The Impact of External Sources on the Friedkin–Johnsen Model

Charlotte Out*
University of Cambridge
Cambridge, United Kingdom
ceo33@cam.ac.uk

Sijing Tu*
KTH Royal Institute of Technology
Stockholm, Sweden
sijing@kth.se

Stefan Neumann†
TU Wien
Vienna, Austria
stefan.neumann@tuwien.ac.at

Ahad N. Zehmakan†
Australian National University
Canberra, Australia
ahadn.zehmakan@anu.edu.au

Abstract

To obtain a foundational understanding of timeline algorithms and viral content in shaping public opinions, computer scientists started to study augmented versions of opinion formation models from sociology. In this paper, we generalize the popular Friedkin–Johnsen model to include the effects of external media sources on opinion formation. Our goal is to mathematically analyze the influence of biased media, arising from factors such as manipulated news reporting or the phenomenon of false balance. Within our framework, we examine the scenario of two opposing media sources, which do not adapt their opinions like ordinary nodes, and analyze the conditions and the number of periods required for radicalizing the opinions in the network. When both media sources possess equal influence, we theoretically characterize the final opinion configuration. In the special case where there is only a single media source present, we prove that media sources which do not adapt their opinions are significantly more powerful than those which do. Lastly, we conduct the experiments on real-world and synthetic datasets, showing that our theoretical guarantees closely align with experimental simulations.

CCS Concepts

• **Theory of computation → Theory and algorithms for application domains**; • **Information systems → Social networks**.

Keywords

Friedkin–Johnsen Model; opinion formation; false balance; social networks

ACM Reference Format:
Charlotte Out, Sijing Tu, Stefan Neumann, and Ahad N. Zehmakan. 2024. The Impact of External Sources on the Friedkin–Johnsen Model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627673.3679780

1 Introduction

Understanding the impact of social media and of conventional media (such as TV or newspapers) on modern societies has been an active research topic of the last decade. This has led to a large body of empirical work, which obtains real-world data from various media sources and analyzes it to obtain insights into societal phenomena [31].

To gain an enhanced theoretical understanding, computer scientists have recently started to study opinion formation models from sociology. They augment these models with abstractions of how online social networks impact the opinion formation process, for instance, when the social network provider (like Facebook or X, previously known as Twitter) aims to minimize the disagreement between the users [13] or in the context of confirmation bias and friend-of-friend recommendations [8].

One shortcoming of previous works is that they usually analyze graphs based on online social networks without any external influence. However, we argue that the discussion on online platforms like X or Reddit is also influenced by external sources, such as conventional media (like newspapers or TV), which are often consumed independent of online social networks.

The influence of these external media sources on the opinion formation process can be particularly stark in the event that the media source is *biased*, i.e., it does not truthfully reflect the average of the opinions. External media sources can be biased, for instance, due to phenomena like *false balance* or *bothsidesism*, where media aim to present both sides of a conflict but unintentionally overrepresent one side of the conflict. This received attention when media gave too much attention to doubters of climate change [9]. Bias can also be the result of intentional manipulations. For example, the Polish PiS party, which formed a majority government from 2015 to 2023, has been blamed for engineering the media coverage to support their own agenda [24]. As exposure to this biased information can lead to negative societal outcomes, including group polarization, intolerance of dissent, and political segregation [36], it is critical to gain an enhanced understanding of the effect of (biased) external media sources on the opinion formation process.

Our contributions. In this paper, we investigate how much impact external sources can have on the opinion-formation process in a social network. In particular, we propose an augmented version of the popular Friedkin–Johnsen (FJ) model [18], in which one or two external sources are added (see Section 3.1 for details). We use our model to study the setting in which the external media sources

*Equal contribution.
†Equal senior authorship.

CIKM '24, October 21–25, 2024, Boise, ID, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679780

are *biased*, i.e., they do not report the average of the individuals' opinions, but their opinion is skewed into a direction. This allows us to analyze how phenomena like false balance or biased media can push a population's average opinion towards a certain direction.

We provide theoretical upper and lower bounds on how much one and two external sources can influence the average opinion in the network (see Theorem 3.3). We then argue that our bounds are tight. More precisely, we show that for regular graphs, our upper and lower bounds match, i.e., that for regular graphs, our analysis is exact. Moreover, we experimentally observe that on real-world social networks (SN), such as Facebook, and synthetic graph models, such as Barabási-Albert (BA) graphs [5], the growth/decrease of the average opinion and the time it takes to reach either radicalized average opinions (close to 0 or close to 1) is similar to the theoretical bounds provided in Theorem 3.3 and Proposition 4.1.

We also study a setting of *repeated* influence from two external sources, in which we consider multiple periods of opinion convergence, after each of which the two external sources update their bias. In the case of two media sources in which one is stronger than the other, we show that in regular graphs even a constant number of periods suffices to radicalize almost all opinions in the network (see Proposition 4.1). Interestingly, we experimentally show that in the regular graphs, this discrepancy in strength between the media sources can be as small as one media source being connected to one more node than the other media source to achieve this radicalization of opinions. For two equally influential external sources, we specify the nodes' final opinions exactly for regular graphs (see Proposition 4.2). In particular, this proposition states that in this setting the total sum of opinions stays unchanged over time, which we additionally verify experimentally.

Furthermore, we give results that differentiate between stubborn and non-stubborn external sources. Here, we say that an external source is *stubborn* if it does not participate in the opinion formation process, and it is *non-stubborn* if it updates its opinion like any other node. We show that non-stubborn external sources can have only very small impact, whereas the impact of stubborn external sources (that we study in the rest of the paper) is significantly higher (see Proposition 5.1). This suggests that it is essential that the media face public scrutiny and peer pressure, to avoid them from (deliberately or unintentionally) biasing a population's average opinion.

1.1 Related Work

Studying opinion formation models and their properties has been an active area of research for at least two decades in the computer science literature. Some of the most well-established models are the threshold model [25, 38], the majority model [10, 23, 42], and the voter model [33]. In this paper, we focus on the popular FJ model.

The paper most closely related to ours is by Gionis, Terzi and Tsaparas [22], who considered the problem of identifying the best k individuals in a network such that if their expressed opinions are fixed to 1, the sum of expressed opinions is maximized; this was motivated, e.g., by marketing campaigns. They found that this problem is NP-hard, but since the objective function is monotone and submodular, this problem admits a greedy $(1-\frac{1}{e})$-approximation algorithm. In this paper, we also consider the sum of opinions in a network and fix the expressed opinions of some nodes. The main difference is that the results in [22] are algorithmic, whereas here we are interested in obtaining analytic bounds on how much the opinions can change. Therefore, the techniques developed by Gionis, Terzi and Tsaparas do not apply in our setting.

Abebe et al. [1] also studied a variation of the FJ model in which each individual has a resistance or stubbornness parameter measuring the individuals' propensity for changing their opinion. They consider the problem of how to change the individual's resistances to maximize (or minimize) the sum of opinions.

Musco, Musco and Tsourakis [29] considered the problem of minimizing the polarization–disagreement index in the FJ model and showed that their objective function is convex which yields an exact polynomial-time algorithm. Zhu, Bao and Zhang [43] studied a similar problem with the goal of adding a small number of edges to the network, and showed that the objective function of their problem is not submodular, although it is monotone. A similar problem based on changing the opinions of a small number of nodes was considered by Makos, Terzi and Tsaparas [27].

Chen and Racz [12] and Gaitonde, Kleinberg and Tardos [19] considered the impact of adversaries on the FJ model, who aim to maximize the polarization and disagreement in the network. They derived analytic bounds on the maximum possible impact of adversaries and also considered the underlying algorithmic problems. This was extended to adversaries with limited information by Tu, Neumann and Gionis [39].

Recently, several works have empirically analyzed the influence of an external media source in different opinion dynamics models. Crokidakis [15] and Nazeri [32] considered the influence of such external mass media by means of Monte Carlo simulations in the two-dimensional Sznajd model [37] and Muslim et al. [30] in the voter model. Pineda and Buendía [34] experimentally analyze the effect of an external media source in the Hegselmann and Krause model [35]. Lastly, Auletta, Coppola and Ferraioli [4] and Candogan [11] considered the setting in which the social media platform itself gives news recommendations to the users, targeting at maximizing user activity on the platform, in the DeGroot model [16].

Apart from external media sources, different forms of (external) bias in opinion dynamics have been considered. The inclusion of stubborn agents (agents with a bias towards a specific opinion) and zealots (agents that never deflect from their initial opinion) has by considered extensively (cf. [3, 20, 28]). Moreover, the scenario in which there is a bias towards a certain (superior) alternative is studied in the majority by Anagnostopoulos [2] and in the voter model by Berenbrink et al. [6]. Lastly, Wilder and Vorobeychik [41] and Corò et al. [14] consider a variant of influence maximization problem in the Independent Cascade and Linear Threshold model respectively, where they have the budget to convince k nodes initially of some preferred opinion. These k nodes in turn start spreading messages which biases the nodes who receive the message towards this preferred opinion.

