares a legitimate opinion rather than a deceptive (or false) opinion with strategy k.

The cost g_ℓ measures attacker i's loss when the defender takes action ℓ where g_ℓ is estimated by the mean similarity between true opinions ω_T and all existing individual user's opinions, ω_j's.

C. Defender Model

A defender, as an OSN administrator, aims to ensure a secure and safe OSN by not tolerating any presence of malicious users propagating disinformation. If the defender receives N_R misconduct reports from legitimate users, the defender can take the following **strategies** whose set is $D = \{a_1^D, a_2^D\}$:

- *Terminating a malicious user* (T; a_1^D) is to remove the account and the corresponding connections with his/her all friends aiming to ensure the safety and security of the given OSN. However, if the suspended user is actually a legitimate user (i.e., false-positive), the user's reputation is ruined. If then, the user can lose his/her social capital due to the removal of all connections with others.
- *Monitoring a suspect user* (M; a_2^D) is to monitor a suspected user with no other actions. If this user is malicious, he/she may keep performing deception attack strategies and can endanger the security and safety of the given OSN. However, if the detected user is a legitimate user (i.e., false-positive), he/she can maintain current relationships with other users and social capital.

The defender selects the strategy with a higher expected payoff value. We quantify the defender's **expected payoff** of strategy ℓ by taking only an attacker's strategies, not a legitimate user's strategies, into considerations because the user's activities are not related to the defender's goal. The defender's expected payoff is given by the weighted sum of utility $u_{\ell k}^D$ as:

$$EP_\ell^D(a^A) = \sum_{k \in A} p_k^A \cdot u_{\ell k}^D, \tag{17}$$

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130

where a^A is an action taken by the attacker with a set of attack strategies A and p_k^A is the probability that the attacker chooses strategy k. The defender can learn p_k^A based on historical reports of reported attackers. If a suspect user is reported by other legitimate users at least N_R times, the defender can make decisions toward this suspect based on the payoffs in Eq. (17). However, if the defender decides the reported suspect as an attacker but it is a legitimate user indeed, it is a false positive and the legitimate user is evicted. The $u_{\ell k}^D$ is the utility of defense strategy ℓ with the attacker's deceptive strategy k. In addition, the defender evaluates the utility of each strategy ℓ by the overall impact on the OSN, as the gain over cost, as:

$$u_{\ell k}^D = \mathrm{ds}(\ell, k, \omega_T, \omega') - c_\ell, \tag{18}$$

where defender's gain $\mathrm{ds}(\ell, k, \omega_T, \omega')$ is the protection of OSN by strategy ℓ, which refers to how closer the affected opinions w''s of all the legitimate users are to true opinion ω_T. This gain calculates two cosine similarities in Eq. (8), for the cases of taking strategy ℓ over not taking it ($\neg\ell$), as follows:

$$\mathrm{ds}(\ell, k, \omega_T, \omega') = \mathrm{s}(\ell, k, \omega_T, \omega') - \mathrm{s}(\neg\ell, k, \omega_T, \omega'), \tag{19}$$

The defender's loss term c_ℓ quantifies the cost for the defender's strategy ℓ as two constants, i.e. $c_T = 0.1$ for $T(a_1^D)$ and $c_M = 0$ for $M(a_2^D)$.

D. User Model

OSN users consume useful information and interact with their friend users. We consider five user types corresponding to the opinion models described in Section III-C, including *Uncertainty-based* (U), *Homophily-based* (H), *Assertion-based* (A), *Herding-based* (HE), and *Encounter-based* (E) user types. The set of users' **strategies** is denoted by $U = \{a_1^U, a_2^U, a_3^U\}$ where each strategy is described as follows:

- *Updating and sharing* (SU; a_1^U) are to update the current opinion based on the received opinion and then share with other friends.
- *Updating* (U; a_2^U) is to update the current opinion based on the received opinion.
- *No updating* (NU; a_3^U) is to ignore a received opinion and keep the current opinion without new updates.

In this game, user i interacts with a friend user j, which can be either a user or an attacker. We assume that user i can be aware of user j's type based on the historical experience on the types of encountered users. That is, user i can estimate the probability of the interacted friend user j being an attacker by $p_{U_i}^{A_j}$, and a legitimate user by $p_{U_i}^{U_j} = 1 - p_{U_i}^{A_j}$ (i.e., not an attacker). Since a specific role of user j has a set of strategies when user j is an attacker or user, user i needs to choose a strategy m with the highest **expected payoff**, by a weighted sum of the utilities when j is either an attacker or a user:

$$EP_m^{U_i}(a^{U_j}) = p_{U_i}^{A_j} \cdot u_m^{U_i A_j} + (1 - p_{U_i}^{A_j}) \cdot u_m^{U_i U_j}, \tag{20}$$

where $u_m^{U_i A_j}$ or $u_m^{U_i U_j}$ is the utility based on user j's type. For an attacker j, the utility $u_m^{U_i A_j}$ is defined by the weighted loss caused by each of attackers' strategy, if user i accepts attacker j's opinion. If i choose NU, as $m = a_3^U$, $u_m^{U_i A_j}$ is 0, in:

$$u_m^{U_i A_j} = \begin{cases} \sum_{k \in A} p_k^{A_j} \cdot -s(m, \omega_F, \omega_i, \omega_j) & \text{if } m = a_1^U \text{ or } a_2^U; \\ 0 & \text{if } m = a_3^U, \end{cases} \tag{21}$$

where $p_k^{A_j}$ refers to user i's belief that attacker j takes strategy k based on the user's historical observations. However, it may not be perfect; thus we considered 90% accuracy of the belief accuracy. The cosine similarity $\mathrm{s}(m, \omega_F, \omega_i, \omega_j)$ by Eq. (8) is between ω_F and the expected opinion of user i encountering attacker j taking deceptive strategy k.

If j is a user type, the utility $u_m^{U_i U_j}$ is defined based on user j's strategy m' in $U_j = \{a_1^{U_j}, a_2^{U_j}, a_3^{U_j}\}$ and user j's OM type. Hence, $u_m^{U_i U_j}$ is given by:

$$u_m^{U_i U_j} = \begin{cases} \sum_{m' \in U_j} p_{m'}^{U_j} \cdot uc_{i'}^{j'} & \text{if user } j \text{ is U-type;} \\ \sum_{m' \in U_j} p_{m'}^{U_j} \cdot hc_{i'}^{j'} & \text{otherwise,} \end{cases} \tag{22}$$

where $p_{m'}^{U_j}$ is the belief of choosing strategy m' for user j. To compute $uc_{i'}^{j'}$ and $hc_{i'}^{j'}$, we first update both users i and j's opinions $\omega_i^{'}$ and $\omega_j^{'}$ by taking strategy m for user i and strategy m' for user j if an update is required for any of m or m'. Afterward, the discounting operators $uc_{i'}^{j'}$ and $hc_{i'}^{j'}$ from Eqs. (6) and (8), respectively, are produced by $\omega_i^{'}$ and $\omega_j^{'}$.

