Resilient Consensus for Expressed and Private Opinions

Yilun Shang

***Abstract*—This paper proposes an opinion formation model featuring both a private and an expressed opinion for a given topic over dynamical networks. Each individual in the network has a private opinion, which is not known by others but evolves under local influence from the expressed opinions of its neighbors, and an expressed opinion, which varies under a peer pressure to conform to the local environment. We design opinion sifting strategies which are purely distributed and provide resilience to a range of adversarial environment involving locally and globally bounded threats as well as malicious and Byzantine individuals. We establish sufficient and necessary graph-theoretic criteria for normal individuals to attain opinion consensus in both directed fixed and time-varying networks. Two classes of opinion clustering problems are introduced as an extension. By designing resilient opinion separation algorithms, we develop necessary and sufficient criteria, which characterize resilient opinion clustering in terms of the ratio of opinions as well as the difference of opinions. Numerical examples including real-world jury deliberations are presented to illustrate the effectiveness of the proposed approaches and test the correctness of our theoretical results.**

***Index Terms*—Social dynamics, resilience, consensus, clustering, social network, multi-agent system.**

I. Introduction

In recent years, the study of opinion formation among individuals and the resulting dynamics it induces in a social network has become a canonical problem in social network analysis, where the phrase "opinion dynamics" encapsulates a wide range of models differentiating in the phenomena of interest including minority opinion spreading, collective decision making, polarization, and emergence of fads etc. In this setting, a common goal is to study the way individuals in a social network exchange their attitudes or opinions to agree on some topic or reach a consensus of opinions [1], [2].

As the classical consensus problems in multi-agent systems, the individuals can only communicate based upon local available information obtained from their neighboring individuals delineated by a social network. Although it is notoriously challenging to characterize and assess the collective behaviors in which human psychological and emotional factors are implicated, various agent-based models of opinion dynamics have been investigated. According to the opinion value held by an individual, opinion dynamics models are categorized as discrete and continuous models. For discrete models, we are inclined to indicate yes or no by using binary values. Well-known examples include the Sznajd model [3], the voter model [4], and Galam's majority-rule model [5]. There are also circumstances where the opinion of an individual is preferably expressed using real numbers and they can smoothly change between the two extremal values. Attitudes in a formal disputation, prices of a commodity, and predictions about certain macroeconomic variables are some of the examples. In the continuous case, two models involving bounded confidence mechanism presented by Deffuant and Weisbuch [6] and Hegselmann and Krause [7], respectively, have attracted considerable research interest. In these models, each individual only discuss with those who have an opinion close to its own within a given threshold, capturing the tendency of homophily in sociology. Pertinent models with hybrid opinion values, e.g. [8], have also been investigated.

The predominant assumption in most existing opinion dynamics including those above is that each individual has a single opinion for a given topic of discussion. However, an individual in reality may hold a private opinion, for the same topic, different to the opinion it expresses due to various reasons such as political correctness or peer pressure [9]. Such discrepancy has been well documented in empirical data and sociopsychological literatures, even linking to major political events including the disintegration of the Soviet Union [10] and the Arab Spring movement [11]. A common reason the discrepancy between private and expressed opinions arises is normative pressures on an individual to conform in a group situation. This often leads to pluralistic ignorance [12], where individuals privately disapprove of a view but publicly go along with it because they believe (sometimes even erroneously) the majority of others accept it. Examples include young Belgian men concerning their attitudes toward communal men's self-description and behavioral intentions [13], and college students concerning their opinions of the average exam study time of their peers [14]. One of the goals of this paper is hence to build a tractable agent-based model to accommodate both an expressed and a private opinion of an individual and study the evolution of opinions based on individuals' available local information in the network.

Beyond the implicit discrepancy between expressed and private opinions, a more tangible and ever-increasing threat to the decision making in social networks comes from the existence of misbehaving individuals [15], partly due to the pervasive applications of social network service. For example, there can be stubborn individuals in a company boardroom or in a discussion group, who influence others but would not vary their own opinions [16], and malicious users in a public forum

Y. Shang is with the Department of Computer and Information Sciences, Faculty of Engineering and Environment, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK (e-mail: shylmath@hotmail.com).

Manuscript received November 12, 2018; revised xxxx xx, xxxx. This work was supported in part by NNSFC, and UoA Flexible fund from Faculty of Engineering and Environment at Northumbria University.

or an e-commerce site, who intentionally present their opinions giving wrong information in anticipation of manipulating the behavior of the entire group [17]. While most of existing works on opinion evolution are built on the assumption that the network is situated in a benign environment possibly with a handful of stubborn individuals, the system and control community has been seeking intensively for resilient solutions to consensus problems in multi-agent systems against more realistic cyber-physical attacks in the past decade. One of such attacks initiated by Byzantine agents [18], [19], meaning that these agents may arbitrarily collude with others and send different values to different neighbors, is extremely harmful yet common in the setting of social networks. Based upon non-local information of the network, resilient coordination protocols have been designed in [20], [21] to relieve Byzantine attacks, where the misbehaving agents can be determined provided the communication network is sufficiently connected. A set of local filtering algorithms is put forward in [22], [23] to alleviate the influence of misbehaving agents, where each normal agent in the network discards the extremal values as compared to its own value. Resilient learning-based protocols are introduced in [24] to find optimal solutions to consensus problems in the presence of malicious attackers and uncertainties in system. By utilizing mobile detectors, it is shown in [25] that consensus can be maintained against Byzantine agents whose number is not restrained by the network connectedness. However, to our knowledge, only single state value is assumed in the prior works, which are thus unable to capture individuals in social networks who have private opinions deviating from the opinions they express.

A. Contributions

In this paper we aim to unfold the influence of private and expressed opinions as well as that of normal and misbehaving individuals on opinion evolution building on social dynamics and agent based modeling literature. This is achieved by model construction, convergence analysis, and clustering study, which we will detail below.

Firstly, this paper presents an agent based model where each agent, i.e. individual, in the network possesses both an expressed and a private opinion regarding a given topic. An agent's private opinion evolves under social influence from the expressed opinions of the agent's neighbors, while the agent determines its own expressed opinion under a pressure to conform to the group opinion in its neighborhood. The psychological rationale behind has been firstly revealed by the celebrated experiments on conformity by Asch [26] and is analytically validated by Ye et al. [9] in studying the evolution of discrepancy between expressed and private opinions. However, the model in [9] required the global network knowledge and did not offer resilience against misbehavior. The consensus was achieved by unpacking ergodicity and the row-stochastic matrix of the entire influence network. Our model, on the other hand, is purely distributed in the sense that only local information is needed, and the approach adopted is totally different.

Secondly, inspired by the Weighted-Mean Subsequence Reduced algorithm [22], [23], we propose distributed sifting algorithms, in which normal agents remove the most extreme expressed opinions in their neighborhood compared to their private opinions at each iteration. We analyze the resilience capabilities of these algorithms under a range of threats including malicious and Byzantine behaviors, and under the assumption of either the total number of misbehaving agents in the network or the number of them in the neighborhood of each normal agent being bounded by a fixed number R. Sufficient and necessary graph-theoretic conditions are provided to guarantee resilient consensus of expressed and private opinions when the underlying network is modeled as either a directed fixed network or a time-varying network. Moreover, our model allows that both the number and the identity of misbehaving agents are not made available to the normal agents, which is highly desirable in the real-world scenarios as such information is typically unavailable.

Finally, we design local strategies that provide resilience to malicious and Byzantine agents and meanwhile give rise to clustering and coexistence of expressed and private opinions instead of reaching a common value. Drawing on the methods of scaled consensus and formation generation, we characterize the sufficient and necessary conditions for opinion separation in terms of quotient, meaning that the agents' opinions approach dictated ratios in the asymptote, and in terms of difference, meaning that the agents' opinions reach assigned differences as time goes to infinity. Opinion clustering phenomenon has been observed extensively in many social networks and described by other opinion dynamics models [27]–[30] including the Deffuant-Weisbuch model, where multiple opinion clusters emerge with probability one if the confidence bound is below a critical value. The unique feature of expressed and private opinions, nevertheless, has been overlooked in these works.

We mention that the idea of accommodating additional source of opinion has been embodied in [31], where by adding some random edges an informal network is introduced to complement the formal network based on observations in social organization structure. The efficacy of informal network is numerically demonstrated to facilitate the consensus process over the formal network. Nevertheless, each individual essentially has a single opinion value.

B. Organization

Section 2 introduces mathematical preliminaries and formulates the problem. We provide convergence analysis for resilient consensus when malicious agents and Byzantine agents exist in the network in Section 3. Resilient opinion clustering and separation are investigated in Section 4. Simulations examples are presented in Section 5 and concluding remarks are drawn in Section 6.

II. Preliminaries and problem formulation

A. Graph theory

Denote by $\mathbb{N}$ the set of non-negative integers. The interaction between n agents in a social network is described by a digraph or directed graph, denoted by $G(t) = (V, E(t))$, where $V = \{v_1, \cdots, v_n\}$ represents the vertex set describing the agents of the network, and $E(t) \subseteq V \times V$ consists of the arcs or directed edges present at time $t \in \mathbb{N}$. We consider a partition of the vertex set, $V = N \cup M$, where the subset N contains the set of normal agents while the subset M contains the group of misbehaving ones, whose identities are not available a priori to any normal agent. The directed edge $(v_i, v_j) \in E(t)$ means that agent v_j has access to the information of agent v_i. The (in-)neighbors of agent v_i are defined by the set $\mathcal{N}_i(t) = \{v_j : (v_j, v_i) \in E(t)\}$. We set the extended neighborhood $\overline{\mathcal{N}}_i(t) = \{v_i\} \cup \mathcal{N}_i(t)$ to include the agent v_i itself. We will suppress the dependence on t for time-invariant networks in the aforementioned notations. In time-dependent networks, we also omit t for simplicity when the meaning is clear from the context.

The following notions of reachable sets and network robustness have been investigated intensively in [22], [23], [32]. Reachable sets and network robustness play a vital part in resilience control and they have an intimate tie with classic connectivity concept in graph theory.

Definition 1. (reachable sets) Let $R, Q \in \mathbb{N}$. A set $S \subseteq V$ is said to be R-reachable if there is an agent $v_i \in S$ satisfying the condition $|\mathcal{N}_i \backslash S| \geq R$, in which $|\cdot|$ represents the size of a set. Furthermore, S is said to be (R, Q)-reachable if $|\{v_i \in S : |\mathcal{N}_i \backslash S| \geq R\}| \geq Q$.

Definition 2. (network robustness) Let $R, Q \in \mathbb{N}$. If for any pair of nonempty disjoint subsets of V, at least one set of them is R-reachable, then the directed graph G is called R-robust. Moreover, G is said to be (R, Q)-robust if for any pair of nonempty, disjoint subsets $S_1, S_2 \subseteq V$, it follows that (a) $|\{v_i \in S_1 : |\mathcal{N}_i \backslash S_1| \geq R\}| = |S_1|$, or that (b) $|\{v_i \in S_2 : |\mathcal{N}_i \backslash S_2| \geq R\}| = |S_2|$, or that (c) $|\{v_i \in S_1 : |\mathcal{N}_i \backslash S_1| \geq R\}| + |\{v_j \in S_2 : |\mathcal{N}_j \backslash S_2| \geq R\}| \geq Q$.

From the definition, it is clear that R-reachability and $(R, 1)$-reachability are equivalent essentially. We will see that these robustness concepts are essential in characterizing the performance of local sifting algorithms.

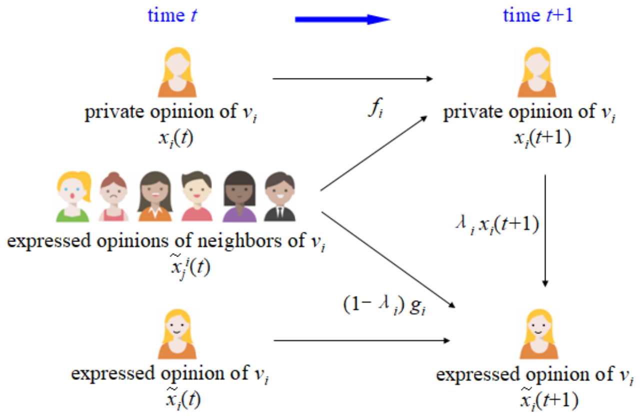

Fig. 1. Schematic illustration of opinion evolution for a normal agent v_i.

B. Opinion model

Suppose that a group of n agents, $\{v_1, \cdots, v_n\}$, form a directed social network $G = (V, E)$ admitting $V = N \cup M$, where as defined above N encapsulates all the normal agents while M is composed of misbehaving ones. Let $\mathbb{R}$ be the set of real numbers. On a given topic, the private and expressed opinions of agent $v_i \in V$ at time t is represented by $x_i(t) \in \mathbb{R}$ and $\tilde{x}_i(t) \in \mathbb{R}$, respectively. We aim to address the following resilient opinion consensus problem in the presence of misbehaving agents.

