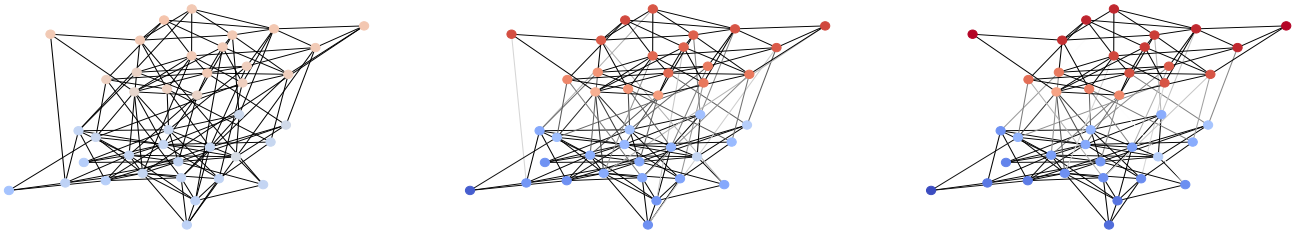
Understanding Filter Bubbles and Polarization in Social Networks

Uthsav Chitra
uchitra@cs.princeton.edu
Princeton University, Dept. of Computer Science

Christopher Musco
cmusco@cs.princeton.edu
Princeton University, Dept. of Computer Science



(a) Example synthetic social network graph.

(b) Graph after network administrator changes just 20% of edge weight.

(c) Graph after network administrator changes just 30% of edge weight.

Figure 1: Social network graphs after converging to equilibrium in the Friedkin-Johnsen opinion dynamics model. Node colors represent the opinions of individuals on an issue: dark red nodes have opinion close to 1, while dark blue nodes have opinion close to −1. The weight of an edge (i.e, strength of social connection between two individuals) is expressed by its shade. In the middle and right networks, we introduce a *network administrator* who is allowed to make small changes to the network, and is incentivized to connect users with content that is similar to their opinion. After reweighting edges by just a small amount (i.e. filtering social content), the network administrator's actions increase a standard measure of opinion polarization in these graphs by 180% and 260%, respectively. This illustrates the formation of a "filter bubble" in a social network.

ABSTRACT

Recent studies suggest that social media usage — while linked to an increased diversity of information and perspectives for users — has exacerbated user polarization on many issues. A popular theory for this phenomenon centers on the concept of "filter bubbles": by automatically recommending content that a user is likely to agree with, social network algorithms create echo chambers of similarly-minded users that would not have arisen otherwise [55]. However, while echo chambers have been observed in real-world networks, the evidence for filter bubbles is largely post-hoc.

In this work, we develop a mathematical framework to study the filter bubble theory. We modify the classic Friedkin-Johnsen opinion dynamics model by introducing another actor, a *network administrator*, who filters user content by making small changes to the edge weights of a social network (for example, adjusting a news feed algorithm to change the level of interaction between users).

On real-world networks from Reddit and Twitter, we show that when the network administrator is incentivized to reduce disagreement among users, even relatively small edge changes can result in the formation of echo chambers in the network and increase user polarization. We theoretically support this observed sensitivity of social networks to outside intervention by analyzing synthetic graphs generated from the *stochastic block model*. Finally, we show that a slight modification to the incentives of the network administrator can mitigate the filter bubble effect while minimally affecting the administrator's target objective, user disagreement.

WISDOM'19, August 4, 2019, Anchorage, Alaska

1 INTRODUCTION

The past decade has seen an explosion in social media use and importance [59]. Online social networks, which enable users to instantly broadcast information about their lives and opinions to a large audience, are used by billions of people worldwide. Social media is also used to access news [57], review products and restaurants, find health and wellness recommendations [60], and more.

Social networks, along with the world wide web in general, have made our world more connected. It has been widely established that social networks and online media increase the diversity of information and opinions that individuals are exposed to [15, 45, 47]. In many ways, the widespread adoption of online social networks has resulted in significant positive progress towards fulfilling Facebook's mission of "bringing the world closer together".

1.1 The puzzle of polarization

Surprisingly, while they enable access to a diverse array of information, social networks have also been widely associated with *increased polarization* in society across many issues [31], including politics [3, 6, 21], science [51], and healthcare [41]. Somehow, despite the exposure to a wide variety of opinions and perspectives,

individuals form polarized clusters, unable to reach consensus with one another. In politics, increased polarization has been blamed for legislative deadlock, erratic policies, and decreased trust and engagement in the democratic process [12, 46].

There have been many efforts to understand this seemingly counterintuitive phenomenon of increased societal polarization. Classical psychological theory asserts that polarization arises from "biased assimilation" [48], i.e. individuals are more likely to trust and share information that already aligns with their views. Isolated examples of intense polarization, such as the 2016 US presidential election and Brexit, can be partially explained by historical, cultural, and ideological factors [37]. However, when such examples are considered in bulk, it becomes clear that changes in social dynamics arising from the increased use of social media must constitute a major contributing factor to the phenomenon of polarization [6, 43].

1.2 Filter bubbles

An influential idea put forward by Eli Pariser suggests an important mechanism that may explain why social media increases societal polarization [55]. According to Pariser, preferential attention to viewpoints similar to those already held by an individual is *explicitly encouraged* by social media companies: to increase metrics like engagement and ad revenue, recommendation systems connect users with information already similar to their current beliefs.

Such recommendations can be direct: friend or follow suggestions on platforms like Facebook or Twitter. Or they can be more subtle: chronological "news feeds" on social media have universally been replaced with individually filtered and sorted feeds which connect users with posts that they are most likely to engage with [30]. By recommending such content, social network companies create "echo chambers" of similar-minded users. Owing to their root cause – the external filtering of content shown to a user – Pariser called these echo chambers *filter bubbles*.

The danger of filter bubbles was recently highlighted by Apple CEO Tim Cook in a commencement speech at Tulane University [28]. Filter bubbles have been blamed for the spread of fake news during the Brexit referendum and the 2016 U.S. presidential election [43], protests against immigration in Europe [35], and even measles outbreaks in 2014 and 2015 [41]. In each of these incidents, instead of bringing diverse groups of users together, social media has reinforced differences between groups and wedged them apart.