2 Preliminaries

Graph Notation. Throughout this paper, we let $G = (V, E, w)$ be an undirected weighted graph which represents a social network. We set $n = |V|$ and $w\colon E \longrightarrow \mathbb{R}_{>0}$. For a node $i \in V$, $N(i) := \{v \in V : \{v, i\} \in E\}$ is the *neighborhood* of i, and $d_i := \sum_{j \in N(i)} w_{ij}$ is the *degree* of i. We let $d_{\max}$ and $d_{\min}$ denote the maximum degree and minimum degree in G, respectively. The weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ is defined as $W_{ij} = w_{ij}$ for all $i, j \in V$. We let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be the diagonal degree matrix given by $D_{ii} = \sum_{j\in N(i)} w_{ij}$ and 0 off the diagonal. We let $\mathbf{L} := \mathbf{D} - \mathbf{W}$ denote the graph Laplacian. We let $\mathbf{I} \in \mathbb{R}^{n\times n}$ be the identity matrix, and $\mathbf{e} \in \mathbb{R}^n$ be the vector with 1 in each entry. All logarithms are natural logarithms unless mentioned otherwise.

Friedkin–Johnsen opinion dynamics. We study opinion dynamics based on the Friedkin–Johnsen (FJ) model [18]. The dynamics are specified by a graph $G = (V, E, w)$, where each node $i \in V$ has a fixed (private) internal opinion $s_i \in [0, 1]$ and a (public) expressed opinion $z_i^{(t)} \in [0, 1]$, which depends on the time $t \in \mathbb{N}_0$. Intuitively, one can think of $[0, 1]$ as an interval of opinions where 0 and 1 are the two extreme viewpoints, for instance, 0 corresponds to the viewpoint that climate change is the most pressing issue of the world and 1 corresponding to denying climate change. It will be convenient for us to consider the vectors $\mathbf{s} \in [0, 1]^n$ and $\mathbf{z}^{(t)} \in [0, 1]^n$ of innate and expressed opinions. Starting with $\mathbf{z}^{(0)} = \mathbf{s}$, at each time step t all nodes $i \in V$ update their expressed opinions by taking the weighted average of their neighbors' expressed opinions, as well as their own innate opinion:

$$z_i^{(t+1)} = \frac{s_i + \sum_{j\in N(i)} w_{ij} z_j^{(t)}}{1 + \sum_{j\in N(i)} w_{ij}}. \tag{2.1}$$

It is worth noticing that in this update rule, we implicitly assume that each node's innate opinion has unit weight 1. The above update rule can be equivalently expressed as:

$$\mathbf{z}^{(t+1)} = (\mathbf{I} + \mathbf{D})^{-1} \left(\mathbf{s} + \mathbf{W}\mathbf{z}^{(t)}\right). \tag{2.2}$$

It is well-known that for $t \to \infty$, the vector of expressed *equilibrium* opinions is given by

$$\mathbf{z}^* := \lim_{t\to\infty} \mathbf{z}^{(t)} = (\mathbf{I} + \mathbf{L})^{-1}\mathbf{s}. \tag{2.3}$$

Interestingly, the sum of innate and of equilibrium opinions is the same (Lemma 2.1). This is well-known in the literature.

Lemma 2.1. *It holds that* $\mathbf{e}^\intercal \mathbf{z}^* = \mathbf{e}^\intercal \mathbf{s}$.

Convergence properties. We will study several convergence properties of our model, where we apply some additional notation and several lemmas from [7] for our proofs. Specifically, we will use the characterizations of *nonhomogeneous first-order matrix difference equations* (Definition 1) and properties of *M-matrices* (Definition 2).

Definition 1. *A* nonhomogeneous first-order matrix difference equation *is of the form* $\mathbf{x}^{t+1} = \mathbf{H}\mathbf{x}^t + \mathbf{c}$, *where* $\mathbf{H}$ *is an* $n \times n$ *matrix,* $\mathbf{c}$ *is an* $n \times 1$ *constant vector, and* $\{\mathbf{x}^t\}_{t=1}^{\infty}$ *is an infinite sequence of* $n \times 1$ *vectors.*

The following lemma characterizes the solution and convergence properties of nonhomogenous first-order matrix difference equations [7, Lemma 3.6].

Lemma 2.2. *Let* $\mathbf{A} = \mathbf{M} - \mathbf{N} \in \mathbb{R}^{n\times n}$ *be such that both* $\mathbf{A}$ *and* $\mathbf{M}$ *are nonsingular matrices. Let* $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ *and* $\mathbf{c} = \mathbf{M}^{-1}\mathbf{b}$. *The vector* $\{\mathbf{x}^t\}_{t=1}^{\infty}$ *of the nonhomogenous first-order matrix difference equation* $\mathbf{x}^{t+1} = \mathbf{H}\mathbf{x}^t + \mathbf{c}$ *converges. Moreover,* $\lim_{t\to\infty} \mathbf{x}^t = \mathbf{A}^{-1}\mathbf{b}$.

Definition 2 (M-matrix). *Let* $\mathbf{A}$ *be an* $n\times n$ *matrix where* $a_{i,j} \leq 0$ *for all* $i \neq j$. *Then* $\mathbf{A}$ *is an* M*-matrix if it is positive semidefinite, i.e., for any vector* $\mathbf{x}$, *it holds that* $\mathbf{x}^\intercal \mathbf{A}\mathbf{x} \geq 0$.

Note that the graph Laplacian matrix $\mathbf{L}$ is an M-matrix, as it is positive semidefinite [40, Proposition 1] and its off-diagonal elements are non-positive. Next, we provide two properties of a nonsingular M-matrices [7, Theorem 2.3].

Lemma 2.3. *The following two properties hold:* ***(a)*** *If* $\mathbf{A}$ *is an* M*-matrix, then* $\mathbf{A} + \mathbf{D}$ *is a nonsingular* M*-matrix for each positive diagonal matrix* $\mathbf{D}$. ***(b)*** *If* $\mathbf{A}$ *is a nonsingular* M*-matrix, then it is inverse-positive; that is* $\mathbf{A}^{-1}$ *exists and* $\mathbf{A}^{-1} \geq 0$, *where the inequality holds elementwise.*

3 Stubborn media sources with one period

In this section, we examine the impact of stubborn media sources on opinion formation among individuals in social networks. These media sources are considered stubborn as they keep their (expressed) opinions fixed throughout an entire period of opinion dynamics, i.e., they do not adhere to the FJ-dynamics in Eq. (2.1). We first give an overview of the model and introduce some further notation in Section 3.1.

3.1 Our model

We consider the scenario in which two competing external media sources M and M' are added to the social network. In real-world scenarios, M and M' could be the media sources with opposing political standings, like CNN and Fox News.

Formally, we let $G = (V, E, w)$ be an undirected weighted graph, and M and M' be two external media sources. Each node $i \in V$ is either connected to M or to M', i.e., we either add an edge (i, M) or (i, M') for all $i \in V$. We set the weight of this edge to $\beta(1 + d_i)$, where $\beta \geq 0$ is a model parameter. This means that an external media source contributes a fraction of β to the influence on the users, and each node is exclusively connected to one media source. Additionally, we let $\alpha \in [0, 1]$ denote the fraction of nodes connected to M. Hence, there are αn nodes connected to M, while $(1 - \alpha)n$ nodes are connected to M'. When $\alpha = 0$ or $\alpha = 1$, the nodes are influenced by a single media source, which we refer to as the *single media setting*.

In this section, we assume that the expressed opinions of M and M' are fixed for the whole period, i.e., M and M' do not participate in the opinion dynamics from Eq. (2.1). We use $z_M \in [0, 1]$ to denote the fixed expressed opinion of M and $z_{M'} \in [0, 1]$ to denote the expressed opinion of M'. For convenience, we define a vector ζ such that $\zeta_i = z_M$ if node i is adjacent to M, and $\zeta_i = z_{M'}$ if node i is adjacent to M'.

In our analysis, we use $\hat{\mathbf{z}}^{(t)}$ to denote the vector of expressed opinions at time step t of nodes in the graph G that contains M and M'. We define $\hat{\mathbf{z}}^* = \lim_{t\to\infty} \hat{\mathbf{z}}^{(t)}$. We start by giving a closed-form formulation for the equilibrium expressed opinions $\hat{\mathbf{z}}^*$.

Theorem 3.1. *Let* $G = (V, E, w)$ *be a weighted graph and consider the setting described in Section 3.1, the equilibrium expressed opinions* $\hat{\mathbf{z}}^*$ *can be formulated as:*

$$\hat{\mathbf{z}}^* = ((1 + \beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1}(\mathbf{s} + \beta\,(\mathbf{I} + \mathbf{D})\,\zeta). \tag{3.1}$$

Before providing the proof of this theorem, it is worth noticing that Theorem 3.1 is a generalization of Eq. (2.3): in the absence of external sources, indicated by $\beta = 0$, Theorem 3.1 matches Eq. (2.3). We note that the result of the theorem holds irrespective of the concrete values of z_M and $z_{M'}$.

Intuitively, Theorem 3.1 shows that the changes to the graph topology in the graph by including M and M' can be translated to adapting and slightly reweighting the innate opinions of the nodes in the initial graph, which we also illustrate in Fig. 1.

