The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2398-7294.htm

Evaluation and control of opinion polarization and disagreement: a review

Yuejiang Li and Hong Zhao
Department of Automation, Tsinghua University, Beijing, China

Received 18 February 2021
Revised 4 April 2021
Accepted 4 April 2021

Abstract

Purpose – The purpose of this paper is to review the recent studies on opinion polarization and disagreement.

Design/methodology/approach – In this work, recent advances in opinion polarization and disagreement and pay attention to how they are evaluated and controlled are reviewed.

Findings – In literature, three metrics: polarization, disagreement and polarization-disagreement index are usually adopted and there is a tradeoff between polarization and disagreement. Different strategies have been proposed in literature which can significantly control opinion polarization and disagreement based on these metrics.

Originality/value – This review is of crucial importance to summarize works on opinion polarization and disagreement and to the better understanding and control of them.

Keywords Crowd networks, Opinion dynamics, Polarization and disagreement

Paper type Literature review

1. Introduction

With the advance of communication and networking technology, the interactions among people are unprecedentedly enhanced. People are free to express their own opinions and interact with others through commenting, liking, retweeting on online social network platforms. The increasing interactions sometimes result in fierce online debates (Durmus and Cardie, 2019; Sridhar *et al.*, 2015). There can be great opinion polarization and disagreement in the whole process, which might lead to online bullying (Squicciarini *et al.*, 2015). In addition, some malicious people intend to spread misinformation in online social networks to sow discord in society, for example, during the 2016 presidential elections in the USA (Silva *et al.*, 2020) and the protest in Hong Kong (Zervopoulos *et al.*, 2020). Such opinion polarization, disagreement and discord are harmful to public security. Therefore, it is of great importance to understand how people form opinions, evaluate the level of opinion polarization and disagreement and prevent the harmful influence of such discord.

This research is supported by the National Key Research and Development Program of China (2017YFB1400100).

In literature, the works studying opinion polarization and disagreement can be classified into three categories:

(1) opinion dynamics modeling;

(2) evaluating and analyzing opinion polarization and disagreement; and

(3) controlling opinion polarization and disagreement.

The relationship among the works of three categories is summarized in Figure 1. Evaluating and analyzing opinion polarization and disagreement is based on the modeling of opinion dynamics. With the opinion dynamic models and evaluations of polarization and disagreement, the works in the third category study how to control polarization and disagreement.

1.1 Opinion dynamics modeling

The opinion dynamics models can be classified into two categories based on whether opinions are discrete or continuous in the model. In discrete models, the opinion value of individuals can either be binary, e.g. voting for Republicans or Democrats or ordinal, e.g. the ratings of a movie (scores in {0, 1, 2, 3, 4, 5}). However, in continuous models, the opinion values are real numbers, usually unified in the range [0, 1] or [–1, 1]. The lower and upper bounds represent the extreme opinion, e.g. complete support for Republicans or Democrats, respectively. The opinion values in between can be interpreted as how close/far it is to/from the extreme opinion of upper/lower bound. In the following, we briefly review the most basic discrete models and continuous models, respectively.

In discrete models, individuals are influenced by their neighbors and update their opinions according to certain rules. One seminal model is the voter model (Liggett, 2013), where individuals randomly adopt one of his/she neighbors' opinions. It will reach the opinion consensus state in the voter model, where all individuals hold the same opinion. Sood and Redner find that both the first and second-order of the degree distribution of the network influence the time to reach consensus (Sood and Redner, 2005).

The foundation work of continuous opinion dynamics models is the DeGroot model (DeGroot, 1974). In this model, an individual updates their opinion by averaging his/her neighbors' opinions. By analyzing the equilibrium of such averaging process, this work shows that when the network is connected, it will reach an opinion consensus state. The Hegselmann-Krause (HK) opinion dynamics model is proposed based on the DeGroot model

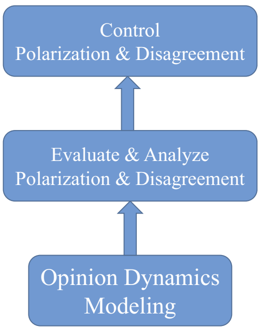


Figure 1. Three categories of works related to opinion polarization and disagreement and the relationships among these works

(Hegselmann and Krause, 2002). It assumes that when the difference between two individuals' opinions is larger than a threshold, these two individuals will ignore each other's opinions when updating. In this model, the opinions will finally converge to different clusters. The opinions in the same cluster are the same, while those from different clusters are different. In this case, opinion polarization exists, that is, individuals hold different opinions at an equilibrium state. The work in Castellano *et al.* (2009) empirically finds that the number of opinion clusters is inversely proportional to the threshold value in the HK model. One of the assumptions in the HK model is that all individuals update their opinions at the same time. Deffuant *et al.* modified the model so that individuals update their opinions in an asynchronous manner (Deffuant *et al.*, 2000). The other important extension of the DeGroot model is the Fredkin-Johnsen (FJ) model (Friedkin and Johnsen, 1990). In this model, each individual is assigned an *internal opinion*, which represents his/her own belief on the topic. When updating opinion, each individual also takes into account his/her internal opinion compared to the DeGroot model. It is showed that, at the equilibrium of the FJ model, all individuals may hold different opinions, which is also an opinion polarization state. Bindel *et al.* explain the dynamics of the FJ model from a game-theoretic perspective. Recently, the FJ model is widely adopted in the analysis of opinion polarization and disagreement (Chen *et al.*, 2018; Matakos *et al.*, 2017; Musco *et al.*, 2018) and opinion maximization (Abebe *et al.*, 2018; Gionis *et al.*, 2013) due to its unique closed-form solution of equilibrium opinions.

