Uncertainty-based False Information Propagation in Social Networks

JIN-HEE CHO, Virginia Polytechnic Institute and State University, USA
SCOTT RAGER, Raytheon BBN Technologies, USA
JOHN O'DONOVAN, University of California, Santa Barbara, USA
SIBEL ADALI, Rensselaer Polytechnic Institute, USA
BENJAMIN D. HORNE, Rensselaer Polytechnic Institute, USA

Many network scientists have investigated the problem of mitigating or removing false information propagated in social networks. False information falls in two broad categories: disinformation and misinformation. Disinformation represents false information that is knowingly shared and distributed with malicious intent. Misinformation in contrast is false information shared unwittingly, without any malicious intent. Many existing methods to mitigate or remove false information in networks concentrate on methods to find a set of seeding nodes (or agents) based on their network characteristics (e.g., centrality features) to treat. The aim of these methods is to disseminate correct information in the most efficient way. However, little work has focused on the role of uncertainty as a factor in the formulation of agents' opinions. Uncertainty-aware agents can form different opinions and eventual beliefs about true or false information resulting in different patterns of information diffusion in networks.

In this work, we leverage an opinion model, called *Subjective Logic* (SL), which explicitly deals with a level of uncertainty in an opinion where the opinion is defined as a combination of belief, disbelief, and uncertainty and the level of uncertainty is easily interpreted as a person's confidence in given belief or disbelief. However, SL considers the dimension of uncertainty only derived from a lack of information (i.e., ignorance), not from other causes such as conflicting evidence. In the era of Big Data where we are flooded with information, conflicting information can increase uncertainty (or ambiguity) and have a greater effect on opinions than a lack of information (or ignorance). In order to enhance the capability of SL to deal with ambiguity as well as ignorance, we propose an SL-based opinion model that includes a level of uncertainty derived from both causes. By developing a variant of the SIR (Susceptible-Infected-Recovered) epidemic model that can change an agent's status based on the state of their opinions, we capture the evolution of agents' opinions over time. We present an analysis and discussion of critical changes in network outcomes under varying values of key design parameters, including the frequency ratio of true or false information propagation, centrality metrics used for selecting seeding false informers and true informers, an opinion decay factor, the degree of agents' prior belief, and the percentage of true informers. We validated our proposed opinion model using both the synthetic network environments and realistic network environments considering a real network topology, user behaviors, and the quality of news articles. The proposed agent's opinion model and corresponding strategies to deal with false information can be applicable to combat the spread of fake news in various social media platforms (e.g., Facebook).

CCS Concepts: • **Computing methodologies** → *Reasoning about belief and knowledge*;

This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

2469-7818/2019/2-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Additional Key Words and Phrases: Subjective logic, opinion, false information, misinformation, disinformation, uncertainty, ambiguity.

ACM Reference Format:
Jin-Hee Cho, Scott Rager, John O'Donovan, Sibel Adalı, and Benjamin D. Horne. 2019. Uncertainty-based False Information Propagation in Social Networks. *ACM Trans. Soc. Comput.* 1, 1, Article 1 (February 2019), 34 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Motivation

Thanks to the advent and proliferation of online social networks (OSNs), we are exposed to vast amounts of information that have often not been verified for their truthfulness. In various types of OSN applications, people publish their own opinions and amplify them by sharing without verification. Unverified rumors or false information can severely damage individuals' reputation or lives; when propagated en-masse, they can influence public opinion and detrimentally affect critical decisions. For example, false information can sway opinion during the process of electing political leaders. Although the concept of 'information warfare' is derived from the military domain, the popularity and proliferation of social network applications or media increased the adverse impact of false information or fake news, leading to the information warfare at individual, societal, and global levels [Brumley et al. 2012; Kopp 2000].

False information is often categorized as *misinformation* or *disinformation*. Although both terms refer to false information, they are distinguished based on the intent behind its creation and propagation. *Disinformation* is false information that disseminated with the explicit purpose of misleading others [Stahl 2006]. In the case of *misinformation*, agents may end up disseminating without such malicious intent due to a mistake regarding its veracity. Hence, the clear distinction can be observed in a sense that misinformers (i.e., people propagating misinformation) can change their false opinions as they obtain more and/or true information over time. On the other hand, disinformers (i.e., people propagating disinformation) never change their opinions, since they intend to change other people's opinions to promote their own agenda. Propagation of false information may introduce the following issues:

- Noisy information, such as unfiltered, unprofessional or unedited information, is disseminated without proper verification in many channels, social networks as well as some media outlets leading to an overload of information [Kumar et al. 2014]. Examples of noise include incomplete, obsolete, and biased information, pranks, contradictions, inappropriately translated information, information modified by unauthorized parties, non-factual information, or scholarly misconduct (e.g., intentionally manipulated scientific data) [Kumar et al. 2014]. Focusing attention on credible information (or valuable information) and filtering out noise is quite challenging. Sometimes, false information continues to be propagated even without clear purpose;
- If individuals without strong opinions on a topic are exposed to false information first, the prior belief formed by this exposure is hard to correct. This is called the "continued influence effect." To correct these beliefs, reasonable explanations should be provided in a timely manner, to prove wrongfulness of the misinformation. One should also propagate the counter-misinformation with care because it can "backfire" if it is disseminated improperly [Lewandowsky et al. 2012]; and

- If people have strong opinions regarding the false information, they are more likely to stick to this opinion in their beliefs and readily reject any opposite information [Lewandowsky et al. 2012].

1.2 Research Questions

To study how people update their opinions based on their interactions and understand how false information can be eradicated or mitigated by propagating true information, this work studies a social network where an agent is modeled after the way humans form opinions based on information that they are exposed to. In this work, an agent forms its opinion based on interactions with other agents and updates its opinions to a certain degree depending on its confidence in and perception of the other agent with whom it is interacting. In this work, we are particularly interested in modeling the dimension of uncertainty that can be derived from both a lack of information (called *ignorance*) and conflicting evidence (called *ambiguity*), which can play a significant role in making a final decision to believe in true or false information. In this work, we would like to answer the following key research questions:

- How can we formulate an agent's opinion based on a perceived level of uncertainty and upon interactions with other agents?
- How does conflicting evidence affect an agent's opinion in terms of the level of perceived uncertainty?
- How can an agent's prior belief or disbelief impact its decision to believe in true or false information?
- What are the key factors that can help to identify, mitigate or potentially eradicate false information in a social network?

1.3 Key Contributions

This work has the following key contributions:

- We develop an opinion model based on Subjective Logic (SL) which is used by agents in forming and updating their opinions upon interacting with other agents. In SL, an opinion is formulated to explicitly deal with uncertainty representing an agent's confidence in a given proposition. However, the current form of SL only considers the uncertainty derived from a lack of information to support a belief, called *ignorance*, not from conflicting evidence, called *ambiguity*. Hence, even if conflicting evidence is received, uncertainty in SL decreases. This situation leads to a sufficient level of belief and disbelief with almost zero uncertainty, but an agent cannot decide which information to believe (i.e., false information or true information) because the amount of information received at two extremes is almost the same. We provide an operator that can adjust an agent's opinion based on the level of observed ambiguity associated with it (see Eqs. (8)-(10)).
- We investigate the effect of agents' prior opinion (belief or disbelief) on the resulting number of agents who believe in true information. We capture the number of agents believing in true information or false information based on a variant of the Susceptible-Infected-Recovered (SIR) epidemic model [Newman 2010]. In this model, agents' decisions to believe true or false information are determined by the expected belief or disbelief with an estimated uncertainty. We develop a variant of the SIR model where an agent's status can change as long as its uncertainty level does not reach zero and the level of its belief or disbelief can still be changed with new information. Hence, unlike the original SIR model, the agent will not be stuck in the recovered state unless its uncertainty reaches zero. Note that in the traditional SIR

model following a life cycle of susceptible-infected-recovered, agents do not go back to their previous states.
- We investigate the effect of different types of agents propagating false information, including disinformers and misinformers, on mitigating or eradicating false information in a network.
- We identify the key factors that significantly increase the effect of removing or mitigating false information through extensive simulation experiments. We conduct a comprehensive sensitivity analysis to investigate the effect of varying the key design parameter values, including the frequency ratio of true or false information propagation, centrality metrics used for selecting seeding false informers (either agents propagating disinformation or misinformation) and true informers (agents propagating true information), opinion decay factor, the degree of agents' prior belief, and the percentage of true informers.
- We validated the proposed opinion models based on both synthetic and realistic network based simulations. Validation was performed using an agent-based simulation as a proxy. This provides high confidence in utilizing the proposed opinion models to predict real world opinion evolution. To create a realistic simulation of a social network for our agent-based proxy, we used a Facebook network topology, survey-based user behaviors in information processing, and an annotated corpus of fake news. To our knowledge, this is the first work that validates an abstract subjective belief model in such a realistic network scenario.

This work is significantly extended from our prior work addressing this research problem in [Cho et al. 2017] in terms of the following:

- An agent's prior belief (or disbelief) is modeled as a probability based on a Gaussian distribution in this paper, whereas its probability in [Cho et al. 2017] was given as a constant in $[0, 1]$. This allows different agents to have different levels of prior belief based on the given distribution.
- To reflect a more realistic scenario for agents' opinion exchanges, the influence probability from one agent to another agent to determine whether these two agents will interact with each other to update their opinions is formulated based on the influence levels of these two agents. This is shown in Eq. (12), and is called *influence-based activation*. Influence of an agent is simply captured based on the agent's degree in the network. In the previous work [Cho et al. 2017], we simply assume a perfect influence probability with 1, which may not be realistic. Note that Eq. (12) is used when the results are analyzed under synthetic simulation environments (Sections 5.4.1, 5.4.2, 5.5.1, and 5.5.2). But when the results are analyzed based on a realistic, proxy agent simulation model, we use a user's sharing behavior to replace the activation rate, as described in Section 5.3.
- A set of misinformers (MIs) is considered in addition to a set of disinformers (DIs) and a set of true informers (TIs) for agents propagating true information. Recall that DIs have bad intent to influence others using false information and do not change their opinions upon exchanging opinions with others. On the other hand, MIs mistakenly propagate false information and so they may change their opinions while interacting with others. In our previous work [Cho et al. 2017], we only considered a set of DIs and TIs as the seeding agents propagating false or true information. In this work, we analyze the effect of varying the values of key design parameters on performance in two network environments: a network with TIs, DIs, and doubters (Ds for agents with lack of confidence and being influenced mainly by TIs or DIs) and a network with TIs, MIs, and Ds. In this work, the initial seeding DIs or MIs form their opinions based on the SL's mapping rule with high belief and close-to-zero values of disbelief and uncertainty as explained in Section 4.2. But again we allow MIs to update their opinions

while DIs do not change their opinions in order to address their unique characteristics of propagating false information either by mistake or with bad intent, respectively.
- More comprehensive simulation experiments are conducted in order to investigate the effect of the following key design parameters: (i) the frequency of true information propagation; (ii) the opinion decay factor over time; (iii) the use of a prior belief (or disbelief) and a different degree when a prior belief is used; (iv) the percentage of TIs; and (v) the centrality metrics used for selecting seeding TIs, DI, or MIs. Our experiments are conducted based on both a synthetic simulation setting and realistic, proxy agent-based simulation setting. As mentioned in Section 1.3, we considered realistic, proxy agent-based simulation setting to model a real network, user behavior, and quality of news articles, and we validated our opinion model based on the realistic environment to provide high confidence in the validity of the proposed opinion model.

1.4 Paper Structure

The remainder of this paper is structured as follows:

- Section 2 provides an overview of the state-of-the-art research in the area of false information propagation, and discusses relevant background on Subjective Logic (SL).
- Section 3 presents and discusses the details of our SL-based opinion model as a variant of SL's consensus operator.
- Section 4 describes the agent features, types and epidemic status.
- Section 5 explains performance metrics, experimental setup, and data used for a realistic network setting. In addition, a set of simulation results based on both synthetic and realistic experimental settings are presented, along with analysis and discussion of their overall trends.
- Section 6 summarizes our key results and provides several pathways for future work.

2 RELATED WORK

In this section, we discuss various approaches to mitigating or eliminating false information in social networks and provide background on SL which was used for formulating agents' opinions in this work.

2.1 Countering False Information Propagation

We discuss the existing approaches from two perspectives: a *network analysis-based approach* (NA) and a *feature-based approach* (FA).

2.1.1 Network Analysis-based Approach (NA). Most NA aims to stop false information propagation by selecting a set of counter-misinformation nodes (i.e., nodes to disseminate true information against misinformation).

Information Maximization (IM) techniques have been used to solve misinformation problems because the selection of a set of nodes to stop/mitigate misinformation propagation is similar to the problem of selecting influential nodes to maximize performance [Campan et al. 2017]. Many IM algorithms have been developed and applied in various contexts such as the spread of technology, the adoption of products with the effect of word of mouth in markets, or game theoretic strategies [Kempe et al. 2003]. Further, the cascading effect of purchasing products by customers is also investigated in terms of maximizing influence of customers who are also likely to influence other customers' purchasing behavior [Domingos and Richardson 2001]. Kempe et al. [2003] prove that the problem of selecting the most influential nodes, called the *influence maximization problem*, is NP-hard. In addition, they show that many heuristic, greedy algorithms to solve this influence problem guarantee provably within 63% of optimal solutions, which is an *approximation algorithm*

with performance guarantee close to $1 - 1/e$. They show that node selection based on network structure (i.e., degree centrality or degree distance) outperforms existing heuristic approaches in selecting influential nodes. Leskovec et al. [2007] propose an outbreak detection algorithm by selecting nodes in a network for quickly detecting the spread of information or a virus. They present a sensor placement algorithm that produces close-to-optimal solutions as a general methodology.

Domingos and Richardson [2001] model a social network based on Markov random fields where a customer's purchasing behavior is affected by his/her desire to buy a product as well as the influence of other customers around the customer to identify ideal customers for marketing. Chen et al. [2009] propose a greedy, heuristic algorithm, the so called *degree discount heuristics*, that outperforms the existing counterpart [Kempe et al. 2003; Leskovec et al. 2007], in terms of both influence spread and running time. Recently, Liu et al. [2014] propose the time constrained IM problem by selecting a set of nodes that can maximize influence over a certain period of time. They employ the concept of influence spreading path to improve algorithmic scalability. Tong et al. [2016] propose an adaptive seeding strategy, selecting a small set of seeding nodes to propagate information, considering network dynamics, so called *uncertainness*, caused by the change of network topology or the probability of activating a node where no perfect probability for activation exists in practice.