At least . . . that's the theory. While Pariser's ideas make logical sense, the magnitude of the "filter bubble effect" has been disputed or questioned for lack of evidence [7, 14, 42, 54, 62, 65].

1.3 Our contributions

The goal of this paper is to better understand filter bubbles, and ultimately, to place Pariser's theory on firmer ground. We do so by developing a mathematical framework for studying the effect of filter bubbles on polarization in social networks, relying on well-established analytical models for *opinion dynamics* [23].

Such models provide simple rules that capture how opinions form and propagate in a social network. The network itself is typically modeled as a weighted graph: nodes are individuals and social connections are represented by edges, with higher weight for relationships with increased interaction. We specifically work with the well-studied Friedkin-Johnsen opinion dynamics model, which models an individual's opinion on an issue as a continuous value between -1 and 1, and assumes that, as time progresses, individuals update their opinions based on the average opinion of their social connections [32]. The Friedkin-Johnsen model has been used successfully to study polarization in social networks [11, 18, 19, 53].

Our contribution is to modify the model by adding an external force: a **network administrator** who filters social interaction between users. Based on modern recommendation systems [4], the network administrator makes small changes to edge weights in the network, which correspond to slightly increasing or decreasing interaction between specific individuals (e.g. by tuning a news feed algorithm). The administrator's goal is to connect users with content they likely agree with, and therefore increase user engagement. Formally, we model this goal by assuming that the network administrator seeks to minimize a standard measure of *disagreement* in the social network. As individuals update their opinions according to the Friedkin-Johnsen dynamics, the administrator repeatedly adjusts the underlying network graph to achieve its own goal.

Using our model, we establish a number of experimental and theoretical results which suggest that content filtering by a network administrator can significantly increase polarization, even when changes to the network are highly constrained (and perhaps unnoticeable by users). First, we apply our augmented opinion dynamics to real-world social networks obtained from Twitter and Reddit. When the network administrator changes only 40% of the edge weight in the network, polarization increases by more than a factor of 40×. These results are striking—they suggest that social networks are very sensitive to influence by filtering. As illustrated in Figure 1, even minor content filtering by the network administrator can create significant "filter bubbles", just as Pariser predicted [55].

Next, to better understand the sensitivity of social networks to filtering, we study a standard generative model for social networks: the stochastic block model [1]. We show that, with high probability, any network generated from the stochastic block model is in a state of **fragile consensus**: that is, under the Friedkin-Johnsen dynamics, the network will exhibit low polarization, but can become highly polarized after only a minor adjustment of edge weights. Our findings give theoretical justification for why a network administrator can greatly increase polarization in real-world networks.

Finally, ending on an optimistic note, we experimentally show that a simple modification to the incentives of the network administrator mitigates the filter bubble effect. Surprisingly, our proposed solution also minimally affects user disagreement, the objective of the network administrator by at most 5%.

1.4 Prior work

Minimizing polarization in social networks. There has been substantial recent work on using opinion dynamics models to study polarization in social networks. [50] first defines polarization in the Friedkin-Johnsen model, and gives an algorithm for reducing polarization in social networks. [53] and [19] give methods for finding network structures which minimize different functions involving polarization and disagreement. Our work differs from these results in that we study network modifications which *increase* polarization, rather than decreasing it. Moreover, we study how such modifications arise even when the network administrator is not explicitly incentivized to change polarization.

Other opinion dynamics models and metrics have also been used to study network polarization. [5] gives an algorithm for mitigating filter bubbles in an influence maximization setting. [34] studies "controversy" in the Friedkin-Johnsen model, a metric related to polarization, and [33] gives an algorithm for reducing controversy.

Modeling filter bubbles and recommendation systems. Biased assimilation, which is when users gravitate towards viewpoints similar to their own, has been argued as one cause of increased polarization in social networks. By generalizing the classic DeGroot model [26] of opinion formation, [22] provides theoretical support for the biased assimilation phenomenon and analyzes the interaction of three recommendation systems on biased assimilation. [27] models biased assimilation in social networks using a variant of the Bounded Confidence Model (BCM) [38], an opinion dynamics model that does not assume a latent graph structure between users. Most similar to our work, [35] creates a variant of the BCM that models biased assimilation, homophily, and algorithmic filtering, and shows how echo chambers can arise as a result of these factors. [17] studies the more general problem of how recommendation systems increase homogeneity of user behavior.

1.5 Notation and Preliminaries

We use bold letters to denote vectors. The i^{th} entry of vector $\mathbf{a}$ is denoted a_i. For a matrix A, A_{ij} is the entry in the i^{th} row and j^{th} column. For a vector $\mathbf{a} \in \mathbb{R}^n$, let diag($\mathbf{a}$) return an $n \times n$ diagonal matrix with the i^{th} diagonal entry equal to a_i. For a matrix $A \in \mathbb{R}^{n \times d}$, let rowsum($A$) return a vector whose i^{th} entry is equal to the sum of all entries in A's i^{th} row. We use $I_{n \times n}$ to denote a dimension n identity matrix, and $\mathbf{1}_n$ to denote the all ones column vector, with the subscript omitted when dimension is clear from context.

Every real symmetric matrix $A \in \mathbb{R}^{n \times n}$ has an orthogonal eigendecomposition $A = U \Lambda U^T$ where $U \in \mathbb{R}^{n \times n}$ is orthonormal (i.e $U^T U = U U^T = I$) and Λ is diagonal, with real valued entries $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ equal to A's eigenvalues. We say a symmetric matrix is positive semidefinite (PSD) is all of its eigenvalues are non-negative (i.e. $\lambda_1 \geq 0$). We use $\preceq$ to denote the standard Loewner ordering: $M \succeq N$ indicates that $M - N$ is PSD. For a square matrix M, $\|M\|_2$ denotes the spectral norm of M and $\|M\|_F$ denotes the Frobenius norm. For a vector v, $||v||_2$ denotes the L^2 norm.