Proof of Theorem 3.1. Let $\hat{D}_{ii} = D_{ii} + \beta(1 + D_{ii})$. By expanding Eq. (2.1), node i's expressed opinion at time step $t + 1$ can be formulated as follows:

$$\begin{aligned}\hat{z}_i^{(t+1)} &= \frac{s_i + \sum_{j \in N(i)} w_{i,j}\hat{z}_j^{(t)} + \beta(1 + \sum_{j \in N(i)} w_{i,j})\zeta_i}{1 + \sum_{j \in N(i)} w_{i,j} + \beta(1 + \sum_{j \in N(i)} w_{i,j})} \\ &= \frac{\sum_{j \in N(i)} w_{i,j}\hat{z}_j^{(t)}}{1 + \hat{D}_{ii}} + \frac{s_i}{1 + \hat{D}_{ii}} + \frac{\beta(1 + D_{ii})\zeta_i}{1 + \hat{D}_{ii}}.\end{aligned}$$

Writing this in matrix notation, we get:

$$\hat{\mathbf{z}}^{(t+1)} = (\mathbf{I} + \hat{\mathbf{D}})^{-1}\mathbf{W}\hat{\mathbf{z}}^{(t)} + (\mathbf{I} + \hat{\mathbf{D}})^{-1}(\mathbf{s} + (\hat{\mathbf{D}} - \mathbf{D})\zeta).$$

Note that this is a nonhomogeneous first-order matrix difference equation (see Definition 1), and hence we can apply Lemma 2.2 to check whether $\hat{\mathbf{z}}^{(t)}$ converges and obtain the value of $\lim_{t\to\infty} \hat{\mathbf{z}}^{(t)}$.

We apply Lemma 2.2 with $\mathbf{H} = (\mathbf{I}+\hat{\mathbf{D}})^{-1}\mathbf{W}$ and $\mathbf{c} = (\mathbf{I}+\hat{\mathbf{D}})^{-1}(\mathbf{s} + (\hat{\mathbf{D}} - \mathbf{D})\zeta)$. Additionally, we set $\mathbf{M}$ to the common term of $\mathbf{c}$ and $\mathbf{H}$, i.e., we let $\mathbf{M} = \mathbf{I} + \hat{\mathbf{D}}$; hence, $\mathbf{N} = \mathbf{W}$, $\mathbf{b} = \mathbf{s} + (\hat{\mathbf{D}} - \mathbf{D})\zeta$ and $\mathbf{A} = \mathbf{I} + \hat{\mathbf{D}} - \mathbf{W}$.

Next, we check whether $\mathbf{A}$ and $\mathbf{M}$ are singular. We notice that $\mathbf{A}$ is a nonsingular M-matrix, as $\mathbf{L}$ is an M-matrix and $\mathbf{A} = \mathbf{I} + \hat{\mathbf{D}} - \mathbf{W} = (1 + \beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L}$ is the sum of an M-matrix and a positive diagonal matrix. Hence by Lemma 2.3, $\mathbf{A}$ is a nonsingular M-matrix. Furthermore, $\mathbf{M}$ is nonsingular as it is a positive diagonal matrix. Hence, we can directly apply Lemma 2.2, and get:

$$\begin{aligned}\hat{\mathbf{z}}^* = \mathbf{A}^{-1}\mathbf{b} &= (\mathbf{I} + \hat{\mathbf{D}} - \mathbf{W})^{-1}[\mathbf{s} + (\hat{\mathbf{D}} - \mathbf{D})\zeta] \\ &= ((1 + \beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1}(\mathbf{s} + \beta(\mathbf{I} + \mathbf{D})\zeta). \qquad \square\end{aligned}$$

Competing media sources. For the rest of this paper, we assume that the two media sources exert opposing influences on the average opinions of network users: M aims to increase the average opinion of network users (i.e., bring it closer to 1), while M' aims to decrease the average opinion (bring it closer to 0). To this end, we consider a bias parameter $\gamma \in (0, 1)$ and let $\bar{s} = \frac{\mathbf{e}^\top \mathbf{s}}{n}$ denote the average innate opinion of the network users. We then set the expressed opinion z_M of media source M to $z_M = \min\{(1 + \gamma)\bar{s}, 1\}$, and the expressed opinion $z_{M'}$ of media source M' to $z_{M'} = (1-\gamma)\bar{s}$. It is important to note that there are instances where $z_M = (1 + \gamma)\bar{s}$ might be greater than 1, thus exceeding the maximum value; we therefore add the minimum constraint. We do not need to set the similar constraint on $z_{M'}$, as $(1 - \gamma)\bar{s}$ is always at least 0.

3.2 Sum of opinions

Our goal is to bound how much impact the external sources M and M' can have on the average opinion in the network. To this end, we derive a bound on how much the total sum of opinions in the network *with* external media sources differ from the total sum of opinions in the network *without* external media sources.

Recall that above we set $z_M = \min\{(1 + \gamma)\bar{s}, 1\}$. We obtain different results for the two cases and refer to them as *non-truncated* and *truncated* opinions.

Before we present our main results of these two cases, let us introduce Lemma 3.2. Our main technical results in this section are built on the top of this lemma since it allows us to derive a bound on the sum of opinions that purely relies on the maximum and minimum degree of the graph. We provide the proof of this lemma in Section 3.3.

Lemma 3.2. *Let $G = (V, E, w)$ be a weighted graph, with degree matrix $\mathbf{D}$ and Laplacian matrix $\mathbf{L}$. Let $\beta \in [0, 1]$, and $d_{\min}$ and $d_{\max}$ be the minimum and maximum degree of G, respectively. Then it holds that,*

(1) $\mathbf{e}^\top((1 + \beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1} \leq \frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top$.
(2) $\mathbf{e}^\top((1 + \beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1} \geq \frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top$.

The inequality holds element-wise. Furthermore, this holds with equality when the graph is d-regular.

Non-truncated opinions. First, we present a result in Theorem 3.3 if $z_M = (1 + \gamma)\bar{s}$, i.e., the opinion of z_M was not truncated.

Theorem 3.3. *Let $G = (V, E, w)$ be a weighted graph and consider the setting described in Section 3.1 with $z_M = (1 + \gamma)\bar{s}$ and $z_{M'} = (1 - \gamma)\bar{s}$. Then the sum of equilibrium expressed opinions can be bounded as,*

$$\begin{aligned}\mathbf{e}^\top\hat{\mathbf{z}}^* &\leq \frac{1+(d_{\max}+1)\beta((2\alpha-1)\gamma+1)}{\beta(d_{\min}+1)+1}\mathbf{e}^\top\mathbf{s}, \\ \mathbf{e}^\top\hat{\mathbf{z}}^* &\geq \frac{1+(d_{\min}+1)\beta((2\alpha-1)\gamma+1)}{\beta(d_{\max}+1)+1}\mathbf{e}^\top\mathbf{s}.\end{aligned}$$

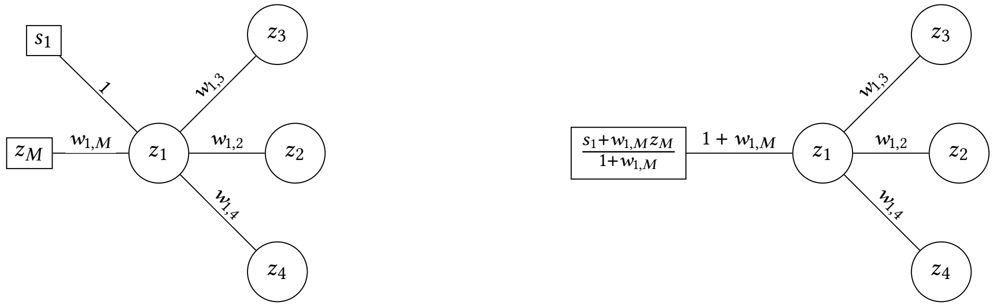
To better illustrate the bounds, Corollary 3.4 states the bound for d-regular graphs, for which it is *tight.*

Corollary 3.4. *Let $G = (V, E, w)$ be a weighted d-regular graph, i.e., $d_{\max} = d_{\min} = d$ and $z_M = (1 + \gamma)\bar{s}$ and $z_{M'} = (1 - \gamma)\bar{s}$. Then the sum of equilibrium expressed opinions is given by:*

$$\mathbf{e}^\top\hat{\mathbf{z}}^* = \left(1 + \gamma \cdot \frac{\beta(d + 1)(2\alpha - 1)}{\beta(d + 1) + 1}\right)\mathbf{e}^\top\mathbf{s}.$$

Observe that the sum of opinions increases (decreases) when $\alpha > \frac{1}{2}$ ($\alpha < \frac{1}{2}$). Furthermore, when $\beta(d + 1)$ is sufficiently larger than 1 then the theorem states that $\mathbf{e}^\top\hat{\mathbf{z}}^* = (1 + \Omega(\gamma(2\alpha - 1)))\mathbf{e}^\top\mathbf{s}$, i.e., under these conditions, the sum of opinions is only controlled by the bias parameter γ and how much the stronger media source dominates (controlled by α).