Influence Limitation (IL) algorithms have been used to counter misinformation based on Information Maximization (IM) approaches. Budak et al. [2011] deal with information of competing campaigns in online social networks with the aim of limiting misinformation propagation where a certain set of nodes propagate 'bad' campaigns while a certain set of nodes disseminate 'good' campaigns. They solve this IL problem by proposing a way to select a set of initial seeding nodes propagating 'good' or 'bad' campaigns, which is an NP-hard problem. This work uses a perfect influence probability (i.e., if v is a direct neighbor of u, v will be always influenced by u's opinion with probability 1), called *high effectiveness property*, allowing the problem to be solved as a submodular problem. This does not reflect reality since many real world problems may not be submodular. Raj H and Narahari [2012] relax the assumption of the high effectiveness property that makes the influence limitation problem submodular in [Budak et al. 2011]. They propose a Shapley value (SV)-based IL algorithm to limit misinformation propagation.

Time constraints of misinformation propagation are also studied. Litou et al. [2016] propose an information propagation model, called the *Dynamic Linear Threshold* (DLT) model, to distinguish credible information from misinformation. This work models an individual user's different information propagation time window and susceptibility to new information on whether to propagate it over a network or not. Nguyen et al. [2012] study how to decontaminate the spread of misinformation in social networks with a small fraction of decontaminating nodes where misinformation is spread out with a set of I. This work finds the so called β_T^I, the smallest fraction of decontaminating nodes, *node protectors*, to make misinformation spread out to $1 - \beta_T^I$. This work improves [Budak et al. 2011] by not assuming the high effectiveness property. Krishnamurthy and Hamdi [2013] examine a misinformation propagation or data incest issue caused by the accidental multiple reuse of data. They propose optimal and sub-optimal algorithms to remove misinformation with a graph theoretic setting based on necessary and sufficient conditions of the network topology representing information flow.

Diffusion models are the methods by which members of a society adopt new behaviors. *Linear Threshold Model* [Granovetter 1978; Kempe et al. 2003; Macy 1991] considers the influence of collective behavior on a person's adoption behavior as follows: a node v adopts a certain behavior (e.g., opinion, product, innovation, information, service) based on a certain threshold of neighboring nodes' adoption behaviors. v's neighbors can be weighted by $b_{v,w}$ where the sum of the weights

of v's neighbors, w's, equals 1. v adopts the certain behavior (i.e., active) when $\sum_{w \in W} b_{v,w} \geq \theta_v$ where w is an active neighbor in the set W of v's neighbors, and θ_v is v's adoption threshold.

Independent Cascading Model derives from interaction processes in particle systems [Durrett 1988]. Based on this particle interaction process, Kempe et al. [2003] explain the diffusion process in which a node v is activated by its neighbor w based on a success probability, $p_{w,v}$, at a discrete time with an independent activation order of its neighbors w's. If v has multiple activated neighbors, the neighboring nodes' attempts to activate v are sequentially performed in an arbitrary order. Goldenberg et al. [2001] use this cascading model in explaining the speed of accepting a new product in the area of marketing.

Axelrod [1997] proposes an agent-based adaptive model to explain the effect of convergent social influence. Similar people tend to be aggregated by forming a community and a person is more likely to be influenced by its neighbors. Axelrod [1997] investigates culture as an individual's attribute. This work examines how local convergence based on interactions among similar agents can introduce global polarization. Hughes et al. [2015] study how an individual's different characteristics in information sharing impact decision making performance in terms of correctness. This work models information based on the ratio of noise and valuable information in order to investigate how an agent can distinguish noisy information from valuable information in order improve decision making. This information model is used in our simulation model as perceived information credibility affects an agent's decision on whether to accept or reject received information (i.e., either false or true information).

The spread of misinformation is also modeled based on epidemic models. Zhao et al. [2011] extend the SIR (Susceptible-Infected-Recovered) model by incorporating a 'forgetting' factor to investigate the rumor propagation process based on the degree, infection rate, and forgetting rate. The authors used LiveJournal, an online social blogging platform, to model the average degree of the network. Zhao et al. [2012] develop a Susceptible-Infected-Hibernator-Removed (SIHR) model that extends the SIR model by adding a direct link from ignorants (i.e., uninfected nodes by a rumor) to stiflers (i.e., stifling nodes against a rumor), which represent a new type of people, hibernators. They further extend SIHR by considering different types of networks, such as a random graph or scale-free graph, to investigate the effect of forgetting and remembering rates on the final size of a rumor spreading [Zhao et al. 2013]. The key difference between Zhao's works [Zhao et al. 2013, 2012, 2011] and our work is that our work developed a variant of the SIR model where the transition from one state to the other by each agent is determined based on the agent's level of uncertainty, not determined by given infection and forgetting rates with the degrees of nodes based on a Poisson distribution as used in [Zhao et al. 2013, 2012, 2011]. In addition, our work uses true informers propagating true information while Zhao's works [Zhao et al. 2013, 2012, 2011] model "stiflers", who are individuals refusing the spread of a rumor. This implies that the true informers are more actively countering false information (either misinformation or disinformation) while the stiflers can mitigate the influence of the rumor but cannot actively counter the spread of the rumor by influencing other agents' opinions in disbelieving the rumor (or false information).

2.1.2 Feature-based Approach (FA). FA is to identify the key features of false information or the sources (i.e., users) of false information. Rumors can be spread quickly based on a user's behavioral features or information features. Wang et al. [2015] study the impact of uncertainty on information forwarding behavior for gossip diffusion in social networks. They use a mixed logic model for a user to determine a particular strategy and an approximation method from mean field theory for the information diffusion process. However, in contrast to our approach, uncertainty is simply considered as a design parameter, rather than deriving it from network, user, or information characteristics. Liang et al. [2015] take the rumor identification approach based on machine learning

techniques in terms of feature design and selection. They identify unique behavior of the rumor publisher and rumor post, compared to normal users and posts, respectively.

Kumar et al. [2014] develop an efficient misinformation detection algorithm based on message attributes derived from cognitive psychology. They aim to identify the deception cues by considering message consistency (i.e., consistent with other messages), message coherence (i.e., coherence within the message), credibility of a source, and general acceptability of messages in online social networks. Qazvinian et al. [2011] propose rumor detection mechanisms by considering features based on content, network, and microblog platforms and validate the performance of the mechanisms using 10,000 tweets which are manually annotated.

Castillo et al. [2011] also analyze microblog postings of trending topics to identify features of rumors in Twitter by classifying credible or non-credible information. Collard et al. [2015] study the features of causing rumor propagation based on two possible psychological causes, scarcity due to a lack of information or profusion due to too much information. Bessi et al. [2015] analyze a sample of 1.2M Facebook users that show the tendency of scientific and conspiracy news consumption by Italian users. Their results show that users are more likely to be engaged when they have same consumption patterns (i.e., homophily) with their friends. In addition, there exists polarization on unverified rumors, conspiracy stories, or intentional satirical false claims. Kopp et al. [2018] investigated the effects of four information theoretic deception models including degradation, corruption, denial, and subversion, in which each form of deception aims to alter the perception of a victim [Kopp 2000]. The authors modeled an integrated deception model considering these four deception elements and validated the diffusion of agents' behaviors in believing fake news, compared to the diffusion empirically observed from the state-of-the-art work [Subramanian 2017; Wardle 2017].

Unlike the existing approach in both NA and FA, our work examines uncertainty of opinions in analyzing the propagation of false and true information considering agents' prior belief and their opinion updates upon receiving conflicting information. In addition, our work investigates what centrality metrics can impact more of information propagation in terms of the perspectives of both false and true informers.

2.2 Subjective Logic

Belief theory has been used to represent the aspect of *subjectivity* in belief. Its origin is discussed in *Dempster-Shafer Theory* (DST) [Shafer 1976]. In [Shafer 1976], instead of using the additivity principle of the probability theory (i.e., the sum of probabilities on all pairwise disjoint possibilities always equals one), it allows observers to assign belief masses to any subset of a state space itself. The well-known criticism against Dempster's rule is that it produces counter-intuitive results when the two argument beliefs have strong conflicts. To resolve this issue, many researchers have proposed solutions for conflicting evidence. Smets and Kennes [1994] propose *Transferrable Belief Model* (TBM) to elaborate DST for resolving the conflicting evidence.

In a same context, SL is proposed to represent an opinion based on three dimensions, *belief*, *disbelief*, and *uncertainty*, where SL is derived by incorporating the perspectives from both probability models and logic [Jøsang 2016]. In SL, the proposed fusion operators consider the nature of situations called *cumulative* and *averaging* fusion operators. The *cumulative rule of combination* is applied in situations where independent belief functions are combined as a function of evidence combination. The *averaging rule of combination* is for situations that require combining dependent belief functions into a function of the average of the evidence. SL has also been utilized in modeling uncertain belief under incomplete, unavailable information [Cho and Swami 2014].

In this work, we adopt SL to explicitly deal with the uncertainty an agent perceives in updating its opinion defined with belief, disbelief, and uncertainty. However, in SL, the degree of uncertainty

is mainly considered in terms of the amount of information available to the agent. As more information becomes available, regardless of whether it is conflicting or not, uncertainty will decrease in SL [Jøsang 2001]. Recently, to deal with ambiguity derived from conflicting evidence, the trust revision algorithm was developed by discounting information based on the estimates of the information sources [Jøsang 2016]. That is, the trust revision algorithm [Jøsang 2016] can slow down the reduction of uncertainty, but uncertainty continues to decrease and reaches zero even with the same amount of belief and disbelief. In such a situation, the model would not provide useful support for the agent to make a decision under the same amount of information supporting the two extremes. To refine the limitation in SL, Wang and Singh [2010] also consider conflicting evidence in mapping the amount of evidence with the three dimensions of an opinion in SL. This method is to formulate an opinion in (b, d, u) format when given a certain amount of evidence. However, our work is different from [Wang and Singh 2010] because our main concern is how to update an agent's opinion after two agents interact with where their opinions may be conflicting to each other. That is, Wang and Singh [2010] give an approach in incorporating evidence (i.e., positive or negative evidence) itself while our work solves how to incorporate multiple agents' opinions which may be conflicting. Further, unlike both Wang and Singh [2010] and Jøsang [2016] which use SL's uncertainty that keeps continuously decreasing upon receiving any evidence even if the decreasing speed is slowed down, uncertainty in our opinion model fluctuates after an agent's opinion is updated upon interactions with other agents. Therefore, our proposed new operator reflects the increase of uncertainty after receiving conflicting opinions in addition to slowing down the reduction of uncertainty based on similarity-based opinion updates. Besides, to the best of our knowledge, this work is the first that uses SL to solve a false information propagation problem where the opinion model deals with uncertainty derived from both a lack of evidence (i.e., ignorance) and conflicting opinions (i.e., ambiguity).

3 OPINION MODEL

In SL, a binomial *opinion* is represented by three dimensions: *belief* (b), *disbelief* (d), and *uncertainty* (u) [Jøsang 2016]. A single opinion on a given proposition is represented by:

$$b, d, u \in [0, 1]^3, \quad b + d + u = 1 \tag{1}$$

3.1 Opinion Formation

We consider a binomial opinion representing *pro* (i.e., belief) or *con* (i.e., disbelief) for a given proposition. For example, an agent believes given information as truth regardless of whether the information is actually true or not. The agent may have degrees of belief (i.e., agree) and/or disbelief (i.e., disagree) towards a given proposition with some degree of uncertainty. Agent i's opinion on proposition A is denoted as w_i^A. For simplicity, we omit A and use w_i to represent an agent i's opinion as:

$$w_i = \{b_i, d_i, u_i, a_i\} \tag{2}$$

where a_i is the *base rate* which normally represents i's general background knowledge towards a given proposition (or topic) [Jøsang 2016]. The base rate can be partially objective or partially subjective because individuals' observations are not the same. The base rate can be interpreted as the prior belief agent i has on a given proposition derived from past experiences (e.g., frequency-based probability) or a cognition related to a bias, personality traits, and/or information processing types in the decision making process. In this work, either interpretation may be applicable. We examine its impact on decision making under uncertainty.

The base rate, a_i, affects expectation probability (i.e., a probability that an agent is expected to make a decision) in either belief or disbelief [Jøsang 2016], denoted by E_{b_i} or E_{d_i}, respectively, where they are given by:

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

Note that $E_{b_i} + E_{d_i} = 1$ as $b_i + d_i + u_i = 1$. Since decisions are often made under uncertainty based on prior beliefs in many real life scenarios, we investigate the effect of a_i on the spread of information diffusion (either true or false) in our agent model.

A person's acceptance towards given information is affected by various factors, including personality (e.g., agreeableness, open-mindedness, stubbornness), impact of neighbors (e.g., herding), homophily (e.g., like-mindedness), competence (e.g., domain knowledge), or confidence (e.g., certainty about its own opinion) [Cho and Swami 2014; Newman 2010].

In SL, an agent forms its opinion based on the amount of directly observed evidence. The following mapping rule is used to initialize agents' opinions:

$$b = \frac{r}{r + s + W}, d = \frac{s}{r + s + W}, u = \frac{W}{r + s + W}. \tag{4}$$

where r is positive evidence (i.e., agree) and s is negative evidence (i.e., disagree) for a particular proposition. For simplicity, we dropped the subscript i denoting the agent. W indicates the amount of uncertainty which can be affected by the inherent errors that can be introduced by the environment itself (e.g., imperfect observability). When $W = 0$, b is a natural estimate of the fractional evidence in favor of the proposition.

3.2 Opinion Update

We model the spread of information in a society (or a network) based on pairwise interactions between individuals. An interaction between two agents, i and j, at time t, represents the spread of evidence from j to i and results in agent i adopting a new opinion for time $t + 1$. There are two steps to agents' opinion updates based on their interactions. First, an opinion update is primarily based on the similarity of the two opinions, with opposite opinions having less influence on one another. This can sometimes lead to ambiguous opinions with high amounts of belief and disbelief. We account for this effect by adding a *conflict measure* term that translates ambiguity to an increase in uncertainty. In addition to the interaction update, we also model forgetfulness via an opinion decay function. It increases opinion uncertainty in the absence of interaction. We discuss these three aspects of opinion update below.