1.2 Contribution and organization

Opinion dynamics models have been studied in academia for decades. The models we reviewed above are the most foundation works and there are many variants of them in literature including stubborn individuals (Wai *et al.*, 2016), noise effect (Su *et al.*, 2017) and external sources (Majmudar *et al.*, 2020), etc. There are some great works that systematically summarize and review the opinion dynamics models (Anderson and Ye, 2019; Noorazar, 2020; Proskurnikov and Tempo, 2017, 2018). However, little efforts have been made to review works on the evaluation and analysis of opinion polarization and disagreement, as well as the control of them. Due to the recent harmful events caused by discord, in this work, we aim at reviewing the recent advances in the study and control strategies of opinion polarization and disagreement.

As most related works are based on the Fredkin-Johnsen (FJ) model, we first briefly introduce this model in Section 2. Then, we review how polarization and disagreement are quantified in literature and the relationships among them in Section 3. Next, the works about controlling polarization and disagreement are reviewed in Section 4. The conclusions and discussion are summarized in Section 5.

2. Preliminaries

In this section, we briefly introduce the FJ opinion dynamics model in Bindel *et al.* (2015) In the FJ model, there is a network with n individual which can be modeled as a graph $G(V, E)$. V is the set of nodes representing individuals. E is the edge set and for $(i, j) \in E$, the meaning is that individual j can influence i. The adjacency matrix $\boldsymbol{W}$ can be used to model the network whose entry-i, j is W_{ij}, the influence weight of j on i, if edge $(i, j) \in E$. Otherwise, $W_{ij} = 0$. In literature, the network structure is usually assumed to be undirected, that is, $W_{ij} = W_{ji}$. Another important concept about the network is the graph laplacian $\boldsymbol{L}$. Let $\boldsymbol{D}$ be the diagonal matrix $diag(d_1, \cdots, d_n)$, where $d_i = \sum_j W_{ij}$ is the degree of individual-i. Then, the definition of graph Laplacian is:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}. \tag{1}$$

Both graph Laplacian $\boldsymbol{L}$ and the adjacency matrix $\boldsymbol{W}$ can characterize the network structure because, given one of them, the other one can be derived.

In the FJ model, individual-i is assigned with an internal opinion s_i, which shows his/her beliefs about the discussed topic. Let $\boldsymbol{s} = [s_1, \cdots, s_n]^T$ be the internal opinions of all individuals. The opinion formation process is divided into time step and the internal opinions $\boldsymbol{s}$, is assumed to be a constant. The express opinion of individuals at time step t is $\boldsymbol{z}(t) = [z_1(t), \cdots, z_n(t)]^T$, where entry $z_i(t)$ is the express opinion of individual-i at time t. Let N_i be the set of individual-i's neighbors. In this model, individual-i updates his/her express an opinion as:

$$z_i(t+1) = \frac{s_i + \sum_{j \in N_i} W_{ij} z_j(t)}{1 + \sum_{j \in N_i} W_{ij}}. \tag{2}$$

That is, the express opinion for individual-i at the next time step $t+1$ is the weighted average of his/her internal opinion and the express opinions of his/her neighbors at this time step t.

The opinion formation process evolves and reaches the equilibrium, where no individuals' opinions change anymore. Let $\mathbf{z}$ be the equilibrium opinions. According to the definition of equilibrium, we have:

$$\boldsymbol{z} = (\boldsymbol{L} + \boldsymbol{I})^{-1} \cdot \boldsymbol{s}, \tag{3}$$

where $\boldsymbol{I}$ is the identity matrix with size $n \times n$ and $\boldsymbol{L}$ is the graph Laplacian of the network. From (3), we know that the equilibrium opinions depend on the internal opinions $\boldsymbol{s}$ and the graph structure $\boldsymbol{L}$ (or $\boldsymbol{W}$). In addition, the choice of initial express opinions $\boldsymbol{z}(0)$ does not influence the equilibrium. Friedkin and Johnsen show that the opinions at equilibrium can be different, that is, there can be opinion polarization at equilibrium. Next, we review works that evaluate and analyze opinion polarization and disagreement based on the FJ model. The notations in this review are summarized in Table 1.

Symbol	Meaning
Opinion related	
s, s_i	All internal opinions, internal opinion of individual-i
$\mathbf{z}(t)$, $z_i(t)$	All express opinions at time t, express opinion of individual-i at time t
z, z_i	All equilibrium opinions, equilibrium opinion of individual-i
$\overline{\mathbf{s}}$, $\overline{\mathbf{z}}$	Mean-centered internal opinions and equilibrium opinion
Network related	
W	The adjacency matrix of the network
L	The graph Laplacian of the network
d_i	The degree of individual-i
Metric related	
$\mathcal{P}$	Opinion polarization
$\mathcal{D}$	Opinion disagreement
PDI(μ)	polarization-disagreement index with tradeoff factor μ
PDI	PDI with tradeoff factor 1

Table 1. Notations

3. Evaluation and analysis of polarization and disagreement

In this section, we first review how opinion polarization, disagreement and other related metrics are defined and evaluated in the literature. Then, we summarize the analyzes about opinion polarization and disagreement.

3.1 Quantifying polarization and disagreement

There are some works that analyze the opinion polarization and disagreement based on the Fredkin-Johnsen (FJ) opinion dynamics model (Chen *et al.*, 2018; Dandekar *et al.*, 2013; Musco *et al.*, 2018). Let **z** be the opinions at equilibrium in the FJ model and:

$$\bar{z} = \frac{1}{n}1^T\boldsymbol{z} \tag{4}$$

be the mean of equilibrium opinions. The mean-centered equilibrium opinions are:

$$\overline{\boldsymbol{z}} = \boldsymbol{z} - \bar{z}\cdot 1 = \left(\boldsymbol{I} - \frac{1}{n}11^T\right)\cdot\overline{\boldsymbol{z}}. \tag{5}$$

Similarly, the mean-centered internal opinions $\overline{\mathbf{s}}$ is defined as:

$$\overline{\boldsymbol{s}} = \boldsymbol{s} - \bar{s}\cdot 1 = \left(\boldsymbol{I} - \frac{1}{n}11^T\right)\cdot\overline{\boldsymbol{s}}, \text{ where} \tag{6}$$

$$\bar{s} = \frac{1}{n}1^T\boldsymbol{s} \tag{7}$$

is the average value of internal opinions. It is shown in Musco *et al.* (2018) that:

$$\overline{\boldsymbol{z}} = (\boldsymbol{L}+\boldsymbol{I})^{-1}\cdot\overline{\boldsymbol{s}}. \tag{8}$$