Definition 3. (resilient consensus) In the network G, the normal agents in N are said to achieve resilient opinion consensus in the presence of misbehaving agents in M if $\lim_{t\to\infty} x_i(t) - x_j(t) = 0$ and $\lim_{t\to\infty} \tilde{x}_i(t) - \tilde{x}_j(t) = 0$ for all $v_i, v_j \in N$ and all initial conditions $\{x_i(0)\}_{i=1}^n$ and $\{\tilde{x}_i(0)\}_{i=1}^n$.

We regard the private opinion $x_i(t)$ as the agent's true opinion, while the expressed opinion $\tilde{x}_i(t)$ can be different from its true opinion due to varied reasons such as political correctness and peer pressure. Opinion consensus here requires both expressed and private opinions to reach a consensus over the entire network (ideally without gap between private and expressed opinions; c.f. Remark 3). The dynamical model of each normal agent $v_i \in N$ is described as

$$x_i(t+1) = f_i\left(x_i(t), \{\tilde{x}_j^i(t) : v_j \in \mathcal{N}_i(t)\}\right) \tag{1}$$

and

$$\begin{aligned}\tilde{x}_i(t) =& \lambda_i x_i(t) \\ &+ (1-\lambda_i) g_i\left(\{\tilde{x}_j^i(t-1) : v_j \in \overline{\mathcal{N}}_i(t-1)\}\right),\end{aligned} \tag{2}$$

where $\tilde{x}_j^i(t) \in \mathbb{R}$ is the opinion value communicated to agent v_i from agent v_j at time step t, and $\tilde{x}_j^i(t) = \tilde{x}_j(t)$ for all $v_j \in N$, meaning normal agents always send their real expressed opinions to their neighbors. We also assume $\tilde{x}_j^j(t) = \tilde{x}_j(t)$ for $v_j \in N$. Misbehaving agents, on the other hand, may send arbitrary values to their neighbors. The function f_i to be designed later describes the influence on the private opinion of v_i at time $t+1$ from its own private opinion and its neighbors' expressed opinions at time t. The parameter $\lambda_i \in [0, 1]$ characterizes the resilience to pressure to conform to its local environment encoded by g_i. Agent v_i is maximally resilient if $\lambda_i = 1$, and minimally resilient if $\lambda_i = 0$. As is common in agent based consensus problems [1], both functions f_i and g_i will be designed as some weighted average functions (see (3) and (4) below). It is natural (but not necessary for our theorems below) that we assume that $x_i(0) = \tilde{x}_i(0)$ for all $1 \leq i \leq n$, meaning that the initial expressed opinions are equivalent to initial private opinions for all agents in the network. Obviously, the expressed and private opinions coincide for each agent when $\lambda_i = 1$ for all $i = 1, \cdots, n$. The influences that act to change agent v_i's private and expressed opinions are illustrated in Fig. 1.

In (1) and (2), f_i and g_i delineate the update functions for normal agent v_i. These functions will be instantiated later so that the normal agents can reach the group's goal resisting the compromise of misbehaving agents, whose number and identity are not available to the normal agents. Misbehaving agents, on the other hand, can exert arbitrary strategies and different rules which are beyond the reach of the normal agents. Specifically, we will examine two kinds of misbehaving agents, namely, malicious and Byzantine agents in this paper.

Definition 4. (malicious agent) We call an agent $v_i \in M$ malicious if it exerts some distinct communication rule $\hat{g}_i$ in (2) at some time $t \in \mathbb{N}$.

Definition 5. (Byzantine agent) We call an agent $v_i \in M$ Byzantine if it exerts some distinct communication rule $\hat{g}_i$ in (2) at some time $t \in \mathbb{N}$, or it does not communicate the same opinion $\tilde{x}_i^j$ to all of its out-neighbors v_j, i.e., $v_i \in \mathcal{N}_j$, at some time $t \in \mathbb{N}$.

A malicious agent is misbehaving due to, for example, stubborn traits, and it sends information in a broadcast manner [8], [16]. Byzantine agents on the other hand are often thought of as one of the most dangerous attackers [18], [23], [25], who typically possess a thorough intelligence of the entire network and hence can potentially collude with other Byzantine agents to manipulate the network by sending wrong information in a point-to-point manner. By the above definitions, both malicious and Byzantine agents can renew their opinions in an arbitrary way at every time step, and hence all malicious agents turn out to be Byzantine, but not vice versa. It is worth noting that we do not care private opinions of misbehaving agents, i.e., not implicate f_i in (1), in the Definitions 4 and 5 because misbehaving agents influence neighbors only via their expressed opinions.

We will investigate two types of opinion consensus models according to the number and location of the misbehaving agents. The first one is called R-globally bounded model. In this model, the total number of misbehaving agents in the set M is upper bounded by a constant $R \in \mathbb{N}$. Another model is R-locally bounded model, where $|\mathcal{N}_i \cap M| \leq R$ for every $v_i \in N$. Every normal agent has at most R misbehaving neighbors in the R-locally bounded model. In both models, misbehaving agents, be they malicious or Byzantine, pose a treat to the group decision making through preventing other agents from reaching common opinions or driving their opinions into a biased or even detrimental situation. Therefore, it is desirable to exercise caution and adopt resilient consensus strategies.

C. Resilient consensus strategy

Base upon nearest-neighbor interaction, we here adopt the distributed local sifting algorithms for the expressed and private opinions of each normal agent $v_i \in N$. As a misbehaving agent may affect both x_i and $\tilde{x}_i$ through f_i and g_i, respectively, at every time step, our strategy essentially goes beyond the Weighted-Mean Subsequence Reduced algorithms detailed in, e.g., [18], [22], [23].

Our sifting algorithm can be performed in three steps, executed synchronously for all agents at each time step $t \in \mathbb{N}$. Fix $R \in \mathbb{N}$. First, each normal agent $v_i \in N$ collects the expressed opinions $\{\tilde{x}_j^i(t)\}$ from its neighbors, and creates an ordered list array for $\{\tilde{x}_j^i(t)\}_{v_j \in \mathcal{N}_i}$ arranging from largest to smallest. Second, the largest R opinions that are strictly greater than $x_i(t)$ in the above array are deleted (if there are fewer than R greater opinions than $x_i(t)$, all of those opinions are discarded). The similar sifting process is exerted to the smaller opinions. The set of agents that are removed by agent v_i at time t is signified by a set $\mathcal{R}_i(t)$. Third, each $v_i \in N$ updates its opinion using the following $f_i(\cdot)$ and $g_i(\cdot)$, respectively, in (1) and (2):

$$\begin{aligned} x_i(t+1) =& a_{ii}(t)x_i(t) \\ &+ \sum_{v_j \in \mathcal{N}_i(t)\backslash\mathcal{R}_i(t)} a_{ij}(t)\tilde{x}_j^i(t), \quad t \in \mathbb{N} \end{aligned} \tag{3}$$

and

$$\begin{aligned} \tilde{x}_i(t) =& \lambda_i x_i(t) + (1-\lambda_i) \\ &\cdot \sum_{v_j \in \overline{\mathcal{N}}_i(t-1)\backslash\mathcal{R}_i(t-1)} b_{ij}(t-1)\tilde{x}_j^i(t-1), \\ &t \in \mathbb{N}\backslash\{0\}, \end{aligned} \tag{4}$$

where $\{a_{ij}(t)\}$ are the weights instantiating f_i satisfying the following three conditions for every $t \in \mathbb{N}$: (Af) $a_{ij}(t) = 0$ if $v_j \notin \overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, (Bf) there is a constant number $\alpha \in (0,1)$ independent of t, such that $a_{ij}(t) \geq \alpha$ for any $v_j \in \overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, and (Cf) $\sum_{v_j \in \overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)} a_{ij}(t) = 1$; and similarly $\{b_{ij}(t)\}$ are the weights instantiating g_i satisfying the following three conditions for every $t \in \mathbb{N}$: (Ag) $b_{ij}(t) = 0$ if $v_j \notin \overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, (Bg) there is a constant number $\beta \in (0,1)$ independent of t, such that $b_{ij}(t) \geq \beta$ for any $v_j \in \overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, and (Cg) $\sum_{v_j \in \overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)} b_{ij}(t) = 1$. Moreover, we assume $\lambda_i \in (0,1]$, meaning naturally that the private opinion of agent i has an influence on its expressed opinion. Since there are finite agents within the network G, we obtain a constant $\lambda > 0$ satisfying $\lambda_i \geq \lambda > 0$ for every $v_i \in N$.

Remark 1. The time-shift $t-1$ in the summation term of (4) is necessary since otherwise both sides of the equation rely on $\tilde{x}_i(t)$, leading to an inconsistent equation. Given $x_i(0)$ and $\tilde{x}_i(0)$ for all $v_i \in V$, at each time step $t = 0$, (3) comes into effect; at subsequent time step $t \geq 1$, (4) comes into effect followed by (3). The strategy is consistent with the description in Fig. 1.

Remark 2. Note that the weights a_{ij} and b_{ij} in (3) and (4) can be arbitrarily chosen provided the corresponding conditions hold. A typical choice could be $a_{ij}(t) = (|\mathcal{N}_i(t)| + 1 - |\mathcal{R}_i(t)|)^{-1}$ and $b_{ij}(t-1) = (|\overline{\mathcal{N}}_i(t-1)| - |\mathcal{R}_i(t-1)|)^{-1}$ so that the weights for all neighbors are equal. In this case, the expressed opinion $\tilde{x}_i$ takes a similar form as in [9], but the global average of all agents in the network is used in [9] (which is essential for the validity of the analysis therein) rather than the adaptive local average adopted here.

The above algorithm has low complexity while relatively accurate to capture relevant social phenomena with minimum parameters. It is purely distributed and only local information available to each agent is used. No prior awareness of the identity of misbehaving agents or the architecture of network is assumed available to normal agents. Moreover, our algorithm will handle the situation where the roles of normal agents and misbehaving agents change. When a normal agent misbehaves at some point, it may apply a strategy freely deviated from the sifting strategy; if a misbehaving agent becomes normal, it will then pick up the sifting strategy. See Example 2 in Section V for an illustration.

In the rest of the paper, we will refer to the above algorithm as the opinion sifting strategy with parameter R or simply R-sifting strategy. A flowchart and the complexity analysis is provided in Supplementary Material.

III. Resilience against misbehaving agents

In this section we study resilient opinion consensus problem ad present the convergence analysis for the opinion dynamics model in the presence of both malicious and Byzantine agents. In each case, we provide resilience results for both globally bounded threats and locally bounded threats. To begin with, define $\Phi(t) := \max_{v_i \in N} x_i(t)$ and $\phi(t) := \min_{v_i \in N} x_i(t)$ respectively as the maximum and minimum private opinions for normal agents. Similarly, let $\tilde{\Phi}(t) := \max_{v_i \in N} \tilde{x}_i(t)$ and $\tilde{\phi}(t) := \min_{v_i \in N} \tilde{x}_i(t)$ be the maximum and minimum expressed opinions, respectively, for normal agents. Moreover, we set $\Phi^*(t) = \max\{\Phi(t), \tilde{\Phi}(t)\}$ and $\phi^*(t) = \min\{\phi(t), \tilde{\phi}(t)\}$. It is not difficult to see that for all $v_i \in N$ and $t \in \mathbb{N}$, $x_i(t+1) \in [\phi^*(t), \Phi^*(t)]$ and $\tilde{x}_i(t+1) \in [\phi^*(t), \Phi^*(t)]$ hold in both R-globally and R-locally bounded models with malicious/Byzantine agents because opinions in the update rules (3) and (4) are convex combinations of values within the interval $[\phi^*(t), \Phi^*(t)]$. This indicates that the domain $[\phi^*(0), \Phi^*(0)]$ is an invariant set meaning that the expressed and private opinions of normal agents will stay in this interval for all time t.

A. Convergence analysis with malicious agents

In the sequel, we establish some necessary and sufficient criteria for opinion consensus in the globally and locally bounded models with malicious agents. We first deal with time-invariant networks $G = (V, E)$. The results for time-varying networks $G(t) = (V, E(t))$ are addressed as a corollary.

Theorem 1. (opinion consensus in R-globally bounded model with malicious agents) *Suppose that we have a time-invariant network characterized by a directed graph $G = (V, E)$, where each normal agent updates its private and expressed opinions according to the opinion R-sifting strategy. Then, in the R-globally bounded model having malicious agents, resilient opinion consensus can be reached if and only if G is $(R+1, R+1)$-robust.*

Proof. (Necessity) Suppose on the contrary that G is not an $(R+1, R+1)$-robust network. Therefore, there exist disjoint nonempty sets $S_1, S_2 \subseteq V$ such that none of the criteria (a)-(c) in Definition 2 hold. Define $X_{S_l}^{R+1} = \{v_i \in S_l : |N_i \backslash S_l| \geq R+1\}$ for $l = 1, 2$. Fix $c_1 < c_2$. Let $x_i(0) = \tilde{x}_i(0) = c_1$ for any agent $v_i \in S_1$, and $x_i(0) = \tilde{x}_i(0) = c_2$ for any agent $v_i \in S_2$. For all the other agents v_i in the network, set $x_i(0) = \tilde{x}_i(0) \in (c_1, c_2)$.