1.6 Road Map

Section 2 Introduce preliminaries on Freidkin-Johnsen opinion dynamics, which form a basis for modeling filter bubble emergence.
Section 3 Introduce our central "network administrator dynamics" and establish experimentally that content filtering can significantly increase polarization in social networks.
Section 4 Explore these findings theoretically by showing that stochastic block model graphs exhibit a "fragile consensus" which is easily disrupted by outside influence.
Section 5 Discuss a small modification to the content filtering process that can mitigate the effect of filter bubbles while still being beneficial for the network administrator.
Section 6 Briefly discuss future directions of study.

2 MODELING OPINION FORMATION

One productive approach towards understanding the dynamics of consensus and polarization in social networks has been to develop simple mathematical models to explain how information and ideas spread in these networks.

While there are a variety of models in the literature, we use the Friedkin-Johnsen opinion dynamics model, which has been used to study polarization in recent work [19, 50, 53].

2.1 Friedkin-Johnsen Dynamics

Concretely, the Friedkin-Johnsen (FJ) dynamics applies to any social network that can be modeled as an undirected, weighted graph G. Let $\{v_1, \ldots, v_n\}$ denote G's nodes and for all $i \neq j$, let $w_{ij} \geq 0$ denote the weight of undirected edge (i, j) between nodes v_i and v_j. Let $d_i = \sum_{j \neq i} w_{ij}$ be the degree of node v_i.

The FJ dynamics model the propagation of an opinion on an issue during a discrete set of time steps $t \in 0, 1, \ldots, T$. The issue may be specific (Do you believe that humans contribute to climate change?) or it may encode a broad ideology (Do your political views align most with conservative or liberal politicians in the US?).

In either case, the FJ dynamics assume that the issue has exactly two poles, with an individual's opinion encoded by a continuous real value in $[-1, 1]$. -1 and 1 represent the most extreme opinions in either direction, while 0 represents a neutral opinion. Each node v_i holds an *innate* (or internal) opinion $s_i \in [-1, 1]$ on the issue. The innate opinion vector $\mathbf{s} = [s_1, \ldots, s_n]$ does not change over time. It can be viewed as the opinion an individual would hold in a social vacuum, with no outside influence from others. The value of s_i might depend on the background, geographic location, religion, race, or other circumstances about individual i.

In addition to an innate opinion, for every time t, each node is associated with an "expressed" or "current" opinion $z_i^{(t)} \in [-1, 1]$, which changes over time. Specifically, the FJ dynamics evolves according to the update rule:

$$z_i^{(t)} = \frac{s_i + \sum_{j \neq i} w_{ij} z_j^{(t-1)}}{d_i + 1}. \tag{1}$$

That is, at each time step, each node adopts a new expressed opinion which is the average of its own innate opinion and the opinion of its neighbors. For a given graph G and innate opinion vector $\mathbf{s}$, it is well known that the FJ dynamics converges to an equilibrium set of opinions [11], which we denote

$$\mathbf{z}^* = \lim_{t \to \infty} \mathbf{z}^{(t)}.$$

It will be helpful to express the FJ dynamics in a linear algebraic way. Let $A \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G, with $A_{ij} = A_{ji} = w_{ij}$ and let D be a diagonal matrix with $D_{ii} = d_i$. Let $L = D - A$ be the graph Laplacian of G. Then we can see that (1) is equivalent to

$$\mathbf{z}^{(t)} = (D + I)^{-1}(A\mathbf{z}^{(t-1)} + \mathbf{s}), \tag{2}$$

where we denote $\mathbf{z}^{(t)} = [z_1^{(t)}, \ldots, z_n^{(t)}]$. From this expression, it is not hard to check that

$$\mathbf{z}^* = (L + I)^{-1}\mathbf{s}. \tag{3}$$

Alternative Models. The Friedkin-Johnsen opinion dynamics model is a variation of DeGroot's classical model for consensus formation in social network [26]. The distinguishing characteristic of the FJ model is the addition of the *innate opinions* encoded in $\mathbf{s}$. Unlike the DeGroot model, which always converges in a consensus when

G is connected (i.e., $z_i^* = z_j^*$ for all i, j) , innate opinions allow for a richer set of equilibrium opinions. In particular, $\mathbf{z}^*$ will typically contain opinions ranging continuously between -1 and 1.

Compared to DeGroot, the FJ dynamics more accurately model a world where an individual's opinion (e.g. on a political issue) is not shaped solely by social influence, but also by an individual's particular background, beliefs, or life circumstances. FJ dynamics are often studied in economics and game theory as an example of a game with price of anarchy greater than one [11]. Other variations on the model include additional variables [39] — for example, allowing the "stubbornness" of an individual to vary [2, 18], or adding additional terms to Equation (1) to indicate that an individual cares about the average network opinion as well as their neighbors' opinions [29].

There also exist many models for opinion formation that fall outside of DeGroot's original framework. Several models involve *discrete* instead of continuously valued opinions. We refer to reader to the overview and discussion of different proposals in [23]. In this paper, we focus on the original FJ dynamics, which are already rich enough to provide several interesting insights on the dynamics of polarization, filter bubbles, and echo chambers.

2.2 Polarization, Disagreement, and Internal Conflict

The fact that $\mathbf{z}^*$ does not always contain a single consensus opinion makes the FJ model suited to understanding how polarization arises on specific issues. Formally, we define polarization as the variance of a given set of opinions.

Definition 2.1 (Polarization, $\mathcal{P}_{\mathbf{z}}$)**.** For a vector of n opinions $\mathbf{z} \in [-1, 1]^n$, let $\text{mean}(\mathbf{z}) = \frac{1}{n}\sum_{j=1}^n z_j$ be the mean opinion in $\mathbf{z}$.

$$\mathcal{P}_{\mathbf{z}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} (z_i - \text{mean}(\mathbf{z}))^2.$$

$\mathcal{P}_{\mathbf{z}}$ ranges between 0 when all opinions are equal and n when half of the opinions in $\mathbf{z}$ equal 1 and half equal -1. $\mathcal{P}_{\mathbf{z}}$ was first proposed as a measure of polarization in [50], and has since been used in other recent work studying polarization in FJ dynamics [19, 53]. While we focus on Definition 2.1, we refer the interested reader to [34] for discussion of alternative measures of polarization.