3.2.1 Homophily-based Opinion Consensus. Homophily, or like-mindedness, significantly affects the way people update opinions [Li et al. 2011]. In this work, the similarity of two agents' opinions, denoted by s_i^j, is computed based on *cosine similarity* [Tan et al. 2005] of the two opinion vectors i and j in terms of their belief and disbelief, (b_i, d_i) and (b_j, d_j), respectively. The calculation of s_i^j based on the cosine similarity is detailed as:

$$s_i^j = \frac{b_i b_j + d_i d_j}{\sqrt{b_i^2 + d_i^2}\sqrt{b_j^2 + d_j^2}} \tag{5}$$

Note that agents' uncertainty levels are not considered to capture the similarity because having similar levels of uncertainty is not relevant to opinion updates. The cosine similarity of the opinions of i and j is symmetric with $s_i^j = s_j^i$.

We use s_i^j, as a discounting operator [Jøsang 2016], to determine the degree to which agent i accepts agent j's opinion. Given two vectors of opinions, $w_i = \{b_i, d_i, u_i\}$ and $w_j = \{b_j, d_j, u_j\}$,

agent i's trust opinion in j's opinion, s_i^j, is given by $w_{i\otimes j} = \{b_{i\otimes j}, d_{i\otimes j}, u_{i\otimes j}\}$ where each element is estimated by:

$$b_{i\otimes j} = s_i^j b_j,\ d_{i\otimes j} = s_i^j d_j,\ u_{i\otimes j} = 1 - s_i^j(1-u_j)\ . \tag{6}$$

Here, $u_{i\otimes j}$ is simply derived by $u_{i\otimes j} = 1 - b_{i\otimes j} - d_{i\otimes j}$ where $b_j + d_j + u_j = 1$. For simplicity, we omit the time step notation, but both sides of the equation refer to time step t.

We use SL's *consensus* operator [Jøsang 2016] for an agent's opinion update upon receiving new information. The updated opinion of agent i after interaction with agent j is denoted as $w_i \oplus b_{i\otimes j} = \{b_i \oplus b_{i\otimes j}, d_i \oplus b_{i\otimes j}, u_i \oplus b_{i\otimes j}\}$ and each element is given by:

$$\begin{aligned} b_i \oplus b_{i\otimes j} &= \frac{b_i(1 - s_i^j(1-u_j)) + s_i^j b_j u_i}{\beta}\ , \\ d_i \oplus d_{i\otimes j} &= \frac{d_i(1 - s_i^j(1-u_j)) + s_i^j d_j u_i}{\beta}\ , \\ u_i \oplus u_{i\otimes j} &= 1 - \frac{(1 - u_i - s_i^j u_j)}{\beta}\ . \end{aligned} \tag{7}$$

where $\beta = u_i + 1 - s_i^j(1-u_j) - u_i(1 - s_i^j(1-u_j)) = 1 - s_i^j(1-u_i)(1-u_j)$ and $\beta \neq 0$ is assumed. $u_i \oplus u_{i\otimes j}$ is the same as $1 - (b_i \oplus b_{i\otimes j} + d_i \oplus d_{i\otimes j})$ and $u_i \oplus u_{i\otimes j}$ is simplified based on Eq. (6). Again, we omit the time step notation; here, the left side represents $w_i(t+1)$ while the right side uses the opinions at t such as $w_i(t)$ and $w_j(t) = w_{i\otimes j}(t)$.

3.2.2 Opinion Adjustment to Deal with Conflicting Evidence. In SL's original consensus operator [Jøsang 2016], regardless of the differences between two opinions, agents' opinions are updated by increasing either belief or disbelief while continuously decreasing uncertainty. Even if Eq. (7) uses a discounter s_i^j to weight an agent's opinion more heavily from neighbors with similar opinions than those with dissimilar opinions, it still strictly decreases uncertainty, albeit slower than it would have without discounting. The uncertainty in SL mainly reduces when more evidence is received regardless of whether or not evidence is conflicting to previously received information. This effect is counter-intuitive to real life phenomena because receiving more information can create confusion when that information is supporting views of opposite extremes.

To capture uncertainty introduced by conflicting evidence, we propose an agent's opinion update that penalizes belief or disbelief values when conflicting information is received. In particular, we use the term *ambiguity* to refer to uncertainty derived from conflicting evidence. After agent i updates its opinion based on an interaction with agent j, as specified in Eq. (7), it then adjusts its opinion based on the distance between its own belief and disbelief. That is, if an agent has an opinion based on receiving the same amount of evidence supporting belief and disbelief ($b \approx d$), uncertainty should increase because conflicting evidence increases ambiguity. To this end, agent i with conflicting evidence adjusts its opinion on top of the updated opinion, $w_i = \{b_i, d_i, u_i\}$, based on Eq. (7). i's adjusted opinion, denoted by $w_i' = \{b_i', d_i', u_i'\}$, is formulated by weighting the distance between i's belief and i's disbelief with c_i, a conflict measure we compute, as follows:

$$b_i' = c_i b_i,\ d_i' = c_i d_i,\ u_i' = 1 - c_i(1-u_i)\ . \tag{8}$$

u_i' can be derived based on $u_i' = 1 - (b_i' + d_i')$ where $b_i' + d_i' + u_i' = 1$. c_i is defined by:

$$c_i = \frac{|E_{b_i} - E_{d_i}|}{E_{b_i} + E_{d_i}} = |2b_i + 2a_i u_i - 1|\ . \tag{9}$$

where c_i is derived based on $b_i + d_i + u_i = 1$ and Eq. (3). Note that we omitted the time step for simplicity. Both the left and right opinions are from the same time step (i.e., $t+1$ in Eq. (7)). In Eq.

(8), c_i implies that when i has a similar level of expected belief (E_{b_i}) and expected disbelief (E_{d_i}), its belief (b_i') and disbelief (d_i'), as shown in Eq. (8), are significantly decreased. On the other hand, if E_{b_i} and E_{d_i} are significantly different, the original opinion based on the consensus operator, Eq. (7), is mostly kept intact. c_i can be equivalently represented by $|2b_i + 2a_iu_i - 1|$, as shown in Eq. (9). This form reflects that when b_i or d_i is relatively high, uncertainty, u_i, is low. On the other hand, when b_i and d_i are low, u_i must be high. Therefore, b_i and d_i are not heavily penalized simply when uncertainty is high; rather the penalty is higher when b_i and d_i are close to each other and u_i is relatively lower. This approach matches our desired real world scenario, because when u_i is high due to a lack of sufficient amount of information, new information can impact b_i and d_i. When uncertainty is low, however, and conflicting information is received, both b_i and d_i are reduced and replaced with rising u_i. Note that agent i's opinion can be updated by Eqs. (7) and (8) only when $u_i > 0$ (i.e., $\beta \neq 0$).

If agents do not have any prior belief towards belief (a_i) or disbelief ($1 - a_i$), an agent simply uses b_i and d_i to obtain c_i, instead of E_{b_i} and E_{d_i}, respectively. Thus, c_i is given by:

$$c_i = \frac{|b_i - d_i|}{b_i + d_i} = \frac{|2b_i + u_i - 1|}{1 - u_i} . \tag{10}$$

Eq. (10) clearly shows that when b_i and d_i are almost the same with small u_i, a higher penalty is applied than the case of b_i and d_i with high u_i. For simplicity, in Eqs. (8), (9), and (10), we omit notation for the time step. Here, both sides of the equations are at time step t.

An agent's opinion update can be summarized by two steps: (i) an agent updates its opinion based on Eq. (7) using a similarity-based discounting operator in Eq. (6); and (ii) the agent adjusts its opinion based on the distance between its belief and disbelief based on Eq. (9) (with prior belief/disbelief) or Eq. (10) (without prior belief/disbelief).

3.2.3 Opinion Decaying. Unless an agent receives new information by interacting with other agents, its opinion decays over time based on a decay factor, γ, over belief and disbelief while uncertainty increases in proportion to γ. For example, many bounded cognition models of human cognition incorporate functions for forgetting information over time. We model the decayed opinion by:

$$\begin{aligned} b_i &= (1-\gamma)b_i, \\ d_i &= (1-\gamma)d_i , \\ u_i &= 1-(1-\gamma)(b_i+d_i) \\ &= u_i + \gamma(1-u_i) . \end{aligned} \tag{11}$$

Note that u_i is simply derived based on $1 - b_i - d_i$ where $b_i + d_i + u_i = 1$. Different from the opinion update by Eqs. (7) and (8) which allow the opinion update only for $\beta > 0$ and $u_i > 0$, respectively, the opinion decay based on Eq. (11) occurs at every time step. Therefore, even if u_i reaches 0, over time it can increase (i.e., $u_i > 0$) and accordingly the agent i can update its opinion upon receiving new information from its neighbors. For simplicity, we omitted the time step notation, but the left side is at time $t + 1$ while the right side is at t.

4 AGENT MODEL

This work considers an online social network as either an undirected graph or a directed graph, $\mathcal{G}$, where vertices, v_i's, are agents i's (e.g., users) in the set of $\mathcal{V}$ and the edges, e_{ij}'s (i.e., 1 for an edge and 0 for no edge), represent the relationships in the set $\mathcal{E}$. Agent i's neighbors refer to other agents directly connected to i. For information propagation, we set an initial number of seeding agents propagating false information, called *false informers* while setting an initial number

of seeding agents disseminating true information, *true informers*, to counter the propagation of the false information. Note that for the false informers, we consider two types: (1) *disinformers* propagating false information with a bad intent and not being influenced by others' opinions; and (2) *misinformers* propagating false information mistakenly and thus being influenced by others' opinions over time. We discuss more details of an agent's characteristics, types and epidemic process of information diffusion as below.

4.1 Agent Features

An agent learns information through interacting with other agents or from a source exogenous to the network like absorbing information broadcast by public media, for example. Upon receiving any information, the agent processes the given information based on its own ability to judge credibility of information (i.e., domain expertise), cognitive capability of information processing (i.e., learning or analytical capability), and/or prior belief (e.g., bias). We consider an agent's individual characteristics as follows:

- *Prior belief* (a_i): An agent i can have a prior belief on a given topic, impacting a decision to believe a given piece of information. The agent's prior belief is modeled based on its base rate, a_i, in Eq. (2). As shown in Eq. (3), an agent's expectation probability uses the base rate, a_i, to consider uncertainty in belief or disbelief and determines the epidemic status in the diffusion of false information as shown in Section 4.3. We treat this prior belief as a bias favoring false information where the prior disbelief is for disbelieving in false information (i.e., believing in true information).
- *Centrality degree* (d_i): An agent is characterized by its influence based on its location in a network, representing its importance or influence in the network. We use agent i's degree as a measure of its influence over other agents. d_i is applied in the rate agent i propagates its opinion over a network, as detailed in Eq. (12).

4.2 Agent Types

In an agent's opinion towards a given proposition, $w = \{b, d, u\}$, b is an agent's belief, agreeing with false information, d is the agent's disbelief, disagreeing in false information, and u is the degree of uncertainty an agent perceives by neither believing nor disbelieving in false information (i.e., I don't know). An agent's initial opinion is set based on the mapping rule in Eq. (4) with different values of r, s, and W, depending on the type of agent. We consider the following four types of agents:

- *Disinformers* (DIs) disseminate false information with bad intent or to achieve a selfish purpose. The opinion of this agent type is initialized with $(r, s, W) = (n, 1, 1)$ where $n >> 1$ (e.g., 1000), leading to $\{b, d, u\} = \{\frac{n}{n+2}, \frac{1}{n+2}, \frac{1}{n+2}\}$ implying that the agent highly agrees with false information (i.e., $b \rightarrow 1$) while it has low disbelief (i.e., $d \rightarrow 0$) and low uncertainty (i.e., $u \rightarrow 0$). DIs do not change their opinion while influencing others' opinions.
- *Misinformers* (MIs) mistakenly propagate false information without any purpose or bad intent. Hence, unlike DIs, their opinions can be changed by being influenced by opinions of other agents they interact with. The initial opinion of MIs is set as same as DIs'.
- *True Informers* (TIs) disseminate true information to counter false information. This agent starts its opinion with $(r, s, W) = (1, n, 1)$ where $n >> 1$, leading to $\{b, d, u\} = \{\frac{1}{n+2}, \frac{n}{n+2}, \frac{1}{n+2}\}$. This means that the agent disagrees with the false information while agreeing with true information. This is represented by low belief (i.e., $b \rightarrow 0$), high disbelief (i.e., $d \rightarrow 1$), and low uncertainty (i.e., $u \rightarrow 0$). Similar to FI, a TI also does not change its opinion while influencing other agents' opinion.

- *Doubters* (Ds) have low confidence (i.e., $u \rightarrow 1$) in their own opinion by not initially agreeing or disagreeing with given false information (i.e., $b \rightarrow 0$ and $d \rightarrow 0$). This agent type is initialized with opinion $(r, s, W) = (1, 1, n)$, leading to $\{b, d, u\} = \{\frac{1}{n+2}, \frac{1}{n+2}, \frac{n}{n+2}\}$, implying low confidence in a given proposition due to lack of information (i.e., ignorance).

If an opinion's uncertainty is close to 0, it implies that an agent has full of confidence in its opinion. In contrast, if the uncertainty is close to 1, the agent is very unsure of its own opinion but is open-minded to change its opinion based on received information. Note that agents can update their opinions upon interacting with other agents as long as their opinions' uncertainty does not reach 0. This means that the agents whose uncertainty levels reach 0 stop updating their opinions.

For our experimental analysis in Section 5, we use an initial number of seeding nodes representing DIs, MIs, TIs, and Ds, denoted by s_{f_d}, s_{f_m}, s_t, and s_d, respectively.

4.3 Epidemic Status of Agents

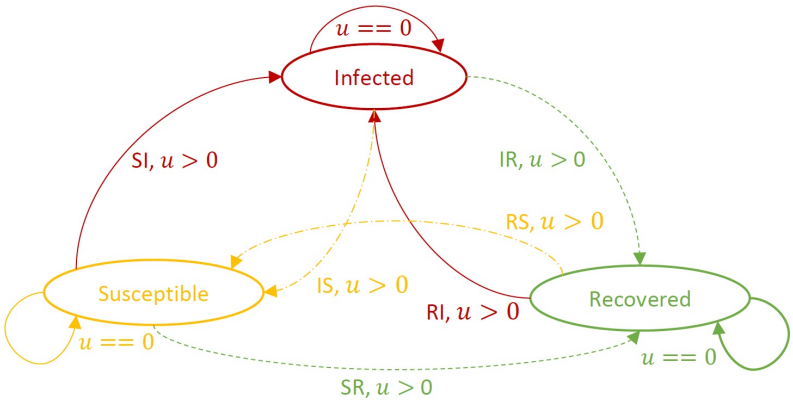


Fig. 1. Epidemic process of information diffusion where u refers to the dimension of uncertainty in an opinion based on Subjective Logic.