Since $|X_{S_1}^{R+1}| + |X_{S_2}^{R+1}| \leq R$, we assume that all agents in the sets $X_{S_1}^{R+1}$ and $X_{S_2}^{R+1}$ are malicious and that they tend to preserve their expressed opinions unchanged. There is at least one normal agent in both S_1 and S_2 respectively (because $|X_{S_1}^{R+1}| < |S_1|$ and $|X_{S_2}^{R+1}| < |S_2|$), say, $v_1 \in S_1 \cap N$ and $v_2 \in S_2 \cap N$. Since v_1 has at most R neighbors not in S_1, we have $x_1(t+1) = x_1(t) = c_1$ for $t \in \mathbb{N}$ invoking the opinion R-sifting strategy. Furthermore, $\tilde{x}_1(t) = \lambda_1 x_1(t) + (1-\lambda_1)c_1 = c_1$ for all t. Likewise, we obtain $x_2(t) = \tilde{x}_2(t) = c_2$ for all t. Thus, consensus cannot be achieved among normal agents in G. The necessity is proved.

(Sufficiency) In light of the comments at the outset of this section, we may assume that $\rho_{\Phi^*} := \lim_{t\to\infty} \Phi^*(t) \geq \rho_{\phi^*} := \lim_{t\to\infty} \phi^*(t)$, since both $\Phi^*(t)$ and $\phi^*(t)$ are monotone and bounded with respect to t. If $\rho_{\Phi^*} = \rho_{\phi^*}$, then both expressed and private opinions will reach consensus eventually. In what follows, we assume that $\rho_{\Phi^*} > \rho_{\phi^*}$ and show that this inequality does not hold by the method of contradiction.

Take $\varepsilon_0 > 0$ satisfying $\rho_{\Phi^*} - \varepsilon_0 > \rho_{\phi^*} + \varepsilon_0$. For $t \in \mathbb{N}$ and any $\varepsilon_i > 0$, we will need the following definitions of four sets $X_\Phi(t, \varepsilon_i) := \{v_i \in V : x_i(t) > \rho_{\Phi^*} - \varepsilon_i\}$, $X_{\tilde{\Phi}}(t, \varepsilon_i) := \{v_i \in V : \tilde{x}_i(t) > \rho_{\Phi^*} - \varepsilon_i\}$, $X_\phi(t, \varepsilon_i) := \{v_i \in V : x_i(t) < \rho_{\phi^*} + \varepsilon_i\}$, and $X_{\tilde{\phi}}(t, \varepsilon_i) := \{v_i \in V : \tilde{x}_i(t) < \rho_{\phi^*} + \varepsilon_i\}$. These sets may contain both normal and malicious agents. By the definition of ε_0, $X_\Phi(t, \varepsilon_0) \cap X_\phi(t, \varepsilon_0) = \emptyset$ and $X_{\tilde{\Phi}}(t, \varepsilon_0) \cap X_{\tilde{\phi}}(t, \varepsilon_0) = \emptyset$.

Choose $\varepsilon \in (0, \varepsilon_0)$ such that

$$\varepsilon < \min\left\{ \frac{\lambda^{|N|}\alpha^{|N|}\varepsilon_0}{1-\lambda^{|N|}\alpha^{|N|}}, \frac{(1+\lambda)^{|N|}\beta^{|N|}(1-\beta(1+\lambda))\varepsilon_0}{1-\beta^{|N|}(1+\lambda)^{|N|}} \right\}. \quad (5)$$

Recall in (Bg) the lower bound β can be made arbitrarily small. Let t_ε be the time step satisfying $\Phi^*(t) < \rho_{\Phi^*} + \varepsilon$ and $\phi^*(t) > \rho_{\phi^*} - \varepsilon$ for all $t \geq t_\varepsilon$. Next, we examine the the pairs of nonempty disjoint sets $X_\Phi(t_\varepsilon, \varepsilon_0)$ and $X_\phi(t_\varepsilon, \varepsilon_0)$, and $X_{\tilde{\Phi}}(t_\varepsilon, \varepsilon_0)$ and $X_{\tilde{\phi}}(t_\varepsilon, \varepsilon_0)$ separately.

(I) $X_\Phi(t_\varepsilon, \varepsilon_0)$ and $X_\phi(t_\varepsilon, \varepsilon_0)$. Notice that the network G is $(R+1, R+1)$-robust with at most R malicious agents. There must be a normal agent in the union $X_\Phi(t_\epsilon, \varepsilon_0) \cup X_\phi(t_\epsilon, \varepsilon_0)$, which has at least $R+1$ neighbors not in its set. Without loss of generality (W.l.o.g.), assume that $v_i \in X_\Phi(t_\varepsilon, \varepsilon_0) \cap N$ has at least $R+1$ neighbors not in $X_\Phi(t_\varepsilon, \varepsilon_0)$. Since these neighbors' private opinions are less than or equal to $\rho_{\Phi^*} - \varepsilon_0$ and at least one of these opinions (say, that of v_j) will be used by v_i, we obtain by using (4),

$$\begin{aligned} \tilde{x}_i(t_\varepsilon+1) &\leq \lambda_j(\rho_{\Phi^*} - \varepsilon_0) + (1-\lambda_j)\tilde{\Phi}(t_\varepsilon) \\ &\leq \lambda_j(\rho_{\Phi^*} - \varepsilon_0) + (1-\lambda_j)(\rho_{\Phi^*} + \varepsilon) \\ &\leq \rho_{\Phi^*} - \lambda_j\varepsilon_0 + \varepsilon(1-\lambda_j). \end{aligned} \quad (6)$$

Using (3), (6), and noting that $x_i(t_\varepsilon)$ and the expressed opinions of v_i's neighbors that can be used at time step t_ε are up bounded by $\Phi^*(t_\varepsilon)$, we have

$$\begin{aligned} x_i(t_\varepsilon+1) &\leq (1-\alpha)\Phi^*(t_\varepsilon) + \alpha(\rho_{\Phi^*} - \lambda_j\varepsilon_0 + \varepsilon(1-\lambda_j)) \\ &\leq (1-\alpha)(\rho_{\Phi^*} + \varepsilon) + \alpha(\rho_{\Phi^*} - \lambda_j\varepsilon_0 + \varepsilon(1-\lambda_j)) \\ &\leq \rho_{\Phi^*} - \alpha\lambda_j\varepsilon_0 + \varepsilon(1-\alpha\lambda_j) \\ &\leq \rho_{\Phi^*} - \alpha\lambda\varepsilon_0 + \varepsilon(1-\alpha\lambda). \end{aligned} \quad (7)$$

The expression (7) is also valid for the private opinion of any normal agent not in $X_\Phi(t_\varepsilon, \varepsilon_0)$ since such agent will adopt its own private opinion $x_i(t_\varepsilon)$, and $x_i(t_\varepsilon) \le \rho_{\Phi^*} - \varepsilon_0$. Hence,

$$\begin{aligned} x_i(t_\varepsilon+1) &\le (1-\alpha)\Phi^*(t_\varepsilon) + \alpha(\rho_{\Phi^*} - \varepsilon_0) \\ &\le (1-\alpha)(\rho_{\Phi^*} + \varepsilon) + \alpha(\rho_{\Phi^*} - \varepsilon_0) \\ &\le \rho_{\Phi^*} - \alpha\varepsilon_0 + \varepsilon(1-\alpha) \\ &\le \rho_{\Phi^*} - \alpha\lambda\varepsilon_0 + \varepsilon(1-\alpha\lambda), \end{aligned} \tag{8}$$

where in the last inequality we used $\lambda < 1$. Analogously, if $v_i \in X_\phi(t_\varepsilon, \varepsilon_0) \cap N$ which has at least $R+1$ neighbors not in $X_\phi(t_\varepsilon, \varepsilon_0)$, we derive a parallel inequality

$$x_i(t_\varepsilon+1) \ge \rho_{\phi^*} + \alpha\lambda\varepsilon_0 - \varepsilon(1-\alpha\lambda), \tag{9}$$

which is also valid for the normal agents outside the set $X_\phi(t_\varepsilon, \varepsilon_0)$ similarly.

Define $\varepsilon_1 = \alpha\lambda\varepsilon_0 - \varepsilon(1-\alpha\lambda)$ such that $0 < \varepsilon < \varepsilon_1 < \varepsilon_0$. Recall that $X_\Phi(t_\varepsilon+1, \varepsilon_1)$ and $X_\phi(t_\varepsilon+1, \varepsilon_1)$ are disjoint sets. Since at least one of the normal agents in $X_\Phi(t_\varepsilon, \varepsilon_0)$ decrements its private opinion to $\rho_{\Phi^*} - \varepsilon_1$ or lower, or at least one of the normal agents in $X_\phi(t_\varepsilon, \varepsilon_0)$ increments its private opinion to $\rho_{\phi^*} + \varepsilon_1$ or higher, the above comments indicate that $|X_\Phi(t_\varepsilon+1, \varepsilon_1) \cap N| < |X_\Phi(t_\varepsilon, \varepsilon_0) \cap N|$ or $|X_\phi(t_\varepsilon+1, \varepsilon_1) \cap N| < |X_\phi(t_\varepsilon, \varepsilon_0) \cap N|$ holds. We now recursively define $\varepsilon_j = \alpha\lambda\varepsilon_{j-1} - (1-\alpha\lambda)\varepsilon$ for each $j \ge 1$ and notice that $\varepsilon_j < \varepsilon_{j-1}$. The above derivations are applicable to every time step $t_\varepsilon + j$ as long as there still exist normal agents in $X_\Phi(t_\varepsilon+j, \varepsilon_j)$ and $X_\phi(t_\varepsilon+j, \varepsilon_j)$. Since there are $|N|$ normal agents in the whole network G, there exists some $T \le |N|$ such that either $X_\Phi(t_\varepsilon+T, \varepsilon_T) \cap N$ or $X_\phi(t_\varepsilon+T, \varepsilon_T) \cap N$ is an empty set. If the former case is true, the private opinions of normal agents at step $t_\varepsilon + T$ are no greater than $\rho_{\Phi^*} - \varepsilon_T$; if the latter case is true, the private opinions of normal agents at step $t_\varepsilon + T$ are no less than $\rho_{\phi^*} + \varepsilon_T$. Therefore, $\varepsilon_T = \alpha\lambda\varepsilon_{T-1} - \varepsilon(1-\alpha\lambda) = \alpha^T\lambda^T\varepsilon_0 - \varepsilon(1-\alpha^T\lambda^T) \ge \alpha^{|N|}\lambda^{|N|}\varepsilon_0 - \varepsilon(1-\alpha^{|N|}\lambda^{|N|}) > 0$ according to the choice of ε in (5).

(II) $X_{\tilde\Phi}(t_\varepsilon, \varepsilon_0)$ and $X_{\tilde\phi}(t_\varepsilon, \varepsilon_0)$. Since the network G is $(R+1, R+1)$-robust with no more than R malicious agents, there is a normal agent in the union $X_{\tilde\Phi}(t_\epsilon, \varepsilon_0) \cup X_{\tilde\phi}(t_\epsilon, \varepsilon_0)$, which has at least $R+1$ neighbors not in its set. W.l.o.g., we assume that $v_i \in X_{\tilde\Phi}(t_\varepsilon, \varepsilon_0) \cap N$ has at least $R+1$ neighbors not in $X_{\tilde\Phi}(t_\varepsilon, \varepsilon_0)$. Since these neighbors' expressed opinions do not exceed $\rho_{\Phi^*} - \varepsilon_0$ and at least one of these opinions will be used by v_i, we obtain by using (4),

$$\begin{aligned} \tilde{x}_i(t_\varepsilon+1) \le& \lambda_i(\rho_{\Phi^*} + \varepsilon) \\ &+ (1-\lambda_i)((1-\beta)(\rho_{\Phi^*} + \varepsilon) + \beta(\rho_{\Phi^*} - \varepsilon_0)) \\ \le& \rho_{\Phi^*} - \varepsilon_0\beta(1+\lambda_i) + \varepsilon(1-\beta+\lambda_i\beta) \\ \le& \rho_{\Phi^*} - \varepsilon_0\beta(1+\lambda) + \varepsilon, \end{aligned} \tag{10}$$

where we have used the condition $\lambda \le \lambda_i \le 1$ and noted that each normal agent's expressed opinion is characterized as a convex combination of the expressed opinions of its neighbors having coefficients no less than β. The expression (10) is also applicable to the expressed opinion of any normal agent not in $X_{\tilde\Phi}(t_\varepsilon, \varepsilon_0)$ as such agent will adopt its own expressed opinion $\tilde{x}_i(t_\varepsilon)$, and $\tilde{x}_i(t_\varepsilon) \le \rho_{\Phi^*} - \varepsilon_0$. Analogously, if $v_i \in X_{\tilde\phi}(t_\varepsilon, \varepsilon_0) \cap N$ which has at least $R+1$ neighbors not in $X_{\tilde\phi}(t_\varepsilon, \varepsilon_0)$, we arrive at a parallel inequality

$$\tilde{x}_i(t_\varepsilon+1) \ge \rho_{\phi^*} + \varepsilon_0\beta(1+\lambda) - \varepsilon, \tag{11}$$

which is also applicable to the normal agents not in the set $X_{\tilde\phi}(t_\varepsilon, \varepsilon_0)$ similarly.