The proposed three-player game is constructed by a series of repeated subgames, each of which is a game of incomplete and imperfect information. When each player plays, the opponent can be either one type or the other (e.g., a user or an attacker when a user plays). Since the opponent's type is unknown in advance, each player will estimate its belief on the type of the opponent as described in this section. This is well-aligned with real-world scenarios because real situations are always filled with many aspects of uncertainties. To better reflect the imperfect observability of each player in its beliefs towards the opponents, we also modeled each player's limited observability with 90% accuracy.

However, in game theory, Nash Equilibrium (NE) solutions are used to provide players' best strategies assuming that each player has a correct belief about of its opponent's move. We also formulated an incomplete information game with NE solutions that are compared against the strategy selections by the players under uncertainty, which are described in this section. In the supplement document, Appendix A elaborates the Nash game with three specific examples of how NE solutions can maximize the benefits of all players from the normal-form game trees and normal-form payoffs matrix, along with the corresponding detailed explanations on them. We elaborated the interactions between the three key players in a single-shot interaction in Fig. 1.

V. Experimental Setup

In this section, we describe the datasets, metrics, and environmental setup used for the simulation experiment settings.

A. Datasets

From two real Twitter datasets, we obtained active accounts that have tweeting or retweeting behaviors. These datasets

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130

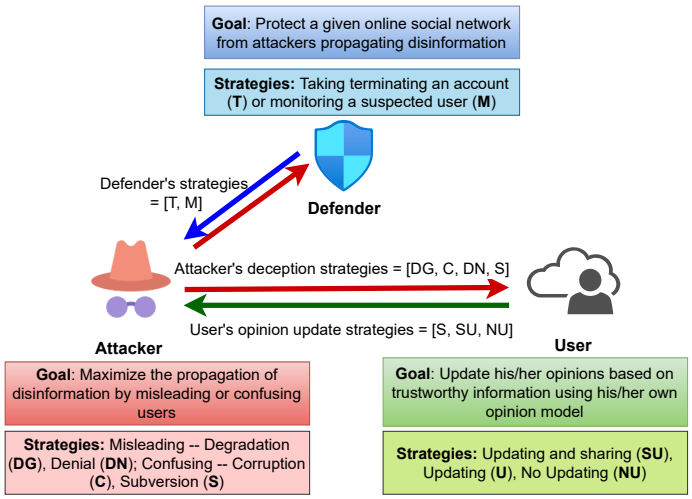


Fig. 1. The pairwise interactions of three agent roles, attackers, users, and a defender in our networked game.

have a broad range of metadata, including user profiles, followers lists, tweets activities and contents. Thus, we extracted complex social behaviors features, such as favorite tweets, activity networks, and tweeting frequencies, to initialize individual sharing likelihood as P_j^f and P_j^p. We describe the two Twitter datasets as below:

- *1KS-10KN* [25, 26]: It contains 10,766 active users, with 9,766 legitimate users of 7,744 average followers and 1,000 spammers with 2,520 average followers.
- *Cresci15* [27]: It has 2,664 active accounts, including 718 fake bots. The 1,946 legitimate users have 398 followers on average while the 718 attackers have 554 average followers.

B. Metrics

We used the following metrics to evaluate the performance of the considered opinion models:

- *Opinions of agents* ($\omega_i = (b_i, d_i, u_i, a_i)$): SL opinions cover four dimensions, including belief (b_i), disbelief (d_i), uncertainty (u_i), and base rate (a_i) as introduced in Section III-A. This metric can show the trends of opinions being diverged or converged as more user interactions are performed. Hence, this metric will allow us to observe the extent of opinion polarization in an OSN.
- *Probability distribution of best-taken strategies*: This metric measures the frequency of a player's chosen strategies with highest payoffs during all repeated subgames. This metric enhances our understanding of each player's preferences in both our proposed game under uncertainty and a Nash game.
- *Structural social capital* (stc_i): We chose betweenness centrality [60] to measure each user i's structural social capital, denoted by bw_i. Structure social capital (STC) measures how a person is connected with other people in a social network. STC is often captured as the degree of bridging in social capital, which is measured by betweenness [54]. We define a user's STC based on the sum of betweenness of the user's friends where we use a normalized STC, denoted by stc_i as a real number ranged $[0, 1]$ by:

$$\mathrm{stc}_i = \exp\left(\frac{-1}{\sum_{k \in F_i} \mathrm{bw}_k}\right), \tag{23}$$

TABLE II
KEY PARAMETERS, THE EXPLANATIONS, AND DEFAULT VALUES

Param.	Explanations	Default Value
P_{true}	Fraction of true informers	0.1
P_{false}	Fraction of false informers	0.1
ξ	Threshold for uncertainty maximization	0.05
ϕ_i^1	Criterion to request or accept a friend	Normal(0.1, 0.1)
ϕ_i^2	Upper bound to unfriend in uncertainty-based OM	0.5
P_i^f	Likelihood of agent i's feeding behavior	0.142 (mean)
P_i^p	Likelihood of agent i's posting behavior	0.186 (mean)
ω_T	A true opinion with high belief and base rate	$(1, 0, 0, 1)$
ω_F	A false opinion with high disbelief	$(0, 1, 0, 0)$
c_ℓ	Cost of a defender's taken strategies	T: 0.1, M:0
N	Size of experiment sample	1,000
I	Number of interactions to choose a strategy	200
ρ, μ, σ	Tolerance of a malicious user to generate a report	Normal(0.5, 0.05)
N_R	Number of malicious reports to alert the defender	3
$p_{U_i}^{A_j}$	Expectation of j as an attacker by nature	0.1

where this metric reveals how individual users' disinformation propagation and information processing methods can affect the degree of the STC. This can help identify key factors of creating or breaking the STC (i.e., bridging) when people exchange their opinions and update them. We use Freeman's betweenness centrality [60] for bw_i, which measures the degree of the shortest paths between two pairs of nodes going through given node i, representing node i's betweenness. If a tiny amount of users have high betweenness, it can represent the unbalanced distribution of network influence or power.