We model the evolution of false information propagation by developing a variant of the SIR (Susceptible-Infected-Recovered) model [Newman 2010]. The three states in the SIR model are defined based on the conditions associated with agents' expected belief or disbelief probabilities, E_{b_i} and E_{d_i}, as follows:

- *Susceptible* (S): An agent is not sure of its status in terms of whether it believes in false information or not. Agents in S have opinions with $E_b \leq 0.5$ and $E_d \leq 0.5$;
- *Infected* (I): An agent believes false information to be true with $E_b > 0.5$; and
- *Recovered* (R): An agent believes false information to be false (or believes in true information) with $E_d > 0.5$.

Among the four types of agents, DIs and TIs will not change their status and be fixed with I and R, respectively. On the other hand, MIs and Ds will change their status based on their updated opinions where all Ds start from the state S and all MIs start from the state I. The epidemic state transitions possible for Ds and MIs are summarized in Table 1. Note that when $u > 0$, an agent can move from its current status to any other status based on each transition condition.

Agent i will adopt an opinion propagated by agent j which is one of i's neighboring agents. We consider j's influence among i's friends by capturing the relative importance of j over i's friends. This rate is called an *activation rate*, act_{ij}, that refers to the rate agent j adopts agent i's opinion to

update its own opinion and is formulated by:

$$act_{ij} = \frac{d_i}{\max\limits_{k \in F_j} d_k} \tag{12}$$

where d_i is i's degree and k is an agent ID directly connected to j (i.e., j's friend) in F_j, a set of j's friends where the denominator in the above equation represents the maximum influence among j's friends.

Note that when DIs (or MIs) and TIs propagate their respective false or true information to their neighbors, the neighbors will update their opinions based on Eqs. (7) and (8) and propagate *their own updated opinions* (not the original information) to their neighbors as well. The opinion propagation stops after each doubter propagates its opinions to its neighbors in the network. Since each agent will propagate its own opinion which is more likely to be different from other agents', some agents will receive multiple opinions related to the same information (i.e., original information propagated by FIs or TIs) but each opinion itself is not necessarily same.

Table 1. Transitions of SIR Epidemic Process by Doubters (Ds).

Status Change	Susceptible (S)	Infected (I)	Recovered (R)
SI	Current: $E_b \leq 0.5, E_d \leq 0.5, u > 0$	Next: $E_b > 0.5$	
IS	Next: $E_b \leq 0.5, E_d \leq 0.5$	Current: $E_b > 0.5, u > 0$	
SS	Current & Next: $E_b \leq 0.5, E_d \leq 0.5, u == 0$		
IR		Current: $E_b > 0.5, u > 0$	Next: $E_d > 0.5$
RI		Next: $E_b > 0.5$	Current: $E_d > 0.5, u > 0$
II		Current & Next: $E_b > 0.5, u == 0$	
RS	Next: $E_b \leq 0.5, E_d \leq 0.5$		Current: $E_d > 0.5, u > 0$
SR	Current: $E_b \leq 0.5, E_d \leq 0.5, u > 0$		Next: $E_d > 0.5$
RR			Current & Next: $E_d > 0.5, u == 0$

All possible epidemic transitions from one state to another are summarized in Table 1. Note that when u reaches at zero, the agent will stay in the current status without further changes (e.g., no update at $u = 0$ or $\beta = 0$) in Eq. (7), implying that its opinion is fixed after $u = 0$. However, again since an agent's opinion decays over time based on Eq. (11) and its opinion is adjusted based on observed ambiguity based on Eq. (8) which allows $u > 0$, the agent's opinion can be updated again based on Eq. (7).

5 NUMERICAL RESULTS AND ANALYSIS

In this section, we describe metrics used for our experiments and the details of our experimental setup. Further, we analyze the experimental results and discuss their overall trends.

5.1 Metrics

The following metrics are used for our experiments:

- *Agent's opinion*: An agent's opinion vector (b, d, u) is represented in the scale of $[0, 1]$ where $b + d + u = 1$.

- *Histogram of Agents' opinions*: The frequency of Ds' opinions is captured in a histogram showing agents' belief, disbelief, and uncertainty.
- *Instantaneous status in the SIR model* ($\mathcal{S}(t), \mathcal{I}(t), \mathcal{R}(t)$): These measure the fraction of doubters, Ds, in the state of susceptible ($\mathcal{S}(t)$), infected ($\mathcal{I}(t)$), or recovered ($\mathcal{R}(t)$) over the total number of Ds (s_d). They are given by:

$$\begin{aligned} \mathcal{S}(t) &= \frac{\sum_{i,E_{b_i}(t)\leq 0.5, E_{d_i}(t)\leq 0.5}^{s_d} D_i}{s_d}, \\ \mathcal{I}(t) &= \frac{\sum_{i,E_{b_i}(t)\geq 0.5}^{s_d} D_i}{s_d}, \\ \mathcal{R}(t) &= \frac{\sum_{i,E_{d_i}(t)\geq 0.5}^{s_d} D_i}{s_d}. \end{aligned} \tag{13}$$

  where $E_{b_i}(t)$ and $E_{d_i}(t)$ are expected belief and disbelief of doubter i at time t and D_i returns 1 when doubter i meets the conditions for the expected belief and disbelief; otherwise it returns 0. Lower $\mathcal{S}(t), \mathcal{I}(t)$ and higher $\mathcal{R}(t)$ are more desirable. We mainly use $\mathcal{R}(t)$ because it is a critical metric measuring the fraction of agents believing in true information over false information.
- *Average fraction of recovered agents* ($\mathcal{R}$): This estimates the time-averaged fraction of the recovered agents across the T times of information propagation where all agents disseminate their opinions to their neighbors per time step t for $t = 1 - T$, $\mathcal{R}$ is computed by:

$$\mathcal{R} = \frac{\sum_{t=1}^{T} \mathcal{R}(t)}{T} \tag{14}$$

  Higher $\mathcal{R}$ is more desirable.
- *Time for Recovered Agents to Govern over Infected Agents* ($\mathcal{T}_R$): This captures the time $\mathcal{T}_R$ that the number of recovered agents exceeds that of infected agents where $I(t) < R(t)$ for $t' < t \leq T$ and $I(t') \geq R(t')$, leading to $T_R = t$. Lower $\mathcal{T}_R$ is more desirable.

5.2 Experimental Setup

For this study, we use an ego-Facebook network [Leskovec 2012] that provides an undirected network with one connected component as described in Table 3. We have two scenarios: (1) a network with TIs, DIs, and Ds; and (2) a network with TIs, MIs, and Ds. In order to clearly investigate the effect of either DIs or MIs, we don't consider a network with all four types of agents.

To initiate information propagation by false informers or FIs (either DIs or MIs) and TIs, respectively, we seed 1 % of the total nodes (i.e., P_{TI} or $P_{MI} = P_{DI}$), which is 11 nodes for each (i.e., s_{f_m} or $s_{f_d} = s_t = 11$) and model the rest of the nodes (i.e., $s_d = 1011$) as Ds. We consider opinion decay factor γ set to 0.05, meaning that an opinion is decayed with the weight, $(1 - \gamma)$. An agent's prior belief is randomly generated with mean $a_\mu = 0.5$ and standard deviation $a_{sd} = 0.1$ implying that the agent's prior disbelief is $1 - r(a_\mu, a_{sd})$ where $r(a_\mu, a_{sd})$, which returns a random prior belief with the given mean and standard deviation. The key design parameters, their meanings and default values used for our experiments are summarized in Table 2.

For our experiments, in addition to examining the effect of opinion evolution with or without prior belief, we conduct sensitivity analysis by varying the values of the following design parameters: (1) frequency of false information propagation per false information propagation, denoted by m; (2) fraction of true informers (TIs) over the total number of agents, N, denoted by P_{TI}; (3) opinion decay factor, denoted by γ; and (4) mean of prior belief, denoted by a_μ.

Table 2. Key parameters, their meanings, and default values

Param.	Meaning	Default Value
N	Total number of agents	based in a real Facebook network
n	A large number of evidence in applying a mapping rule of Eq. (4) to assign agents' initial opinions in Section 4.2	1000
T	Total session time for time step $t = 1 - T$	100
γ	Opinion decay factor to decay opinion weighted by $(1 - \gamma)$ in Eq. (11)	0.05 or vary
P_{TI}	Fraction of true informers (TIs) among the total number of agents, N	0.01 or vary
P_{DI}	Fraction of disinformers (DIs) among the total number of agents, N	0.01
P_{MI}	Fraction of misinformers (MIs) among the total number of agents, N	0.01
s_{f_m}	Number of misinformers (MIs) assigned in a network	$P_{MI} \times N$ when $s_{f_d} = 0$; 0 otherwise
s_{f_d}	Number of misinformers (DIs) assigned in a network	$P_{DI} \times N$ when $s_{f_m} = 0$; 0 otherwise
s_t	Number of true informers (TIs) assigned in a network	$P_{TI} \times N$
s_d	Number of doubters (Ds) assigned in a network	$N - (s_{f_m} + s_{f_d} + s_t)$
a_μ	Mean given to generate a random number in $[0, 1]$ with a standard deviation, a_{sd}	0.5 or vary
a_{sd}	Standard deviation given to generate a random number in $[0, 1]$ with a mean prior belief, a_μ	0.1

Table 3. Network characteristics

N	1033	Avg. degree	51.785	Avg. path length	2.949
$\|\mathcal{E}\|$	26747	Modularity	0.54	Avg. clustering coeff.	0.534

We also investigate the effect of different centrality types of seeding agents, TIs or DIs (or MIs), when DIs (or MIs) or TIs are fixed at random selection, respectively. For the centrality metrics used to select those seeding agents, TIs or FIs (either DIs or MIs), we used the following metrics: (1) high-degree; (2) closeness; (3) betweenness; and (4) pagerank. As these centrality metrics are well-known, we omit the description of each metric. Interested readers can refer to [Newman 2010] for the details.

5.3 Network Settings

In Section 5.2, we described the conditions for our synthetic network setting, where users' behaviors and the initial false or true information are modeled based on synthetic data generated. To be specific, when no real settings are used, in order to demonstrate results under more flexible and diverse ranges of key design parameters, we strictly follow the experimental setup described in Section 5.2. Note that a user shares its opinion based on the activation rate in Eq. (12) and all its neighbors receiving its opinion will update their opinions based on the received opinion. In addition, we strictly follow the initial opinion setup assumed in Section 4.2 in order to initialize their opinions based on SL.

In addition to our synthetic network settings, we run experiments using real data. In general, it is not feasible to collect data on network-specific and information-specific (in this case false news)

Algorithm 1 Opinion formation based on the evidence set

procedure $[b, d, u]$ = FORMULATEOPINION(*node_type*, **evidence**)
  evidence $\leftarrow$ a $3 \times N_m$ matrix where the numbers of positive, negative, and uncertain evidence is in row while the number of article statements is N_m
  $b \leftarrow$ belief in false information
  $d \leftarrow$ disbelief in false information (or belief in true information)
  $u \leftarrow$ uncertainty by disbelieving either true or false information (i.e., don't know or confused)
  $r \leftarrow$ # of evidence supporting false information
  $s \leftarrow$ # of evidence supporting true information
  $W \leftarrow$ # of evidence supporting neither true or false information
  if *node_type* is a false informer **then**
    for $i = 1$ to N_m **do**
    end for
    if **evidence**$(i).r > 0$ **then** ▷ believing false information
      $r = r +$ **evidence**$(i).r$ ▷ **evidence**$(i).r$ is evidence supporting false information in i-th evidence
      $s = s +$ **evidence**$(i).s$ ▷ **evidence**$(i).s$ is evidence supporting true information in i-th evidence
      $W = W +$ **evidence**$(i).w$ ▷ **evidence**$(i).w$ is evidence supporting neither true nor false information in i-th evidence
    end if
  else ▷ when a node type is a true informer
    for $i = 1$ to N_m **do**
      if **evidence**$(i).s > 0$ **then** ▷ believing true information
        $r = r +$ **evidence**$(i).r$ ▷ **evidence**$(i).r$ is evidence supporting false information in i-th evidence
        $s = s +$ **evidence**$(i).s$ ▷ **evidence**$(i).s$ is evidence supporting true information in i-th evidence
        $W = W +$ **evidence**$(i).w$ ▷ **evidence**$(i).w$ is evidence supporting neither true nor false information in i-th evidence
      end if
    end for
  end if
  $b \leftarrow \frac{r}{r+s+W}$
  $d \leftarrow \frac{s}{r+s+W}$
  $u \leftarrow \frac{W}{r+s+W}$
end procedure

opinion/behavior dynamics. Therefore, as a proxy, we use multiple independent data to simulate a social network consisting of agents characterized by certain behavioral propensities. The three data types used to implement a realistic network setting in this work are as follows:

- *Network topology*: We use the same network topology as before, which is a Facebook ego-network [Leskovec 2012]. Again the network characteristics are described in Table 3 and its topology and corresponding degrees distribution are shown in Fig. 2 (a) and (b).

Table 4. Mapping annotated grades to evidence for SL-based opinion formation.

Annotated grade	Evidence supporting true news (s)	Negative supporting false news (r)	Uncertain (w)
true	4	0	0
mostly true	3	0	1
half true	2	0	2
barely true	0	2	2
false	0	3	1
pants on fire	0	4	0

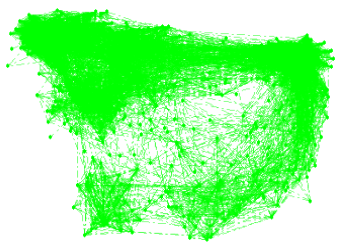
- *User behaviors*: In order to better simulate user behaviors, we collected social media sharing and reading behaviors using an Amazon Mechanical Turk survey. This survey asks each human subject for news-specific sharing and reading behavior as follows:
  - For the *sharing* behavior, we asked 'When you use social media, how often do you share news?'. Each user could select one of the following answers: always/most of the time, half of the time, sometimes, and never. Survey participants are filtered using several validity checks based on the literature to ensure answer quality. For simulation purposes, these are translated into a scale of [0, 1], corresponding to 1, 0.5, 0.25, and 0.1, respectively. This sharing behavior is used as a probability for an agent to share its opinion with its direct neighbors. However, a neighboring agent will update its opinion based on the agent's reading behavior, as described below.
  - For the *reading* behavior, we asked 'How often do you read news?'. Each user could select one of the following answers: multiple times per day, daily, weekly/monthly, and never. Similar to the sharing behavior, these are translated into a scale of [0, 1], corresponding to 1, 0.5, 0.25, and 0.1, respectively. This probability is used to determine whether an agent will accept its neighbor's sent opinion and update its opinion accordingly.