Define $\tilde\varepsilon_1 = \varepsilon_0\beta(1+\lambda) - \varepsilon$ such that $0 < \varepsilon < \tilde\varepsilon_1 < \varepsilon_0 := \tilde\varepsilon_0$. Recall that $X_{\tilde\Phi}(t_\varepsilon+1, \tilde\varepsilon_1)$ and $X_{\tilde\phi}(t_\varepsilon+1, \tilde\varepsilon_1)$ are disjoint sets. Since at least one of the normal agents in $X_{\tilde\Phi}(t_\varepsilon, \tilde\varepsilon_0)$ decrements its private opinion to $\rho_{\Phi^*} - \tilde\varepsilon_1$ or lower, or at least one of the normal agents in $X_{\tilde\phi}(t_\varepsilon, \tilde\varepsilon_0)$ increments its private opinion to $\rho_{\phi^*} + \tilde{\varepsilon_1}$ or higher, the above comments indicate that $|X_{\tilde\Phi}(t_\varepsilon+1, \tilde\varepsilon_1) \cap N| < |X_{\tilde\Phi}(t_\varepsilon, \tilde\varepsilon_0) \cap N|$ or $|X_{\tilde\phi}(t_\varepsilon+1, \tilde\varepsilon_1) \cap N| < |X_{\tilde\phi}(t_\varepsilon, \tilde\varepsilon_0) \cap N|$ holds. We recursively define $\tilde\varepsilon_j = \beta(1+\lambda)\tilde\varepsilon_{j-1} - \varepsilon$ for each $j \ge 1$ and notice that $\tilde\varepsilon_j < \tilde\varepsilon_{j-1}$. This derivation is also applicable to every time step $t_\varepsilon + j$ provided there still exist normal agents in $X_{\tilde\Phi}(t_\varepsilon+j, \tilde\varepsilon_j)$ and $X_{\tilde\phi}(t_\varepsilon+j, \tilde\varepsilon_j)$. Since there are $|N|$ normal agents in the whole network G, there exists some $\tilde{T} \le |N|$ such that either $X_{\tilde\Phi}(t_\varepsilon+\tilde{T}, \tilde\varepsilon_{\tilde{T}}) \cap N$ or $X_{\tilde\phi}(t_\varepsilon+\tilde{T}, \tilde\varepsilon_{\tilde{T}}) \cap N$ becomes empty. If the former case is true, the expressed opinions of normal agents at step $t_\varepsilon + \tilde{T}$ are no greater than $\rho_{\Phi^*} - \tilde\varepsilon_{\tilde{T}}$; if the latter case is true, the expressed opinions of normal agents at step $t_\varepsilon + \tilde{T}$ are no less than $\rho_{\phi^*} + \tilde\varepsilon_{\tilde{T}}$. Therefore, we have

$$\begin{aligned} \tilde\varepsilon_{\tilde{T}} &= \beta(1+\lambda)\tilde\varepsilon_{\tilde{T}-1} - \varepsilon \\ &= \beta^{\tilde{T}}(1+\lambda)^{\tilde{T}}\tilde\varepsilon_0 - \varepsilon\frac{1-\beta^{\tilde{T}}(1+\lambda)^{\tilde{T}}}{1-\beta(1+\lambda)} \\ &\ge \beta^{|N|}(1+\lambda)^{|N|}\tilde\varepsilon_0 - \varepsilon\frac{1-\beta^{|N|}(1+\lambda)^{|N|}}{1-\beta(1+\lambda)} > 0 \end{aligned} \tag{12}$$

by the choice of ε in (5) and taking $\beta(1+\lambda) < 1$ (since β can be made arbitrarily small).

Combining the above comments, any normal agent v_i in the network at time step $t_\varepsilon + \max\{T, \tilde{T}\}$ has opinion $\max\{x_i(t), \tilde{x}_i(t)\}$ either at most $\rho_{\Phi^*} - \min\{\varepsilon_T, \tilde\varepsilon_{\tilde{T}}\}$ or no less than $\rho_{\phi^*} + \min\{\varepsilon_T, \tilde\varepsilon_{\tilde{T}}\}$, leading to contradiction with the definitions of the limits ρ_{Φ^*} and ρ_{ϕ^*}. □

Remark 3. We have shown in the sufficiency of Theorem 1 that, if G is $(R+1, R+1)$-robust, both the private and expressed opinions of normal agents in the network G converge to the same limit, which is stronger than what is required in Definition 3. This means the discrepancy between expressed and private will ultimately vanish, avoiding the harmful "spiral of silence" phenomenon observed in some social networks [33]. Persistence to the individuals' initial opinions is found to be a possible cause of non-vanishing discrepancy between expressed and private opinions in [9] and even contributes to shaping the final opinion configuration [34]. In our framework, such "stubbornness" to the initial opinions has been considered as a malicious behavior and is aimed to be overcome. Therefore, we are able to show the strong result of vanishing discrepancy, which interestingly agrees with recent report on Twitter [35].

Note that the set $\mathcal{R}_i(t)$ in (3) and (4) is time-varying. Hence, the network structure G in Theorem 1 is no longer fixed essentially. Moreover, for time-dependent communication topologies, we have the corollary stated as follows. The proof is given in Supplementary Material.

Corollary 1. *Suppose that we have a time-varying network characterized by a directed graph $G(t) = (V, E(t))$, where each normal agent updates its private and expressed opinions according to the opinion sifting strategy with parameter R. Signified by $\{t_k\}$ the time steps in which $G(t)$ is $(R+1, R+1)$-robust. Then, in the R-globally bounded model having malicious agents, resilient opinion consensus can be reached if $|\{t_k\}| = \infty$ and there is a constant $\tau \in \mathbb{N}$ such that $|t_{k+1} - t_k| \leq \tau$ for all k.*

In the case of locally bounded models, in which misbehaving agents are much more prevalent but the number of them are still bounded in each normal agent's neighborhood, an amenable way to deal with malicious agents is to characterize the network structure by the notion of R-robustness.

Theorem 2. (opinion consensus in R-locally bounded model with malicious agents) *Suppose that we have a time-invariant network characterized by a directed graph $G = (V, E)$, where each normal agent updates its private and expressed opinions according to the opinion R-sifting strategy. Then, in the R-locally bounded model having malicious agents, resilient opinion consensus can be reached when G is $2R+1$-robust. Moreover, G is $R+1$-robust when resilient opinion consensus in the R-locally bounded model with malicious agents can be reached.*

Proof. (Necessity) Assume to the contrary that G is not $R+1$-robust. There exist disjoint nonempty sets $S_1, S_2 \subseteq V$ such that each agent in these two sets has at most R neighbors not in the set. Suppose that there exist normal agents in both S_1 and S_2. Fix $c_1 < c_2$. Let $x_i(0) = \tilde{x}_i(0) = c_1$ for any agent $v_i \in S_1$, and $x_i(0) = \tilde{x}_i(0) = c_2$ for any agent $v_i \in S_2$. For all the other agents, set $x_i(0) = \tilde{x}_i(0) \in (c_1, c_2)$. It is obvious that the expressed or private opinions of agents in S_1 and S_2 will not achieve consensus under the resilient opinion R-sifting strategy as they never refer to any opinions not in their own sets. Thus, resilient consensus among normal agents in G cannot be reached.

(Sufficiency) We proceed in a similar line as in Theorem 1. Suppose that $\rho_{\Phi^*} := \lim_{t\to\infty} \Phi^*(t)$ and $\rho_{\phi^*} := \lim_{t\to\infty} \phi^*(t)$. In the sequel, we will prove $\rho_{\Phi^*} = \rho_{\phi^*}$ by contradiction. For this purpose, suppose that $\rho_{\Phi^*} > \rho_{\phi^*}$. Choose $\varepsilon_0 > 0$ so that $\rho_{\Phi^*} - \varepsilon_0 > \rho_{\phi^*} + \varepsilon_0$. For $t \in \mathbb{N}$ and $\varepsilon_i > 0$, we consider four sets defined by $X_\Phi(t, \varepsilon_i) := \{v_i \in V : x_i(t) > \rho_{\Phi^*} - \varepsilon_i\}$, $X_{\tilde{\Phi}}(t, \varepsilon_i) := \{v_i \in V : \tilde{x}_i(t) > \rho_{\Phi^*} - \varepsilon_i\}$, $X_\phi(t, \varepsilon_i) := \{v_i \in V : x_i(t) < \rho_{\phi^*} + \varepsilon_i\}$, and $X_{\tilde{\phi}}(t, \varepsilon_i) := \{v_i \in V : \tilde{x}_i(t) < \rho_{\phi^*} + \varepsilon_i\}$. By the definition of ε_0, $X_\Phi(t, \varepsilon_0) \cap X_\phi(t, \varepsilon_0) = \emptyset$ and $X_{\tilde{\Phi}}(t, \varepsilon_0) \cap X_{\tilde{\phi}}(t, \varepsilon_0) = \emptyset$. Set $\varepsilon \in (0, \varepsilon_0)$ such that

$$\varepsilon < \left\{ \frac{\lambda^{|N|}\alpha^{|N|}\varepsilon_0}{1 - \lambda^{|N|}\alpha^{|N|}}, \frac{(1+\lambda)^{|N|}\beta^{|N|}(1 - \beta(1+\lambda))\varepsilon_0}{1 - \beta^{|N|}(1+\lambda)^{|N|}} \right\}. \tag{13}$$

Let t_ε be the time step satisfying $\Phi^*(t) < \rho_{\Phi^*} + \varepsilon$ and $\phi^*(t) > \rho_{\phi^*} - \varepsilon$ for any $t \geq t_\varepsilon$.

Recall that the pair of sets $X_\Phi(t_\varepsilon, \varepsilon_0) \cap N$ and $X_\phi(t_\varepsilon, \varepsilon_0) \cap N$ is nonempty and disjoint. Since G is $2R+1$-robust, at least one of this couple of sets is $2R+1$-reachable. We suppose, w.l.o.g., that $X_\Phi(t_\varepsilon, \varepsilon_0) \cap N$ is $2R+1$-reachable, and hence there is an agent $v_i \in X_\Phi(t_\varepsilon, \varepsilon_0) \cap N$ having no less than $2R+1$ neighboring agents not in its set. As there are no more than R malicious agents within $\mathcal{N}_i$, v_i will refer to no less than one of its normal neighbors' private opinions not in $X_\Phi(t_\varepsilon, \varepsilon_0) \cap N$ under the resilient opinion R-sifting strategy. Accordingly, proceeding as in the proof of Theorem 1, we arrive at $x_i(t_\varepsilon + 1) \leq \rho_{\Phi^*} - \alpha\lambda\varepsilon_0 + \varepsilon(1 - \alpha\lambda)$. This remains valid for the renewed private opinion of each normal agent not in $X_\Phi(t_\varepsilon, \varepsilon_0) \cap N$ as such an agent adopts its own private opinion in the renewal procedure. Analogously, if $v_i \in X_\phi(t_\varepsilon, \varepsilon_0) \cap N$ has more than or equal to $2R+1$ neighbors not in its set, we have a similar inequality $x_i(t_\varepsilon + 1) \geq \rho_{\phi^*} + \alpha\lambda\varepsilon_0 - \varepsilon(1 - \alpha\lambda)$, which is also applicable to the normal agents not in $X_\phi(t_\varepsilon, \varepsilon_0) \cap N$. Now by defining $\varepsilon_1 = \alpha\lambda\varepsilon_0 - (1 - \alpha\lambda)\varepsilon$ which satisfies $0 < \varepsilon < \varepsilon_1 < \varepsilon_0$, we can utilize the same proof in Theorem 1 (by setting recursively ε_j for $j \geq 1$ and (13)) to conclude that there exists $T \leq |N|$ such that either $X_\Phi(t_\varepsilon + T, \varepsilon_T) \cap N$ or $X_\phi(t_\varepsilon + T, \varepsilon_T) \cap N$ is empty and ε_T is positive.

Similar arguments can be applied for the two nonempty and disjoint sets $X_{\tilde{\Phi}}(t_\varepsilon, \tilde{\varepsilon}_0) \cap N$ and $X_{\tilde{\phi}}(t_\varepsilon, \tilde{\varepsilon}_0) \cap N$ as far as expressed opinions are concerned. We then are able to conclude that there exists $\tilde{T} \leq |N|$ such that either $X_{\tilde{\Phi}}(t_\varepsilon + \tilde{T}, \tilde{\varepsilon}_{\tilde{T}}) \cap N$ or $X_{\tilde{\phi}}(t_\varepsilon + \tilde{T}, \tilde{\varepsilon}_{\tilde{T}}) \cap N$ is empty and $\tilde{\varepsilon}_{\tilde{T}}$ is positive. Combining the above aspects for private and expressed opinions, we similarly derive the contradiction and hence conclude the sufficiency. □

Note that there is a gap between the necessary criterion and the sufficient criterion in Theorem 2. It would be interesting to explore whether the above graph theoretical conditions are tight. The following result for time-dependent networks can be shown in the same manner as Corollary 1.