This result is significant for two reasons: (1) Lemma 2.1 asserts that *without* the intervention of external media sources, the total sum of expressed and innate opinions of the network users always stays the same, i.e., $\mathbf{e}^\top\mathbf{z}^* = \mathbf{e}^\top\mathbf{s}$. Hence, the theorem characterizes the power of the external sources and shows that, under the parameter settings from above, the external sources bias the average opinion by a factor of $1 + \Omega(\gamma(2\alpha - 1))$. (2) The result of the theorem is enabled by the fact that the media sources M and M' are stubborn, i.e., their expressed opinions are fixed. In Section 5 we show that if there is only a single media source (i.e., $\alpha = 1$), which can only control its innate opinion but has to update its expressed



(a) An example of how z_1 gets influence from different nodes,including M, in our model.

(b) An equivalent influence on z_1, by merging the influence of node M with the node 1's innate opinion s_1.

Figure 1: Two equivalent ways to present the influence of a stubborn media source M and its neighbors on node 1, at each time step. The nodes represent the innate or expressed opinions; we use circles to present nodes' expressed opinions and use boxes to annotate fixed innate opinions. We assume that $N(1)=\{2,3,4\}$ and $w_{1,M}=\beta(1+\sum_{i=1}^{3}w_{1,i})$.

opinion based on the update rule in Eq. (2.1), the bias on the sum of opinions is much weaker (it only contributes a factor of $1+\frac{1+\gamma}{n}$ rather than $1+\Omega(\gamma)$).

Proof of Theorem 3.3. First we give a lower bound on the sum of expressed opinions,

$$\begin{aligned}\mathbf{e}^\top\hat{\mathbf{z}}^* &\overset{(a)}{=} \mathbf{e}^\top\left((1+\beta)\cdot\mathbf{I}+\beta\cdot\mathbf{D}+\mathbf{L}\right)^{-1}\left(\mathbf{s}+\beta(\mathbf{I}+\mathbf{D})\zeta\right)\\ &\overset{(b)}{\geq} \frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top\left(\mathbf{s}+\beta(\mathbf{I}+\mathbf{D})\zeta\right)\\ &\geq \frac{1}{\beta(d_{\max}+1)+1}\left(\mathbf{e}^\top\mathbf{s}+(d_{\min}+1)\beta\mathbf{e}^\top\zeta\right)\\ &\overset{(c)}{=} \frac{1+(d_{\min}+1)\beta((2\alpha-1)\gamma+1)}{\beta(d_{\max}+1)+1}\mathbf{e}^\top\mathbf{s},\end{aligned}$$

where in Step (a) we used Theorem 3.1, in Step (b) we used Lemma 3.2 (2), and Step (c) follows from how we set ζ (recall that $\zeta_i=z_M=(1+\gamma)\cdot\frac{\mathbf{e}^\top\mathbf{s}}{n}$ if i is adjacent to M, and $\zeta_i=z_{M'}=(1-\gamma)\cdot\frac{\mathbf{e}^\top\mathbf{s}}{n}$ if i is adjacent to M'). To obtain the upper bound stated in Theorem 3.3, we perform the same calculation, except that now we use Lemma 3.2 (1) in Step (b). □

Truncated opinions. Now, we consider the case when $(1+\gamma)\bar{\mathbf{s}}>1$, and hence we truncate z_M by setting $z_M=1$. Nonetheless, we still set $z_{M'}=(1-\gamma)\frac{\mathbf{e}^\top\mathbf{s}}{n}$. Note that this setting is interesting, since now M induces a smaller bias than before, and we need to understand how much this boosts the impact of M'. For d-regular graphs, we obtain Proposition 3.5, which gives us a closed form solution for $\mathbf{e}^\top\hat{\mathbf{z}}^*$.

Proposition 3.5. *Let $G=(V,E,w)$ be a weighted d-regular graph. Suppose $z_M=1$ and $z_{M'}=(1-\gamma)\bar{\mathbf{s}}$. Then the sum of equilibrium expressed opinions is*

$$\mathbf{e}^\top\hat{\mathbf{z}}^*=\frac{(1+\beta(1+d)(1-\alpha)(1-\gamma))\mathbf{e}^\top\mathbf{s}+\alpha\beta(1+d)n}{1+\beta(1+d)}. \tag{3.2}$$

Observe that the proposition implies that $\mathbf{e}^\top\hat{\mathbf{z}}^*$ increases linearly as a function of α (since $\mathbf{e}^\top\mathbf{s}\leq n$), and it decreases linearly as a function of the media bias γ. This dependency on γ stems from the fact that we truncate z_M, while $z_{M'}$ decreases as γ increases.

Furthermore, in Corollary 3.6 we derive a lower bound on $\mathbf{e}^\top\hat{\mathbf{z}}^*$ that is *independent* of d and β that purely relies on the innate opinions, α and γ. Note that this provides a general bound on the power of the bias of M' if $z_M=1$, since we provide a lower bound on $\mathbf{e}^\top\hat{\mathbf{z}}^*$.

Corollary 3.6. *Suppose G is d-regular. By rearranging Eq. (3.2), we obtain the following lower bound on $\mathbf{e}^\top\hat{\mathbf{z}}^*$: $\mathbf{e}^\top\hat{\mathbf{z}}^*>\mathbf{e}^\top\mathbf{s}(1-\gamma+\alpha\gamma)$.*

Proof of Proposition 3.5. We use Theorem 3.1 and simplify the expression:

$$\begin{aligned}\mathbf{e}^\top\hat{\mathbf{z}}^* &= \mathbf{e}^\top\left((1+\beta)\mathbf{I}+\beta\mathbf{D}+\mathbf{L}\right)^{-1}(\mathbf{s}+\beta(\mathbf{I}+\mathbf{D})\zeta)\\ &\overset{(a)}{=} \frac{1}{1+\beta(1+d)}\mathbf{e}^\top(\mathbf{s}+\beta(\mathbf{I}+\mathbf{D})\zeta)\\ &\overset{(b)}{=} \frac{\mathbf{e}^\top\mathbf{s}+\beta(1+d)(\alpha n+(1-\alpha)(1-\gamma)\mathbf{e}^\top\mathbf{s})}{1+\beta(1+d)}\\ &= \frac{1+(1-\alpha)(1-\gamma)\beta(1+d)}{1+\beta(1+d)}\mathbf{e}^\top\mathbf{s}+\frac{\alpha n\beta(1+d)}{1+\beta(1+d)},\end{aligned}$$

where Step (a) holds by Lemma 3.2, which we prove below. Step (b) holds by plugging $z_M=1$ and $z_{M'}=(1-\gamma)\frac{\mathbf{e}^\top\mathbf{s}}{n}$ into ζ, since $\mathbf{e}^\top\zeta=\alpha n\cdot 1+(n-\alpha n)\frac{\mathbf{e}^\top\mathbf{s}}{n}(1-\gamma)=\alpha n\cdot 1+(1-\alpha)(1-\gamma)\mathbf{e}^\top\mathbf{s}$. □

Proof of Corollary 3.6. We start by applying Proposition 3.5:

$$\begin{aligned}&\frac{1+(1-\alpha)(1-\gamma)\beta(1+d)}{1+\beta(1+d)}\mathbf{e}^\top\mathbf{s}+\frac{\alpha n\beta(1+d)}{1+\beta(1+d)}\\ &\overset{(a)}{>} \frac{1+(1-\alpha)(1-\gamma)\beta(1+d)+\alpha\beta(1+d)}{1+\beta(1+d)}\mathbf{e}^\top\mathbf{s}\\ &\overset{(b)}{\geq} \frac{(1-\gamma+\alpha\gamma)+(1-\gamma+\alpha\gamma)\beta(1+d)}{1+\beta(1+d)}\mathbf{e}^\top\mathbf{s}\\ &=(1-\gamma+\alpha\gamma)\mathbf{e}^\top\mathbf{s}.\end{aligned}$$

Note that Step (a) is obtained from $n>\mathbf{e}^\top\mathbf{s}$, and Step (b) is obtained from $\alpha\leq 1$, hence $1-\gamma+\alpha\gamma=1+\gamma(\alpha-1)\leq 1$. By substituting back into the formula, we obtain

$$\mathbf{e}^\top\hat{\mathbf{z}}^*>\mathbf{e}^\top\mathbf{s}(1-\gamma+\alpha\gamma),$$

which is what we claimed in the corollary. □

3.3 Proof of Lemma 3.2

First, we prove that it holds element-wise that $\mathbf{e}^\top \geq \frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$, and $\mathbf{e}^\top \leq \frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$. After that, we show that $(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$ is a non-singular M-matrix; this implies that $(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)^{-1}$ only contains nonnegative elements (see Lemma 2.3). In the end, we show that by combining these two properties, the lemma follows.