- *Trust* (T_i): Trust measures the level of relational social capital (RST), which is well-aligned with the bonding, among users. User i's trust, T_i, is quantified by how much other friend users, j's, trust user i and given by [61]:

$$\mathrm{T}_i = \frac{1}{2|\mathrm{F}_i|} \sum_{j \in \mathrm{F}_i} (\mathrm{T}_{ji}^f + \mathrm{T}_{ji}^p), \tag{24}$$

where F_i is a set of user i's friends. T_{ji}^x is user j's trust in user i in activity x, including feeding (f) and posting (p) behaviors, derived from their social interactions, such as sharing information. The T_{ji}^x is calculated based on x activities, including the number of feeding (f) or posting (p) interactions, I_{ji}^x, between user i and user j by:

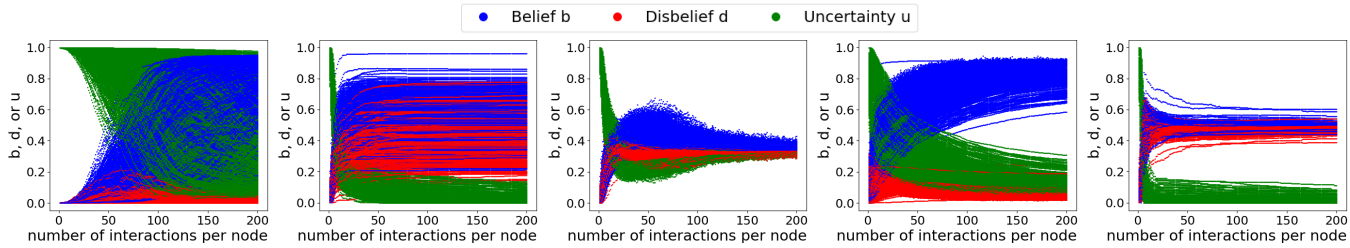
$$\mathrm{T}_{ji}^x = \frac{\mathrm{I}_{ji}^x}{\max(\mathrm{I}_{jk}^x \text{ for } k \in \mathrm{F}_j)}, \quad x \in \{f, p\}. \tag{25}$$

- *Network communities*: Communities are generated by three popular graph partitioning algorithms, including Kernighan–Lin bipartition algorithm [18], greedy modularity maximization algorithm [17], and label propagation-based algorithm [19].
- *Polarization scores*: Polarization refers to increasing differences in social, political, or attitudinal aspects between groups [62]. Accordingly, a polarization score is an indicator of how different two groups are in those various characteristics of people in the groups. We consider methods measuring network polarization scores, including modularity [17], community boundary connectivity [20], random walk controversy [21], and community performance [22].

C. Environment Setup

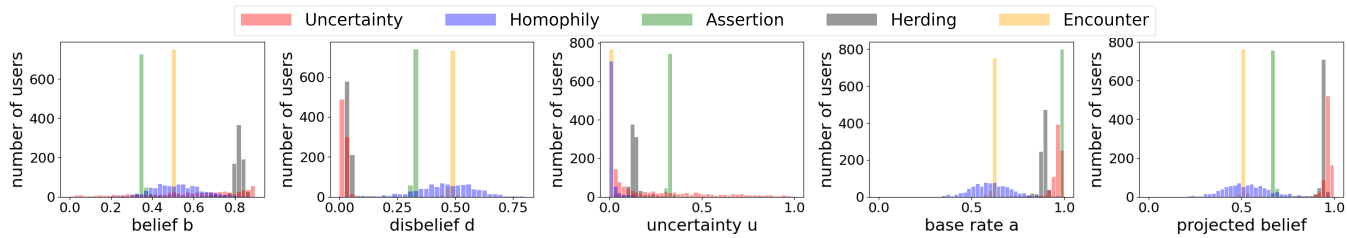
We consider an OSN with N=1,000 users randomly selected from each of *Cresci15* and *1KS-10KN* datasets for simulations.

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130



(a) Uncertainty-based OM (b) Homophily-based OM (c) Assertion-based OM (d) Herding-based OM (e) Encounter-based OM

Fig. 2. The evolution of SL-based opinions of all legitimate users over 200 interactions in belief (b, blue), disbelief (d, red), and uncertainty (u, green) under the dataset 1KS-10KN [25, 26].



(a) b (b) d (c) u (d) a (e) $P(b)$

Fig. 3. The histograms of three opinion masses (i.e., (b, d, u)), base rate a, and the projected belief $P(b)$ for all normal users after 200 interactions: Each subfigure shows the histograms for the five OMs in a given element under the dataset 1KS-10KN [25, 26].

We assign the users with the top 20% degrees as true informers and false informers (or attackers), with P_{false}=P_{true}=10% users for each. The rest of the users (i.e., 80%) belong to a same type, i.e. one of the U, H, A, HE, or E types described in Section IV-D. Each individual has a uniform distribution of possible actions for the three player types. This strategy distribution can be refreshed according to Dirichlet distribution [63] where the new outcome observed from each subgame contributes to a unit of supporting evidence. The top 20 topics of a user are generated by applying the latent Dirichlet allocation (LDA) algorithm to a collection of his tweets and retweets. Based on lacking most followers from the real followers list, users connect to others with the highest topic similarity score [64] to form the initial friending network. We exclude an attacker-attacker relationship to maximize the coverage of disinformation propagation. The experiment is repeated by 100 runs and each run covers $I = 200$ interactions. Table II summarizes the key parameters and default values.

In the first interaction, each player starts a subgame by:

- **1-A**: An attacker takes Subversion (S) action by disseminating false opinions ω_F to one neighbor user.
- **1-B**: For all legitimate users, user i first chooses friend user j based on user j's information sharing tendency P_{ij} in Eq. (11). After then, user i randomly selects a strategy, SU, U, or NU. If SU or U is chosen to update i's opinions, user i will use the OM in Section III-C, corresponding to his type, and choose a strategy based on the payoff described in Section IV-D.
- **1-C**: If user i accepts attacker j's opinion and accordingly updates ω_i, attacker j will modify ω_i with j's chosen deception strategy to share the deceptive opinion with another friend user in the next interaction, aiming to increase the uncertainty of other users' opinions in the OSN.
- **1-D**: Each user decides to add a new friend or remove an interacting friend in this interaction based on Section III-D.

This repeated game continues until the Ith interaction as:

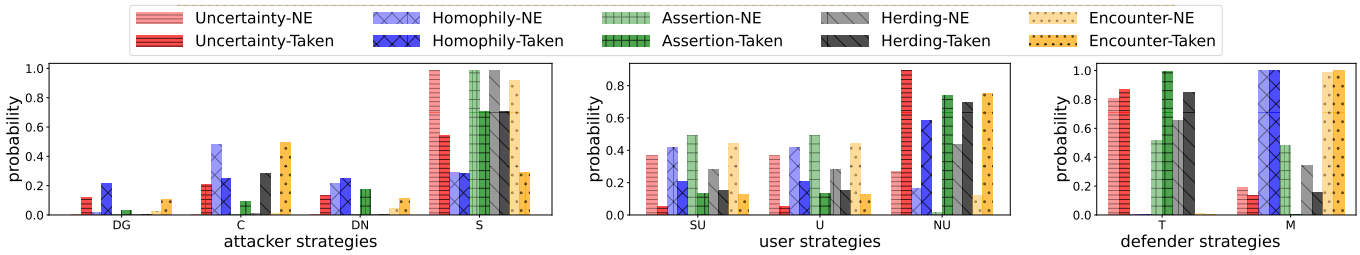
- **2-A**: All legitimate users follow step 1-B except for deciding a best strategy based on the expected payoffs of available strategies in Eq. (20). User i can report to the defender (i.e., an OSN service provider) that the interacting friend user j is malicious if their opinion difference based on Eq. (12) is larger than the tolerance threshold ρ. To model the different ρ by individual users, ρ follows a Gaussian curve with mean μ and standard deviation σ. Depending on the level of an individual user's ρ, false positives can be generated.
- **2-B**: Each attacker selects a friend based on Eq. (11) and decides a best deceptive strategy based on the expected payoffs of the attack strategies in Eq. (14) to spread a deceptive opinion by taking step 1-C.
- **2-C**: When the defender realizes that a suspect user receives at least N_R misconduct reports, the defender can determine the best strategy based on the expected payoffs of the available strategies in Eq. (17).
- **2-D**: Each user repeats step 1-D to maintain the friend network.