Corollary 2. *Suppose that we have a time-varying network characterized by a directed graph $G(t) = (V, E(t))$, where each normal agent updates its private and expressed opinions according to the opinion sifting strategy with parameter R. Signified by $\{t_k\}$ the time steps in which $G(t)$ is $2R+1$-robust. Therefore, in the R-locally bounded model having malicious agents, resilient opinion consensus can be reached if $|\{t_k\}| = \infty$ and there is a constant $\tau \in \mathbb{N}$ such that $|t_{k+1} - t_k| \leq \tau$ for all k.*

B. Convergence analysis with Byzantine agents

In this subsection, we investigate necessary and sufficient criteria for opinion consensus in the globally and locally bounded models when there are Byzantine agents in the network. Recall that Byzantine agents have the ability to communicate disparate opinions to different neighbors at any time step, and they are much more difficult to cope with. Define $G_N(t) = (N, E_N(t))$ to be the subnetwork of $G(t) = (V, E(t))$ that is induced by the set of normal agents N, where $E_N(t)$ consists of all directed edges among the normal agents at time step t. The following result deals with fixed network structure in the globally bounded model.

Theorem 3. (opinion consensus in R-globally bounded model with Byzantine agents) *Suppose that we have a* *time-invariant network characterized by a directed graph $G = (V, E)$, where each normal agent updates its private and expressed opinions according to the opinion R-sifting strategy. Then, in the R-globally bounded model having Byzantine agents, resilient opinion consensus can be reached if and only if G_N is $R+1$-robust.*

Proof. (Necessity) Suppose that G_N is not $R+1$-robust. There must exist a pair of nonempty and disjoint sets $S_1, S_2 \subseteq N$ which is not $R+1$ reachable. Each agent in this couple of sets has no more than R normal neighbors not in the set. Fix $c_1 < c_2$. Let $x_i(0) = \tilde{x}_i(0) = c_1$ for any $v_i \in S_1$, and $x_i(0) = \tilde{x}_i(0) = c_2$ for any $v_i \in S_2$. For all the other agents v_i in the network, set $x_i(0) = \tilde{x}_i(0) \in (c_1, c_2)$. Suppose that all Byzantine agents always communicate the expressed opinion c_1 to each agent v_i in S_1, and the expressed opinion c_2 to each agent v_i in S_2 at each time step t. By using the resilient opinion R-sifting strategy, agents in S_1 and S_2 will never adopt opinions not in their own sets. Thus, consensus is not reached among normal agents (in N). The necessity if proved.

(Sufficiency) Similarly, we define that $\rho_{\Phi^*} := \lim_{t\to\infty} \Phi^*(t)$ and $\rho_{\phi^*} := \lim_{t\to\infty} \phi^*(t)$. Assume that $\rho_{\Phi^*} > \rho_{\phi^*}$. Choose $\varepsilon_0 > 0$ satisfying $\rho_{\Phi^*} - \varepsilon_0 > \rho_{\phi^*} + \varepsilon_0$. For $t \in \mathbb{N}$ and $\varepsilon_i > 0$, we define four sets $Y_\Phi(t, \varepsilon_i) := \{v_i \in N : x_i(t) > \rho_{\Phi^*} - \varepsilon_i\}$, $Y_{\tilde{\Phi}}(t, \varepsilon_i) := \{v_i \in N : \tilde{x}_i(t) > \rho_{\Phi^*} - \varepsilon_i\}$, $Y_\phi(t, \varepsilon_i) := \{v_i \in N : x_i(t) < \rho_{\phi^*} + \varepsilon_i\}$, and $Y_{\tilde{\phi}}(t, \varepsilon_i) := \{v_i \in N : \tilde{x}_i(t) < \rho_{\phi^*} + \varepsilon_i\}$. According to the definition of ε_0, $Y_\Phi(t, \varepsilon_0) \cap Y_\phi(t, \varepsilon_0) = \emptyset$ and $Y_{\tilde{\Phi}}(t, \varepsilon_0) \cap Y_{\tilde{\phi}}(t, \varepsilon_0) = \emptyset$. Fix $\varepsilon \in (0, \varepsilon_0)$ such that (5) is satisfied. Define t_ε as the time step such that $\Phi^*(t) < \rho_{\Phi^*} + \varepsilon$ and $\phi^*(t) > \rho_{\phi^*} - \varepsilon$ for all time step $t \geq t_\varepsilon$.

Since G_N is $R+1$-robust with no more than R Byzantine agents, there exits an agent in $Y_\Phi(t_\varepsilon, \varepsilon_0)$ or $Y_\phi(t_\varepsilon, \varepsilon_0)$ that has more than or equal to $R+1$ normal neighboring agents not in its set. W.l.o.g, we assume that $v_i \in Y_\Phi(t_\varepsilon, \varepsilon_0)$ has at least $R+1$ normal neighboring agents not in $Y_\Phi(t_\varepsilon, \varepsilon_0)$. With the same argument as in Theorem 1, we establish the inequality $x_i(t_\varepsilon + 1) \leq \rho_{\Phi^*} - \alpha\lambda\varepsilon_0 + (1 - \alpha\lambda)\varepsilon$. This expression also holds for the renewed private opinion of every normal agent node not in $Y_\Phi(t_\varepsilon, \varepsilon_0)$. Likewise, if $v_i \in Y_\phi(t_\varepsilon, \varepsilon_0)$ which has at least $R+1$ normal neighbors not in $Y_\phi(t_\varepsilon, \varepsilon_0)$, we obtain similarly $x_i(t_\varepsilon + 1) \geq \rho_{\phi^*} + \alpha\lambda\varepsilon_0 - (1 - \alpha\lambda)\varepsilon$, which is also applicable to the normal agents not in $Y_\phi(t_\varepsilon, \varepsilon_0)$.

Define $\varepsilon_1 = \alpha\lambda\varepsilon_0 - (1 - \alpha\lambda)\varepsilon$ such that $0 < \varepsilon < \varepsilon_1 < \varepsilon_0$. Notice that the sets $Y_\Phi(t_\varepsilon + 1, \varepsilon_1)$ and $Y_\phi(t_\varepsilon + 1, \varepsilon_1)$ are disjoint. The comments in the above paragraph imply that $|Y_\Phi(t_\varepsilon+1, \varepsilon_1)| < |Y_\Phi(t_\varepsilon, \varepsilon_0)|$ or $|Y_\phi(t_\varepsilon+1, \varepsilon_1)| < |Y_\phi(t_\varepsilon, \varepsilon_0)|$ holds. We can recursively define $\varepsilon_j = \alpha\lambda\varepsilon_{j-1} - (1-\alpha\lambda)\varepsilon$ for every $j \geq 1$ and recall that $\varepsilon_j < \varepsilon_{j-1}$. The above discussion is applicable to each time step $t_\varepsilon + j$ provided $Y_\Phi(t_\varepsilon + j, \varepsilon_j)$ and $Y_\phi(t_\varepsilon + j, \varepsilon_j)$ are non-empty. Since G_N contains $|N|$ normal agents, there is some $T \leq |N|$ such that either $Y_\Phi(t_\varepsilon + T, \varepsilon_T)$ or $Y_\phi(t_\varepsilon + T, \varepsilon_T)$ is an empty set. On the other hand, $\varepsilon_T = \alpha\lambda\varepsilon_{T-1} - (1-\alpha\lambda)\varepsilon \geq \alpha^{|N|}\lambda^{|N|}\varepsilon_0 - (1-\alpha^{|N|}\lambda^{|N|})\varepsilon > 0$ according to the choice of ε.

Similar arguments can be applied for the two nonempty and disjoint sets $Y_{\tilde{\Phi}}(t_\varepsilon, \tilde{\varepsilon}_0)$ and $Y_{\tilde{\phi}}(t_\varepsilon, \tilde{\varepsilon}_0)$ as far as expressed opinions are concerned (by Theorem 1). We then are able to conclude that there exists $\tilde{T} \leq |N|$ such that either $Y_{\tilde{\Phi}}(t_\varepsilon + \tilde{T}, \tilde{\varepsilon}_{\tilde{T}})$ or $Y_{\tilde{\phi}}(t_\varepsilon + \tilde{T}, \tilde{\varepsilon}_{\tilde{T}})$ is empty and $\tilde{\varepsilon}_{\tilde{T}}$ is positive. Combining the above aspects for private and expressed opinions, we similarly derive the contradiction. The sufficiency is then proved. □

The above Remark 3 can also be applied here as the sufficiency gives a stronger result which indicates that the discrepancy between private and expressed opinions is ultimately vanishing. When moving to time-varying networks, we have the following result.

Corollary 3. *Suppose that we have a time-varying network characterized by a directed graph $G(t) = (V, E(t))$, where each normal agent updates its private and expressed opinions according to the opinion sifting strategy with parameter R. Signified by $\{t_k\}$ the time steps in which $G(t)$ is $2R+1$-robust. Then, in the R-globally bounded model having Byzantine agents, resilient opinion consensus can be reached if $|\{t_k\}| = \infty$ and there is a constant $\tau \in \mathbb{N}$ such that $|t_{k+1} - t_k| \leq \tau$ for all k.*

Proof. Suppose that $G(t)$ is $2R+1$-robust, then $G_N(t)$ is $R+1$ robust. This is can be seen since there exist at most R Byzantine agents in the whole network G. Thanks to Theorem 3, we prove the corollary by invoking a similar argument as that in Corollary 1. □

We next turn to the locally bounded models when there are Byzantine agents deploying across a fixed network topology.

Theorem 4. (opinion consensus in R-locally bounded model with Byzantine agents) *Suppose that we have a time-invariant network characterized by a directed graph $G = (V, E)$, where each normal agent updates its private and expressed opinions according to the opinion R-sifting strategy. Then, in the R-locally bounded model having Byzantine agents, resilient opinion consensus can be reached if and only if G_N is $R+1$-robust.*

Proof. (Necessity) The exact proof of necessity of Theorem 3 is applied here.

(Sufficiency) Reasoning similarly as in Theorem 3, we assume that $\rho_{\Phi^*} := \lim_{t\to\infty} \Phi^*(t)$ and $\rho_{\phi^*} := \lim_{t\to\infty} \phi^*(t)$. Suppose that $\rho_{\Phi^*} > \rho_{\phi^*}$. Select $\varepsilon_0 > 0$ satisfying $\rho_{\Phi^*} - \varepsilon_0 > \rho_{\phi^*} + \varepsilon_0$. For $t \in \mathbb{N}$ and $\varepsilon_i > 0$, we define four sets as in Theorem 3. By the definition of ε_0, $Y_\Phi(t, \varepsilon_0)$ and $Y_\phi(t, \varepsilon_0)$ are disjoint; and $Y_{\tilde{\Phi}}(t, \varepsilon_0)$ and $Y_{\tilde{\phi}}(t, \varepsilon_0)$ are also disjoint. Fix $\varepsilon_0 > \varepsilon > 0$ following (5). Denote by t_ε be the time step such that $\Phi^*(t) < \rho_{\Phi^*} + \varepsilon$ and $\phi^*(t) > \rho_{\phi^*} - \varepsilon$ for all time step $t \geq t_\varepsilon$.

Since G_N is $R+1$-robust, there exists an agent in $Y_\Phi(t_\varepsilon, \varepsilon_0)$ or $Y_\phi(t_\varepsilon, \varepsilon_0)$ that has at least $R+1$ normal neighboring agents not in its set. We assume, w.l.o.g., that $v_i \in Y_\Phi(t_\varepsilon, \varepsilon_0)$ has at least $R+1$ normal neighbors not in $Y_\Phi(t_\varepsilon, \varepsilon_0)$. Noting that there are at most R Byzantine agents in $\mathcal{N}_i$, v_i will adopt at least one of its normal neighbors' expressed opinions not in $Y_\Phi(t_\varepsilon, \varepsilon_0)$ under the opinion R-sifting strategy. Based upon the same reasoning as in Theorem 1, we obtain $x_i(t_\varepsilon+1) \leq \rho_{\Phi^*} - \alpha\lambda\varepsilon_0 + (1 - \alpha\lambda)\varepsilon$. This is also valid for the renewed private opinion of each normal agent not in $Y_\Phi(t_\varepsilon, \varepsilon_0)$. Likewise, if $v_i \in Y_\phi(t_\varepsilon, \varepsilon_0)$ which has at least $R+1$ normal neighbors not in $Y_\phi(t_\varepsilon,\varepsilon_0)$, we arrive at $x_i(t_\varepsilon+1)\ge\rho_\phi+\alpha\lambda\varepsilon_0-(1-\alpha\lambda)\varepsilon$, which is also applicable to the normal agents not in $Y_\phi(t_\varepsilon,\varepsilon_0)$.