Under the FJ model, the polarization of the equilibrium set of opinions has a simple closed form. In particular, let $\overline{\mathbf{s}} = \mathbf{s} - \mathbf{1}\cdot\text{mean}(\mathbf{s})$ be the mean centered set of innate opinions on a topic, and define $\overline{\mathbf{z}}$ similarly. Using that $\mathbf{1}$ is in the null-space of any graph Laplacian L, it is easy to check (see [53] for details) that $\text{mean}(\mathbf{z}) = \text{mean}(\mathbf{s})$ and thus $\overline{\mathbf{z}}^* = (L + I)^{-1}\overline{\mathbf{s}}$. It follows that:

$$\mathcal{P}_{\mathbf{z}^*} = \overline{\mathbf{s}}^T (L + I)^{-2} \overline{\mathbf{s}}. \tag{4}$$

In addition to polarization, we define two other quantities of interest involving opinions in a social network. Both have appeared repeatedly in studies involving the FJ dynamics [2, 19, 53].

The first quantity measures how much node i's opinion differs from those of its neighbors.

Definition 2.2 (Local Disagreement, $\mathcal{D}_{G,\mathbf{z},i}$)**.** For $i \in 1, \ldots, n$, a vector of opinions $\mathbf{z} \in [-1, 1]^n$, and social network graph G,

$$\mathcal{D}_{G,\mathbf{z},i} \stackrel{\text{def}}{=} \sum_{j \in 1,\ldots n, j \neq i} w_{ij}(z_i - z_j)^2.$$

We also define an aggregate measure of disagreement.

Definition 2.3 (Global Disagreement, $\mathcal{D}_{G,\mathbf{z}}$)**.** For a vector of opinions $\mathbf{z} \in [-1, 1]^n$, and social network graph G,

$$\mathcal{D}_{G,\mathbf{z}} \stackrel{\text{def}}{=} \frac{1}{2} \cdot \sum_{i=1}^{n} \mathcal{D}_{G,\mathbf{z},i}.$$

The factor of $1/2$ is included so that each edge (i, j) is only counted once. When G has graph Laplacian L, it can be checked (see e.g. [53]) that $\mathcal{D}_{G,\mathbf{z}} = \mathbf{z}^T L\mathbf{z} = \overline{\mathbf{z}}^T L\overline{\mathbf{z}}$.

Disagreement measures how misaligned each node's opinion is with the opinions of its neighbors. We are also interested in how misaligned a node's expressed opinion is with its innate opinion.

Definition 2.4 (Local Internal Conflict, $\mathcal{I}_{\mathbf{z},\mathbf{s},i}$)**.** For $i \in 1, \ldots, n$, a vector of expressed opinions $\mathbf{z} \in [-1, 1]^n$, and a vector of innate opinions $\mathbf{s} \in [-1, 1]^n$,

$$\mathcal{I}_{\mathbf{z},\mathbf{s},i} \stackrel{\text{def}}{=} (z_i - s_i)^2.$$

We also define an aggregate measure of internal conflict.

Definition 2.5 (Global Internal Conflict, $I_{\mathbf{z},\mathbf{s}}$)**.** For a vector of expressed opinions $\mathbf{z} \in [-1, 1]^n$, and a vector of innate opinions $\mathbf{s} \in [-1, 1]^n$,

$$\mathcal{I}_{\mathbf{z},\mathbf{s}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathcal{I}_{\mathbf{z},\mathbf{s},i} = \|\mathbf{z} - \mathbf{s}\|_2^2.$$

Since $\text{mean}(\mathbf{z}) = \text{mean}(\mathbf{s})$, we equivalently have $\mathcal{I}_{\mathbf{z},\mathbf{s}} = \|\overline{\mathbf{z}} - \overline{\mathbf{s}}\|_2^2$.

We can rewrite both the Friedkin-Johnsen update rule and equilibrium opinion vector as solutions to optimization problems involving minimizing disagreement and internal conflict.

Claim 2.1. *The Friedkin-Johnsen dynamics update rule (Equation 1) is equivalent to*

$$z_i^{(t)} = \arg\min_z \mathcal{D}_{G,\mathbf{z},i} + \mathcal{I}_{\mathbf{z},\mathbf{s},i}. \tag{5}$$

The equilibrium opinion vector $\mathbf{z}^*$ (Equation 3) is equivalent to

$$\mathbf{z}^* = \arg\min_z \mathcal{D}_{G,\mathbf{z}} + \mathcal{I}_{\mathbf{z},\mathbf{s}}. \tag{6}$$

It was also observed in [19] that polarization, disagreement, and internal conflict obey a "conservation law" in the Friedkin-Johnsen dynamics.

Claim 2.2 (Conservation law). *For any graph G with Laplacian L, innate opinions $\mathbf{s} \in [-1, 1]^n$, and equilibrium opinions $\mathbf{z}^*$,*

$$\mathcal{P}_{\mathbf{z}^*} + 2 \cdot \mathcal{D}_{G,\mathbf{z}^*} + \mathcal{I}_{\mathbf{z}^*,\mathbf{s}} = \overline{\mathbf{s}}^T \overline{\mathbf{s}}. \tag{7}$$

Now, combining Equations (6) and (7) tells us that $\mathbf{z}^*$, the equilibrium solution of the Friedkin-Johnsen dynamics, maximizes polarization plus disagreement.

$$\mathbf{z}^* = \arg\max_z \mathcal{P}_z + \mathcal{D}_{G,\mathbf{z}}. \tag{8}$$

Now suppose we add another actor, whose goal is to minimize disagreement, to the model. Informally, since the users of the network are maximizing polarization + disagreement, and this other actor is minimizing disagreement, one would expect polarization to increase. This intuitive observation motivates the network administrator dynamics, described below, as a vehicle for the emergence of filter bubbles in a network.