  This survey provided us with 1,406 participants after validity filtering and included a wide range of demographics (including age and education). Hence, the probability distributions derived from this newly collected behavior data should provide a more realistic setting for our simulation. The use of an agent's real behavioral tendency is expected to impact on the performance because the results based on synthetic network environments assume that an agent will share its opinion based on the activation probability in Eq. (12) while its neighbors opinion update can always happen with the perfect probability (i.e., 1).
- *Annotated fake/real news*: We used the annotated fake/real news based on [Wang 2017]. This information provides scores for a given statement based on the following scale from the most real information to the most fake information: true, mostly true, half true, barely true, false, and pants on fire. This follows the labeling convention set by the online fact-checker politifact.com. We use this information to model the opinions of false informers (i.e., disinformers and misinformers) and true informers. News may not perfectly support one extreme over the other extreme. This is different from how we modeled false informers and true informers in Section 4.2. To model false information (i.e., fake news) and true information (i.e., true news) based on a real network scenario, we use the annotation with a grade, representing the highest level of truthfulness to the highest level of untruthfulness (i.e., *true* to *pants on fire*). We assume that a false informer will only consider false information while a true informer will only consider true information. Agents will form their opinions based on the grades given by applying SL's mapping rule. How each piece of evidence in SL (i.e., positive, negative, uncertain) is obtained based on the real fake/real news is explained

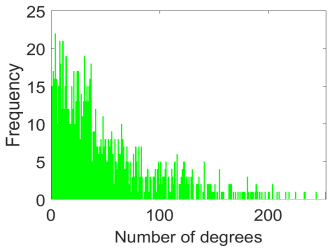
in Table 4. In addition, Algorithm 1 details how each true or false informer's initial opinion is set. Based on the fake/real articles used [Wang 2017], a false informer's initial opinion is set to $(b, d, u) = (0.700893, 0, 0.299107)$ while a true informer's initial opinion is set to $(b, d, u) = (0, 0.720434, 0.279566)$.

Fig. 2 shows the information used for our experiments in Sections 5.4.3 and 5.5.3, including Facebook network topology, annotated news quality, and user behavioral data based on surveying human subjects. That is, we show the network topology, the degree distribution of the network, the histogram of annotated news quality, and the distribution of 16 user types with respect to % of population.

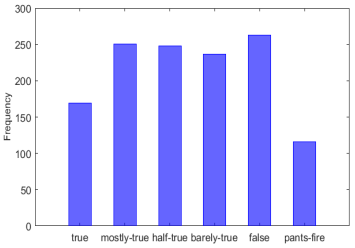
In the next sections, we discuss the simulation results with synthetic network settings (Sections 5.4.1, 5.4.2, 5.5.1, 5.5.2) and realistic network settings (Sections 5.4.3 and 5.5.3) under two scenarios: (1) a network consisting of agents with TIs, DIs, and Ds; and (2) a network consisting of agents with TIs, MIs, and Ds.



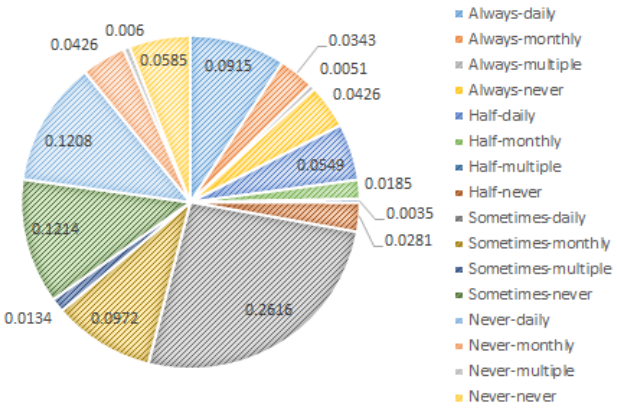
(a) Facebook network topology



(b) Degree distribution of (a)



(c) News quality histogram

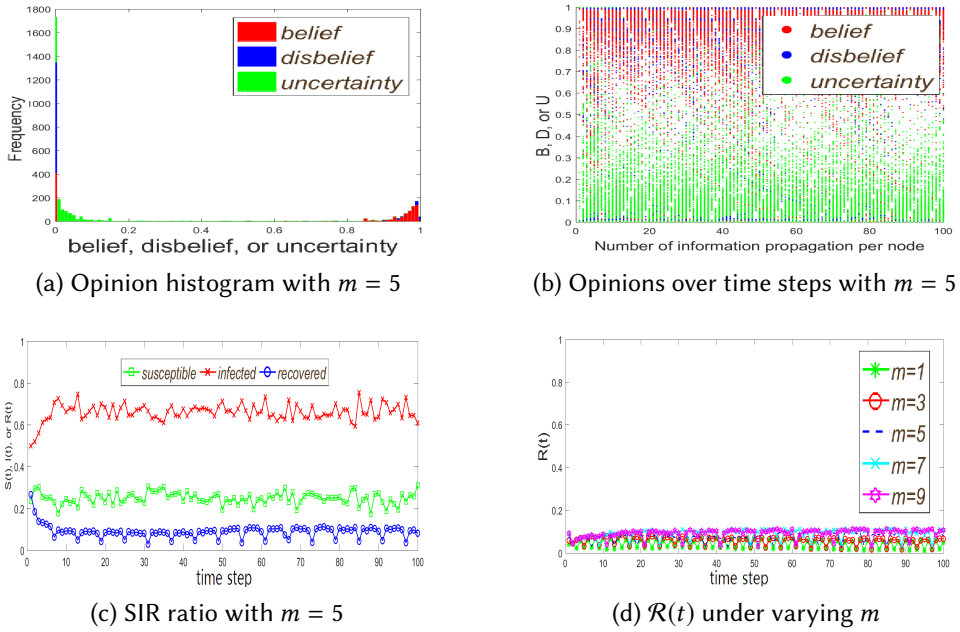


(d) News sharing frequency based on news reading frequency

Fig. 2. Description of information used to simulate a realistic network environment.

5.4 Simulation Results with TIs, DIs, and Ds

In this section, we discuss the effect of key design parameter values when doubters (Ds) use prior belief and when they don't, where a network is composed of true informers (TIs), disinformers (DIs), and doubters (Ds). Recall that TIs and DIs only influence Ds' opinions while they are not influenced by other agents' opinions.



(a) Opinion histogram with $m = 5$

(b) Opinions over time steps with $m = 5$

(c) SIR ratio with $m = 5$

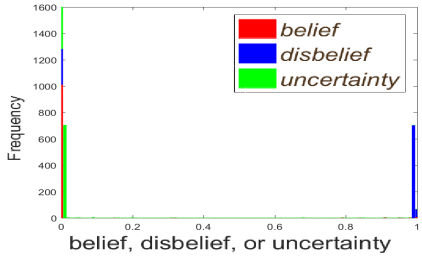
(d) $\mathcal{R}(t)$ under varying m

Fig. 3. Opinion histogram, opinion evolution, SIR ratio, and the fraction of recovered agents ($\mathcal{R}(t)$) without prior belief.

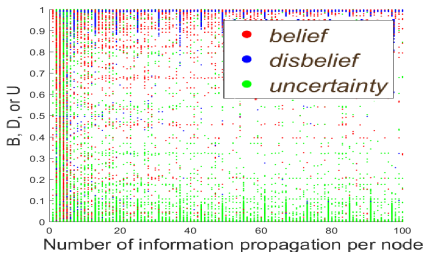
5.4.1 Results with Doubters without Prior Belief. Fig. 3 (a)-(c) shows the opinion evolution and corresponding SIR ratio when $m = 5$, propagating false information once and propagating true information five times, where doubters (Ds) do not use prior belief for their decisions. Results for $m < 5$ are very similar, so we omit them here. As evident in these, when m is sufficiently large, belief in false information is still dominant with a fairly high level of uncertainty. As shown in Fig. 3 (c), infected agents are clearly dominant even if more true information is propagated than false information. Fig. 3 (d) shows the instantaneous status of the fraction of recovered agents ($\mathcal{R}(t)$) without using prior belief. Although $\mathcal{R}(t)$ is slightly higher for larger m values than for smaller m values, $\mathcal{R}(t)$ is very small without showing much distinction for each m. From this result, we can conclude that in this network condition of TIs, DIs, and Ds (and agents do not use their prior belief), it is hard to mitigate or remove false information in the network. Since we do not observe any significant impact of varying m and other design parameter values (i.e., decay factor γ and fraction of TIs, P_{TI}, or DIs, P_{DI}), we omit showing the results from trials varying those parameters here.

Results Summary. Under a network with TIs, DIs, and Ds, where Ds do not use their prior belief in updating their opinions, mitigating or stopping false information information can rarely happen.

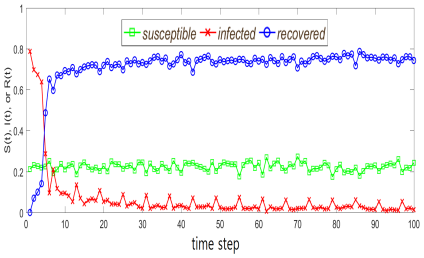
5.4.2 Results with Doubters with Prior Belief. Fig. 4 (a)-(c) shows opinion evolution when $m = 5$ when Ds use prior belief (i.e., $a_\mu = 0.5$ and $a_{sd} = 0.1$) for their decisions. Unlike Fig. 3 (a)-(c), we can clearly observe the effect of larger m with the dominance of recovered agents $\mathcal{R}(t)$ from $t = 5$. Further, we notice a lower level of uncertainty with prior belief compared to the level of uncertainty without prior belief. From these results, we can conclude that using prior belief can reduce the level of perceived uncertainty in agents' opinions. Fig. 4 (d) shows the effect of varying m on $\mathcal{R}(t)$ when Ds use prior belief with $a_\mu = 0.5$ and $a_{sd} = 0.1$. In particular, when $m \geq 5$ and $t \geq 50$, $\mathcal{R}(t)$ performs similarly maintaining the highest performance.



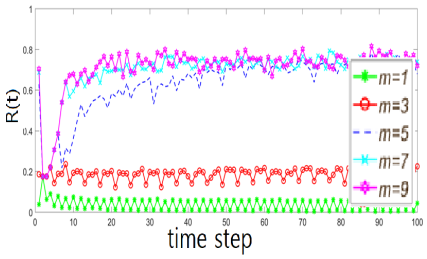
(a) Opinion histogram with $m = 5$



(b) Opinions over time steps with $m = 5$

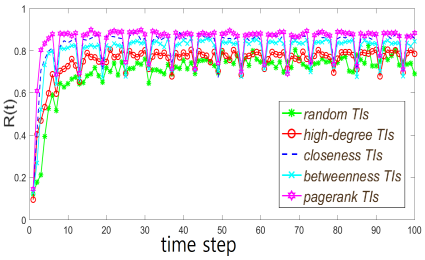


(c) SIR ratio with $m = 5$

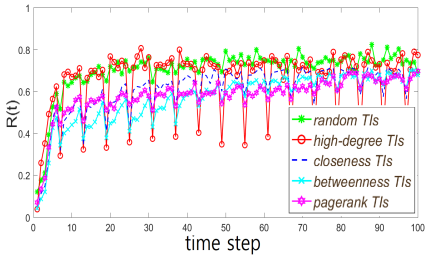


(d) $\mathcal{R}(t)$ w.r.t. varying m

Fig. 4. Opinion histogram, opinion evolution, SIR ratio, and the fraction of recovered agents ($\mathcal{R}(t)$) with prior belief when the network consists of TIs, DIs, and Ds and synthetic network environments are used.



(a) $\mathcal{R}(t)$ with varying the types of TIs and random DIs



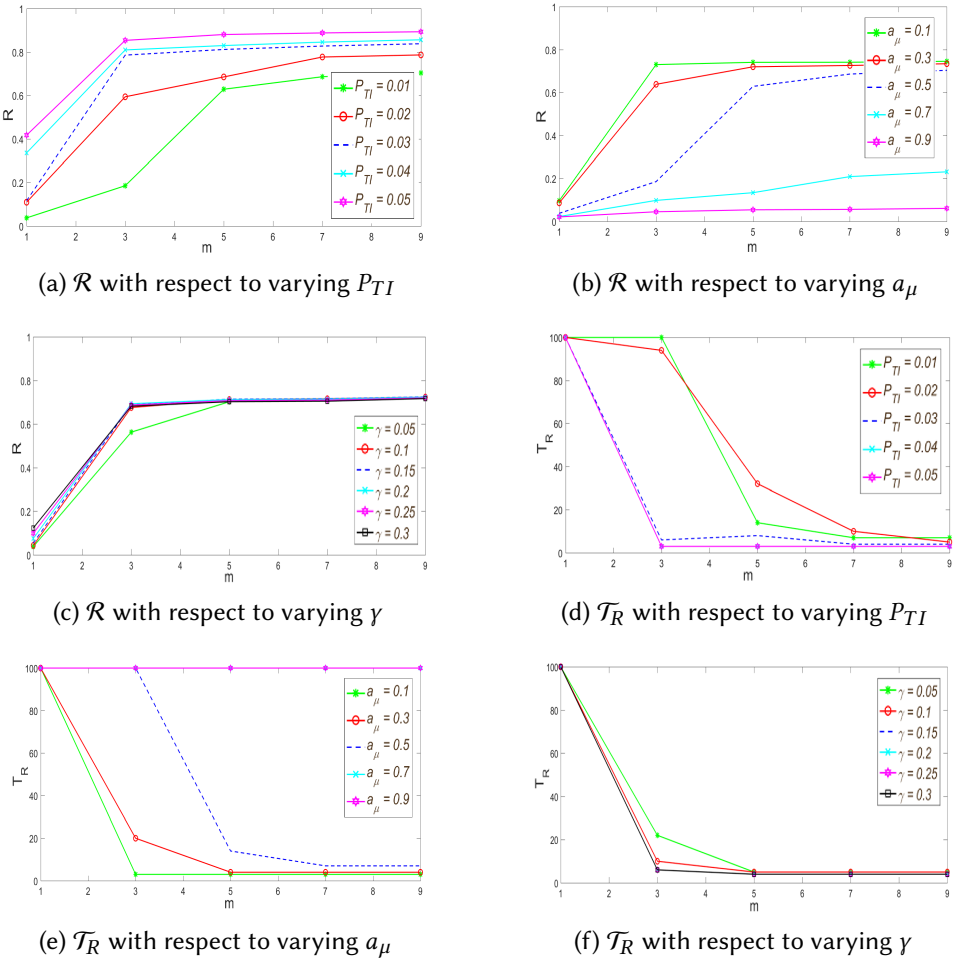
(b) $\mathcal{R}(t)$ with varying the types of DIs and random TIs

Fig. 5. Fraction of recovered agents ($\mathcal{R}(t)$) with prior belief with respect to various centrality types of TIs or DIs where a network consists of agent types of TIs, DIs, and Ds and synthetic network environments are used.