Define $\varepsilon_1=\alpha\lambda\varepsilon_0-(1-\alpha\lambda)\varepsilon$, such that $0<\varepsilon<\varepsilon_1<\varepsilon_0$. The same reasoning in the proof of Theorem 3 (by recursively defining $\varepsilon_j=\alpha\lambda\varepsilon_{j-1}-(1-\alpha\lambda)\varepsilon$ for each $j\ge 1$) leads to $\varepsilon_T>0$ for some $T\le|N|$. Similarly, we get $\tilde{\varepsilon}_{\tilde{T}}>0$ for some $\tilde{T}\le|N|$ when expressed opinion update is taken into consideration. Combining the above aspects for private and expressed opinions, we similarly derive the contradiction which concludes the sufficiency. □

Using similar arguments as in Corollary 3, we can establish the following result for time-varying networks in the locally bounded model.

Corollary 4. *Suppose that we have a time-varying network characterized by a directed graph $G(t)=(V,E(t))$, where each normal agent updates its private and expressed opinions according to the opinion sifting strategy with parameter R. Signified by $\{t_k\}$ the time steps in which $G(t)$ is $2R+1$-robust. Then, in the R-locally bounded model having Byzantine agents, resilient opinion consensus can be reached if $|\{t_k\}|=\infty$ and there is a constant $\tau\in\mathbb{N}$ such that $|t_{k+1}-t_k|\le\tau$ for all k.*

IV. Resilient opinion clustering

In this section, we design opinion sifting strategies to allow opinion clustering and co-existence of different opinions in the long term in the presence of both malicious and Byzantine agents. We will differentiate two situations, where opinions may approach assigned ratios or differences. In each case, necessary and sufficient criteria are discussed under globally and locally bounded misbehaving conditions.

A. Opinion separation in terms of quotient

Recall that the opinion network $G=(V,E)$ with $V=N\cup M$ consisting of n interacting agents. Given scalar number $\gamma_i\ne 0$ for every agent $v_i\in V$, we define the opinion separation in terms of quotient as follows.

Definition 6. (opinion separation in terms of quotient) The normal agents in N are said to achieve resilient opinion separation in terms of quotient with respect to $(\gamma_1,\cdots,\gamma_n)$ in the presence of misbehaving agents in M if $\lim_{t\to\infty}\gamma_i x_i(t)-\gamma_j x_j(t)=0$ and $\lim_{t\to\infty}\gamma_i\tilde{x}_i(t)-\gamma_j\tilde{x}_j(t)=0$ for all $v_i,v_j\in N$ and all initial conditions $\{x_i(0)\}_{i=1}^n$ and $\{\tilde{x}_i(0)\}_{i=1}^n$.

If we set $\gamma_1=\gamma_2=\cdots=\gamma_n=1$ above, then we reproduce the resilient opinion consensus in Definition 3. In general, we have $x_i/x_j\to\gamma_j/\gamma_i$ and $\tilde{x}_i/\tilde{x}_j\to\gamma_j/\gamma_i$ as t tends to infinity. Here, by introducing different non-zero scales (can be positive or negative) we are able to allow different expressed and private opinions when t tends to infinity. Opinion separation in terms of quotient is very useful in social opinion networks. In the antagonistic or competitive scenario, an agent may simply reject whatever its competitors support and advocate whatever its competitors object. This can be appropriately modeled by setting $\gamma_i=1$ while $\gamma_j=-1$ for a pair of rivals v_i and v_j. In the literature of control theory, asymptotic approaching to prescribed ratios has been extensively under the name of scaled consensus; see e.g. [36]–[38]. However, only single state value for each agent is considered in the existing literature. To realized resilient opinion separation, we need to modify the proposed strategy in Section 2 to accommodate $\{\gamma_i\}_{i=1}^n$. For $R\in\mathbb{N}$, we refer to the following algorithm as the opinion separation strategy with parameters $\{\gamma_i\}_{i=1}^n$ and R.

Our separation algorithm can be performed in three steps, executed synchronously for all agents at each time step $t\in\mathbb{N}$. First, each normal agent $v_i\in N$ collects the expressed opinions $\{\tilde{x}_j^i(t)\}$ from its neighbors, and creates an ordered array for $\{\gamma_j\tilde{x}_j^i(t)\}_{v_j\in\mathcal{N}_i}$ arranging from largest to smallest. Second, the largest R opinions that are strictly greater than $\gamma_i x_i(t)$ in the above array are deleted (if there are fewer than R greater opinions than $\gamma_i x_i(t)$, all of those opinions are discarded). The similar sifting process is exerted to the smaller opinions. The set of nodes that are discarded by agent v_i at time t is signified by a set $\mathcal{R}_i(t)$. Third, each $v_i\in N$ updates its opinion applying the functions $f_i(\cdot)$ and $g_i(\cdot)$, respectively, in (1) and (2):

$$\begin{aligned}x_i(t+1)=&\operatorname{sgn}(\gamma_i)\Bigg[a_{ii}(t)\gamma_i x_i(t)\\&+\sum_{v_j\in\mathcal{N}_i(t)\backslash\mathcal{R}_i(t)}a_{ij}(t)\gamma_j\tilde{x}_j^i(t)\Bigg],\quad t\in\mathbb{N}\end{aligned}\tag{14}$$

and

$$\begin{aligned}\tilde{x}_i(t)=&\lambda_i x_i(t)+(1-\lambda_i)\\&\cdot\sum_{v_j\in\overline{\mathcal{N}}_i(t-1)\backslash\mathcal{R}_i(t-1)}b_{ij}(t-1)\frac{\gamma_j}{\gamma_i}\tilde{x}_j^i(t-1),\\&t\in\mathbb{N}\backslash\{0\},\end{aligned}\tag{15}$$

where $\operatorname{sgn}(\cdot)$ is the standard signum function, $\{a_{ij}(t)\}$ are the weights instantiating f_i satisfying the following three conditions for every $t\in\mathbb{N}$: (Af') $a_{ij}(t)=0$ if $v_j\notin\overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, (Bf') there exists a constant number $\alpha\in(0,1)$ independent of t, such that $|\gamma_i|a_{ij}(t)\ge\alpha$ for any $v_j\in\overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, and (Cf') $\sum_{v_j\in\overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)}|\gamma_i|a_{ij}(t)=1$; and similarly $\{b_{ij}(t)\}$ are the weights instantiating g_i satisfying the following three conditions for every $t\in\mathbb{N}$: (Ag) $b_{ij}(t)=0$ if $v_j\notin\overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, (Bg) there exists a constant number $\beta\in(0,1)$ independent of t, such that $b_{ij}(t)\ge\beta$ for any $v_j\in\overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)$, and (Cg) $\sum_{v_j\in\overline{\mathcal{N}}_i(t)\backslash\mathcal{R}_i(t)}b_{ij}(t)=1$. Moreover, as in the opinion sifting strategy we assume $\lambda_i\in(0,1]$, and hence there exists a constant $\lambda>0$ satisfying $\lambda_i\ge\lambda>0$ for every $v_i\in N$.

Theorem 5. (resilient opinion separation in terms of quotient) *Suppose that we have a time-invariant network characterized by a directed graph $G=(V,E)$, where each normal agent updates its private and expressed opinions according to the opinion separation strategy with parameters $\{\gamma_i\}_{i=1}^n$ and R. Then, (i) in the R-globally bounded model having malicious agents, resilient opinion separation in terms of quotient can be achieved if and only if G is $(R+1,R+1)$-robust; (ii) in the R-locally bounded model having malicious agents, resilient opinion separation in terms of quotient can be achieved if G is $2R+1$-robust; and G is $R+1$-robust if resilient opinion separation in terms of quotient in the R-locally bounded model with malicious agents can be achieved; (iii) in the R-globally* *bounded model having Byzantine agents, resilient opinion separation in terms of quotient can be achieved if and only if G_N is $R+1$-robust; (iv) in the R-locally bounded model having Byzantine agents, resilient opinion separation in terms of quotient can be achieved if and only if G_N is $R+1$-robust.*

Proof. Define $\Phi(t) := \max_{v_i\in N}\gamma_i x_i(t)$ and $\phi(t) := \min_{v_i\in N}\gamma_i x_i(t)$ respectively as the maximum and minimum private opinions for normal agents. Further, let $\tilde{\Phi}(t) := \max_{v_i\in N}\gamma_i\tilde{x}_i(t)$ and $\tilde{\phi}(t) := \min_{v_i\in N}\gamma_i\tilde{x}_i(t)$ be the maximum and minimum expressed opinions, respectively, for normal agents. By setting $\Phi^*(t) = \max\{\Phi(t),\tilde{\Phi}(t)\}$ and $\phi^*(t) = \min\{\phi(t),\tilde{\phi}(t)\}$, we can follow the similar arguments in Theorems 1-4 to show the statements (i)-(iv), respectively. We omit the detailed proof due to the limitation of space. □

Remark 4. It might be tempting to introduce another set of parameters, say, $\{\tilde{\gamma}_i\}_{i=1}^n$, to adjust the expressed opinions in Definition 6 in hope of showing different limit ratios for expressed opinions. However, this would be technically infeasible in the analysis of Eqs. (14) and (15), and theoretically enigmatic since γ_i can be regarded as a social psychological trait associated with each individual, which should remain the same for both private and expressed opinions for a particular individual [39]. As a result, the sufficiency parts of Theorem 5 is stronger than what is required in Definition 6 (c.f. Remark 3). Moreover, results for time-dependent network $G(t) = (V, E(t))$ can also be derived similarly.

B. Opinion separation in terms of difference

In addition to the parameters $\{\gamma_i\}_{i=1}^n$, let $h = (h_1,\cdots,h_n)\in\mathbb{R}^n$. We give the following definition for resilient opinion separation in terms of difference.

Definition 7. (opinion separation in terms of difference) The normal agents in N are said to achieve resilient opinion separation in terms of difference with respect to $(\gamma_1,\cdots,\gamma_n)$ and h in the presence of misbehaving agents in M if $\lim_{t\to\infty}\gamma_i x_i(t)-\gamma_j x_j(t) = h_i - h_j$ and $\lim_{t\to\infty}\gamma_i\tilde{x}_i(t)-\gamma_j\tilde{x}_j(t) = h_i - h_j$ for all $v_i, v_j\in N$ and all initial conditions $\{x_i(0)\}_{i=1}^n$ and $\{\tilde{x}_i(0)\}_{i=1}^n$.

From Definition 7, we see that the opinion separation in terms of difference is achieved if there exists vectors $\xi,\tilde{\xi}\in\mathbb{R}^n$ such that for each $v_i\in N$, $\gamma_i x_i(t)$ tends to $h_i+\xi$ and $\gamma_i\tilde{x}_i(t)$ tends to $h_i+\tilde{\xi}$ as time goes to infinity. This ideally reflects many moderation processes in group decision making where individuals may have different drifts based upon the default opinion in question. We design the following opinion separation strategy with parameters $\{\gamma_i\}_{i=1}^n$, h and R.

The separation algorithm again can be performed in three steps, executed synchronously for all agents at each time step $t\in\mathbb{N}$. First, each normal agent $v_i\in N$ collects the expressed opinions $\{\tilde{x}_j^i(t)\}$ from its neighbors, and creates an ordered array for $\{\gamma_j\tilde{x}_j^i(t)-h_j\}_{v_j\in\mathcal{N}_i}$ from largest to smallest. Second, the largest R opinions that are strictly greater than $\gamma_i x_i(t)-h_i$ in the above array are deleted (if there are fewer than R grater opinions than $\gamma_i x_i(t)-h_i$, all of those opinions are discarded). The similar sifting process is exerted to the smaller opinions. The set of agents that are discarded by agent v_i at time t is signified by $\mathcal{R}_i(t)$. Third, each $v_i\in N$ updates its opinion using the functions $f_i(\cdot)$ and $g_i(\cdot)$, respectively, in (1) and (2):

$$\begin{aligned}x_i(t+1) =& h_i/\gamma_i + \mathrm{sgn}(\gamma_i)\Bigg[a_{ii}(t)(\gamma_i x_i(t)-h_i)\\ &+\sum_{v_j\in\mathcal{N}_i(t)\backslash\mathcal{R}_i(t)} a_{ij}(t)(\gamma_j\tilde{x}_j^i(t)-h_j)\Bigg],\\ &t\in\mathbb{N}\end{aligned}\tag{16}$$

and

$$\begin{aligned}\tilde{x}_i(t) =& h_i/\gamma_i + \lambda_i(x_i(t)-h_i/\gamma_i)+(1-\lambda_i)\\ &\cdot\sum_{v_j\in\overline{\mathcal{N}}_i(t-1)\backslash\mathcal{R}_i(t-1)} b_{ij}(t-1)(\tilde{x}_j^i(t-1)-h_j/\gamma_j),\\ &t\in\mathbb{N}\backslash\{0\},\end{aligned}\tag{17}$$

where $\{a_{ij}(t)\}$ are the weights instantiating f_i satisfying the same three conditions (Af'), (Bf'), and (Cf') for every $t\in\mathbb{N}$; and $\{b_{ij}(t)\}$ are the weights instantiating g_i satisfying the same three conditions (Ag), (Bg), and (Cg) for every $t\in\mathbb{N}$. We again assume $\lambda_i\in(0,1]$, and hence there exists a constant $\lambda>0$ satisfying $\lambda_i\ge\lambda>0$ for every $v_i\in N$.