Polarization. The polarization is defined as:

$$\begin{aligned}\mathcal{P} &= \sum_i (z_i - \bar{z})^2 = \overline{\boldsymbol{z}}^T\overline{\boldsymbol{z}} = \overline{\boldsymbol{s}}^T(\boldsymbol{L}+\boldsymbol{I})^{-2}\overline{\boldsymbol{s}} \\ &= \boldsymbol{s}^T(\boldsymbol{L}+\boldsymbol{I})^{-1}\left(\boldsymbol{I} - \frac{1}{n}11^T\right)(\boldsymbol{L}+\boldsymbol{I})^{-1}\boldsymbol{s},\end{aligned} \tag{9}$$

which is the variance of equilibrium opinions. From the definition, we can see that polarization measure how equilibrium opinions deviate from the average (Musco *et al.*, 2018).

Disagreement. Different from polarization, disagreement quantifies the extent to which the express opinions of neighbors are different from each other (Chen *et al.*, 2018). First, the local disagreement on edge $(i, j) \in E$ is defined as:

$$d(i,j) = W_{ij} \cdot (z_i - z_j)^2 = W_{ij} \cdot \left((z_i - \bar{z}) - (z_j - \bar{z})\right)^2 = W_{ij} \cdot (\bar{z}_i - \bar{z}_j)^2. \tag{10}$$

The above equation also shows that local disagreement on edge (i, j) can be calculated through either equilibrium opinions $\mathbf{z}$ or the mean-centered ones $\bar{\mathbf{z}}$. Then, the disagreement of the whole network is defined as the sum of all local disagreement on edges, that is:

$$\mathcal{D} = \sum_{(i,j)\in E} d(i,j). \tag{11}$$

It is shown in (Musco *et al.*, 2018) that:

$$\mathcal{D} = \bar{\boldsymbol{z}}^T \boldsymbol{L} \bar{\boldsymbol{z}} = \bar{\boldsymbol{s}}^T (\boldsymbol{L}+\boldsymbol{I})^{-1} \boldsymbol{L} (\boldsymbol{L}+\boldsymbol{I})^{-1} \bar{\boldsymbol{s}} = \boldsymbol{s}^T (\boldsymbol{L}+\boldsymbol{I})^{-1} \boldsymbol{L} (\boldsymbol{L}+\boldsymbol{I})^{-1} \boldsymbol{s}. \tag{12}$$

Polarization-disagreement index (PDI). This metric combines both opinion polarization and disagreement in a weighted average manner, that is:

$$PDI(\mu) = \mathcal{P} + \mu \cdot \mathcal{D}, \tag{13}$$

where the hyperparameter μ represents the importance of opinion disagreement to PDI comparing with opinion polarization. In this review, we denote *PDI* as *PDI*(1), where opinion polarization and disagreement contribute equally. For *PDI*, we have:

$$PDI = \mathcal{P} + \mathcal{D} = \bar{\boldsymbol{z}}^T (\mathbf{L}+\mathbf{I}) \bar{\boldsymbol{z}} = \bar{\boldsymbol{s}}^T (\boldsymbol{L}+\boldsymbol{I})^{-1} \bar{\boldsymbol{s}} = \boldsymbol{s}^T \left(\boldsymbol{I} - \frac{1}{n} 11^T\right) (\boldsymbol{L}+\boldsymbol{I})^{-1} \left(\boldsymbol{I} - \frac{1}{n} 11^T\right) \boldsymbol{s}. \tag{14}$$

Remarks on polarization and disagreement. The polarization, disagreement-related metrics defined above are summarized in Table 2. From the definition, we can see that they are all quadratic forms of internal opinions $\boldsymbol{s}$ (or mean-centered express opinion $\bar{\boldsymbol{z}}$ and mean-centered internal opinion $\bar{\boldsymbol{s}}$). In addition, as shown in Gaitonde *et al.* (2020), these three quadratic forms are all positive semi-definite, and thus, they are all convex functions with respect to $\boldsymbol{s}$ (or $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{z}}$).

Items	Through $\bar{\boldsymbol{z}}$	Through $\bar{\boldsymbol{s}}$	Through s
Polarization: $\mathcal{P}$	$\bar{\boldsymbol{z}}^T \bar{\boldsymbol{z}}$	$\bar{\boldsymbol{s}}^T (\boldsymbol{L}+\boldsymbol{I})^{-2} \bar{\boldsymbol{s}}$	$\boldsymbol{s}^T (\boldsymbol{L}+\boldsymbol{I})^{-1} \left(\boldsymbol{I} - \frac{1}{n} 11^T\right) (\boldsymbol{L}+\boldsymbol{I})^{-1} \boldsymbol{s}$
Disagreement: $\mathcal{D}$	$\bar{\boldsymbol{z}}^T \boldsymbol{L} \bar{\boldsymbol{z}}$	$\bar{\boldsymbol{s}}^T (\boldsymbol{L}+\boldsymbol{I})^{-1} \boldsymbol{L} (\boldsymbol{L}+\boldsymbol{I})^{-1} \bar{\boldsymbol{s}}$	$\boldsymbol{s}^T (\boldsymbol{L}+\boldsymbol{I})^{-1} \boldsymbol{L} (\boldsymbol{L}+\boldsymbol{I})^{-1} \boldsymbol{s}$
PDI: $\mathcal{P}+\mathcal{D}$	$\bar{\boldsymbol{z}}^T (\boldsymbol{L}+\boldsymbol{I}) \bar{\boldsymbol{z}}$	$\bar{\boldsymbol{s}}^T (\boldsymbol{L}+\boldsymbol{I})^{-1} \bar{\boldsymbol{s}}$	$\boldsymbol{s}^T \left(\boldsymbol{I} - \frac{1}{n} 11^T\right) (\boldsymbol{L}+\boldsymbol{I})^{-1} \left(\boldsymbol{I} - \frac{1}{n} 11^T\right) \boldsymbol{s}$