Note that the roles of attackers and users do not change. In particular, regardless of whether the users hold true information or disinformation, they will propagate their opinions based on their sharing behaviors and update their opinions based on their preferences via various opinion models.

VI. Simulation Results & Analysis

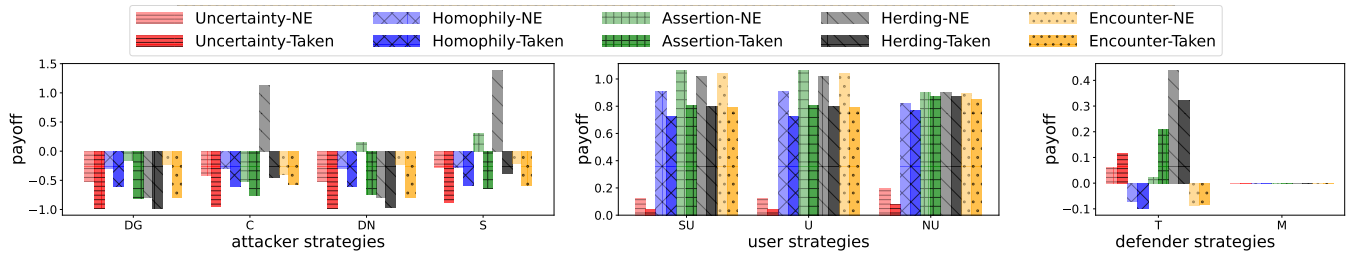
We presented the results under *1KS-10KN* [25, 26] and discussed their underlying trends in this section. Under *Cresci15* [27], we observed the overall trends are significantly similar. Due to the space constraint, we placed the results using the *Cresci15* in Appendix B of the supplement document. By default, we presumed 10% true informers, 10% false informers, and 80% of the agents (i.e., legitimate users) all

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130



(a) Attackers (b) Users (c) Defender

Fig. 4. The probability distributions of the taken strategies by each player type based on NE solutions and the solutions by our proposed game under the dataset 1KS-10KN [25, 26].



(a) Attackers (b) Users (c) Defender

Fig. 5. The comparison of the average payoffs of the strategies taken by each player type based on the NE and our proposed games under the dataset 1KS-10KN [25, 26].

used the same opinion model where each opinion model (OM) is detailed in Section III-C.

A. Uncertain Opinions in Disinformation Propagation

Fig. 2 illustrates the trends of opinion dynamics under the five OMs throughout the 200 interactions in each subfigure. The three opinion dimensions (b, d, u) as belief, disbelief, and uncertainty, are plotted by contrasting colors. As expected, different opinion OMs introduce distinct impacts on the dynamics of opinions. From the view of belief b supporting true information, opinions from U-type and HE-type OMs have the top two highest b in Figs. 2(a) and 2(d). U-type users can form the highest uncertainty u as well. In Fig. 2(a), the increase of accepting true information via more interactions with friends is mainly caused by the use of uncertainty when two users update their opinions based on uncertainties from both opinions. Moreover, U-type users tend to neglect the attackers' noisy, deceptive, and uncertain opinions to amplify false information or disturb true information. In Fig. 2(b), H-type users have opinions with low uncertainties and less compromise with other opinions, indicating high opinion polarization. On the other hand, all opinions from both the A-type and E-type OMs can form a consensus with low beliefs and low uncertainties while E-type OM in Fig. 2(e) has a similar level of b and d. A-type OM in Fig. 2(c) maintains equal b, d, and u.

Fig. 3 examines the histograms for five OMs in the same range by analyzing all users' opinions after all interactions. Besides the four opinion components, we plot projected beliefs, $P(b)$'s, to know each user's decision preference. Those results are correlated with the final states in Fig. 2 and confirm the results of the U-type OM: the highest b in Fig 3(a) and $P(b)$ in Fig 3(e), lowest d in Fig 3(b), and high a in Fig 3(d) with the initial value of 0.5. Thus, disinformation is significantly mitigated by the U-type OM as well as U-type users' ability to report malicious users propagating highly uncertain opinions to the defender. H-type users all have b, d, and a around 0.5 and reduce uncertainty in the starting interactions. The results from the H-type OM suggest that users, who rely on homophily may trust disinformation more. This is because they cannot identify noisy and uncertain opinions if the observed opinion difference is less than ϕ^1.

B. Strategy Selection, Payoffs, and Nash Equilibrium

Fig. 4 shows how the selected strategies in our game-theoretic framework differ from the NE solutions under the five OMs. Recall that NEs are derived based on the assumption of perfect observations so that the players can intelligently achieve mutual benefits based on their accurate beliefs towards the opponents. The probability distributions of taken strategies in NE and our game are inconsistent in both attackers and users, as shown in Figs. 4(a) and 4(b), respectively. This is because players can only partially observe the opponents' previous actions in this repeated game.

In Fig. 4(a), we can observe that attackers' NE solutions for U, A, HE, and E-type users are dominated by S while those for H-type users favor C 40% more. The homophily-based OM produces more similar probability distributions of the attackers' strategies taken under both our and NE games than other OMs. Under the taken strategies from our game, the attackers in the H-type OM network have an equal chance for each strategy while the attackers in all other OM networks prefer S and C. Legitimate users' NE solutions in Fig. 4(b) for all OMs have a significant increase of U and SU to accept other opinions more, compared to the distributions of U and SU in our game. This comparison suggests that users in our designed model have more resistance to accept other risky or uncertain opinions when handling disinformation.

Our game model has the opposite situations for the users, where all OMs choose NU at least at 50% probability in

Fig. 4(b). Besides, U-type and H-type users have the least and most motivation to update opinions, respectively. The defender's choices from both NE and our game model fit well except for the assertion-based OM in Fig. 4(c). Under both conditions, the defenders in the U-type and HE-type networks are more likely to remove malicious accounts, while T is rarely selected in H-type and E-type networks. However, the defender in the A-type user network has a relatively equal chance of taking T or M strategy from NE solutions rather than having a strong bias to T in our game.