Theorem 6. (resilient opinion separation in terms of difference) *Suppose that we have a time-invariant network characterized by a directed graph $G=(V,E)$, where each normal agent updates its private and expressed opinions according to the opinion separation strategy with parameters $\{\gamma_i\}_{i=1}^n$, h and R. Then, (i) in the R-globally bounded model having malicious agents, resilient opinion separation in terms of difference can be achieved if and only if G is $(R+1, R+1)$-robust; (ii) in the R-locally bounded model having malicious agents, resilient opinion separation in terms of difference can be achieved if G is $2R+1$-robust; and G is $R+1$-robust if resilient opinion separation in terms of difference in the R-locally bounded model with malicious agents can be achieved; (iii) in the R-globally bounded model having Byzantine agents, resilient opinion separation in terms of difference can be achieved if and only if G_N is $R+1$-robust; (iv) in the R-locally bounded model having Byzantine agents, resilient opinion separation in terms of difference can be achieved if and only if G_N is $R+1$-robust.*

Proof. Denote by $\tilde{y}_j^i(t) = \tilde{x}_j^i(t)-h_j/\gamma_j$, $y_j(t) = x_j(t)-h_j/\gamma_j$, and $\tilde{y}_j(t) = \tilde{x}_j(t)-h_j/\gamma_j$ for $v_i, v_j\in V$. The update rules (16) and (17) can be recast as

$$\begin{aligned}y_i(t+1) =& \mathrm{sgn}(\gamma_i)\Bigg[a_{ii}(t)\gamma_i y_i(t)\\ &+\sum_{v_j\in\mathcal{N}_i(t)\backslash\mathcal{R}_i(t)} a_{ij}(t)\gamma_j\tilde{y}_j^i(t)\Bigg],\quad t\in\mathbb{N}\end{aligned}\tag{18}$$

and

$$\begin{aligned}\tilde{y}_i(t) =& \lambda_i y_i(t)+(1-\lambda_i)\\ &\cdot\sum_{v_j\in\overline{\mathcal{N}}_i(t-1)\backslash\mathcal{R}_i(t-1)} b_{ij}(t-1)\frac{\gamma_j}{\gamma_i}\tilde{y}_j^i(t-1),\\ &t\in\mathbb{N}\backslash\{0\},\end{aligned}\tag{19}$$

respectively, for $v_i\in N$.

It follows from Theorem 5 that resilient opinion separation in terms of difference is achieved for $\{y_i(t), \tilde{y}_i(t)\}_{v_i \in N}$, which is equivalent to having vectors $\xi, \tilde{\xi} \in \mathbb{R}^n$ such that $\lim_{t\to\infty} \gamma_i x_i(t) = h_i + \xi$ and $\lim_{t\to\infty} \gamma_i \tilde{x}_i(t) = h_i + \tilde{\xi}$ for every $v_i \in N$. The proof is complete. □

Theorem 6 deals with fixed network topology $G = (V, E)$ and the case for time-varying network $G(t) = (V, E(t))$ can be derived analogously as in Sections 3 and 4.1.

Remark 5. As a final remark for the theoretical analysis, we emphasize that in the above theoretical results (Theorems 1-6) the upper bound R of the misbehaving nodes is not an extra limitation. Given any number R, consensus behavior can be expected if the communication network is appropriately robust.

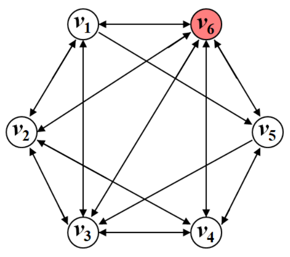

Fig. 2. Communication network G with six agents for Example 1.

V. Simulation Results

In this section, we illustrate the effectiveness of the proposed resilient consensus and separation strategies via several numerical examples.

Example 1. Suppose we have a network $G = (V, E)$ over $n = 6$ agents with $V = N \cup M$, where $N = \{v_1, \cdots, v_5\}$ and $M = \{v_6\}$ as shown in Fig. 2. It is straightforward to check that the digraph G is 3-robust and G_N is 2-robust. Throughout the example, we assume the normal agents have initial private opinions $x_1(0) = -7$, $x_2(0) = -6$, $x_3(0) = 6$, $x_4(0) = 1$, and $x_5(0) = -3$. Agent v_6 is a malicious agent who has initial expressed opinion $\tilde{x}_6(0) = -5$ and updates its expressed opinion following $\tilde{x}_6(t+1) = (\tilde{x}_1(t) + \tilde{x}_2(t) + \tilde{x}_3(t) + \tilde{x}_4(t) + \tilde{x}_5(t))/5 - 0.2t + \sin(t)$. Agent v_6 is aware of the whole network and tries to misguide the group opinion.

The opinion evolution for the initial configuration where $\tilde{x}_i(0) = x_i(0)$ for $1 \le i \le 5$ is shown in Fig. 3(a). Here, we choose $a_{ii}(t) = 2/3$, $a_{ij}(t) = (|\mathcal{N}_i(t)| - |\mathcal{R}_i(t)|)^{-1}/3$ in (3); $\lambda_i = 1/4$ and $b_{ij}(t-1) = (|\mathcal{N}_i(t-1)| + 1 - |\mathcal{R}_i(t-1)|)^{-1}$ in (4). We observe from Fig. 3(a) that the malicious agent is unable to prevent the collective consensus of the network as the normal agents apply the opinion sifting strategy with parameter $R = 1$, which agrees with our theoretic prediction, i. e., Theorems 2, 3, and 4. To further illustrate the evolution of discrepancy between private and expressed opinions, we define the quantity

$$\Delta_i(t) = |x_i(t) - \tilde{x}_i(t)|, \quad t \in \mathbb{N}, \tag{20}$$

for $v_i \in N$. It follows from Fig. 3(b) that all Δ_i's initially equal zero followed by a growth and ultimately die out after around $t = 20$, reaching global consensus over the whole network. In the inset of Fig. 3(b) we illustrate the situation when ordinary consensus protocol without sifting is applied. In this case, the interaction between the misbehaving agent v_6 and the other normal agents diverges the system.

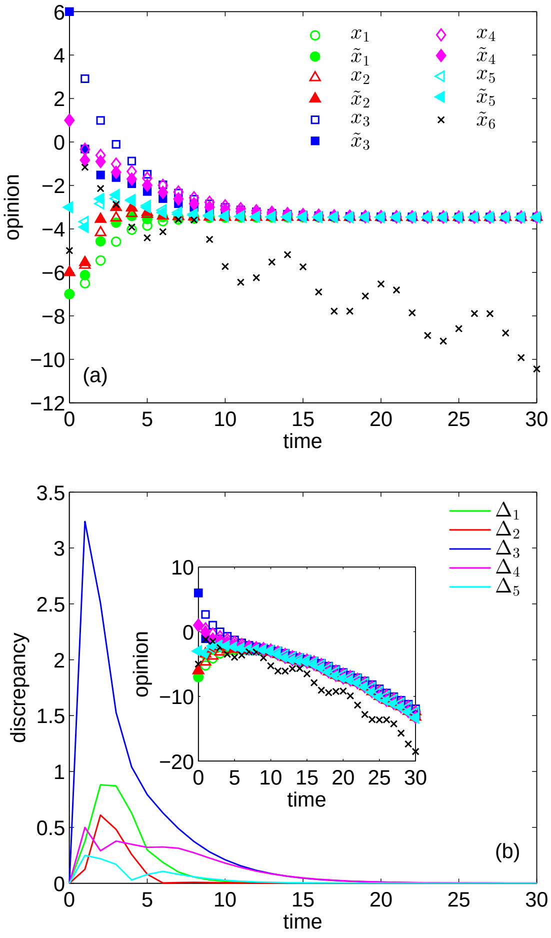

Fig. 3. Opinion consensus over network G with a misbehaving agent v_6 for (a) identical initial conditions for expressed and private opinions and (b) the evolution of discrepancy between expressed and private opinions. Inset of (b): opinion divergence under ordinary protocol without sifting procedure (Legends as in panel (a)).

In Fig. 4, we consider a situation where the discrepancy between private and expressed opinions exists in the beginning. We assume $\tilde{x}_1(0) = 3$, $\tilde{x}_2(0) = 2$, $\tilde{x}_3(0) = -2$, $\tilde{x}_4(0) = 4$, $\tilde{x}_5(0) = -1$. From Fig. 4(a) and (b), we see that the normal agents mange to reach consensus despite the large initial discrepancy between their respective private and expressed opinions. Remarkably, the speed to consensus seems to be only affected negligibly as compared to Fig. 3. Moreover, we observe from the inset of Fig. 4(b) that consensus fails when ordinary consensus protocol without sifting is in use.

Next, we study the opinion clustering in terms of quotient by introducing the parameters $\gamma_1 = \gamma_2 = 2$, and $\gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = -1$. We choose $a_{ii}(t) = 2|\gamma_i|^{-1}/3$ and $a_{ij}(t) = (|\mathcal{N}_i| - |\mathcal{R}_i(t)|)^{-1}|\gamma_i|^{-1}/3$ in (14). The trajectories of opinions are shown in Fig. 5(a). We observe that the opinions are separated into two clusters such that v_1 and v_2 converge to about 0.5 while v_3, v_4 and v_5 tend to reach about -1. This verifies our theoretical result, i. e., Theorem 5, and shows that

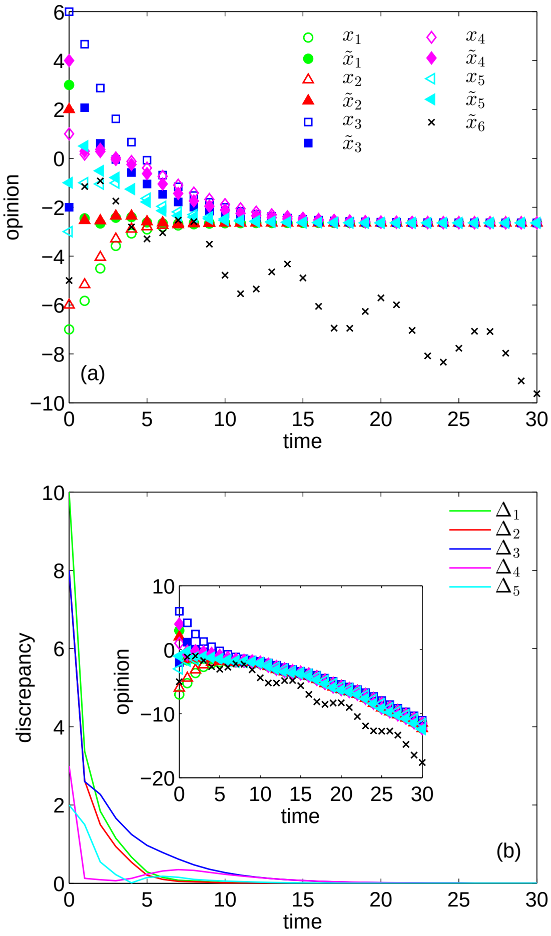

Fig. 4. Opinion consensus over network G with a misbehaving agent v_6 for (a) disparate initial conditions for expressed and private opinions and (b) the evolution of discrepancy between expressed and private opinions. Inset of (b): opinion divergence under ordinary protocol without sifting procedure (Legends as in panel (a)).

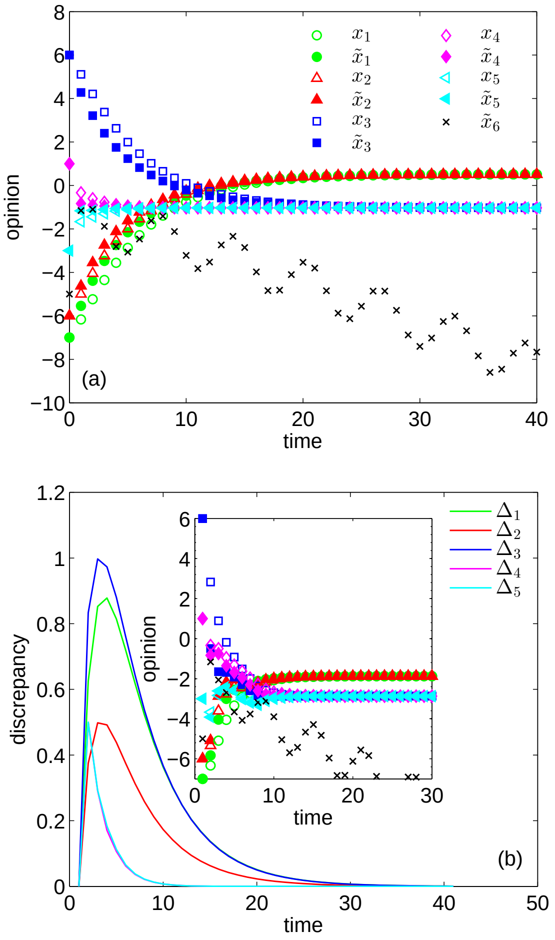

Fig. 5. Opinion separation in terms of quotient over network G with a misbehaving agent v_6 for (a) identical initial conditions for expressed and private opinions and (b) the evolution of discrepancy between expressed and private opinions. Inset of (b): opinion separation in terms of difference over the same network (Legends as in panel (a)).

the opinions approach predetermined ratio using the opinion separation strategy with given parameters $\{\gamma_i\}_{i=1}^{6}$ and $R=1$. The discrepancy of private and expressed opinions is shown in Fig. 5(b) in line with theoretical results. As a comparison, we plot in the inset of Fig. 5(b) the opinion clustering in terms of difference following our separation strategy with $\gamma_i=1$ for all $1\le i\le 6$, $h_1=h_2=1$, and $h_i=0$ for $3\le i\le 6$. The result shows that v_1 and v_2 reach the opinion around -2 while the other normal agents attain the opinion around -3 as one would expect from Theorem 6. It is worth mentioning that although the misbehaving agent v_6 fails to thwart the group's consensus or clustering, it still influences the trajectories and the final opinions of the normal agents since it is fully engaged in the interaction among these agents.