We note that $\mathbf{e}^\top \geq \frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$, as,

$$\begin{aligned}
&\frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)\\
&\overset{(a)}{=}\frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top\,(\mathbf{I}+(\mathbf{D}+\mathbf{I})\beta)\\
&\overset{(b)}{\leq}\frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top\mathbf{I}\,(1+\beta(d_{\max}+1))=\mathbf{e}^\top,
\end{aligned}$$

where Step (a) holds since $\mathbf{e}^\top\mathbf{L}=\mathbf{0}$ and Step (b) holds since $\mathbf{D}+\mathbf{I}\leq(d_{\max}+1)\mathbf{I}$.

Similarly, to show that $\mathbf{e}^\top \leq \frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$, we proceed as follows:

$$\begin{aligned}
&\frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)\\
&\overset{(a)}{=}\frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top(\mathbf{I}+(\mathbf{D}+\mathbf{I})\beta)\\
&\overset{(b)}{\geq}\frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top\mathbf{I}(1+\beta(d_{\min}+1))=\mathbf{e}^\top,
\end{aligned}$$

where Step (a) follows from the fact that $\mathbf{e}^\top\mathbf{L}=\mathbf{0}$, and Step (b) holds because $\mathbf{D}+\mathbf{I}\geq(d_{\min}+1)$.

Next, we notice that $(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$ is a non-singular M-matrix, as it is the sum of the M-matrix $\mathbf{L}$ and the positive diagonal matrix $(\mathbf{I}+(\mathbf{D}+\mathbf{I})\beta)$, according to Lemma 2.3, it is a non-singular M-matrix.

Before we finish the proof, we briefly prove an observation. Consider vectors $\mathbf{a}\geq\mathbf{b}$, where the inequality holds element-wise, and a vector $\mathbf{c}$ with non-negative entries. Now observe that $\mathbf{a}^\top\mathbf{c}=\sum_i a_ic_i\geq\sum_i b_ic_i=\mathbf{b}^\top\mathbf{c}$.

Finally, we prove the lemma by combining the previous results. To show this, we let x_j denote the j'th entry of $\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)^{-1}$ and we let y_j denote the j'th entry of $\frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top$. Then we set $\mathbf{a}^\top=\mathbf{e}^\top$, $\mathbf{b}^\top=\frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$, and we let $\mathbf{c}$ be the j'th column vector of $(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)^{-1}$. Notice that by Lemma 2.3 (see above), $\mathbf{c}$ is a vector with nonnegative entries. Now we obtain that,

$$\begin{aligned}
x_j=\mathbf{e}^\top\mathbf{c}&\geq\frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)\mathbf{c}\\
&=\frac{1}{\beta(d_{\max}+1)+1}\mathbf{e}^\top\mathbf{d}=y_j,
\end{aligned}$$

where $\mathbf{d}$ is the indicator vector with a 1 in the j'th entry and 0 in all other entries. Since this holds for all j, this implies our first bound.

Similarly, letting $\mathbf{a}^\top=\frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)$, $\mathbf{b}^\top=\mathbf{e}^\top$, and $\mathbf{c}$ any column of $(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)^{-1}$, we can prove $\mathbf{e}^\top(\mathbf{I}+\mathbf{L}+(\mathbf{D}+\mathbf{I})\beta)^{-1}\leq\frac{1}{\beta(d_{\min}+1)+1}\mathbf{e}^\top$.

4 Stubborn media sources with multiple periods

In this section, we study the impact of two stubborn media sources on the expressed opinions over a longer time horizon. In Section 3 we considered the expressed opinions $\hat{\mathbf{z}}^*$ after M and M' were added to the graph and the expressed opinions converged. Now we consider multiple *periods* of such convergence steps (after each of which z_M and $z_{M'}$ are updated) to understand how quickly opinions can get radicalized towards an extreme opinion after multiple periods. This setting is applicable when each period of convergence corresponds to a new societal topic and is similar to a setting studied by [19].

Formally, our model is as follows. We consider a sequence of *periods* $t=0,1,2,\ldots$. At the beginning of each period t, individuals set their innate opinions to their expressed opinions from the previous period. Formally, let $\tilde{z}_i^{(t)}$ denote the equilibrium expressed opinion of node i at the end of period t. Then in period $t+1$, all nodes i set their innate opinions to $s_i^{(t+1)}=\tilde{z}_i^{(t)}$. After updating the nodes' innate opinions, the opinions of the external sources are also updated, now defined as $z_M^{(t+1)}:=\min\{(1+\gamma)\bar{\mathbf{s}}^{(t+1)},1\}$, and $z_{M'}^{(t+1)}:=(1-\gamma)\bar{\mathbf{s}}^{(t+1)}$, where $\bar{\mathbf{s}}^{(t+1)}=\frac{\mathbf{e}^\top\mathbf{s}^{(t+1)}}{n}$ is the average innate opinion at the start of period $t+1$. Then we run the model from Section 3 to obtain $\tilde{z}_i^{(t+1)}$. In our analysis, we restrict ourselves to regular graphs and assume $\alpha\geq1/2$. We will provide results for $\alpha>\frac{1}{2}$ and for $\alpha=\frac{1}{2}$.

Unequally strong media sources. First, we assume that $\alpha>\frac{1}{2}$, i.e., that source M is connected to strictly more nodes than M'. We first compute the minimum number of periods it takes to obtain $z_M^{(t)}=1$, which we consider as a criterion for radicalization, since in this case the average opinion in the network is at least $1/(1+\gamma)$. We denote this number of periods by ℓ^*.

Proposition 4.1. *Let G be a d-regular graph and let γ, $\tilde{\mathbf{z}}^{(0)}$ α and ℓ^* be as defined above. We assume that $\alpha>1/2$. Then,*

$$\ell^*=\frac{\log\left(\frac{n}{\mathbf{e}^\top\mathbf{s}(1+\gamma)}\right)}{\log\left(1+\gamma\cdot\frac{(d+1)\beta(2\alpha-1)}{(d+1)\beta+1}\right)}.$$

Observe that if for the initial innate opinions at period 0 it holds that $\mathbf{e}^\top\mathbf{s}=\Omega(n)$ and $\gamma\cdot\frac{(d+1)\beta(2\alpha-1)}{(d+1)\beta+1}$ is a constant (which are reasonable assumptions for most scenarios), then $\ell^*=O(1)$. That is, it takes a constant number of periods to radicalize the sum of opinions in the network.

Proof of Proposition 4.1. By Corollary 3.4 we note that,

$$\mathbf{e}^\top\hat{\mathbf{z}}^*=\left(1+\frac{(d+1)\beta\gamma(2\alpha-1)}{(d+1)\beta+1}\right)\mathbf{e}^\top\mathbf{s}.$$

Now, observe that

$$\mathbf{e}^\top\tilde{\mathbf{z}}^{(k)}=\left(1+\frac{(d+1)\beta\gamma(2\alpha-1)}{(d+1)\beta+1}\right)^k\mathbf{e}^\top\mathbf{s}.$$

As we assume that $\alpha>1/2$, we are interested in finding the period for which,

$$(1+\gamma)\bar{\tilde{\mathbf{z}}}^{(k)}=(1+\gamma)\cdot\frac{\mathbf{e}^\top\tilde{\mathbf{z}}^{(k)}}{n}=1.$$

Now elementary calculations show that,

$$(1+\gamma)\cdot\frac{\mathbf{e}^\top\tilde{\mathbf{z}}^{(k)}}{n}=1 \iff k=\frac{\log\left(\frac{n}{\mathbf{e}^\top\mathbf{s}(1+\gamma)}\right)}{\log\left(1+\frac{(d+1)\gamma(2\alpha-1)}{(d+1)\beta+1}\right)}. \qquad \square$$

Next, let us consider multiple periods where the expressed opinion of M is truncated, i.e., $z_M = 1$. Here, even though $\alpha > 0.5$, $\mathbf{e}^\top\tilde{\mathbf{z}}$ might decrease as $z_{M'}$ is still $(1-\gamma)\bar{\mathbf{s}}^{(t+1)}$ but $z_{M'} = 1$ (instead of $(1+\gamma)\bar{\mathbf{s}}^{(t+1)}$). However, by setting $\alpha > 0.5$, $\hat{\mathbf{z}}^* = \tilde{\mathbf{z}}^{(t+1)}$ and $\mathbf{s} = \tilde{\mathbf{z}}^{(t)}$ in Corollary 3.6 we observe that the normalized average opinion cannot drop below $\frac{1}{(1+\gamma)^2}$.

Equally strong media sources. Next, we consider the case when $\alpha = \frac{1}{2}$, i.e., when both media sources are equally strong. We give a closed-form solution of the final expressed opinions after an infinite number of periods. Here, we denote the final expressed opinions by $\tilde{\mathbf{z}}^{(\infty)} = \lim_{t\to\infty}\tilde{\mathbf{z}}^{(t)}$. Furthermore, we let $\zeta^{(0)}$ denote the vector which contains the initial opinions of the sources that the nodes are connected to; more formally, we set $\zeta_i^{(0)} = s_M^{(0)}$ if node i is connected to source M and $\zeta_i^{(0)} = s_{M'}^{(0)}$ if node i is connected to source M'. Then we obtain the following result.