The payoff values in Fig. 5 can help understand the differences of best strategies between each user type and between NE solutions and the solutions taken by the players. The ranking of payoffs values for strategies under each player type (i.e., an attacker, user, or defender) in Fig. 5 correlates well with the best strategies chosen in Fig. 4. This result confirms the preferences of the best strategies for a specific player type. In Fig. 5(a), the payoffs from NE solutions are all higher than the corresponding strategies taken. Also, the payoffs from the strategies chosen in our game are all negative values, whereas the payoffs from NE-based strategies taken by A-type and HE-type users are positive values. The higher NE payoffs are also found for the user strategies in Fig. 5(b) and some of the defender's strategies, such as the defender in U-type, A-type, and E-types OMs in Fig. 5(c). In Fig. 5(b), the U-type users have much lower payoffs compared to other OMs because, in Eq. (22), the utility of user j depends on j's type. If user j is a U-type user, it will use $uc^j_{im'}$, which is quite different from the utility $hc^j_{im'}$ for all other type users j's. In Fig. 5(c), the payoffs of strategy M for all the OM types are zero due to the way we estimate the payoffs in Eq. (18).

C. Effect of Disinformation Propagation on Polarization

Fig. 6 compares network topology changes between the communities of the initial state and those after disinformation propagation under the five different opinion models (OMs). We used three community detection algorithms, as described in Section V-B, to generate network communities. We also demonstrated the community topologies by separating individual communities and placing users within the same communities closer. Each user has his/her opinion by updating their opinions through the interactions with other users after disinformation propagation. Hence, we plot the projected belief $P(b_i)$ of the individual user using a color bar, by showing complete belief supporting true information ($P(b_i) = 1$) as blue and perfect disbelief supporting disinformation ($P(b_i) = 0$) as red. We showed the initial uncertain opinions ($P(b_i) = 0.5$) in green in Figs. 6(a), 6(g), and 6(m). Although various community detection algorithms generate different communities, they all have the same trends to reveal users' opinion polarization caused by disinformation propagation. Most U-type users form their opinions as true information. On the contrary, H-type users in two communities believe either highly true or highly false information from all three community detection algorithms. This implies that disinformation can increase the polarization of the network if users update opinions based on like-minded attitudes, such as homophily. Herding-based OM also forms communities that have high beliefs similar to U-type OM. However, the lighter blue color in Figs. 6(e), 6(k), and 6(q) mean that the projected belief is lower than U-type communities. Both A-type and E-type user networks generate polarized communities although the polarization is not as strong as the H-type user network because the the node color shows higher beliefs (more blue-wish) in the A-type and E-type OM networks than the H-type user network.

Fig. 7 demonstrates the polarization measures under each of the three community algorithms. Polarization scores from the network influenced by U-type users are all lowered except for the community boundary scores within label propagation communities. The lower polarization denotes that in processing disinformation, U-type users reduce more gaps between communities. Polarization scores from the H-type users network in Fig. 7(c) increase under all four methods compared to the initial polarization scores. In the other two communities, the polarization scores increase under two methods while decreasing under the other two methods in Figs. 7(a) and 7(b). However, the H-type users have the highest polarization scores compared with other OMs. Higher polarization means a more polarized network when diffusing disinformation in the network of U-type users than in the network of other user types.

D. Effect of Disinformation Propagation on Social Capital

Fig. 8 plots the distribution of all users' bridging social capital in the five OMs before and after 200 user interactions upon disinformation propagation. Since most betweenness values are below 0.001, we only plot the ones large than 0.001 for better visualization. In all the five user types, we can observe the remarkable decrease of betweenness after 200 interactions in Fig. 8(f), showing the structural social capital (STC) of all the nodes in the initial and five OMs networks. The lower betweenness indicates weaker structural holes because the network after disinformation propagation is connected less than the original network. For H-type users in Fig. 8(b), the users have the largest betweenness compared to other types. This implies that H-type users have more partial network power and influence by having more structural holes with higher STC.

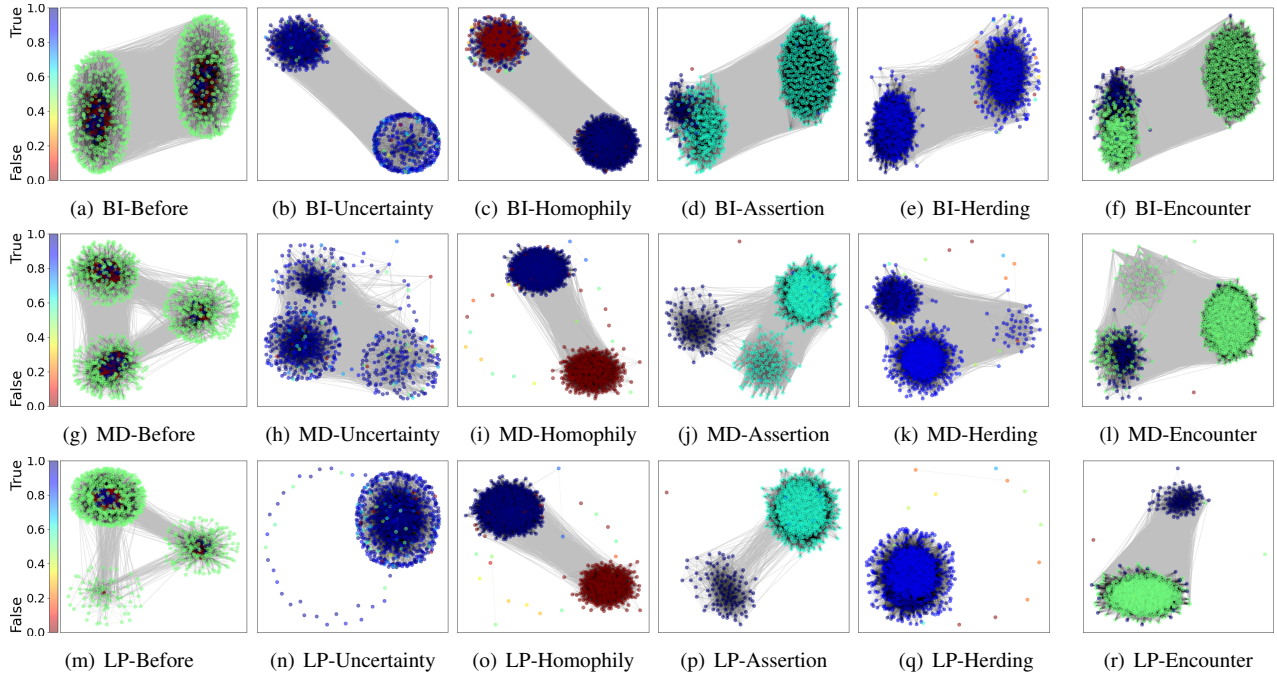
Fig. 9 plots the distribution of all users' trust, as relational social capital (RC), in the five OMs after 200 user interactions upon disinformation propagation. All the five OMs show a lower RC after the interactions in Fig. 9(f). Similar to the STC, trust values in H-type networks are the highest among all five OM networks. This implies that in the H-type network, users with high homophily gather more tightly. Hence, it may be hard for the H-type users to be with other users with a lack of similarity, leading to high opinion and network polarization.