Example 2. In this example, we study the Zachary's karate club, which characterizes the friendship between members of a university karate club in 1977 [40]. A core subgraph of Zachary's karate club, referred to as rich-core G^{Karate}, is determined in [41]; see Fig. 6. It is direct to check that G^{Karate} is an undirected 3-robust graph. When $0\le t\le 10$, we assume that $B=\{v_9\}$ is the only misbehaving agent whose expressed opinion follows the dynamics $\tilde{x}_9(t+1)=(\tilde{x}_2(t)+\tilde{x}_7(t)+\tilde{x}_8(t))/3+t\cos(t)$. When $t\ge 11$, v_9 becomes normal and $B=\{v_8\}$ with $\tilde{x}_8(t+1)=(\tilde{x}_5(t)+\tilde{x}_6(t)+\tilde{x}_7(t)+\tilde{x}_9(t))/4-0.1t-\sin(t^2)$. The initial opinion configuration is taken as $x_1(0)=-\tilde{x}_1(0)=-3$, $x_2(0)=-\tilde{x}_2(0)=-4$, $x_3(0)=-\tilde{x}_3(0)=2$, $x_4(0)=-\tilde{x}_4(0)=1$, $x_5(0)=-\tilde{x}_5(0)=-5$, $x_6(0)=-\tilde{x}_6(0)=3.5$, $x_7(0)=-\tilde{x}_7(0)=-1.5$, $x_8(0)=-\tilde{x}_8(0)=5.5$, and $\tilde{x}_9(0)=0$.

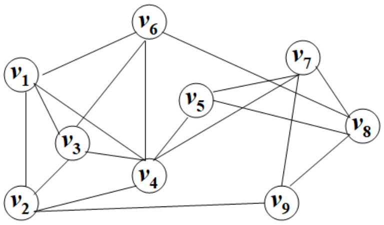

Fig. 6. Rich-core of Zachary's karate club, G^{Karate}, with one misbehaving agent being v_9 when $t\in[0,10]$ and switched to v_8 when $t\ge 11$.

The parameters in protocols (3) and (4) are taken as $a_{ii}(t)=0.9$, $a_{ij}(t)=(|\mathcal{N}_i(t)|-|\mathcal{R}_i(t)|)^{-1}/10$, $\lambda_i=0.6$, $b_{ij}(t-1)=$

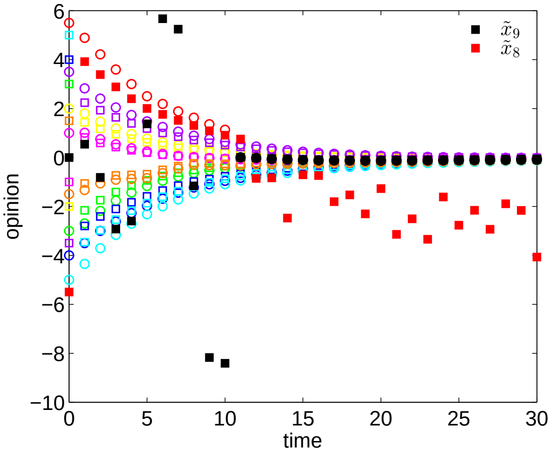

Fig. 7. Opinion consensus over G^{Karate} with one misbehaving agent v_9 when $t \in [0, 10]$ and v_8 when $t \geq 11$. Private opinions (circles) and corresponding expressed opinions (squares) for each agent have the same color. Hollow symbols are for normal agents, while solid symbols are for misbehaving agents.

$(|\mathcal{N}_i(t-1)|+1-|\mathcal{R}_i(t-1)|)^{-1}$ for all i and j. It is interesting to see from Fig. 7 that when v_9 finally "cleans up its act", consensus can still be achieved embracing v_9 as one would expect. This again is in line with our main theorems.

Real-world examples of opinion formation during jury deliberation in Los Angeles County Superior Court as well as a mock jury simulation in Tennessee are provided in Supplementary Material. Building upon real-world data, we illustrate that the proposed sifting algorithms enable consensus for both expressed and private opinions for general interpersonal interaction networks confirming empirical observations.

VI. Conclusion

In this paper, the problem of resilient opinion consensus involving both private and expressed opinions with misbehaving agents has been considered. We have proposed a unique opinion dynamics model accommodating both an individual's private opinion, which is not disclosed to others but evolves under local influence from the expressed opinions of its neighbors, and it's expressed opinion, which evolves under a peer pressure to conform to the local environment. Based on the introduced resilient opinion sifting strategies, we have established sufficient and necessary graph-theoretical conditions to guarantee opinion consensus in a variety of adversarial environments involving local and global threats as well as malicious and Byzantine agents. As a further extension, two types of opinion clustering problems are discussed and corresponding sufficient and necessary conditions are presented to characterize opinion separation in terms of quotient and difference. Both directed fixed networks and time-varying networks have been investigated. Future work includes the design of asynchronous opinion sifting strategies for normal agents and also the analysis of the possible influence of time delays.

References

[1] Y. Dong, M. Zhan, G. Kou, Z. Ding, and H. Liang, "A survey on the fusion process in opinion dynamics," *Inf. Fusion*, vol. 43, pp. 57–65, 2018.

[2] A. V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamical social networks. Part II," *Annu. Rev. Control*, vol. 45, pp. 166–190, 2018.

[3] K. Sznajd-Weron and J. Sznajd, "Opinion evolution in closed community," *Int. J. Mod. Phys. C*, vol. 11, pp. 1157–1165, 2000.

[4] P. Clifford and A. Sudbury, "A model for spatial conflict," *Biometrika*, vol. 60, pp. 581–588, 1973.

[5] S. Galam, "Minority opinion spreading in random geometry," *Eur. Phys. J. B*, vol. 25, pp. 403–406, 2002.

[6] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," *Advs. Complex Syst.*, vol. 3, pp. 87–98, 2000.

[7] R. Hegselmann and U. Krause, "Opinion dynamics and bounded confidence: models, analysis and simulation," *J. Art. Soc. Soc. Simul.*, vol. 5, pp. 1–33, 2002.

[8] Y. Shang, "Hybrid consensus for averager-copier-voter networks with non-rational agents," *Chaos, Solitons & Fractals*, vol. 110, pp. 244–251, 2018.

[9] M. Ye, Y. Qin, A. Govaert, B. D. O. Anderson, and M. Cao, "An influence network model to study discrepancies in expressed and private opinions," *Automatica*, vol. 107, pp. 371–381, 2019.

[10] T. Kuran, "Sparks and prairie fires: a theory of unanticipated political revolution," *Public Choice*, vol. 61, pp. 41–74, 1989.

[11] J. Goodwin, "Why we were surprised (again) by the Arab Spring," *Swiss Polit. Sci. Rev.*, vol. 17, pp. 452–456, 2011.

[12] C. L. Munsch, J. R. Weaver, J. K. Bosson, and L. T. O'Connor, "Everybody but me: pluralistic ignorance and the masculinity contest," *J. Soc. Issues*, vol. 74, pp. 551–578, 2018.

[13] S. V. Grootel, C. V. Laar, L. Meeussen, T. Schmader, and S. Sczesny, "Uncovering pluralistic ignorance to change men's communal self-descriptions, attitudes, and behavioral intentions," *Front. Psychol.*, vol. 9, art. 1344, 2018.

[14] S. G. Buzinski, J. Clark, M. Cohen, B. Buck, and S. P. Roberts, "Insidious Assumptions: how pluralistic ignorance of studying behavior relates to exam performance," *Teach. Psychol.*, vol. 45, pp. 333-339, 2018.

[15] N. Laleh, B. Carminati, and E. Ferrari, "Risk assessment in social networks based on user anomalous behaviors," *IEEE Trans. Depend. Secure Comput.*, vol. 15, pp. 295–308, 2018.

[16] V. Amelkin, F. Bullo, and A. K. Singh, "Polar opinion dynamics in social networks," *IEEE Trans. Autom. Contr.*, vol. 62, pp. 5650–5665, 2017.

[17] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: a network-based spam detection framework for reviews in online social media," *IEEE Trans. Inf. Foren. Secur.*, vol. 12, pp. 1585–1595, 2017.

[18] Y. Shang, "Resilient consensus of switched multi-agent systems," *Syst. Contr. Lett.*, vol. 122, 12–18, 2018.

[19] Y. Shang, "Consensus of hybrid multi-agent systems with malicious nodes," *IEEE Trans. Circuits Syst. Express Briefs*, doi:10.1109/TCSII.2019.2918752.

[20] S. Sundaram and C. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Trans. Autom. Contr.*, vol. 56, pp. 1495–1508, 2011.

[21] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: a system theoretic approach," *IEEE Trans. Autom. Contr.*, vol. 57, pp. 90–104, 2012.

[22] H. Zhang and S. Sundaram, "Robustness of information diffusion algorithms to locally bounded adversaries," *Proc. 2012 American Control Conference*, Montreal, Canada, pp. 5855–5861, 2012.

[23] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE J. Select. Areas Commun.*, vol. 31, pp. 766–781, 2013.

[24] R. Moghadam and H. Modares, "Resilient autonomous control of distributed multiagent systems in contested environments," *IEEE Trans. Cybern.*, 10.1109/TCYB.2018.2856089, 2018.

[25] C. Zhao, J. He, and J. Chen, "Resilient consensus with mobile detectors against malicious attacks," *IEEE Trans. Sign. Inf. Proc. Netw.*, vol. 4, pp. 60–69, 2018.

[26] S. E. Asch, "Effects of group pressure upon the modifictaion and distortion of judgments,", in *Groups, Leadership and Men*, Ed. H. Guetzkow, Carnegie University Press, Pittsburgh, pp. 222–236, 1951.

[27] S. Moldovan, E. Muller, Y. Richter, and E. Yom-Tov, "Opinion leadership in small groups," *Int. J. Res. Market.*, vol. 34, pp. 536–552, 2017.

[28] J. Semonsen, C. Griffin, A. Squicciarini, and S. Rajtmajer, "Opinion dynamics in the presence of increasing agreement pressure," *IEEE Trans. Cybern.*, 10.1109/TCYB.2018.2799858, 2018.

[29] C. Antonopoulos and Y. Shang, "Opinion formation in multiplex networks with general initial distributions," *Sci. Rep.*, vol. 8, art. 2852, 2018.

[30] Y. Shang, "On the structural balance dynamics under perceived sentiment," *Bull. Iranian Math. Soc.*, 10.1007/s41980-019-00286-4, 2019.

[31] X. Song, W. Shi, Y. Mao, and C. Yang, "Impact of informal networks on opinion dynamics in hierarchically formal organizaiton," *Physica A*, vol. 436, pp. 916–924, 2015.

[32] H. Zhang, E. Fata, and S. Sundaram, "A notion of robustness in complex networks," *IEEE Trans. Control Netw. Syst.*, vol. 2, pp. 310–320, 2015.

[33] Y. Liu, J. R. Rui, and X. Cui, "Are people willing to share their political opinions on Facebook? Exploring roles of self-presentational concerns in spiral of silence," *Comput. Hum. Behav.*, vol. 75, pp. 294–302, 2017.

[34] Y. Shang, "Deffuant model with general opinion distributions: first impression and critical confidence bound,", *Complexity*, vol. 19, pp. 38–49, 2013.

[35] K. Hayashi, E. Umehara, and Y. Ogawa, "Analysis of twitter messages about the osaka metropolis plan in Japan," *IEEE Int. Conf. Big Data*, pp. 3064–3070, 2017.

[36] S. Roy, "Scaled consensus," *Automatica*, vol. 51, pp. 259–262, 2015.

[37] Y. Shang, "Finite-time scaled consensus through parametric linear iterations," *Int. J. Syst. Sci.*, vol. 48, pp. 2033–2040, 2017.

[38] Y. Shang, "Scaled consensus of switched multi-agent systems," *IMA J. Math. Control Inform.*, vol. 36, pp. 639–657, 2019.

[39] H. Markus and E. Wurf, "The dynamic self-concept: a social psychological perspective," *Ann. Rev. Psychol.*, vol. 38, pp. 299–337, 1987.

[40] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropol. Res.*, vol. 33, pp. 452–473, 1977.

[41] A. Ma and R. J. Mondragón, "Rich-cores in networks," *PLoS ONE*, vol. 10, art. no. e0119678, 2015.