Fig. 5 demonstrates how different centrality types used in selecting TIs or DIs affect the fraction of recovered agents ($\mathcal{R}(t)$) under the presence of TIs, DIs, and Ds in a network with m is fixed at 5. Fig. 5 (a) shows $\mathcal{R}(t)$ when the different centrality types of TIs are used. As higher $\mathcal{R}(t)$ implies better performance, the performance order in $\mathcal{R}(t)$ is: pagerank TIs $\geq$ closeness TIs $\geq$ betweenness TIs $\geq$ high-degree TIs $\geq$ random TIs. Although we didn't show $\mathcal{S}(t)$ in this figure, we also observed lower $\mathcal{S}(t)$ corresponding to higher $\mathcal{R}(t)$, which is reasonable based on the impact of different centrality types of TIs. The result implies that highly influential TIs can clearly mitigate or stop false information propagated by DIs which are randomly selected. Overall the result implies that when a node connected with more nodes through third parties, called a node with high indirect connectedness, it can have more effect than one directly connected with more nodes when multiple rounds of information diffusion are performed, showing a higher ripple effect in the end.

Fig. 5 (b) shows the effect of varying different centrality types used in selecting DIs when Ds use prior belief with respect to $\mathcal{R}(t)$. The result is well aligned with the results observed in Fig. 5 (a).

In terms of DIs' perspective, lower performance in $\mathcal{R}(t)$ represents their higher performance, propagating false information more effectively, because selection based on a given centrality metric can mitigate increase of $\mathcal{R}(t)$. In that sense, higher $\mathcal{R}(t)$ by selecting high-degree DIs reflects poor performance of the DIs as false information propagation is effectively mitigated by the five times of true information propagation. On the other hand, selecting DIs based on pagerank, closeness, or betweenness significantly hurts mitigation of false information, implying that DIs with high pagerank, closeness, or betweenness can more effectively propagate false information, compared to DIs with high-degree. Although we didn't show $\mathcal{I}(t)$ due to space constraint, the trend of $\mathcal{R}(t)$ is well aligned with higher $\mathcal{I}(t)$ with pagerank, closeness, or betweenness while showing lower $\mathcal{I}(t)$ with high-degree and random.



(a) $\mathcal{R}$ with respect to varying P_{TI}

(b) $\mathcal{R}$ with respect to varying a_μ

(c) $\mathcal{R}$ with respect to varying γ

(d) $\mathcal{T}_R$ with respect to varying P_{TI}

(e) $\mathcal{T}_R$ with respect to varying a_μ

(f) $\mathcal{T}_R$ with respect to varying γ

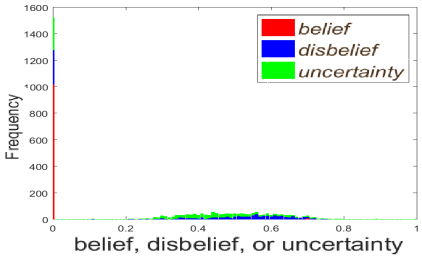
Fig. 6. Effect of varying the percentage of TIs (P_{TI}), the base rate means (a_μ), and decay factor (γ) on $\mathcal{R}$ and $\mathcal{T}_R$ where a network consists of agent types of TIs, DIs, and Ds with prior belief and synthetic network environments are used.

Fig. 6 demonstrates how the values of key design parameters impact the two performance metrics, $\mathcal{R}$ and $\mathcal{T}_R$ discussed in Section 5.1. Fig. 6 (a) and (d) shows the effect of varying P_{TI} on $\mathcal{R}$ and $\mathcal{T}_R$. We notice two key phenomena from the results. First, there exists a critical m for each P_{TI} that can significantly increase performance. Accordingly, when $P_{TI} \geq 0.03$, we can observe constant performance from $m = 3$. Second, when m is sufficiently large, even if P_{TI} is small, fairly high performance in both $\mathcal{R}$ and $\mathcal{T}_R$ is achievable.

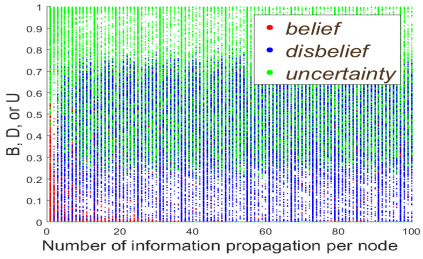
Fig. 6 (b) and (e) shows the effect of Ds' prior beliefs when their mean, a_μ, varies, given standard deviation, $a_{sd} = 0.1$. It is noticeable that when a_μ is no larger than 0.5, meaning that Ds are not biased for false information, a sufficiently large m (i.e., $m \geq 5$) can achieve fairly good performance in both $\mathcal{R}$ and $\mathcal{T}_R$.

Fig. 6 (c) and (f) shows the effect of opinion decay factor, γ, on $\mathcal{R}$ and $\mathcal{T}_R$. Although the effect of γ is not significant, we can notice that at least more opinion decay can help performance in $\mathcal{R}$ and $\mathcal{T}_R$.

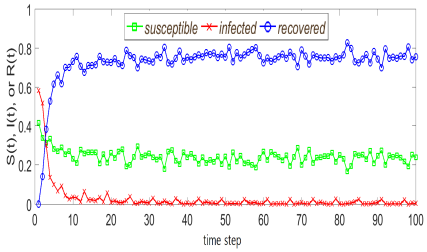
Results Summary. When m is sufficiently large (i.e., $m \geq 5$), the performance in the SIR ratio (based on Eq. (13)) is fairly high. TIs or DIs with high pagerank, closeness, or betweenness show a high effect on $\mathcal{R}$ and $\mathcal{T}_R$ showing better or worse performance corresponding to TIs or DIs, respectively, while TIs or DIs with high-degree can only comparably perform with randomly selected TIs or DIs. In terms of the effect of varying P_{TI} and a_μ, there exists a critical m that makes the performance maintain at the highest performance. The effect of γ is not significant; but lower γ (i.e., not forgetting) at least is not desirable for the performance in $\mathcal{R}$ and $\mathcal{T}_R$.



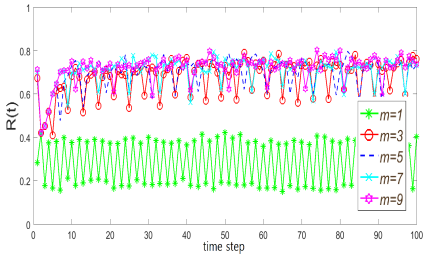
(a) Opinion histogram with $m = 5$ with prior belief/disbelief



(b) Opinions over time step with $m = 5$ with prior belief/disbelief



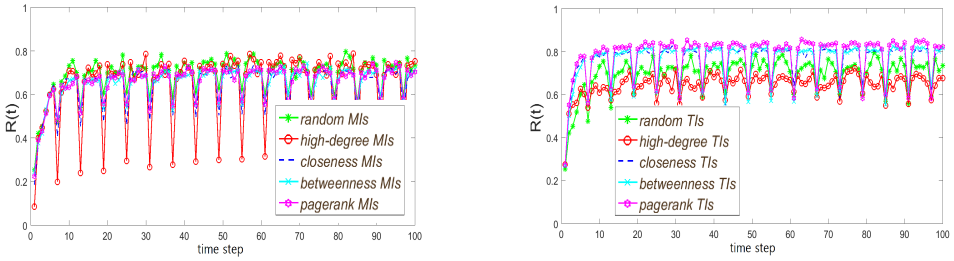
(c) SIR ratio with $m = 5$ with prior belief/disbelief



(d) $\mathcal{R}(t)$ under varying m

Fig. 7. Opinion histogram, opinion evolution, SIR ratio, and the fraction of recovered agents ($\mathcal{R}(t)$) with prior belief or disbelief when the network consists of TIs, DIs, and Ds and the realistic network environments are used.

5.4.3 Results based on Realistic Network Environments. In this section, we show the effect of using prior belief in a network consisting of TIs, DIs, and Ds. Recall that all information used to consider a real network topology, the quality of articles read by users, and the users' behavioral characteristics in information processing (i.e., sharing and reading) are described in Fig. 2. Fig. 7 shows how agents' opinions evolve when agents use prior belief or disbelief, in terms of the opinion histogram, evolution, and the SIR ratio over time when agents do not use prior belief or disbelief. Although the trends observed from the results are similar to the results using synthetic network environments shown in Fig. 4, we can notice a relatively high degree of uncertainty under realistic network environments used. However, the high uncertainty is quickly credited when the agents



(a) $\mathcal{R}(t)$ under varying DIs' centrality (b) $\mathcal{R}(t)$ under varying TIs' centrality

Fig. 8. Fraction of recovered agents ($\mathcal{R}(t)$) with prior belief or disbelief with respect to various centrality types of TIs or DIs where a network consists of agent types of TIs, DIs, and Ds and the realistic network environments are used.

make decisions on whether to believe true or false information because they can use non-biased prior belief based on information processing behaviors by the users in Fig. 2 (c). This result is evident from Fig. 7 (c) showing a quicker governance of true information over false information (i.e., at the third propagation of true information) than under the case with synthetic network environments in Fig. 4 (c) (i.e., at the fifth propagation of true information). In addition, under varying numbers of propagating true information (m) per false information, due to the higher uncertainty observed under realistic network environments, the leverage of the uncertainty as credit based on prior belief or disbelief is more clearly observed, compared to the case under synthetic network environments. Notice that all cases with $m > 1$ reach the highest performance in that agents believe in true information with respect to $\mathcal{R}(t)$.

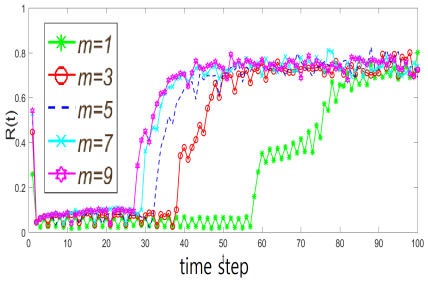
Fig. 8 also shows how the different network positions (based on multiple centrality types) of TIs or DIs affect the fraction of recovered agents ($\mathcal{R}(t)$) believing in true information over time. The trends observed under realistic network environments in Fig. 8 are very similar to those under synthetic network environments in Fig. 5. However, we can see much less sensitivity over different centrality types due to the high uncertainty observed in the evolution of opinions. Recall that the initial opinions of TIs or DIs are not highly biased to respective extremes when their opinions are formulated based on the quality of real articles (due to their inherent noise). However, we can still observe the highly similar trends between the two cases when respective network environment is used in that nodes more indirectly connected with other nodes have more power to affect the evolution of opinions over the network.

Results Summary. The trends of the results using realistic network environments are highly similar to the ones observed under synthetic network environments. However, due to the noisiness of realistic network environments with respect to the quality of article statements, the opinions of seeding agents (i.e., TIs, DIs, or MIs) have relatively high uncertainty compared to those of seeding agents under synthetic network environments. Hence, we observe agents' opinions that have high uncertainty. As agents use their unbiased prior belief or disbelief, their decisions in believing in true information are not affected much by noisy information. The results using realistic network environments confirmed that a node with high indirect connectedness has more power to reach out other nodes more, leading to more nodes to believe in true information.

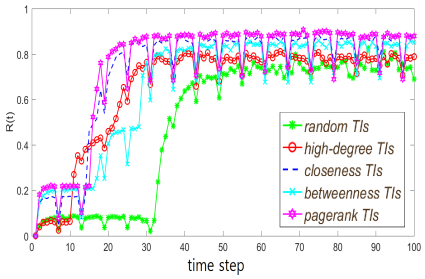
5.5 Simulation Results with TIs, MIs, and Ds

In this section, we discuss the effect of key design parameter values when doubters (Ds) update their opinions with or without using prior belief where a network is composed of true informers (TIs), misinformers (MIs), and doubters (Ds). Recall that TIs only influence other agents' opinions (i.e., Ds and MIs) while they are not influenced by other agents. Note that unlike DIs, MIs mistakenly

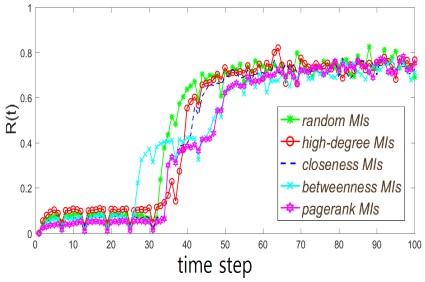
propagate false information initially but then can change their opinions upon interacting with other agents.



(a) $\mathcal{R}(t)$ under varying m



(b) $\mathcal{R}(t)$ with varying the types of TIs and random MIs



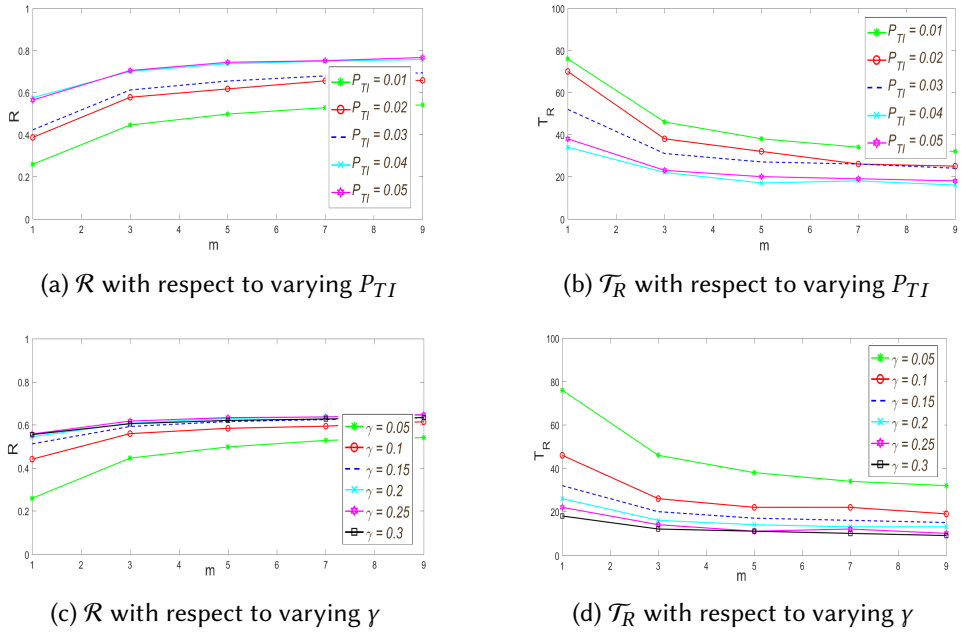
(c) $\mathcal{R}(t)$ with varying the types of MIs and random TIs

Fig. 9. Fraction of recovered agents ($\mathcal{R}(t)$) without prior belief where a network consists of agent types of TIs, MIs, and Ds without prior belief.