Proposition 4.2. *Let $G = (V, E, w)$ be a weighted d-regular graph and let $\alpha = \frac{1}{2}$. Then $\mathbf{e}^\top\tilde{\mathbf{z}}^{(t)} = \mathbf{e}^\top\tilde{\mathbf{s}}^{(0)}$ for all $t \geq 0$ and $\tilde{\mathbf{z}}^{(\infty)} = (\mathbf{I} + \frac{1}{\beta(1+d)}\mathbf{L})^{-1}\zeta^{(0)}$.*

Proposition 4.2 shows that when the media sources are equally strong, the sum of opinions does not change (as one might have expected). Surprisingly, the proposition also implies that $\tilde{\mathbf{z}}^{(\infty)}$ is solely dependent on the sources' initial opinions $\zeta^{(0)}$, the parameter β which determines the impact of the sources on the nodes, and the graph's adjacency matrix $\mathbf{W}$ (which determines the degree d and the Laplacian matrix $\mathbf{L}$). Interestingly, observe that if $\beta(d+1)$ is large, $\tilde{\mathbf{z}}^{(\infty)}$ is very close to $\zeta^{(0)}$.

Proof of Proposition 4.2. We first plug $\alpha = \frac{1}{2}$ into Corollary 3.4, and we notice that $\mathbf{e}^\top\hat{\mathbf{z}}^* = \mathbf{e}^\top\mathbf{s}$. As we always set $\mathbf{s}^{(t+1)}$ to the expressed equilibrium opinions at the end of period t, $\tilde{\mathbf{z}}^{(t)}$, it follows that for any period it holds that $\mathbf{e}^\top\tilde{\mathbf{z}}^{(t+1)} = \mathbf{e}^\top\tilde{\mathbf{z}}^{(t)}$.

Next, consider the t-th period. We set M's opinion to

$$z_M^{(t)} := (1+\gamma)\frac{\mathbf{e}^\top\hat{\mathbf{z}}^{(t)}}{n} = (1+\gamma)\frac{\mathbf{e}^\top\mathbf{s}}{n},$$

and M''s opinion to

$$z_{M'}^{(t)} := (1-\gamma)\frac{\mathbf{e}^\top\hat{\mathbf{z}}^{(t)}}{n} = (1-\gamma)\frac{\mathbf{e}^\top\mathbf{s}}{n}.$$

In other words, $z_M^{(t)}$ and $z_{M'}^{(t)}$ stay the same over all the periods.

Based on our observation, we use the fact that ζ always stays the same over all the periods and re-formulate the linear equation of Theorem 3.1 into the following nonhomogeneous first-order matrix difference equation:

$$\tilde{\mathbf{z}}^{(t+1)} = ((1+\beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1}(\tilde{\mathbf{z}}^{(t)} + \beta\,(\mathbf{I}+\mathbf{D})\,\zeta^{(0)}).$$

We again apply Lemma 2.2 to check whether $\tilde{\mathbf{z}}^{(\infty)} = \lim_{t\to\infty}\tilde{\mathbf{z}}^{(t)}$ exists and obtain the value.

To apply Lemma 2.2, we set $\mathbf{H} = ((1+\beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1}$, $\mathbf{c} = ((1+\beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1}\beta(\mathbf{I}+\mathbf{D})\zeta^{(0)}$, $\mathbf{M} = (1+\beta)\mathbf{I} + \beta\mathbf{D} + \mathbf{L}$, $\mathbf{N} = \mathbf{I}$, $\mathbf{b} = \beta(\mathbf{I}+\mathbf{D})\zeta^{(0)}$, and $\mathbf{A} = \beta\mathbf{I} + \beta\mathbf{D} + \mathbf{L}$.

We observe that both $\mathbf{A}$ and $\mathbf{M}$ are non-singular M-matrices, hence

$$\begin{aligned}\tilde{\mathbf{z}}^{(\infty)} &= \mathbf{A}^{-1}\mathbf{b}\\ &= (\beta\mathbf{I} + \beta\mathbf{D} + \mathbf{L})^{-1}\beta(\mathbf{I}+\mathbf{D})\zeta^{(0)}\\ &= (\mathbf{I} + (\beta(\mathbf{I}+\mathbf{D}))^{-1}\mathbf{L})^{-1}\zeta^{(0)}\\ &= \left(\mathbf{I} + \frac{1}{\beta(1+d)}\mathbf{L}\right)^{-1}\zeta^{(0)}.\end{aligned}$$

Notice that the last equality holds as the graph is a d-regular graph. □

5 Non-stubborn Media Sources

Next, we consider a single *non-stubborn* media source, i.e., now the media source participates in the FJ-dynamics like any other node. Formally, our model is as follows. Initially, we set the innate opinion and the expressed opinion of M to be the same, i.e., $s_M = z_M^{(0)} = (1+\gamma)\bar{\mathbf{s}}$, but now $z_M^{(t)}$ is updated based on Eq. (2.1). Intuitively, one would expect that here M influences the other nodes' opinions much less than in the stubborn setting. Indeed, we show that for any (possibly non-regular) graph, the sum of expressed opinions increases by *at most* a factor of $1 + \frac{1+\gamma}{n}$. This is in stark contrast to our discussion after Corollary 3.4, where we argued that for d-regular graphs with $d(\beta+1) \geq 1$, the sum of expressed opinions increases by *at least* a factor of $1 + \Omega(\gamma)$.

Proposition 5.1. *After the convergence of opinion dynamics, we obtain that $\mathbf{e}^\top\hat{\mathbf{z}}^* \leq \left(1 + \frac{1+\gamma}{n}\right)\mathbf{e}^\top\mathbf{s}$.*

Proof. By Lemma 2.1, the sum of the expressed opinions of all nodes, including the external source, is equal to the sum of innate opinions of all nodes, namely $\mathbf{e}^\top\hat{\mathbf{z}}^* + \hat{z}_M^* = \mathbf{e}^\top\mathbf{s} + s_M$. Re-arranging the equation, $\mathbf{e}^\top\hat{\mathbf{z}}^* = \mathbf{e}^\top\mathbf{s} + s_M - \hat{z}_M^*$. As $\hat{z}_M^* \geq 0$, an upper bound on the sum of expressed opinions is given by $\mathbf{e}^\top\hat{\mathbf{z}}^* \leq \mathbf{e}^\top\mathbf{s} + s_M$. Plugging $s_M \leq (1+\gamma)\mathbf{e}^\top\mathbf{s}/n$ into the formula, we obtain

$$\mathbf{e}^\top\hat{\mathbf{z}}^* \leq \mathbf{e}^\top\mathbf{s} + (1+\gamma)\mathbf{e}^\top\mathbf{s}/n = \left(1 + \frac{1+\gamma}{n}\right)\mathbf{e}^\top\mathbf{s}. \qquad \square$$

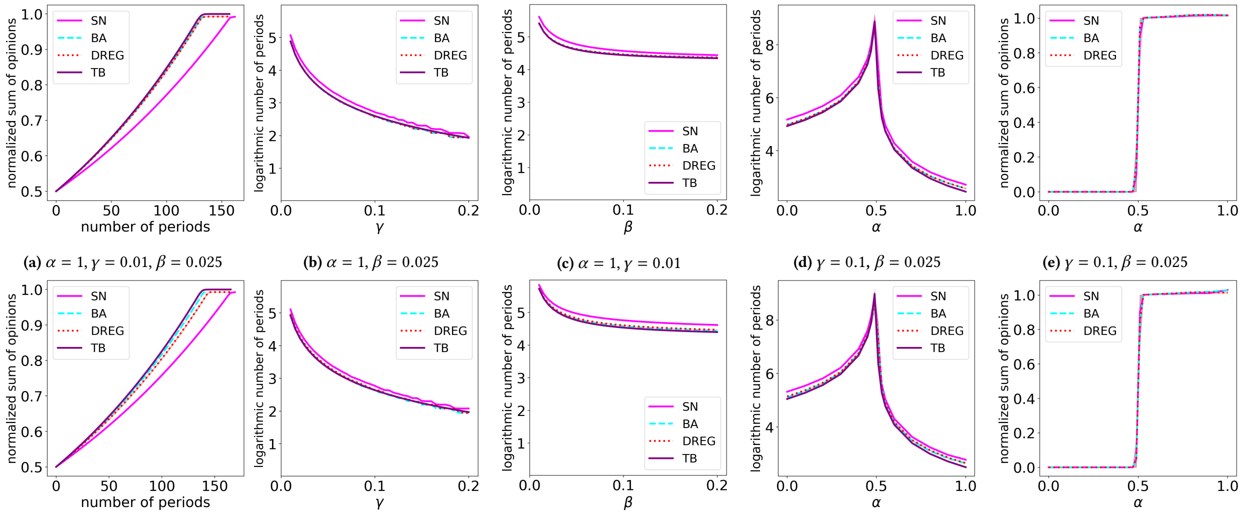
6 Experimental Results

Next, we run experiments to validate how well our theoretical bounds match the behavior in on real-world graph data and synthetic graph models.

6.1 Setup

Real-world Networks. For our experiments, we use publicly available Social Network (SN) data from [26]. Our experiments were conducted on Facebook SN (4039 nodes and 88234 edges) and Wikipedia SN (7115 nodes and 103689 edges) datasets; we abbreviate them as FB and WK, respectively.