3 THE EMERGENCE OF FILTER BUBBLES

We introduce another actor to the Friedkin-Johnsen opinion dynamics, the **network administrator**. The network administrator increases user engagement via personalized filtering, or showing users content that they are more likely to agree with. In the Friedkin-Johnsen model, this corresponds to the network administrator reducing disagreement by making changes to the edge weights of the graph (e.g. users see more content from users with similar opinions, and less content from users with very different opinions).

3.1 Network Administrator Dynamics

Formally, our extension of the Friedkin-Johnsen dynamics has two actors: users, who change their expressed opinions $\mathbf{z}$, and a network administrator, who changes the graph G. The *network administrator dynamics* are as follows.

Network Administrator Dynamics.
Given initial graph $G^{(0)} = G$ and initial opinions $z^{(0)} = s$, in each round $r = 1, 2, 3, \ldots$

- First, the users adopt new expressed opinions $z^{(r)}$. These opinions are the equilibrium opinions (Equation 3) of the FJ dynamics model applied to $G^{(r-1)}$:
$$z^{(r)} = (L^{(r-1)} + I)^{-1} s. \tag{9}$$
Here $L^{(r-1)}$ is the Laplacian of $G^{(r-1)}$.
- Then, given user opinions $z^{(r)}$, the network administrator minimizes disagreement by modifying the graph, subject to certain restrictions:
$$G^{(r)} = \arg\min_{G \in S} \mathcal{D}_{G, z^{(r)}}. \tag{10}$$
S is the constrained set of graphs the network administrator is allowed to change to.

3.1.1 Restricting changes to the graph. S, the set of all graphs the network admin can modify the graph to (Equation 17), should reflect realistic changes that a recommender system would make. For example, if the network admin is unconstrained, then the network admin will simply set $w_{ij} = 0$ for all edges (i, j), as the empty graph minimizes disagreement. This is entirely unrealistic, however, as a social network would never eliminate all connections between users. In our experiments, we define S as follows:

Constraints on the network administrator.
Given $\epsilon > 0$ and initial graph $\overline{G}$ with adjacency matrix $\overline{W}$, let S contain all graphs with adjacency matrix W satisfying:

(1) $||W - \overline{W}||_F < \epsilon \cdot ||\overline{W}||_F$.
(2) $\sum_j W_{ij} = \sum_j (\overline{W})_{ij}$ for all i, i.e. the degree of each vertex should not change.

The first constraint prevents the network administrator from making large changes to the initial graph $\overline{W}$. Here, ϵ represents an L^2 constraint parameter for how much the network administrator can change edge weight in the network. The second constraint restricts the network administrator to only making changes that maintain the total level of interaction for every user. Otherwise, the network administrator could reduce disagreement by decreasing the total edge weight in the graph — corresponding to having people spend less time on the network — which is not realistic.

Note that, since S gives a convex set over adjacency matrices and $\mathcal{D}_{G, z^{(r)}}$ is a convex function (as a function of the adjacency matrix of G), the minimization problem in Equation (17) has a unique solution, eliminating any ambiguity for the network administrator.

3.1.2 Convergence. Although it is not immediately obvious, the Network Administrator Dynamics do converge. In each round, the users are minimizing disagreement + internal conflict (Equation 6), while the network admin is minimizing disagreement (Equation 17). Thus, we can view the Network Administrator Dynamics as alternating minimization on disagreement + internal conflict:

$$\arg\min_{z \in \mathbb{R}^n, W \in S} \mathcal{D}_{G,z} + \mathcal{I}_{z,s}. \tag{11}$$

While $\mathcal{D}_{G,z} + \mathcal{I}_{z,s}$ is not convex in both z and W, it is convex in one variable when the other is fixed. Because our constraints on W are also convex, alternating minimization will converge to a stationary point of $\mathcal{D}_{G,z} + \mathcal{I}_{z,s}$ [9, 10]. Moreover, while the convergence point is not guaranteed to be the global minima of $\mathcal{D}_{G,z} + \mathcal{I}_{z,s}$, we empirically find that alternating minimization converges to a better solution than well-known optimization methods such as sequential quadratic programming [13] and DMCP [58].

3.2 Experiments

Using two real-world networks, we show that content filtering by the network administrator greatly increases polarization.

Datasets. We use two real-world networks collected in [25], which were previously used to study polarization in [53]. We briefly describe the datasets. More details can be found in [25, 53].

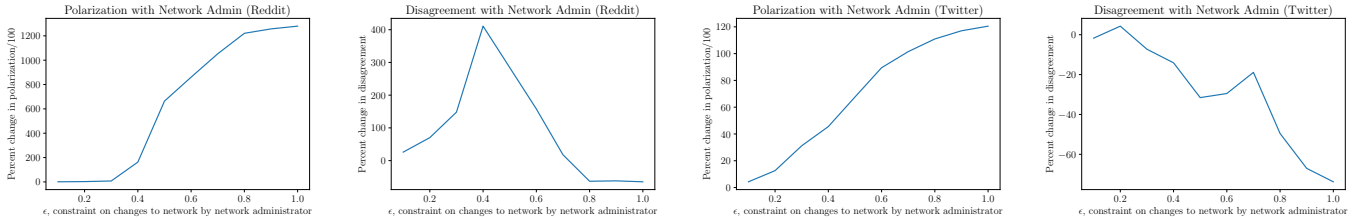
Twitter is a network with $n = 548$ nodes and $m = 3638$ edges. Edges correspond to user interactions. The network depicts the debate over the Delhi legislative assembly elections of 2013.

Reddit is a network with $n = 556$ nodes and $m = 8969$ edges. Nodes are users who posted in the politics subreddit, and there is an edge between two users if there exist two subreddits (other than politics) that both users posted in during the given time period.