VII. Conclusion

Here we conclude our work as below:

- Uncertainty-based OM can assist users in excluding false, contradicting, and uncertain information and to believe and accept true information. However, other OMs based on homophily, assertion, and encounter can easily mislead users to believe false information from Section VI-A.

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130

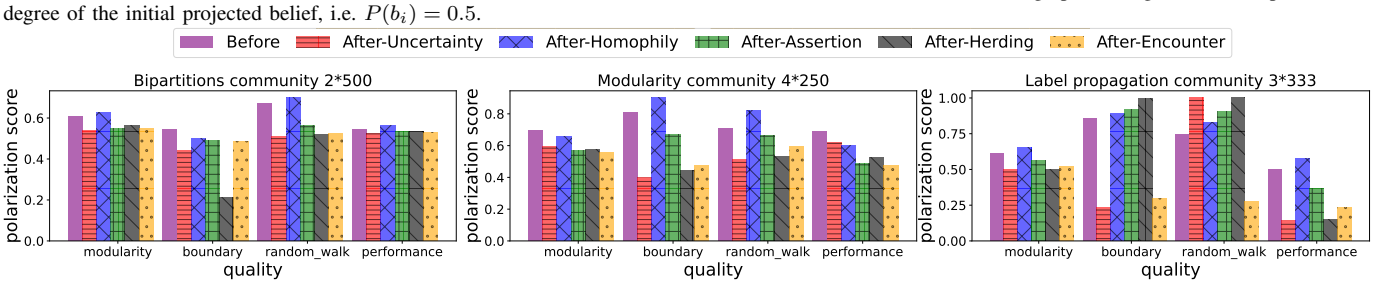


(a) BI-Before (b) BI-Uncertainty (c) BI-Homophily (d) BI-Assertion (e) BI-Herding (f) BI-Encounter

(g) MD-Before (h) MD-Uncertainty (i) MD-Homophily (j) MD-Assertion (k) MD-Herding (l) MD-Encounter

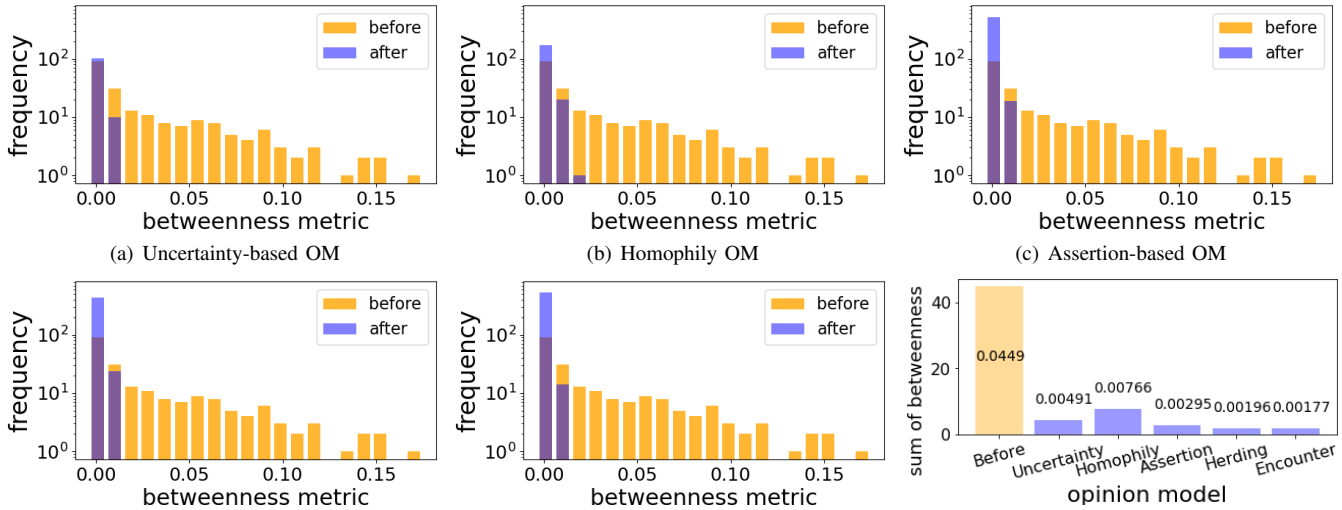
(m) LP-Before (n) LP-Uncertainty (o) LP-Homophily (p) LP-Assertion (q) LP-Herding (r) LP-Encounter

Fig. 6. The community plots of network topology based on three community detection algorithms, bipartitions, modularity, and label propagation, to show networks before and after disinformation propagation under the five OMs under the dataset 1KS-10KN [25, 26]. We observe distinct communities formed depending on a different community detection algorithm used with disparate clusters in the plots. The node colors reflect the projected belief $P(b_i)$ of each node in a color bar, where belief in true information is in blue and belief in disinformation is in red. Green in subgraphs (a), (g), and (m) represents the degree of the initial projected belief, i.e. $P(b_i) = 0.5$.



(a) Two Bipartitions Communities (b) Four Modularity Communities (c) Three Label Propagation Communities

Fig. 7. The polarization scores from the four polarization metrics measuring the graph partition of three community algorithms under the dataset 1KS-10KN [25, 26]: Each metric compares the networks before and after 200 user interactions of users under the five opinion models after disinformation is propagated over an OSN.