Synthetic Graphs. We also conducted experiments on synthetic graphs, namely Barabási-Albert (BA) graphs and d-regular random graphs (DREG), i.e., random graphs with a uniform distribution



(a) $\alpha = 1, \gamma = 0.01, \beta = 0.025$ **(b)** $\alpha = 1, \beta = 0.025$ **(c)** $\alpha = 1, \gamma = 0.01$ **(d)** $\gamma = 0.1, \beta = 0.025$ **(e)** $\gamma = 0.1, \beta = 0.025$

(f) $\alpha = 1, \gamma = 0.01, \beta = 0.035$ **(g)** $\alpha = 1, \beta = 0.0355$ **(h)** $\alpha = 1,\gamma = 0.01$ **(i)** $\gamma = 0.1, \beta = 0.035$ **(j)** $\gamma = 0.1, \beta = 0.035$

Figure 2: In $(a), (b), (c), (d)$ and (e) we consider the FB SN and the BA and DREG graph with comparable parameters to FB. In $(f), (g), (h), (i)$ and (j) we consider the WK SN and the BA and DREG graph with comparable parameters to WK. In (a) and (f) the normalized sum of expressed opinions over multiple periods is depicted. The logarithmic number of periods to reach normalized average opinion $1/(1+\gamma)$ is depicted in (b) and (g) for different values of γ and in (c) and (h) for different values of β. In (d) and (i) the logarithmic number of periods to reach normalized average opinion $1/(1+\gamma)$ or $\epsilon := 10/n$ for different values of α are shown. Lastly, in (e) and (j) the final normalized sum of opinions at the end of the process for different values of α are given. The graph TB depicts the theoretical bound from Corollary 3.4 in (a) and (f) and Proposition 4.1 for $(b), (g), (c), (h)$ and (d) and (i).

over all d-regular graphs on n nodes. We include the BA graph as it is a model to simulate real-world SNs and the DREG since we provide theoretical results, particularly for regular graphs. The parameters in these synthetic graphs were chosen such that they are comparable to the real-world SNs, i.e., such that the (expected) number of nodes/edges is the same as in the aforementioned real-world networks. For example, the DREG graph comparable to the FB SN has 4039 nodes and degree $44 \approx (2 \times 88234)/4039$. All edges in these networks are of weight 1, except from edges connected to the media source (edge (i, M) has weight $\beta(1 + d_i)$ similarly as in the theoretical setup).

Innate Opinions. In our experiments, the innate opinions are chosen from a Gaussian distribution with mean 0.5 and variance 0.2. We choose the mean in this way such that on average the innate opinions are not biased towards 0 or 1. There is nothing unique about our choice for the variance, and our results would hold for different values of the variance as well.

Theoretical Bound. When reporting our results, we sometimes include a plot corresponding to one of our theoretical results; we denote this plot by Theoretical Bound (TB).

Implementation. To compute $\hat{\mathbf{z}}^*$ as in Theorem 3.1, we rely on the algorithm of [21] and its implementation in Laplacians.jl. To generate the synthetic graphs, we rely on the implementations in [17]. Furthermore, our experiments are implemented in Julia and our code is available in the supplementary material.

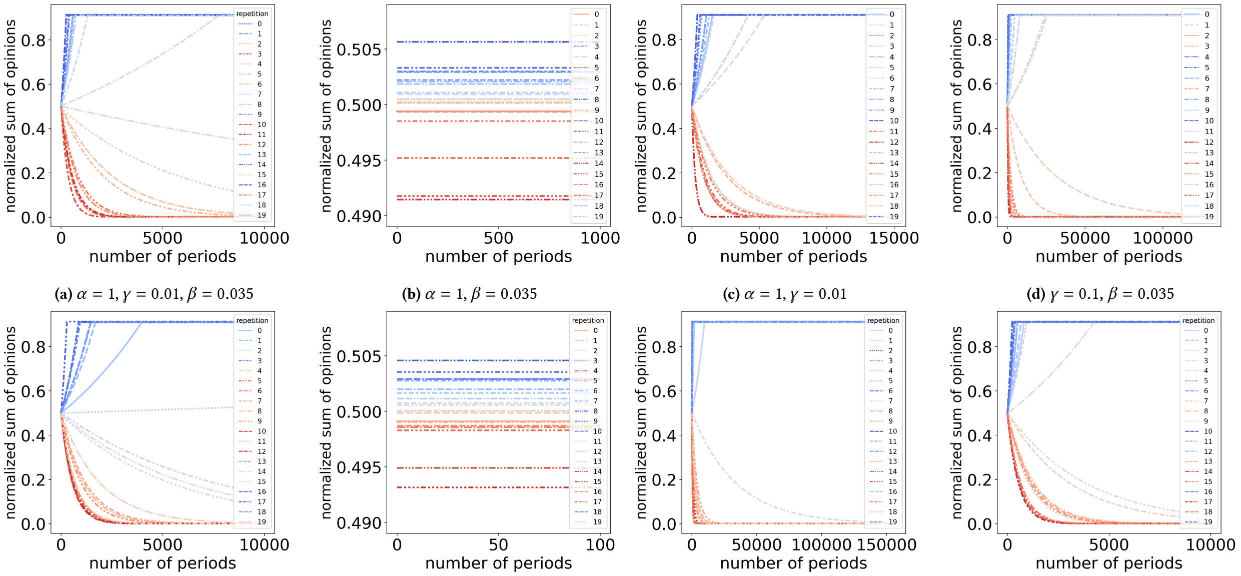
Repetitions. Each experiment is repeated 20 times. In the plots in Fig. 2 the average output with confidence intervals are depicted (note that the repetitions are highly concentrated). In the plots in Fig. 3, each one of the 20 repetitions is plotted individually.

6.2 Findings

Results for a Single Media Source. In line with Theorem 3.3, Fig. 2a and Fig. 2f show that when all nodes are connected to a single source M ($\alpha = 1$) the normalized sum of opinions converges to a value arbitrarily close to 1. Interestingly, the BA graph also follows the theoretical bound for DREG (depicted in purple) quite closely in both networks. Fig. 2b, Fig. 2g and Fig. 2c Fig. 2h indicate that both bias parameter β and external influence magnitude γ are negatively correlated with the number of periods to reach normalized average opinion $1/(1 + \gamma)$ (see also Proposition 4.1). However, the dependence on β appears to be less significant, as we previously observed in Corollary 3.4 for the regular graph.

Results for Multiple Media Sources. In Fig. 2d and Fig. 2i the logarithm of the number of periods it takes to reach normalized average opinion $1/(1 + \gamma)$ for different values of α is depicted. Consistent with Proposition 4.1, we observe that the process takes the longest for values of α close to 0.5 in both networks. The asymmetry in the plot is due to the nature of our model (from $n/2$, it takes less time to reach $n/(1 + \gamma)$ by increasing with a factor of $(1 + \gamma)$ than to reach a very small constant by decreasing by a factor of $1 - \gamma$). In Fig. 2e and Fig. 2j we observe that the normalized sum of opinions converges for values of $\alpha < 0.5$ to 0 and for $\alpha > 0.5$ to 1, as previously proven for the regular graphs in Corollary 3.4.

Next, we further investigate the behavior at the threshold value $\alpha = 0.5$. As the number of nodes in FB and WK SN are 4039 and



(a) $\alpha = 1, \gamma = 0.01, \beta = 0.035$

(b) $\alpha = 1, \beta = 0.035$

(c) $\alpha = 1, \gamma = 0.01$

(d) $\gamma = 0.1, \beta = 0.035$

(e) $\alpha = 0.5, \gamma = 0.1, \beta = 0.5$

(f) $\alpha = 0.5, \gamma = 0.1, \beta = 0.5$

(g) $\alpha \approx 0.5, \gamma = 0.1, \beta = 0.5$

(h) $\alpha \approx 0.5, \gamma = 0.1, \beta = 0.5$

Figure 3: In all the above figures, we plot the normalized sum of opinions when $\alpha = 0.5$ or $\alpha \approx 0.5$. Each one of the 20 repetitions is plotted individually. In (a) and (e) we consider the FB and WK SN respectively with a randomly chosen node removed. Figures (b) and (f) correspond to DREG graphs with $n = 4038$, $d = 44$ and $n = 7114$, $d = 30$ respectively (note that these are comparable parameters to the FB and TW SN with a single node removed, which is the same across repetitions). In (c) and (g) we depict the FB and WK SN respectively. Lastly, in (d) and (h) DREG graphs with comparable parameters to FB and WK SN respectively are shown.

7115 respectively, and thus odd (so α cannot be exactly 0.5 in this case), we delete one node from these graphs uniformly at random to achieve $\alpha = 0.5$. Moreover, we generate DREG graphs on 4038 nodes with $d = 44$ (the same average degree as FB SN) and 7114 and $d = 30$ (same average degree as WK SN).