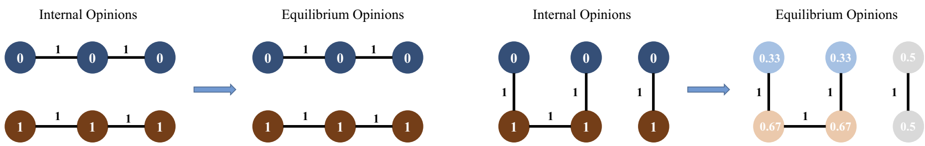
Table 2. Definition of polarization and disagreement

3.2 Analysis of polarization and disagreement

As both opinion polarization and opinion disagreement are harmful to public security, both of them are expected to be weakened. However, this is hard to achieve (Musco *et al.*, 2018). Consider the examples in Figure 2, there are both six individuals in the network. Three of them have internal opinion 0, while the other three of them have opinion 1. Both networks have four edges with weight 1. We can see that the polarization and disagreement at the equilibrium are different in the two networks. The network in Figure 2(a) has lower disagreement and higher polarization, while that in Figure 2(b) has higher disagreement and lower polarization. This is because that the network in Figure 2(a) only connects individuals with the same internal opinion. This forms an "echo chamber" (Jamieson and Cappella, 2008) where individuals only interact with those who have similar opinions and their interactions further enhance their opinions. While the network in Figure 2(b) connects individuals with different internal opinions. According to the FJ model, individuals with different opinions influence each other and their express opinions get closer to others. Therefore, the polarization in this network is small.

Algorithm 1: Opinion dynamics with the network administrator.
Input: Initial graph Laplacian $\hat{\boldsymbol{L}}$;
Initial internal opinion $\hat{\boldsymbol{s}}$;
Repeated round number ROUND.
Output: Expressed opinions after ROUND $\boldsymbol{z}$
for $r = 1, \cdots, ROUND$ **do**
  $\boldsymbol{z} \leftarrow (\boldsymbol{L} + \boldsymbol{I})^{-1} \cdot \boldsymbol{s}$; //Opinion updating
  Solve (15) and obtain new network adjacency matrix $\boldsymbol{W}$; //Weight adjusting.
  Update graph Laplacian $\boldsymbol{L}$ with $\boldsymbol{W}$.
end

Chitra and Musco further analyze opinion polarization and disagreement in real online social networks. They introduce *network administrators* into the FJ opinion dynamics model, whose function is to increase individual engagement via personalized filtering or showing individuals content that they are more likely to agree with. This corresponds to reducing opinion disagreement by adjusting edge weights of the graph in the FJ model (e.g. individuals see more content from others with similar opinions). Their proposed opinion dynamics with network administrator is shown in Algorithm 1. Specifically, the dynamics with network administrator includes multiple rounds. Suppose that the initial graph adjacency matrix is $\hat{W}$ and the internal opinions are $\boldsymbol{s}$. In each round, all individuals first update their opinions according to the FJ model and reach equilibrium $\boldsymbol{z}$. Then, the network



Polarization	$(0-0.5)^2 \times 3 + (1-0.5)^2 \times 3 = 1.5$
Disagreement	$2 \times (0-0)^2 \times 1 + 2 \times (1-1)^2 = 0$
PDI	$1.5 + 0 = 1.5$

(a)

Polarization	$(0.33-0.5)^2 \times 2 + (0.67-0.5)^2 \times 2 + (0.5-0.5)^2 \times 2 = 0.1156$
Disagreement	$2 \times (0.33-0.67)^2 \times 1 + (0.67-0.67)^2 \times 1 + (0.5-0.5)^2 \times 1 = 0.2312$
PDI	$0.1156 + 0.2312 = 0.3468$

(b)

Note: (a) Polarization; (b) Disagreement

Figure 2.
Two examples for polarization and disagreement

administrator adjusts the network structure to minimize disagreement based on the equilibrium opinions, that is:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \mathcal{D} = \bar{\boldsymbol{z}}^T \boldsymbol{L} \bar{\boldsymbol{z}} \\ s.t. \quad & \bar{\boldsymbol{z}} = \left(\boldsymbol{I} - \frac{1}{n} 11^T \right) \cdot \boldsymbol{z} \\ & \boldsymbol{L} \text{ is graph Laplacian of } \boldsymbol{W} \\ & \| \boldsymbol{W} - \hat{\boldsymbol{W}} \|_F \leq \epsilon \cdot \| \hat{\boldsymbol{W}} \|_F \\ & \sum_j W_{ij} = \sum_j \hat{W}_{ij}. \end{aligned} \tag{15}$$

The last two constraints in the above optimization problem ensure that the total change of weights is bounded and the total weights of each individual remain unchanged. The whole process is repeated until it converges.

Chitra and Musco further validate the proposed model with Twitter and Reddit data in De *et al.* (2014). They showed that as ϵ in (15) increases, that is, the network administrator can adjust more weights of the network, the polarization increases surprisingly fast while disagreement shrinks. This observation further validates that there is a tradeoff between opinion polarization and disagreement. In addition, the network administrator in this work acts as recommender systems in online social networks and their recommender behavior (exposing individuals with a similar opinion to each other) can cause "filter bubble" effect (Pariser, 2011), which have been blamed for causing severe opinion polarization in social science and psychology (Bakshy *et al.*, 2015; Garimella *et al.*, 2018; Stroud, 2010).

4. Control of polarization and disagreement

Based on the analyzes of opinion polarization and disagreement, a key problem is how to control them. With the definition of metrics related to polarization and disagreement in Table 2, we can see that controlling polarization and disagreement can be done by tuning the network structure $\boldsymbol{L}$ (or $\boldsymbol{W}$) or the internal opinions $\boldsymbol{s}$.