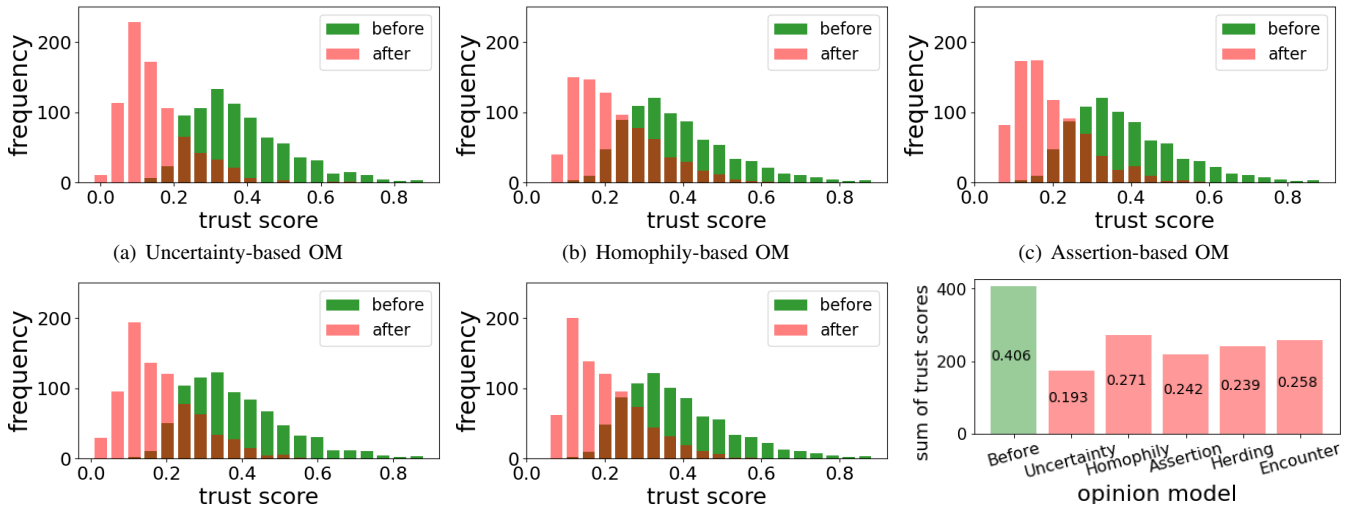


(a) Uncertainty-based OM (b) Homophily OM (c) Assertion-based OM

(d) Herding-based OM (e) Encounter-based OM (f) Sum betweenness

Fig. 8. The betweenness scores before and after 200 user interactions and opinions updates with respect to the five opinion models (OMs) under the dataset 1KS-10KN [25, 26]. In (f), we also indicated the mean betweenness of an individual user in the label on the top of each bar.

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2022.3181130



(a) Uncertainty-based OM (b) Homophily-based OM (c) Assertion-based OM

(d) Herding-based OM (e) Encounter-based OM (f) Sum trust

Fig. 9. The trust scores before and after 200 user interactions under the five opinion models (OMs) under the dataset 1KS-10KN [25, 26]. In (f), we also indicated the mean trust of an individual user in the label on each bar.

- Users using the uncertainty-based OM can also contribute to mitigating disinformation by effectively reporting suspicious friends to defenders, i.e., an OSN administrator. The misconduct reports can help defenders terminate malicious users to block the negative influence of disinformation from Section VI-B.
- The inconsistency of the actions taken in our game and NE game, shown in Figs. 4 and 5, Section VI-B, is because our game model considered more realistic action scenarios where players may not perfectly keep track of correct observations towards the opponents' previous actions.
- All results from network polarization analyses in Section VI-C strongly support that the uncertainty-based opinion model can help users share uniform opinions and become united. On the contrary, the homophily-based opinion model makes users tend to divide into more polarized opinion groups.
- Disinformation propagation can cause the decrease of social capital in terms of both structural social capital and relational social capital. Homophily-type users have more partial network power and gather more tightly than other types from Section VI-D.
- A user's game-theoretic process of disinformation propagated in a network can significantly affect the dynamics of the network in terms of network topology and network influence, which is represented by communities, polarization, and social capital. Compared to all other OMs, homophily-based OM can cause the highest polarization in the network, while uncertainty-based OM can help users access true information best.

References

[1] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[2] J.-H. Cho, "Dynamics of uncertain and conflicting opinions in social networks," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 518–531, 2018.

[3] J.-H. Cho, S. Rager, J. O'Donovan, S. Adali, and B. D. Horne, "Uncertainty-based false information propagation in social networks," *ACM Transactions on Social Computing*, vol. 2, no. 2, Jun. 2019.

[4] J. S. B. Evans, *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc, 1989.

[5] A. G. Greenwald, "The totalitarian ego: Fabrication and revision of personal history." *American Psychologist*, vol. 35, no. 7, p. 603, 1980.

[6] J. Kruger and D. Dunning, "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments." *Journal of Personality and Social Psychology*, vol. 77, no. 6, p. 1121, 1999.

[7] D. Dunning, "The Dunning–Kruger effect: On being ignorant of one's own ignorance," in *Advances in Experimental Social Psychology*. Elsevier, 2011, vol. 44, pp. 247–296.

[8] 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," *Physica A*, vol. 390, pp. 2619–2625, 2011.

[9] L. Zhao, J. Wang, Y. Chen, Q. Wang, J. Cheng, and H. Cui, "SIHR rumor spreading model in social networks," *Physica A*, vol. 391, no. 7, pp. 2444–2453, 2012.

[10] L. Zhao, X. Qiu, X. Wang, and J. Wang, "Rumor spreading model considering forgetting and remembering mechanisms in inhomogeneous networks," *Physica A*, vol. 392, no. 4, pp. 987–994, 2013.

[11] H. Yang, "A consensus opinion model based on the evolutionary game," *Europhysics Letters*, vol. 115, no. 4, p. 40007, 2016.

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

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

[14] 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 Dynamics*, vol. 95, no. 1, pp. 523–539, 2019.

[15] H. Zhang, Y. Li, Y. Chen, and H. V. Zhao, "Smart evolution for information diffusion over social networks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1203–1217, 2020.

[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, 2018.

[17] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[18] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *The Bell System Technical Journal*, vol. 49, no. 2, pp. 291–307, 1970.

[19] G. Cordasco and L. Gargano, "Community detection via semi-synchronous label propagation algorithms," in *2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA)*, 2010, pp. 1–8.

[20] P. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2013.

[21] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *ACM Transactions on Social Computing*, vol. 1, no. 1, pp. 1–27, 2018.

[22] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[23] Z. Wan, Y. Mahajan, B. W. Kang, T. J. Moore, and J.-H. Cho, "A survey on centrality metrics and their network resilience analysis," *IEEE Access*, vol. 9, pp. 104 773–104 819, 2021.

[24] Z. Guo and J.-H. Cho, "Game theoretic opinion models and their application in processing disinformation," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 01–07.

[25] 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*. Springer, 2011, pp. 318–337.

[26] 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 *Proceedings of the 21st International Conference on WWW*, 2012, pp. 71–80.

[27] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake Twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.

[28] G. Szabó and C. Tőke, "Evolutionary prisoner's dilemma game on a square lattice," *Physical Review E*, vol. 58, no. 1, p. 69, 1998.

[29] D. Huang, L. Yang, P. Li, X. Yang, and Y. Y. Tang, "Developing cost-effective rumor-refuting strategy through game-theoretic approach," *IEEE Systems Journal*, 2020.

[30] K. Yoshikawa *et al.*, "A fake news dissemination model based on updating reliability and doubt among individuals," in *The 11th International Conference of Awareness Science and Technology (iCAST)*, 2020, pp. 1–8.