5.5.1 Results with Doubters without Prior Belief. Fig. 9 shows the instantaneous fraction of agents believing in true information, $\mathcal{R}(t)$, when Ds do not use prior belief and a network has three types of agents including TIs, MIs, and Ds. Note that we report our results based on synthetic network environments here. First, we show $\mathcal{R}(t)$ under varying m in Fig. 9 (a). Unlike DIs, since MIs are also influenced by other agents' opinions, $\mathcal{R}(t)$ increases and even higher m expedites the increase of $\mathcal{R}(t)$. This result is very interesting compared to the effect of varying m under a network with TIs, DIs, and Ds without using prior belief, as shown in Fig. 3 (d) because the mitigated opinions of MIs can significantly reduce false information propagation, resulting in strikingly increasing $\mathcal{R}(t)$ over time. In particular, notice the phase transition that exists. At a critical time moment, low recovered agents transition to high recovered agents. Fig. 9 (a) shows how this transition point depends on m.

Fig. 9 (b) and (c) investigates how different centrality types of TIs or MIs affect $\mathcal{R}(t)$ when Ds do not use prior belief with $m = 5$ and a network consists of TIs, MIs, and Ds. Similar to Fig. 9 (a), there exists a critical phase transition moment from low $\mathcal{R}(t)$ to high $\mathcal{R}(t)$. When the type of TI varies with random MIs in Fig. 9 (b), the performance in $\mathcal{R}(t)$ (higher is better) is: pagerank TIs $\geq$ closeness TIs $\geq$ betweenness TIs $\geq$ high-degree TIs $\geq$ random TIs. This implies that true information can be better propagated when nodes are indirectly connected rather than being directly or randomly connected. In Fig. 9 (c), we vary the types of MIs while the type of TIs is fixed with random selection. The performance order in $\mathcal{R}(t)$ is almost reverse to the case in Fig. 9 (b) as: random MIs $\geq$ high-degree MIs $\geq$ closeness MIs $\geq$ betweenness MIs $\geq$ pagerank MIs. This implies that when MIs are selected based on pagerank, betweenness, or closeness, their false information

propagation is more effective and makes it hard for randomly selected TIs to eradicate their false information in the network.



(a) $\mathcal{R}$ with respect to varying P_{TI} (b) $\mathcal{T}_R$ with respect to varying P_{TI}

(c) $\mathcal{R}$ with respect to varying γ (d) $\mathcal{T}_R$ with respect to varying γ

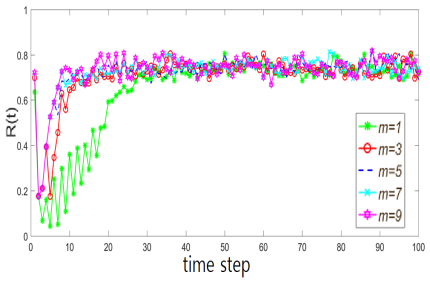
Fig. 10. Effect of varying the percentage of TIs (P_{TI}), the base rate means (a_μ), and decay factor (γ) on $\mathcal{R}$ and $\mathcal{T}_R$ where a network consists of agent types of TIs, MIs, and Ds without prior belief.

Fig. 10 (a) and (b) shows the effect of varying the percentage of TIs over total agents, N, on the fraction of recovered agents $\mathcal{R}$ and the time point that recovered agents dominates over infected agents, $\mathcal{T}_R$. As expected, as P_{TI} increases, higher $\mathcal{R}$ is observed while $\mathcal{T}_R$ decreases, implying that higher P_{TI} shows higher performance by having more recovered agents with a shorter convergence time that governs more recovered agents over infected agents. In addition, as expected, higher m increases $\mathcal{R}$ while decreasing $\mathcal{T}_R$. In particular, notice that when $m > 3$, performance is more like constant rather than a rapid increase of $\mathcal{R}$ or a rapid decrease of $\mathcal{T}_R$. Also when P_{TI} increases from 0.1 to 0.2 or from 0.3 to 0.4, the performance increases significantly in both $\mathcal{R}$ and $\mathcal{T}_R$.

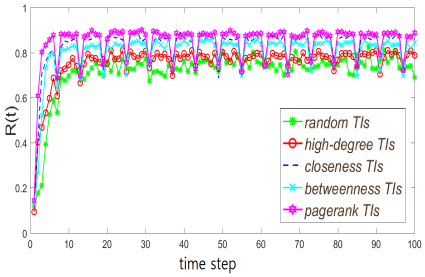
Fig. 10 (c) and (d) demonstrates how an opinion decay factor, γ, affects the performance in $\mathcal{R}$ and $\mathcal{T}_R$ when a network is comprised of TIs, MIs, and Ds and agents do not use prior belief. Recall that when the network consists of TIs, DIs, and Ds as shown in Fig. 6 (c) and (d), the decay factor didn't show much sensitivity regardless of whether agents use prior belief or not. The effect of the decay factor is prominent under this network environment even if agents do not use prior belief. Fig. 10 (c) and (d) proves that higher γ (i.e., more opinion decay) significantly increases $\mathcal{R}$ and decreases $\mathcal{T}_R$, implying more effective in mitigating false information. On the other hand, lower γ strikingly decreases $\mathcal{R}$ and increases $\mathcal{T}_R$, meaning less effective in reducing false information propagation. Since MIs can change their opinions over time even if they first propagate false information, higher opinion decay significantly helps reduce the effect of propagating false information by MIs.

Results Summary. Unlike the results shown in Section 5.4.1, the highly significant effect of varying key design parameters even without using prior belief is observed. First, due to the nature of MIs whose opinions are also influenced by other agents' opinions while their opinions are initiated with high belief in false information, there exists the critical transition from the period governed by the infected agents to the period dominated by the recovered agents. Higher m (i.e., the number of true information propagation per false information propagation) brings the critical

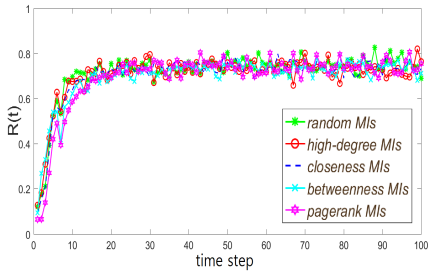
phase transition early while lower m delays the critical transition later. We also investigated the effect of centrality metrics used for selecting TIs or MIs. The results are well aligned with the reasoning obtained from Fig. 5 such that centrality metrics representing high indirect connectedness (i.e., pagerank, betweenness, or closeness) can generate more effective than ones with high direct connectedness (i.e., high-degree). In terms of TIs' perspective, TIs selection with pagerank, closeness, or betweenness is more favored than those based on high-degree or random selection. On the other hand, in terms of MIs' perspective, MIs selection with pagerank, closeness, or betweenness can be more favored. Lastly, a higher percentage of TIs can increase the resulting number of recovered agents and bring about the dominance of the recovered agents more quickly in a network. However, when a sufficiently large m is used, a significantly better performance is achievable with more recovered agents or making the critical phase transition early. Unlike little sensitivities observed in Fig. 6 (e) and (f), the clear sensitivity over a range of varying the decay factor, γ, is observed. Higher γ (i.e., more decaying) increases performance in both $\mathcal{R}$ and $\mathcal{T}_R$.



(a) $\mathcal{R}(t)$ under varying m



(b) $\mathcal{R}(t)$ with varying the types of TIs and random MIs



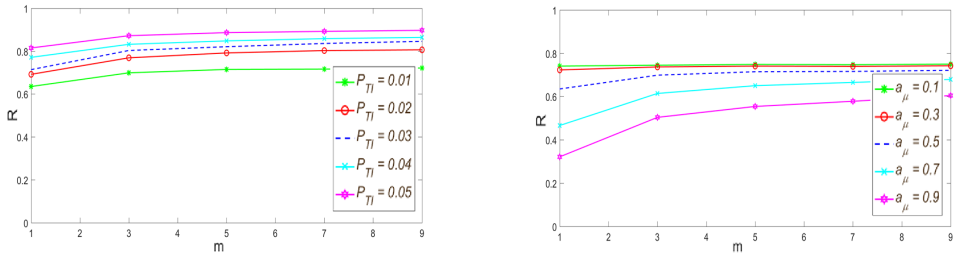
(c) $\mathcal{R}(t)$ with varying the types of MIs and random TIs

Fig. 11. Fraction of recovered agents ($\mathcal{R}(t)$) without prior belief where a network consists of agent types of TIs, MIs, and Ds with prior belief.

5.5.2 Results with Doubters with Prior Belief. In Fig. 11, now we discuss the fraction of recovered agents ($\mathcal{R}(t)$) when a network consists of TIs, MIs, and Ds where MIs and Ds use their prior belief. Unlike Fig. 9, Fig. 11 (a) does not show high sensitivity particularly when $m > 1$. Notice that when $m > 1$, almost 80 % of agents are recovered agents as shown in Fig. 9 (a). Even if there is little sensitivity with respect to varying m, higher performance is constantly observed with using prior belief than without using prior belief.

In Fig. 11 (b) and (c), we show how the centrality metrics used for selecting TIs or MIs affects the performance in $\mathcal{R}(t)$. First, there is higher sensitivity when varying the centrality types of TIs than when varying the centrality types of MIs. In Fig. 11 (b) showing the effect of varying the types of

TIs, susceptible agents move to recovered agents or recovered agents to susceptible agents while the number of infected agents quickly becomes almost zero and maintains as is. This is because MIs can change their opinions and join the pool of Ds' opinion in agreeing with true information over time. The trends observed here are similar to what we observed in Fig. 5 (b). Centrality metrics focusing on indirect connectedness give better chances to reach out to more nodes in a network and result in better performance in propagating true information by TIs. That is, pagerank, closeness, or betweenness indeed shows higher $\mathcal{R}(t)$. Under varying the types of MIs as shown in Fig. 11 (c), we do not see much sensitivity but overall observe quickly increasing $\mathcal{R}(t)$ reaching about 0.8. This would be because MIs can also join the pool of Ds' opinions by changing their opinions, which allow vanishing the initial false information quickly.



(a) $\mathcal{R}$ with respect to varying P_{TI} (b) $\mathcal{R}$ with respect to varying a_μ

(c) $\mathcal{T}_R$ with respect to varying P_{TI} (d) $\mathcal{T}_R$ with respect to varying a_μ

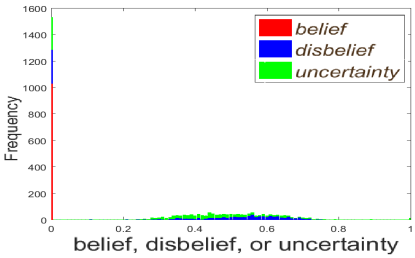
Fig. 12. Effect of varying the percentage of TIs (P_{TI}) and the base rate means (a_μ) on $\mathcal{R}$ and $\mathcal{T}_R$ where a network consists of agent types of TIs, MIs, and Ds with prior belief.

Now we discuss how agents' prior beliefs in a network with TIs, MIs, and Ds affect the performance in $\mathcal{R}$ and $\mathcal{T}_R$ with respect to varying P_{TI}, a_μ, and γ. In Fig. 12 (a) and (d), we examined the effect of varying P_{TI} on $\mathcal{R}$ and $\mathcal{T}_R$. Higher performance is observed when higher P_{TI} is used on $\mathcal{R}$ and $\mathcal{T}_R$. In particular, a steady, constant increasing $\mathcal{R}$ is observed when P_{TI} increases. However, we do not observe significantly higher effect when using higher m, showing almost same $\mathcal{R}$ for $m > 3$. In Fig. 12 (d), as long as $m > 3$, we do not observe significantly high effect of using higher P_{TI} or higher m on $\mathcal{T}_R$. These phenomena can be well explained in a sense that MIs change their opinions over time as they are influenced by other agents' opinions as well.

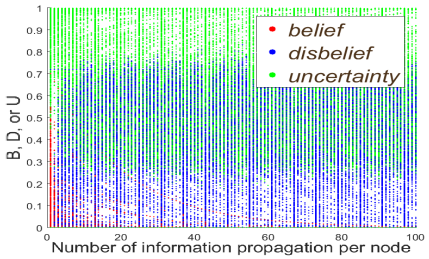
In Fig. 12 (b) and (e), we examined the effect of varying agents' prior belief a_μ on $\mathcal{R}$ and $\mathcal{T}_R$. Recall that a_μ is the base rate used in expected belief where the corresponding $1 - a_\mu$ is used to estimate expected disbelief in Eq. (3). Higher a_μ means that an agent is more biased for belief (i.e., favorably believing in false information) while lower a_μ implies that an agent is more biased for disbelief (i.e., favorably believing in true information). As a_μ becomes lower, higher $\mathcal{R}$ and lower $\mathcal{T}_R$ are observed, implying higher performance. Notice that as long as agents are not unfairly biased in believing in false information (i.e., $a_\mu > 0.5$), a fairly high $\mathcal{R}$ and a fairly low $\mathcal{T}_R$ are observed particularly when $m > 3$.

We also investigated the effect of varying the opinion decay, γ on $\mathcal{R}$ and $\mathcal{T}_R$. However, we rarely observed sensitivity of γ on the two metrics, and thus omitted showing them. Since MIs' opinions change over time by being influenced by other agents' opinions, the impact of their original false information vanishes over time, resulting in less effect of using higher γ.