4.1 Control over network structure

As the example in Figure 2, the network structure has a large impact on polarization and disagreement. As the graph Laplacian $\boldsymbol{L}$ shows how individuals are influenced by each other, tuning $\boldsymbol{L}$ can be explained as interfering with the interactions among individuals. Musco *et al.* first considered minimizing the PDI with the graph Laplacian. They assumed that the total weight of the network is a constant m and the problem was formulated as:

$$\begin{aligned} \min_{\mathbf{L}} \quad & PDI = \bar{\boldsymbol{s}}^T (\boldsymbol{L} + \boldsymbol{I})^{-1} \bar{\boldsymbol{s}} \\ s.t. \quad & Tr(L) = m, \end{aligned} \tag{16}$$

where the constraint means that the total weights of the network remain a constant m. Furthermore, Musco *et al.* show that (16) is a convex optimization problem with respect to $\boldsymbol{L}$. However, if the objective function in (16) is $PDI(\mu)$ where $\mu \neq 1$, the convexity does not hold anymore.

Although (16) is convex, there are $n \times n$ (the size of $\boldsymbol{L}$) variables to be decided, which consumes large memory and time when solving. In addition, if the solution corresponds to a dense network, that is, a lot of entries of $\boldsymbol{L}$ are not zero, this means that the interactions between any two individuals need to be adjusted precisely. This is infeasible in reality under the constraints of limited resources and time. To overcome the above two issues, Musco *et al.* implement the sparse algorithm in Spielman and Srivastava (2011), Spielman and Teng (2011, 2014) to effectively solve (16) and obtain a suboptimal solution, which has much fewer edges. According to their experiments on synthetic networks, the suboptimal solution have only about $\frac{1}{7}$ edges compared to the optimal solution to (16), while the gap between *PDI* calculated by the suboptimal solution and the optimal solution is negligible.

Chen *et al.* argue that it is expected to minimize polarization and disagreement of a certain topic by tuning network structure before this topic begins to be discussed. However, the metrics in Table 2 are all related to internal opinions, which is hard to obtain before the topic begins. To achieve this goal, Chen *et al.* regard the mean-centered internal opinions $\overline{\mathbf{s}}$ as a random vector and define the *average-case conflict risk (ACR)* of metrics in Table 2. Note that all metrics in Table 2 can be expressed in the form of $\overline{\boldsymbol{s}}^T \boldsymbol{M}_* \overline{\boldsymbol{s}}$, where $\mathbf{M}_*$ is the positive semi-definite matrix of metric *, that is, $\boldsymbol{M}_{\mathcal{P}} = (\boldsymbol{L}+\mathbf{I})^{-2}$ for polarization, $\boldsymbol{M}_{\mathcal{D}} = (\boldsymbol{L}+\boldsymbol{I})^{-1}\boldsymbol{L}(\boldsymbol{L}+\boldsymbol{I})^{-1}$ for disagreement and $\boldsymbol{M}_{PDI} = (\boldsymbol{L}+\boldsymbol{I})^{-1}$ for PDI. The ACR assumes that all mean-centered internal opinions are independent and they follow uniform distribution in [–1, 1]. Therefore, $E[\overline{\boldsymbol{s}}\overline{\boldsymbol{s}}^T] = \boldsymbol{I}$. The ACR is defined as the mean of a metric, that is:

$$ACR_* = E\left[\overline{\boldsymbol{s}}^T \boldsymbol{M}_* \overline{\boldsymbol{s}}\right] = E\left[Tr\left(\overline{\boldsymbol{s}}\overline{\boldsymbol{s}}^T \boldsymbol{M}_*\right)\right] = Tr\left(E[\overline{\boldsymbol{s}}\overline{\boldsymbol{s}}^T]\boldsymbol{M}_*\right) = Tr(\boldsymbol{M}_*). \tag{17}$$

Furthermore, Chen *et al.* formulate the problem to minimize ACR_* by controlling the network structure as

$$\begin{aligned} \min_{\boldsymbol{W}} \quad & ACR_* = Tr(\boldsymbol{M}_*) \\ s.t. \quad & 0 \leq \boldsymbol{W} \leq 1 \\ & ||\boldsymbol{W} - \hat{\boldsymbol{W}}||_1 \leq k, \end{aligned} \tag{18}$$

where the first constraint means that each edge weight is in the range [0, 1]. The norm $||\cdot||_1$ in the second constraint is the entry-wise one-norm. Therefore, the second constraint shows that the difference between $\boldsymbol{W}$ and a known adjacency matrix $\hat{\boldsymbol{W}}$ is bounded by k. It is shown in (Chen *et al.*, 2018) that only ACR_{PDI} is convex. Chen *et al.* empirically find that the complete network where all edges' weights are 1 can both minimize ACR_{PDI} and $ACR_{\mathcal{P}}$ in (18). However, the network that can minimize $ACR_{\mathcal{D}}$, it contains sets of disconnected subgraphs which are cliques, trees and chains. Chen *et al.* argue that the disconnected cliques, trees and chain network structure seem to correspond with common management structures in companies: a flat organization corresponds to a clique, while a hierarchical organization corresponds to a tree. From the perspective of companies' interests, it is often assumed to reduce disagreement. Therefore, the learned network structure can provide guidance for companies' team construction.

4.2 Control over internal opinion

In literature, there are some works that assume that the network structure $\boldsymbol{W}$ or ($\boldsymbol{L}$) is known and try to manipulate individuals' internal opinions to control the polarization and disagreement.