In both networks, each user has multiple opinions associated to them, obtained via sentiment analysis on multiple posts. Similar to [53], we average each of these opinions to obtain an equilibrium expressed opinion z_i^* for each user i. Inverting Equation (3) yields innate opinions $\mathbf{s} = (L + I)\mathbf{z}$, which we clamp to $[-1, 1]$. This yields a rough estimate of the innate opinions of each user, and provides a starting point for analyzing the dynamics of polarization.

Results. Figure 2 shows our results applying the network administrator dynamics to the Reddit and Twitter datasets. For both networks, we calculate the increase in polarization after introducing the network administrator dynamics, relative to the polarization of the equilibrium opinions without the network administrator. We plot this polarization increase versus ϵ, the L^2 parameter that specifies how much the network administrator can change the network. We also plot the increase in disagreement versus ϵ.

Once ϵ is large enough, polarization rises greatly in both networks. For example, when $\epsilon = 0.5$, polarization increases by a factor of around **700**× in the Reddit network, and a factor of around



(a) Change in polarization, Reddit network

(b) Change in disagreement, Reddit network

(c) Change in polarization, Twitter network

(d) Change in disagreement, Twitter network

Figure 2: Applying network administrator dynamics to real-world social networks. Details in Section 3.

60× in the Twitter network. While polarization increases in both networks, it is interesting to observe that the Twitter network is more resilient than the Reddit network. Surprisingly, for $\epsilon < 0.7$, disagreement also increases in the Reddit network — so the network administrator does not even reduce user disagreement.

Overall, our experiments illustrate how recommender systems can greatly increase opinion polarization in social networks, and give experimental credence to the theory of filter bubbles [55].

4 FRAGILE CONSENSUS IN SOCIAL NETWORK GRAPHS

Our results in Section 3 establish that polarization in Friedkin-Johnsen opinion models can significantly increase even when the network administrator adjusts just a small amount of edge weight.

To better understand this empirical finding, we present a theoretical analysis of the *sensitivity* of social networks to outside influence. In this work we are most interested in the effect of "filtering" by a network administrator, but our analysis can also be applied to potential influence from advertisers [36, 44] or propaganda [16]. We want to understand how easily such outside influence can affect the polarization of a network.

4.1 The Stochastic Block Model

We consider a common generative model for networks that can lead to polarization: the stochastic block model (SBM) [40].

Definition 4.1 (Stochastic Block Model (SBM))**.** The stochastic block model is a random graph model parametrized by $n > 0$ and $p, q \in [0, 1]$. The model generates a graph G with $2n$ vertices, where the vertex set of G, is partitioned into two sets or "communities", $S = \{v_1, \ldots, v_n\}$ and $T = \{v_{n+1}, \ldots, v_{2n}\}$. Edges are generated as follows. For all $v_i, v_j \in V$:

- If $v_i, v_j \in S$ or $v_i, v_j \in T$, set $w_{ij} = 1$ with probability p, and $w_{ij} = 0$ otherwise.
- If $v_i \in S, v_j \in T$ or $v_i \in T, v_j \in S$, set $w_{ij} = 1$ with probability q, and $w_{ij} = 0$ otherwise.

Also known as "planted partition model", the stochastic block model has as long history of study in statistics, machine learning, theoretical computer science, statistical physics, and a number of other areas. It has been used to study social dynamics, suggesting it as a natural choice for analyzing the dynamics of polarization [8, 49]. We refer the reader to the survey in [1] for a complete discussion of applications and prior theoretical work on the model.

There are many possible variations on Definition 4.1. For example, S and T may differ in size or V may be partitioned into more than two communities. Our specific setup is both simple and well-suited to studying the dynamics of opinions with two poles.

4.2 Opinion Dynamics in the SBM

As in most work on the SBM, we consider the natural setting where $q < p$, i.e. the probability of two nodes being connected is higher when the nodes are in the same community, and lower when they are in different communities. This setting results in a graph G which is "partitioned": G looks like two identically distributed Erdős-Rényi random graphs, connected by a small number of random edges.

We assume nodes in S have innate opinions clustered near -1 (one end of the opinion spectrum), and nodes in T have innate opinions clustered near 1, i.e., nodes with similar innate opinions are more likely to be connected. This property, known as "homophily", is commonly observed in real social networks [22]. Homophily arises because innate opinions are often correlated with demographics like age, geographic location, and education level—demographics which also influence the probability that two nodes are connected.

With the SBM chosen as a model for graphs which resemble real-world social networks, our main question in this section is:

> *How sensitive is the equilibrium polarization of a Friedkin-Johnsen opinion dynamics to changes in the underlying social network graph G, when G is generated from a SBM?*

To answer this question, we analyze how the equilibrium polarization of SBM networks depends on parameters p and q. We show that polarization of the equilibrium opinions decreases *quadratically* with q, which means that even networks with very few edges between S and T have low polarization.

Formally, let $A \in \mathbb{R}^{2n \times 2n}$, $D = \text{diag}(\text{rowsum}(A))$, and $L = D - A$, be the adjacency matrix, diagonal degree matrix, and Laplacian, respectively, of a graph G drawn from the stochastic block model. For simplicity, assume the FJ dynamics with $\mathbf{s}$ set to completely polarized opinions, which perfectly correlate with a node v_i's membership in either $S = \{v_1, \ldots v_n\}$ or $T = \{v_{n+1} \ldots v_{2n}\}$:

$$s_i = \begin{cases} 1 & \text{for } i \in 1, \ldots, n \\ -1 & \text{for } i \in n+1, \ldots, 2n \end{cases} \tag{12}$$

Our main result is:

Theorem 4.1 (Fragile consensus in SBM networks). *Let G be a graph generated by the SBM with $1/n \leq q \leq p$ and $p > c \log^4 n/n$*

for some universal constant c. Let $\mathbf{s}$ *be the innate opinion vector defined in Equation* (12)*, and let* $\mathbf{v}^*$ *be the equilibrium opinion vector according to the FJ dynamics. Then for sufficiently large n,*

$$C\frac{2n}{(2nq+1)^2} \le \mathcal{P}_{\mathbf{v}^*} \le C'\frac{2n}{(2nq+1)^2}$$

with probability 97/100*, for universal constants* C, C' *.*

Note that our assumptions on q and p are mild – we simply need that, in expectation, each node has at least one connection outside of its home community, and $O(\log^4 n)$ connections within its home community. In real-world social networks, the average number of connections should typically exceed these minimum requirements.