[31] D. Zinoviev and V. Duong, "A game theoretical approach to broadcast information diffusion in social networks," in *Proceedings of the 44th Annual Simulation Symposium*, 2011, pp. 47–52.

[32] A. Sonowal *et al.*, "An improved model for dynamic opinion updates in online social networks," in *2020 IEEE 4th CICT*, 2020, pp. 1–6.

[33] M. Zhan, H. Liang, G. Kou, Y. Dong, and S. Yu, "Impact of social network structures on uncertain opinion formation," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 4, pp. 670–679, 2019.

[34] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *PNAS*, vol. 113, no. 3, pp. 554–559, 2016.

[35] A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Homophily and polarization in the age of misinformation," *The Euro. Phys. Jour. Spec. Topics*, vol. 225, no. 10, pp. 2047–2059, 2016.

[36] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Transactions on the Web*, vol. 13, no. 2, pp. 1–22, 2019.

[37] 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+ Society*, vol. 5, no. 2, 2019.

[38] M. T. Al Amin, C. Aggarwal, S. Yao, T. Abdelzaher, and L. Kaplan, "Unveiling polarization in social networks: A matrix factorization approach," in *IEEE INFOCOM*, 2017, pp. 1–9.

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

[40] G. Asmolov. (2018) The disconnective power of disinformation campaigns. [Online]. Available: https://jia.sipa.columbia.edu/disconnective-power-disinformation-campaigns

[41] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, "Social media, political polarization, and political disinformation: A review of the scientific literature," *SSRN*, 2018.

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

[43] G. Verma, A. Swami, and K. Chan, "The impact of competing zealots on opinion dynamics," *Physica A*, vol. 395, pp. 310–331, 2014.

[44] L. Li, A. Scaglione, A. Swami, and Q. Zhao, "Consensus, polarization and clustering of opinions in social networks," *IEEE Journal on Selected Areas in Commu-*

nications, vol. 31, no. 6, pp. 1072–1083, 2013.

[45] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining (First Edition)*, Boston, MA, USA, 2005.

[46] D. J. Langley, M. C. Hoeve, J. R. Ortt, N. Pals, and B. van der Vecht, "Patterns of herding and their occurrence in an online setting," *Journal of Interactive Marketing*, vol. 28, no. 1, pp. 16–25, 2014.

[47] M. Heesacker, R. E. Petty, and J. T. Cacioppo, "Field dependence and attitude change: Source credibility can alter persuasion by affecting message-relevant thinking," *Journal of personality*, vol. 51, no. 4, pp. 653–666, 1983.

[48] S. J. Tobin and M. M. Raymundo, "Persuasion by causal arguments: The motivating role of perceived causal expertise," *Social Cognition*, vol. 27, no. 1, pp. 105–127, 2009.

[49] D. Price, "A general theory of bibliometric and other cumulative advantage processes," *J. American Society for Information Science*, vol. 27, no. 5, pp. 292–306, 1976.

[50] M. O. Jackson and Y. Zenou, "Chapter 3 – games on networks," ser. Handbook of Game Theory with Economic Applications, H. P. Young and S. Zamir, Eds. Elsevier, 2015, vol. 4, pp. 95–163.

[51] S. Tadelis, *Game Theory: An Introduction*. Princeton University Press, 2013.

[52] A. Sanjab, W. Saad, and T. Başar, "A game of drones: Cyber-physical security of time-critical UAV applications with cumulative prospect theory perceptions and valuations," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6990–7006, 2020.

[53] C. F. Camerer, *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2011.

[54] R. S. Burt, *Social Capital: Theory and Research*. New York: Aldine de Gruyter, 2001, pp. 31–56.

[55] J. Venkatanathan, E. Karapanos, V. Kostakos, and J. Gonçalves, "Network, personality and social capital," in *Proceedings of the 4th Annual ACM Web Science Conference*, 2012, pp. 326–329.

[56] C. M. Lakon, D. C. Godette, and J. R. Hipp, "Network-based approaches for measuring social capital," in *Social Capital and Health*. Springer, 2008, pp. 63–81.

[57] E. Y. Li, C. H. Liao, and H. R. Yen, "Co-authorship networks and research impact: A social capital perspective," *Research Policy*, vol. 42, no. 9, pp. 1515–1530, 2013.

[58] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[59] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.

[60] L. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[61] J.-H. Cho, I. Alsmadi, and D. Xu, "Privacy and social capital in online social networks," in *2016 IEEE Global Communications Conf. (GLOBECOM)*, 2016, pp. 1–7.

[62] J. McCoy, T. Rahman, and M. Somer, "Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities," *American Behavioral Scientist*, vol. 62, no. 1, pp. 16–42, 2018.

[63] R. Hankin *et al.*, "A generalization of the Dirichlet distribution," *Journal of Statistical Software*, vol. 33, no. 11, pp. 1–18, 2010.

[64] Z. Wang, J. Liao, Q. Cao, H. Qi, and Z. Wang, "Friendbook: A semantic-based friend recommendation system for social networks," *IEEE Transactions on Mobile Computing*, vol. 14, no. 3, pp. 538–551, Mar. 2015.

Zhen Guo received the MS degree in biological sciences and MS degree in computer science from Fordham University, New York in 2013 and in 2016. From 2016, he has been a PhD candidate in computer sciences at Virginia Polytechnic Institute and State University, Falls Church, Virginia. His recent research interests have been on understanding and combating online social deception by various concepts derived from social and behavioral theories and various AI-based techniques.

Jaber Valinejad (M'19) received his PhD from Virginia Tech's Bradley department of electrical and computer engineering, Greater Washington D.C., USA. He earned a master's degree in computer science from the same institution. He is currently a postdoctoral research fellow in the data and system science lab at Harvard University's Medical School in Cambridge, Massachusetts, USA. He was a fellow of the following NSF-sponsored interdisciplinary programs while earning his Ph.D. at Virginia Tech: 1) disaster resilience and risk management (DRRM); and 2) urban computing. His research focuses on community resilience, computational social science, energy, and socio-technical systems through the use of network science, data science, natural language processing (NLP), social sensing tools (such as Twitter and Google), machine learning, and optimization techniques.

Jin-Hee Cho (M'09; SM'14) is currently an associate professor in the department of computer science at Virginia Tech since 2018. Prior to joining the Virginia Tech, she worked as a computer scientist at the U.S. Army Research Laboratory (USARL), Adelphi, Maryland, since 2009. Dr. Cho has published over 160 peer-reviewed technical papers in leading journals and conferences in the areas of 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. She received the best paper awards in 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, Dr. Cho was selected for the 2013 Presidential Early Career Award for Scientists and Engineers (PECASE). She also a recipient of The 2022 Faculty Fellow Award in the College of Engineering at Virginia Tech. Dr. Cho received MS and PhD degrees in computer science from the Virginia Tech in 2004 and 2008, respectively. She is a senior member of the IEEE and a member of the ACM.