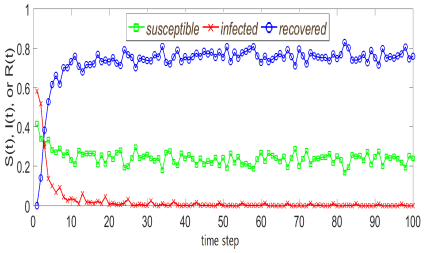
Results Summary. Overall, compared to the results shown in Section 5.5.1 where agents do not use prior belief in a network of TIs, MIs, and Ds, little sensitivity of varying key design parameters is observed although the absolute performance is better when using prior belief than when not using prior belief. For example, under varying m, after a certain time period, no sensitivity of varying m is observed, showing constantly high performance. Selection of TIs based on more indirect connectedness (i.e., pagerank, closeness, or betweenness) is preferred to maximize performance in $\mathcal{R}(t)$. Although little effect is observed when varying γ, the effect of higher $\mathcal{R}$ and lower a_μ is clear under varying P_{TI} and a_μ. However, when P_{TI} is sufficiently high (i.e., $P_{TI} > 0.01$) and a_μ is sufficiently low (i.e., $a_\mu \leq 0.5$), we observe a critical m that can increase performance; but after this critical m, there is no further significant improvement in $\mathcal{R}$ and $\mathcal{T}_R$.



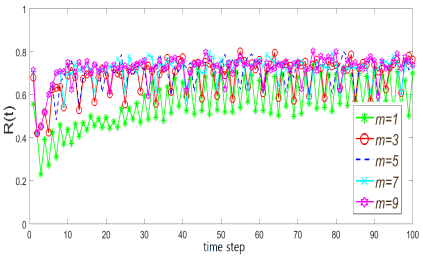
(a) Opinion histogram with $m = 5$ with prior belief/disbelief



(b) Opinions over time step with $m = 5$ with prior belief/disbelief



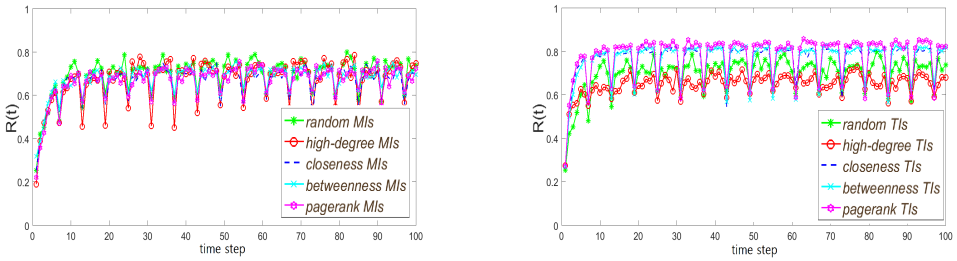
(c) SIR ratio with $m = 5$ with prior belief/disbelief



(d) $\mathcal{R}(t)$ under varying m

Fig. 13. Opinion histogram, opinion evolution, SIR ratio, and the fraction of recovered agents ($\mathcal{R}(t)$) with prior belief or disbelief when the network consists of TIs, MIs, and Ds and the realistic network environments are used.

5.5.3 Results based on Realistic Network Environments. In this section, we investigated the effect of agents' opinion evolution when a network consists of TIs, MIs, and Ds and realistic network environments are used. In Figs. 13 and 14, the trends of the observed results are highly similar to the ones shown in Figs. 7 and 8, respectively. But under this network environment with TIs, MIs, and Ds, due to MIs' open-minded attitude to update their opinions, we observe significantly higher $\mathcal{R}(t)$ even with $m = 1$ in Fig. 13 (d), compared to the results in Fig. 7 (d). In addition, Fig. 14 shows that the minimum $\mathcal{R}(t)$ is much higher than the minimum $\mathcal{R}(t)$ shown in Fig. 8.



(a) $\mathcal{R}(t)$ under varying MIs' centrality (b) $\mathcal{R}(t)$ under varying TIs' centrality

Fig. 14. Fraction of recovered agents ($\mathcal{R}(t)$) with prior belief or disbelief with respect to various centrality types of TIs or DIs where a network consists of agent types of TIs, MIs, and Ds and the realistic network environments are used.

Results Summary. The results are highly similar to the ones observed under a network with TIs, DIs, and Ds. Due to MIs' open-mindedness in updating their opinions, higher $\mathcal{R}(t)$ is observed under a network with TIs, MIs, and Ds than the network with TIs, DIs, and Ds.

6 CONCLUSION

In this work, we studied how to mitigate or eradicate false information in a social network more effectively by developing an opinion model dealing with multiple causes of uncertainty. To be specific, we proposed an opinion model based on Subjective Logic (SL) which can explicitly deal with the dimension of uncertainty in an opinion. However, the current form of SL's opinion only reduces the level of uncertainty proportional to the amount of information received regardless of whether or not the received information is conflicting to each other. In order to study the interactions and evolution of dynamic opinions in social networks, the proposed opinion model enhances SL by introducing the following: (1) an operator to deal with ambiguity derived from conflicting evidence; (2) homophily-based opinion updates; (3) opinion decay over time; and (4) the use of expected belief or disbelief in updating opinions. In addition, we associated agents' opinion status with the three statuses in the SIR model by allowing agents to move from one state to another state as long as their opinions' uncertainty does not reach zero.

Through the extensive simulation experiments using both synthetic and realistic network environments, we obtained the following findings:

- Agents' prior belief in updating their opinions can help reduce the level of uncertainty in their opinions and allows the network to reach a dominance of recovered agents believing in true information, leading to effectively mitigating or eradicating false information in a network.
- Even if a small fraction of agents propagate true information, as long as a sufficient amount of true information is propagated, false information can be fairly and effectively mitigated.
- As long as agents are not more biased for false information, more true information or more true informers can significantly help increase the number of agents believing in true information.
- Higher opinion decays mostly help reduce the governance of false information in a network.
- An agent's indirect connectedness (e.g., pagerank, betweenness, or closeness) in a network can significantly help propagate its information (either true or false information) more effectively, compared to its direct connectedness (e.g., high-degree).

- If agents initially propagating false information can change their opinions (i.e., misinformers), false information propagation can be much more easily mitigated or removed from a network over time even if agents do not use their prior belief in updating their opinions.

Different from the state-of-the-art agent models dealing with dynamics and evolution of opinions, this work considered the two dimensions of uncertainty, ignorance and ambiguity, which can affect the level of uncertainty in agents' opinions. This allows us to study how to counter false information with two different types of false informers. In particular, we used an agent's uncertainty in its own opinion as a vehicle for it to move from one status to another in the SIR model. This design is to reflect more realistic scenarios that a person's confidence can be fluctuating depending on received information, while the existing opinion models keep reducing uncertainty in people's opinions upon receiving more evidence. The proposed agent's opinion model and corresponding strategies to deal with false information can be applicable to combat the spread of fake news in various social media platforms.

Our findings in this study can help to inform operators of social network platforms such as Facebook of ways to encourage their users to mitigate the spread of false information. We propose four key steps towards this goal. First, operators should encourage users to weight their own prior beliefs carefully, with more confidence when consuming new information. Second, users who believe they know the facts should be encouraged to share them. Third, those users who are energized to fight against the spread of false information should be encouraged by the network operators to become more connected, especially to different groups, and to become more vocal about the facts. Finally, operators should promote the idea of intervention, as long as it is polite and constructive: if a user believes that a connection is spreading false information because that connection has been misinformed, the network should encourage that user to convince that person of the facts, in a polite and constructive manner, using whatever supporting evidence is available.

For future research, we plan to extend this work in three directions. First, we will investigate the effect of public media information, in addition to the effect of information exchanges based on the interactions of two agents. In particular, we are interested in investigating how the quality of public media information (i.e., information credibility) can affect the behaviors agents accept or if they filter out the information. Second, we will examine how agents' topic expertise on a given topic proposition can affect their beliefs in given information in the same domain. Lastly, we will consider both DIs and MIs in a network and investigate their interplay in terms of how DIs can leverage the influence of MIs in order to amplify their false information propagation.

REFERENCES

R. Axelrod. 1997. The Dissemination of Culture. *Journal of Conflict Resolution* 41, 2 (1997), 203–226.

A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi. 2015. Viral Misinformation: The Role of Homophily and Polarization. In *Proceedings of the 24th International Conference on World Wide Web (ACM WWW '15 Companion)*. New York, NY, USA, 355–356.

Lachlan Brumley, Carlo Kopp, and Kevin Burt Korb. 2012. Cutting through the tangled web: An information-theoretic perspective on information warfare. *Air Power Australia Analyses* IX (2012), 1–25.

C. Budak, D. Agrawal, and A. E. Abbadi. 2011. Limiting the Spread of Misinformation in Social Networks. In *ACM International World Wide Web Conference*.

A. Campan, A. Cuzzocrea, and T. M. Truta. 2017. Fighting fake news spread in online social networks: Actual trends and future research directions. In *2017 IEEE International Conference on Big Data (Big Data)*. 4453–4457.

C. Castillo, M. Mendoza, and B. Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. 675–684.

W. Chen, Y. Wang, and S. Yang. 2009. Efficient Influence Maximization in Social Networks. In *Proceedings of the 15th ACM SIGKDD*. New York, NY, USA, 199–208.

J.H. Cho, T. Cook, S. Rager, J. O'Donovan, and S. Adali. 2017. Modeling and Analysis of Uncertainty-based False Information Propagation in Social Networks. In *IEEE Global Communications Conference (GLOBECOM 2017)*. Singapore.

J.H. Cho and A. Swami. 2014. Dynamics of Uncertain Opinions in Social Networks. In *IEEE Military Communications Conference (MILCOM)*. Baltimore, MD, USA.

M. Collard, L. Brisson, P. Collard, and E. Stattner. 2015. Rumor spreading modeling: Profusion versus scarcity. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 1547–1554.

P. Domingos and M. Richardson. 2001. Mining the Network Value of Customers. In *Proceedings of the Seventh ACM SIGKDD*. New York, NY, USA, 57–66.

R. Durrett. 1988. *Lecture Notes on Particle Systems and Percolation*. Wadsworth Publishing.

J. Goldenberg, Barak Libai, and Eitan Muller. 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters* 12, 3 (2001), 211–223.

M. Granovetter. 1978. Threshold Models of Collective Behavior. *Amer. J. Sociology* 83, 6 (1978), 1420–1443.

D. Hughes, S. Adalı, J.-H. Cho, and J. A. Mangels. 2015. Individual Differences in Information Processing in Networked Decision Making. (2015).

A. Jøsang. 2001. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 9, 3 (June 2001), 279–311.

A. Jøsang. 2016. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer.

D. Kempe, J. Kleinberg, and E. Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD*. New York, NY, 137–146.

C. Kopp. 2000. Information Warfare: A Fundamental Paradigm of Infowar. *Systems: Enterprise Computing Monthly* (2000), 46–55.

Carlo Kopp, Kevin B. Korb, and Bruce I. Mills. 2018. Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to "fake news". *PLOS ONE* 13 (11 2018), 1–35.

V. Krishnamurthy and M. Hamdi. 2013. Misinformation Removal in Social Networks: Constrained Estimation on Dynamic Directed Acyclic Graphs. *IEEE Journal of Selected Topics in Signal Processing* 7, 2 (April 2013), 333–346.

K. Kumar, P. Krishna, and G. Geethakumari. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences* 4, 1 (2014), 14.

J. Leskovec. 2012. Social circles: Facebook. (2012). https://snap.stanford.edu/data/egonets-Facebook.html

J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. 2007. Cost-effective Outbreak Detection in Networks. In *Proceedings of the 13th ACM SIGKDD*. New York, NY, USA, 420–429.

S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. 2012. Misinformation and Its Correction. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131.

L. Li, A. Scaglione, A. Swami, and Q. Zhao. 2011. Opinion dynamics and learning in social networks. *IEEE Journal on Selected Areas in Communications* 31, 6 (March 2011), 3–49.

G. Liang, W. He, C. Xu, L. Chen, and J. Zeng. 2015. Rumor Identification in Microblogging Systems Based on Users' Behavior. *IEEE Transactions on Computational Social Systems* 2, 3 (Sept 2015), 99–108.

I. Litou, V. Kalogeraki, I. Katakis, and D. Gunopulos. 2016. Real-Time and Cost-Effective Limitation of Misinformation Propagation. In *17th IEEE International Conference on Mobile Data Management*.

B. Liu, G. Cong, Y. Zeng, D. Xu, and Y. M. Chee. 2014. Influence Spreading Path and Its Application to the Time Constrained Social Influence Maximization Problem and Beyond. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2014), 1904–1917.

M. W. Macy. 1991. Chains of Cooperation: Threshold Effects in Collective Action. *American Sociological Review* 56, 6 (1991), 730–747.

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

N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. New York, NY, USA, 213–222.

V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. 2011. Rumor Has It: Identifying Misinformation in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1589–1599.

P. Raj H and Y. Narahari. 2012. Influence Limitation in Multi-Campaign Social Networks: A Shapley Value Based Approach. In *8th IEEE International Conference on Autonomous Science and Engineering*. Seoul, Korea, 735–740.

G. Shafer. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.

P. Smets and R. Kennes. 1994. The transferable belief model. *Artificial Intelligence* 66 (1994), 191–234.

B. C. Stahl. 2006. On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective. *Informing Science Journal* 9 (2006).

S. Subramanian. 2017. Inside the Macedonian Fake-News Complex. (2017). https://www.wired.com/2017/02/veles-macedonia-fake-news/

P.-N. Tan, M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

G. Tong, W. Wu, S. Tang, and D. Z. Du. 2016. Adaptive Influence Maximization in Dynamic Social Networks. *IEEE/ACM Transactions on Networking* PP, 99 (2016), 1–14.

W. Y. Wang. 2017. Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Vancouver, BC, Canada.

Y. Wang and M. P. Singh. 2010. Evidence-based Trust: A Mathematical Model Geared for Multiagent Systems. *ACM Transactions on Autonomous and Adaptive Systems* 5, 4, Article 14 (Nov. 2010), 28 pages.

Y. Wang, A. V. Vasilakos, J. Ma, and N. Xiong. 2015. On Studying the Impact of Uncertainty on Behavior Diffusion in Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 2 (Feb 2015), 185–197.

C. Wardle. 2017. Fake news. It's complicated. (2017). https://firstdraftnews.com/fake-newscomplicated/

L. Zhao, X. Qiu, X. Wang, and J. Wang. 2013. Rumor spreading model considering forgetting and remembering mechanisms in inhomogeneous networks. *Physica A: Statistical Mechanics and its Applications* 392, 4 (2013), 987–994.

L. Zhao, J. Wang, Y. Chen, Q. Wang, J. Cheng, and H. Cui. 2012. SIHR rumor spreading model in social networks. *Physica A: Statistical Mechanics and its Applications* 391, 7 (2012), 2444–2453.

L. Zhao, Q. Wang, J. Cheng, Y. Chen, J. Wang, and W. Huang. 2011. Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal. *Physica A: Statistical Mechanics and its Applications* 390, 13 (2011), 2619–2625.