The normalized sum of opinions over multiple periods for these graphs are depicted in Fig. 3a, Fig. 3e (SN's with a randomly chosen deleted node) and Fig. 3b, Fig. 3f (DREG graphs with comparable parameters to FB and WK SN with a single removed node). We observe that in line with Proposition 4.2, the sum of expressed opinions stays the same in the DREG graphs on 4038 and 7114 nodes but on the FB and WK SN converges to either 0 (this happens in iterations whose graphs are colored red in Fig. 3a, Fig. 3e) or to 1 (this happens in iterations whose graphs are colored blue in Fig. 3a and Fig. 3e). Lastly, we perform the same experiment on the actual FB and WK SN and DREG graphs with comparable parameters to FB and WK SN as depicted in Fig. 3c and Fig. 3g and Fig. 3d, Fig. 3h. Here, due to the odd number of nodes of FB and WK SN, either $\alpha = 0.4999$ or $\alpha = 0.5001$, both with probability 0.5. We thus observe in Fig. 3d that by adding a single node to the DREG graph the final set of opinions also becomes radicalized, even though it takes many periods. This indicates that even a very small difference in the power of the external media source has a significant impact on the final opinion configuration.

Summary. Our experiments have shown that our simulations results and the bounds that we obtained theoretically match very well. This is interesting since some of our theoretical results were mostly derived for d-regular graphs, and the real-world networks are not regular. This empirically indicates that our theoretical results transfer to real-world graphs.

7 Conclusions

In this paper, our goal was to obtain a mathematical understanding of how external sources impact the opinion formation process in social networks. To study this formally, we proposed a generalized version of the popular FJ model and derived analytic bounds on the power of the external sources. Several of our bounds are tight for regular graphs, and we showed experimentally that our theoretical bounds closely match simulation results on real-world datasets.

In the future, it would be interesting to study the more general setup, where an external source can connect to only a given number of nodes and aims to optimize a specific objective, such as maximizing/minimizing the final sum of opinions. Another potential avenue for future research is investigating the impact of external sources on polarization in the network, especially when two external sources attempt to pull the opinions in opposite directions.

Acknowledgments

This research has been funded by the Vienna Science and Technology Fund (WWTF) [Grant ID: 10.47379/VRG23013], the ERC Advanced Grant REBOUND (834862), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Rediet Abebe, Jon Kleinberg, David Parkes, and Charalampos E Tsourakakis. 2018. Opinion dynamics with varying susceptibility to persuasion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1089–1098.

[2] Aris Anagnostopoulos, Luca Becchetti, Emilio Cruciani, Francesco Pasquale, and Sara Rizzo. 2022. Biased opinion dynamics: when the devil is in the details. *Information Sciences* 593 (2022), 49–63.

[3] Vincenzo Auletta, Ioannis Caragiannis, Diodato Ferraioli, Clemente Galdi, and Giuseppe Persiano. 2017. Information retention in heterogeneous majority dynamics. In *Web and Internet Economics: 13th International Conference, WINE 2017, Bangalore, India, December 17–20, 2017, Proceedings 13*. Springer, 30–43.

[4] Vincenzo Auletta, Antonio Coppola, and Diodato Ferraioli. 2021. On the Impact of Social Media Recommendations on Opinion Consensus. In *International Conference of the Italian Association for Artificial Intelligence*. Springer, 263–278.

[5] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[6] Petra Berenbrink, George Giakkoupis, Anne-Marie Kermarrec, and Frederik Mallmann-Trenn. 2016. Bounds on the Voter Model in Dynamic Networks. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

[7] Abraham Berman and Robert J Plemmons. 1994. *Nonnegative matrices in the mathematical sciences*. SIAM.

[8] Nikita Bhalla, Adam Lechowicz, and Cameron Musco. 2023. Local Edge Dynamics and Opinion Polarization. In *WSDM*. ACM, 6–14.

[9] Maxwell T Boykoff and Jules M Boykoff. 2004. Balance as bias: Global warming and the US prestige press. *Global environmental change* 14, 2 (2004), 125–136.

[10] Markus Brill, Edith Elkind, Ulle Endriss, and Umberto Grandi. 2016. Pairwise diffusion of preference rankings in social networks. (2016).

[11] Ozan Candogan, Nicole Immorlica, Bar Light, and Jerry Anunrojwong. 2022. Social learning under platform influence: Consensus and persistent disagreement. *arXiv preprint arXiv:2202.12453* (2022).

[12] Mayee F Chen and Miklos Z Racz. 2020. Network disruption: maximizing disagreement and polarization in social networks. *arXiv preprint arXiv:2003.08377* (2020).

[13] Uthsav Chitra and Christopher Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 115–123.

[14] Federico Corò, Emilio Cruciani, Gianlorenzo D'Angelo, and Stefano Ponziani. 2022. Exploiting social influence to control elections based on positional scoring rules. *Information and Computation* 289 (2022), 104940.

[15] Nuno Crokidakis. 2012. Effects of mass media on opinion spreading in the Sznajd sociophysics model. *Physica A: Statistical Mechanics and its Applications* 391, 4 (2012), 1729–1734.

[16] Morris H. DeGroot. 1974. Reaching a Consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121. http://www.jstor.org/stable/2285509

[17] James Fairbanks, Mathieu Besançon, Schölly Simon, Júlio Hoffiman, Nick Eubank, and Stefan Karpinski. 2021. JuliaGraphs/Graphs.jl: an optimized graphs package for the Julia programming language. https://github.com/JuliaGraphs/Graphs.jl/

[18] Noah E Friedkin and Eugene C Johnsen. 1990. Social influence and opinions. *Journal of Mathematical Sociology* 15, 3-4 (1990), 193–206.

[19] Jason Gaitonde, Jon Kleinberg, and Eva Tardos. 2020. Adversarial perturbations of opinion dynamics in networks. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 471–472.

[20] Serge Galam and Frans Jacobs. 2007. The role of inflexible minorities in the breaking of democratic opinion dynamics. *Physica A: Statistical Mechanics and its Applications* 381 (2007), 366–376.

[21] Yuan Gao, Rasmus Kyng, and Daniel A Spielman. 2023. Robust and Practical Solution of Laplacian Equations by Approximate Elimination. *arXiv preprint arXiv:2303.00709* (2023).

[22] Aristides Gionis, Evimaria Terzi, and Panayiotis Tsaparas. 2013. Opinion maximization in social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 387–395.

[23] Umberto Grandi, Lawqueen Kanesh, Grzegorz Lisowski, Ramanujan Sridharan, and Paolo Turrini. 2023. Identifying and eliminating majority illusion in social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5062–5069.

[24] Dariusz Kalan. 2019. Poland's State of the Media. https://foreignpolicy.com/2019/11/25/poland-public-television-law-and-justice-pis-mouthpiece/. Accessed: 2024-01-12.

[25] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

[26] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

[27] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. 2017. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery* 31 (2017), 1480–1505.

[28] Mauro Mobilia, Anna Petersen, and Sidney Redner. 2007. On the role of zealotry in the voter model. *Journal of Statistical Mechanics: Theory and Experiment* 2007, 08 (2007), P08029.

[29] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. 2018. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 world wide web conference*. 369–378.

[30] Roni Muslim, Rinto Anugraha Nqz, and Muhammad Ardhi Khalif. 2024. Mass media and its impact on opinion dynamics of the nonlinear q-voter model. *Physica A: Statistical Mechanics and its Applications* 633 (2024), 129358.

[31] Seth A Myers, Chenguang Zhu, and Jure Leskovec. 2012. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 33–41.

[32] Kamyar Nazeri. 2018. Effect of Social Media on Opinion Formation. arXiv:1805.08310 [physics.soc-ph]

[33] Petros Petsinis, Andreas Pavlogiannis, and Panagiotis Karras. 2023. Maximizing the probability of fixation in the positional voter model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12269–12277.

[34] M. Pineda and G.M. Buendía. 2015. Mass media and heterogeneous bounds of confidence in continuous opinion dynamics. *Physica A: Statistical Mechanics and its Applications* 420 (2015), 73–84. https://doi.org/10.1016/j.physa.2014.10.089

[35] Hegselmann Rainer and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: models, analysis and simulation. (2002).

[36] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review* 34, 3 (2017), 150–160.

[37] Katarzyna Sznajd-Weron and Jozef Sznajd. 2000. Opinion evolution in closed community. *International Journal of Modern Physics C* 11, 06 (2000), 1157–1165.

[38] Christopher Tran and Elena Zheleva. 2022. Heterogeneous peer effects in the linear threshold model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4175–4183.

[39] Sijing Tu, Stefan Neumann, and Aristides Gionis. 2023. Adversaries with Limited Information in the Friedkin-Johnsen Model. In *KDD*. ACM, 2201–2210. https://doi.org/10.1145/3580305.3599255

[40] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17 (2007), 395–416.

[41] Bryan Wilder and Yevgeniy Vorobeychik. 2018. Controlling Elections through Social Influence. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 265–273.

[42] Ahad N Zehmakan. 2021. Majority opinion diffusion in social networks: An adversarial approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5611–5619.

[43] Liwang Zhu, Qi Bao, and Zhongzhi Zhang. 2021. Minimizing polarization and disagreement in social networks via link recommendation. *Advances in Neural Information Processing Systems* 34 (2021), 2072–2084.