Musco *et al.* propose to control internal opinions to minimize the PDI. We can show from the definition of polarization and disagreement that if all individuals hold the same internal opinion, for example, $\boldsymbol{s} = \boldsymbol{0}$, both polarization $\mathcal{P}$ and disagreement $\mathcal{D}$ reach their minima 0. This trivial solution exists because there are no constraints on the internal opinions. In Musco *et al.* (2018), the author proposes the problem that given an internal opinion $\mathbf{s}$, how to change the internal opinion constrained by a constraint α so that the PDI is minimized? The mathematical formulation is:

$$\begin{aligned} \min_{\mathbf{d}} \quad & (\boldsymbol{s}-\boldsymbol{d})^T\left(\boldsymbol{I}-\frac{1}{n}11^T\right)(\boldsymbol{L}+\boldsymbol{I})^{-1}\left(\boldsymbol{I}-\frac{1}{n}11^T\right)(\boldsymbol{s}-\boldsymbol{d}) \\ s.t. \quad & 0 \le \boldsymbol{d} \le \boldsymbol{s} \\ & 1^T\boldsymbol{d} \le \alpha, \end{aligned} \tag{19}$$

where $\boldsymbol{d}$ is the changing vector to the internal opinions. The above optimization problem is convex (specifically semi-definite programming) and can be solved efficiently with techniques, for example, the interior point method, in polynomial time (Boyd and Vandenberghe, 2004). Musco *et al.* run the above optimization on the synthetic network with the power-law degree distribution (Newman, 2005) and find that the internal opinions which are large tend to be reduced most. In addition, the authors examine the equilibrium opinion after optimization and find that all opinions at equilibrium are close to each other. If the internal opinions follow the power-law distribution before optimization, the equilibrium opinions after optimization tend to be centered around 0. However, if the internal opinion follows the uniform distribution before optimization, the equilibrium opinions after optimization tend to be centered around 0.5.

The above work assumes that all individuals' innate opinions can be manipulated, which is hard to achieve in reality. From the perspective of the adversary, Chen and Racz aim at maximizing polarization and disagreement by only controlling a few individuals' internal opinions. The individuals who are controlled by an adversary is called *target individual*. Given the internal opinions $\boldsymbol{s}$, let $\mathcal{P}(\boldsymbol{s})$ and $\mathcal{D}(\boldsymbol{s})$ be the polarization and disagreement, respectively, according to Table 2. Then, with the known internal opinions $\hat{\mathbf{s}}$ and the graph Laplacian of the network, the maximization problem of the adversary is formulated as:

$$\begin{aligned} \max_{\mathbf{s}} \quad & \mathcal{P}(\boldsymbol{s}) \\ s.t. \quad & ||\boldsymbol{s}-\hat{\mathbf{s}}||_0 = k \\ & 0 \le \boldsymbol{s} \le 1, \text{ and} \end{aligned} \tag{20}$$

$$\begin{aligned} \max_{\mathbf{s}} \quad & \mathcal{D}(\boldsymbol{s}) \\ s.t. \quad & ||\boldsymbol{s}-\hat{\mathbf{s}}||_0 = k \\ & 0 \le \mathbf{s} \le 1. \end{aligned} \tag{21}$$

The objectives of the two optimization problems are maximizing opinion polarization and disagreement, respectively. This is because, from the perspective of the adversaries, they want the society to be in chaos and the opinion polarization and disagreement to be large. The constraints $||\boldsymbol{s}-\hat{\mathbf{s}}||_0 = k$ limit the resources of the adversaries, which means that only k of individuals' internal opinions can be controlled. By solving these two problems, the adversaries could find k targeted individuals and change their internal opinions correspondingly by, for example, persuasion. Chen and Racz derive the following inequalities:

$$\mathcal{P}(\boldsymbol{s}_{\mathcal{P}}) \le \mathcal{P}(\hat{\boldsymbol{s}}) + 3k \text{ and} \tag{22}$$

$$\mathcal{D}(\boldsymbol{s}_{\mathcal{D}}) \le \mathcal{D}(\hat{\boldsymbol{s}}) + 8d_{max}k, \tag{23}$$

where $\boldsymbol{s}_{\mathcal{P}}$ and $\boldsymbol{s}_{\mathcal{D}}$ are the optimal solution to (20) and (21), respectively, and d_{max} is the largest degree of the given network. The above two bounds show that both the increase of polarization and disagreement are bounded linearly by k.

Algorithm 2: Greedy algorithm for maximizing polarization or disagreement.
Input: Initial graph laplacian $\hat{\boldsymbol{L}}$;
Initial internal opinion $\hat{\boldsymbol{s}}$;
Number of target individuals k.
Output: The set of target individuals Ω;
The manipulated internal opinion $\boldsymbol{s}$.

```
s ← ŝ and Ω ← ∅;                                          //Initialization.
for i = 1, ···, k do
   //Find only one individual that can maximize P (or D) in each iteration.
   maxVal ← 0;                                          //The maxima of P (or D)
   index ← 0;                                //The individual that maximize P (or D)
   setVal ← 0;                      //The internal opinion value set to individual index
   for j = 1, ···, n do
      if j ∉ Ω then
      s' = s;
      //Enumerate two extreme opinions for individual j.
      Set s'_j to 0, obtain s'_0, calculate P(s'_0) (or D(s'_0));
      if P(s'_0) ≥ maxVal;                               // D(s'_0) ≥ maxVal for D
         then
            maxVal ← P(s'_0), index ← j and setVal ← 0;
            // maxVal ← D(s'_0) for D
         end
         Set s'_j to 1, obtain s'_1, calculate P(s'_1) (or D(s'_1));
         if P(s'_1) ≥ maxVal;                            // D(s'_1) ≥ maxVal for D
         then
            maxVal ← P(s'_1), index ← j and setVal ← 1;
            // maxVal ← D(s'_0) for D
         end
      end
   end
   Change the index entry of s to setVal;                          //Update s
   Ω ← Ω ∪ {index};                                               //Update Ω
end
```

Due to the convexity of polarization $\mathcal{P}$, disagreement $\mathcal{D}$, Chen and Racz first show that any internal opinion of the target individual must be set to the extreme opinion, that is, either 0 or 1. There are $\binom{n}{k}$ cases to choose k out of n individuals. Setting each chosen individual to