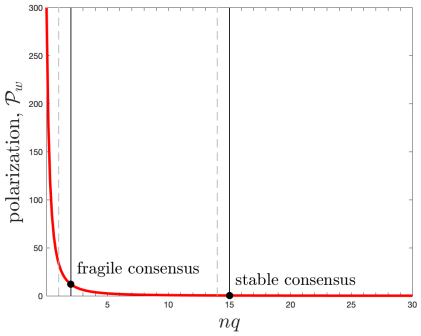


Figure 3: The equilibirum polarization of a SBM social network plotted as a function of nq, i.e. the average number of "out-of-group" edges in the network per node. Polarization falls rapidly with nq, leading to a state of potentially fragile consensus, where removing a small number of edges from a network can vastly increase polarization.

Remarks. Theorem 4.1 leads to two important observations. First, with high probability, the equilibrium polarization of a SBM network is *independent* of p, the probability of generating an "in-group" edge. This is highly counterintuitive: one would expect that increasing p would decrease polarization, as each node would be surrounded by a larger proportion of like-minded nodes.

Second, when nq is sufficiently large, polarization scales as $\sim \frac{2n}{(2nq)^2}$. Since the maximum polarization in a network with $2n$ nodes is $2n$, this says that the polarization of an SBM graph drops quadratically with nq, the expected number of "out-of-group" edges per node. This behavior is visualized in Figure 3.

The second observation suggests an interesting conclusion on social networks that are relatively un-polarized (i.e., are near consensus). In particular, it is possible for these networks to be in a state of **fragile consensus**, meaning that if small number of edges are removed between S and T – for example by a network administrator – polarization can increase rapidly. In fact, this is the case even when edges between S and T are eliminated *randomly*. Doing so produces a new G' also drawn from an SBM, but with parameter $q' < q$. Referring to Figure 3 and Theorem 4.1, G' can have significantly higher polarization than G, even when q' is close to q.

4.3 Expectation Analysis

To prove Theorem 4.1, we apply McSherry's "perturbation" approach for analyzing the stochastic block model [52, 61]. We first bound the polarization of an SBM graph *in expectation*, and then show that the bound carries over to random SBM graphs.

Lemma 4.2. *Let* $\overline{G}$ *be a graph with* $2n$ *vertices and adjacency matrix*

$$\overline{A} = \begin{bmatrix} 0 & p & \dots & p & q & q & \dots & q \\ p & 0 & \ddots & \vdots & q & q & \ddots & \vdots \\ \vdots & \ddots & \ddots & p & \vdots & \ddots & \ddots & q \\ p & \dots & p & 0 & q & \dots & q & q \\ q & q & \dots & q & 0 & p & \dots & p \\ q & q & \ddots & \vdots & p & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & q & \vdots & \ddots & \ddots & p \\ q & \dots & q & q & p & \dots & p & 0 \end{bmatrix}$$

Let $\mathbf{s}$*, as defined in Equation* (12)*, be the innate opinion vector for the network, and let* $\mathbf{w}^*$ *be the resulting equilibrium opinion vector according to the FJ dynamics. Then,*

$$\mathcal{P}_{\mathbf{w}^*} = \frac{2n}{(2nq+1)^2}. \tag{13}$$

Proof. Let $\overline{D}$ and $\overline{L}$ be the diagonal degree matrix and Laplacian of $\overline{G}$, respectively. Since $\mathbf{s}$ is mean centered, we have that $\mathcal{P}_{\mathbf{w}^*} = \mathbf{s}^T(\overline{L}+I)^{-2}\mathbf{s}$. To analyze $\mathcal{P}_{\mathbf{w}^*}$, we need to obtain an explicit representation for the eigendecomposition of $(\overline{L}+I)^{-2}$.

Let $U = [\mathbf{u}^{(1)}, \mathbf{u}^{(2)}]$ where $\mathbf{u}^{(1)} = \frac{1}{\sqrt{2n}}\mathbf{1}_{2n}$ and $\mathbf{u}^{(2)} = \frac{1}{\sqrt{2n}}\mathbf{s}$. We can check that $\overline{A} + pI = U\Lambda U^T$ where $\Lambda = \mathrm{diag}(n(p+q), n(p-q))$. Now, let $\overline{U} = [\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, Z] \in \mathbb{R}^{2n\times 2n}$ where $Z \in \mathbb{R}^{2n\times(2n-2)}$ is a matrix with orthonormal columns satisfying $Z^T\mathbf{u}^{(1)} = \mathbf{0}$ and $Z^T\mathbf{u}^{(2)} = \mathbf{0}$. Such a Z can be obtained by extending $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}$ to an orthonormal basis. Note that $\overline{U}$ is orthogonal, i.e. $\overline{U}\overline{U}^T = \overline{U}^T\overline{U} = I$. Since $\overline{L} + I = (1 + n(p+q))I - (\overline{A} + pI)$, we see that

$$\overline{L} + I = \overline{U}S\overline{U}^T. \tag{14}$$

where $S = \mathrm{diag}([1, 2nq+1, n(p+q)+1, \dots, n(p+q)+1])$.

Since $\overline{U}$ is orthonormal, it follows that $(L+I)^{-2}$ has eigendecomposition $(L+I)^{-2} = \tilde{U}S^{-2}\tilde{U}^T$ is . Moreover, since $\mathbf{s} = \sqrt{2n}\cdot\mathbf{u}^{(2)}$, we have that $\mathbf{s}$ is orthogonal to $\mathbf{u}^{(1)}$ and the columns of Z. Thus,

$$\mathcal{P}_{\mathbf{w}^*} = \mathbf{s}^T(\overline{L}+I)^{-2}\mathbf{s} = \frac{2n}{(2nq+1)^2}. \qquad \square$$

4.4 Perturbation Analysis

With the proof of Lemma 4.2 in place, we prove Theorem 4.1 by appealing to the following standard result on matrix concentration.