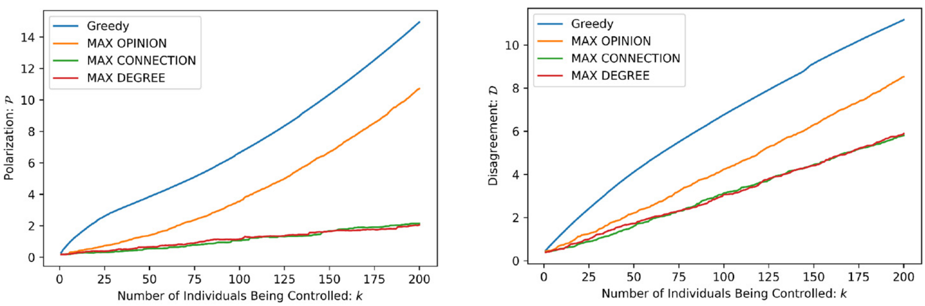
0 or 1 requires 2^k enumerations. Therefore, it requires $\binom{n}{k} \cdot 2^k$ enumerations to decide the optimal solution to (20) and (21) and it is infeasible to use such brute force enumeration method. Chen and Racz propose to use the hill-climbing greedy algorithm in Algorithm 2 to solve the above problems (Domingos and Richardson, 2001; Kempe *et al.*, 2003; Richardson and Domingos, 2002). The greedy algorithm iteratively finds k target individuals, that is, it only finds one individual that can maximize polarization or disagreement and the internal opinion of him/her in each iteration. This process continues until k individuals are found. Compared to the brute force algorithm, it only needs $k \times n \times 2$ enumerations, which significantly reduces the computational complexity. In addition, Chen and Racz also choose the following heuristic methods to decide which k individuals to choose and how their internal opinions should be set:

- MEAN OPINION. Choose the k individuals who have the internal opinions that are the closest to the mean internal opinion.
- MAX CONNECTION. Choose the k individuals who have the most connections with other individuals, that is, the corresponding rows in the adjacency matrix which have the most non zero entries.
- MAX DEGREE. Choose the k individuals who have the largest degree.

With the chosen k individuals, their internal opinions are set to either 0 or 1, respectively, so that polarization or disagreement is maximized. The performances of greedy algorithms and heuristic algorithms are tested with Twitter data in De *et al.* (2014) and shown in Figure 3. The results are implemented from https://github.com/mayeechen/network-disruption. We can see that the greedy algorithm is superior to other heuristic methods on both maximizing polarization and disagreement. The MEAN OPINION algorithm performs the best among heuristic methods. One possible reason is that this algorithm intensionally separates internal opinions which are originally close to each other to different extremes (either 0 or 1). The original "friends" who have similar opinions and interests are provoked by the adversary and their friendships would be broken. In this way, the polarization and disagreement can be greatly enhanced.

5. Conclusion and discussion

In this paper, we review works of evaluating and controlling opinion polarization and disagreement. Based on the Fredkin-Johnsen opinion dynamics model, polarization and



(a) **(b)**

Notes: (a) Polarization; (b) Disagreement

Figure 3.
Polarization and disagreement after manipulating k individuals' internal opinions

disagreement are defined in the literature. Polarization shows how equilibrium opinions deviate from the average, while disagreement measures the total difference of equilibrium opinions between each pair of individuals. There is a tradeoff between polarization and disagreement and the PDI is defined, which is the sum of polarization and disagreement. We also review works of controlling polarization and disagreement by manipulating individuals' internal opinions or network structure. These problems can be efficiently solved by convex optimization or a greedy algorithm.

Although the polarization and disagreement problems have been studied, there are still some issues that need to be investigated. First, most of the control strategies are based on the polarization and disagreement equilibrium opinion. In reality, it is often expected to control polarization and disagreement as soon as possible. Therefore, it is interesting to study how to control polarization and disagreement dynamically. Second, the evaluation and control are based on the Fredkin-Johnsen model. However, ingredients like noise effect, external influence is common in reality. The influence of these ingredients on opinion polarization and disagreement and how to control polarization and disagreement in such settings are also need to be exploited. Last but not least, there might exist, attackers who want to maximize the polarization and disagreement, while some defenders are expected to minimize them. It is also interesting to investigate the evolution of polarization and disagreement in such an adversarial setting.

References

Abebe, R., Kleinberg, J., Parkes, D. and Tsourakakis, C.E. (2018), "Opinion dynamics with varying susceptibility to persuasion", *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1089-1098.

Anderson, B.O. and Ye, M. (2019), "Recent advances in the modelling and analysis of opinion dynamics on influence networks", *International Journal of Automation and Computing*, Vol. 16 No. 2, pp. 129-149.

Bakshy, E., Messing, S. and Adamic, L.A. (2015), "Exposure to ideologically diverse news and opinion on facebook", *Science*, Vol. 348 No. 6239, pp. 1130-1132.

Bindel, D., Kleinberg, J. and Oren, S. (2015), "How bad is forming your own opinion?", *Games and Economic Behavior*, Vol. 92, pp. 248-265.

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press.

Castellano, C., Fortunato, S. and Loreto, V. (2009), "Statistical physics of social dynamics", *Reviews of Modern Physics*, Vol. 81 No. 2, p. 591.

Chen, X., Lijffijt, J. and De Bie, T. (2018), "Quantifying and minimizing risk of conflict in social networks", *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1197-1205.

Dandekar, P., Goel, A. and Lee, D.T. (2013), "Biased assimilation, homophily, and the dynamics of polarization", *Proceedings of the National Academy of Sciences*, Vol. 110 No. 15, pp. 5791-5796.

De, A., Bhattacharya, S., Bhattacharya, P., Ganguly, N. and Chakrabarti, S. (2014), "Learning a linear influence model from transient opinion dynamics", *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 401-410.

DeGroot, M.H. (1974), "Reaching a consensus", *Journal of the American Statistical Association*, Vol. 69 No. 345, pp. 118-121.

Deffuant, G., Neau, D., Amblard, F. and Weisbuch, G. (2000), "Mixing beliefs among interacting agents", *Advances in Complex Systems*, Vol. 03 No. 01n04, pp. 87-98.

Domingos, P. and Richardson, M. (2001), "Mining the network value of customers", *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 57-66.

Durmus, E. and Cardie, C. (2019), "Modeling the factors of user success in online debate", *Proceedings of the World Wide Web Conference (WWW)*, pp. 2701-2707.

Friedkin, N.E. and Johnsen, E.C. (1990), "Social influence and opinions", *The Journal of Mathematical Sociology*, Vol. 15 Nos 3/4, pp. 193-206.

Gaitonde, J., Kleinberg, J. and Tardos, E. (2020), "Adversarial perturbations of opinion dynamics in networks", *Proceedings of ACM Conference on Economics and Computation (EC)*, pp. 471-472.

Garimella, K., De Francisci Morales, G., Gionis, A. and Mathioudakis, M. (2018), "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship", *Proceedings of the World Wide Web Conference (WWW)*, pp. 913-922.

Gionis, A., Terzi, E. and Tsaparas, P. (2013), "Opinion maximization in social networks", *Proceedings of SIAM International Conference on Data Mining (SDM)*, SIAM, pp. 387-395.

Hegselmann, R. and Krause, U. (2002), "Opinion dynamics and bounded confidence models, analysis, and simulation", *Journal of Artificial Societies and Social Simulation*, Vol. 5 No. 3, pp. 1-2.

Jamieson, K.H. and Cappella, J.N. (2008), *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*, Oxford University Press.

Kempe, D., Kleinberg, J. and Tardos, É. (2003), "Maximizing the spread of influence through a social network", *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 137-146.

Liggett, T.M. (2013), *Stochastic Interacting Systems: contact, Voter and Exclusion Processes*, Springer Science and Business Media.

Majmudar, J.R., Krone, S.M., Baumgaertner, B.O. and Tyson, R.C. (2020), "Voter models and external influence", *The Journal of Mathematical Sociology*, Vol. 44 No. 1, pp. 1-11.

Matakos, A., Terzi, E. and Tsaparas, P. (2017), "Measuring and moderating opinion polarization in social networks", *Data Mining and Knowledge Discovery*, Vol. 31 No. 5, pp. 1480-1505.

Musco, C., Musco, C. and Tsourakakis, C.E. (2018), "Minimizing polarization and disagreement in social networks", *Proceedings of World Wide Web Conference (WWW)*, pp. 369-378.

Newman, M.E. (2005), "Power laws, pareto distributions and zipf's law", *Contemporary Physics*, Vol. 46 No. 5, pp. 323-351.

Noorazar, H. (2020), "Recent advances in opinion propagation dynamics: a 2020 survey", *The European Physical Journal Plus*, Vol. 135 No. 6, pp. 1-20.

Pariser, E. (2011), *The Filter Bubble: What the Internet is Hiding from You*, Penguin UK.

Proskurnikov, A.V. and Tempo, R. (2017), "A tutorial on modeling and analysis of dynamic social networks: part I", *Annual Reviews in Control*, Vol. 43, pp. 65-79.

Proskurnikov, A.V. and Tempo, R. (2018), "A tutorial on modeling and analysis of dynamic social networks: part II", *Annual Reviews in Control*, Vol. 45, pp. 166-190.

Richardson, M. and Domingos, P. (2002), "Mining knowledge-sharing sites for viral marketing", *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 61-70.

Silva, M., Giovanini, L., Fernandes, J., Oliveira, D. and Silva, C.S. (2020), "Facebook ad engagement in the russian active measures campaign of 2016", arXiv preprint arXiv:2012.11690.

Sood, V. and Redner, S. (2005), "Voter model on heterogeneous graphs", *Physical Review Letter*, Vol. 94 No. 17, p. 178701.

Spielman, D.A. and Srivastava, N. (2011), "Graph sparsification by effective resistances", *SIAM Journal on Computing*, Vol. 40 No. 6, pp. 1913-1926.

Spielman, D.A. and Teng, S. (2011), "Spectral sparsification of graphs", *SIAM Journal on Computing*, Vol. 40 No. 4, pp. 981-1025.

Spielman, D.A. and Teng, S. (2014), "Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems", *SIAM Journal on Matrix Analysis and Applications*, Vol. 35 No. 3, pp. 835-885.

Squicciarini, A., Rajtmajer, S., Liu, Y. and Griffin, C. (2015), "Identification and characterization of cyberbullying dynamics in an online social network", *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 280-285.

Sridhar, D., Foulds, J., Huang, B., Getoor, L. and Walker, M. (2015), "Joint models of disagreement and stance in online debate", *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 116-125.

Stroud, N.J. (2010), "Polarization and partisan selective exposure", *Journal of Communication*, Vol. 60 No. 3, pp. 556-576.

Su, W., Chen, G. and Hong, Y. (2017), "Noise leads to quasi-consensus of Hegselmann–Krause opinion dynamics", *Automatica*, Vol. 85, pp. 448-454.

Wai, H., Scaglione, A. and Leshem, A. (2016), "Active sensing of social networks", *IEEE Transactions on Signal and Information Processing over Networks*, Vol. 2 No. 3, pp. 406-419.

Zervopoulos, A., Alvanou, A.G., Bezas, K., Papamichail, A., Maragoudakis, M. and Kermanidis, K. (2020), "Hong Kong protests: using natural language processing for fake news detection on twitter", *Proceedings of International Conference on Artificial Intelligence Applications and Innovations*, Springer, pp. 408-419.

Further reading

Chen, M. and Racz, M.Z. (2020), "Network disruption: maximizing disagreement and polarization in social networks", arXiv preprint arXiv:2003.08377.

Chitra, U. and Musco, C. (2020), "Analyzing the impact of filter bubbles on social network polarization", *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 115-123.

Corresponding author

Yuejiang Li can be contacted at: lyj18@mails.tsinghua.edu.cn