Lemma 4.3 (Corollary of Theorem 1.4 in [63]). *Let* A *be the adjacency matrix of a graph drawn from the SBM, and let* $\overline{A} = \mathbb{E}[A]$ *as in Lemma 4.2. There exists a universal constant* c *such that if* $p \ge c\log^4 n/n$*, then with probability* 99/100*,*

$$\|A - \overline{A}\|_2 \le 3\sqrt{pn}.$$

We also require a standard Bernstein inequality (see e.g. [64]):

Lemma 4.4 (Bernstein Inequality). *Let $X_i, \ldots, X_m$ be independent random variables with variances $\sigma_1^2, \ldots, \sigma_m^2$ and $|X_i| \leq 1$ almost surely for all i. Let $X = \sum_{i=1}^m X_i$, $\mu = \mathbb{E}[X]$, and $\sigma^2 = \sum_{i=1}^m \sigma_i^2$.*

$$\Pr[|X - \mu| > \epsilon] \leq e^{\frac{\epsilon^2}{2\sigma^2 + 2\epsilon/3}}$$

Using these two bounds, we can prove:

Lemma 4.5. *Let L be the Laplacian of a graph G drawn from the SBM and let $\overline{L} = \mathbb{E}[L]$. For fixed constant c_0, with probability 98/100,*

$$\|L - \overline{L}\|_2 \leq c_0 \sqrt{pn \log n}.$$

Note that when $p \geq c \log^4 n$, $c_0\sqrt{pn \log n} \leq \frac{c_0}{\sqrt{c}\log^{1.5} n} \cdot pn$, so for sufficiently large n, this lemma implies that $\|L - \overline{L}\|_2 \leq \frac{1}{2}pn$

Proof. Let D be the degree matrix of G and recall that $\mathbb{E}[D] = \overline{D}$. By triangle inequality, $\|L - \overline{L}\|_2 \leq \|D - \overline{D}\|_2 + \|A - \overline{A}\|_2$. By Lemma 4.3, $\|A - \overline{A}\|_2 < 3\sqrt{pn}$. Additionally, $\|D - \overline{D}\|_2$ is bounded by $\max_i |D_{ii} - \overline{D}_{ii}|$. D_{ii} is a sum of Bernoulli random variables with total variance σ^2 upper bounded by $2np$. It follows from Lemma 4.4 and our assumption that $p = \Omega(1/n)$ that for any i, $|D_{ii} - \overline{D}_{ii}| \leq c_1\sqrt{pn \log n}$ with probability $1 - \frac{1}{200n}$ for a fixed universal constant c_1. By a union bound, we have that $\max_i |D_{ii} - \overline{D}_{ii}| \leq c_1\sqrt{pn \log n}$ with probability 99/100 for all i. A second union bound with the event that $\|A - \overline{A}\|_2 < 3\sqrt{pn}$ gives the lemma with $c_0 = 3 + c_1$. □

With Lemma 4.5 in place, we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. We separately consider two cases.

Case 1, $\mathbf{q \geq p/2}$. In this setting, all eigenvalues of $\overline{L} + I$ lie between $pn + 1$ and $2pn + 1$, except for the smallest eigenvalue of 1, which has corresponding eigenvector $\mathbf{u}^{(1)} = \mathbf{1}/\sqrt{2n}$. Since $L\mathbf{u}^{(1)} = \mathbf{0}$, $\mathbf{u}^{(1)}$ is also an eigenvector of $L + I$. Let $P = \mathbf{u}^{(1)}\mathbf{u}^{(1)T}$ be a projection onto this eigenvector. Using that $\mathbf{u}^{(1)}$ is an eigenvalue of both L and $\overline{L}$ and applying Lemma 4.5, we have:

$$(0.5pn + 1)(I - P) \leq (I - P)(L + I)(I - P) \leq (2.5pn + 1)(I - P).$$

Since $(I - P)(L + I)(I - P)$ and $(I - P)$ commute, it follows that $(0.5pn+1)^2(I-P) \leq (I-P)(L+I)^2(I-P) \leq (2.5pn+1)^2(I-P)$. Finally, noting that $(I - P)\mathbf{s} = \mathbf{s}$, $\mathbf{s}^T\mathbf{s} = 2n$, and $M \leq N \Rightarrow N^{-1} \leq M^{-1}$ gives the Theorem for $q \geq p/2$.

Case 2, $\mathbf{q < p/2}$. The small q case is more challenging, requiring a strengthening of Lemma 4.5. This lemma asserts that every eigenvalue of L is within additive error $c_0\sqrt{pn \log n}$ from the corresponding eigenvalue in $\overline{L}$. While strong for $\overline{L}$'s largest eigenvalues of $(p + q)n$, the statement can be weak for L's smallest non-zero eigenvalue of $2nq$. We require a tighter *relative error* bound:

Lemma 4.6. *Assume $1/n \leq q < p/2$. Let $\lambda_2(L)$ be L's smallest non-zero eigenvalue. With probability 99/100, for sufficiently large n,*

$$\frac{1}{2}nq \leq \lambda_2(L) \leq 4nq$$

Due to space constraints, the proof of Lemma 4.6 is omitted here. It can be found in the full version of this paper available at [20].

In addition to Lemma 4.6, we also require the well-known bound.

Lemma 4.7 (Davis-Kahan Theorem [24]). *Let M and H be $m \times m$ symmetric matrices with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_m$, respectively, and eigenvalues $\lambda_1, \ldots, \lambda_m$ and $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$. If $\|M - H\|_2 \leq \epsilon$, then for all i,*

$$(\mathbf{v}_i^T\tilde{\mathbf{v}}_i)^2 \geq 1 - \frac{\epsilon^2}{\min_{j \neq i} |\lambda_i - \lambda_j|^2}$$

Again, let $P = \mathbf{u}^{(1)}\mathbf{u}^{(1)T}$. Let $\tilde{U} \in \mathbb{R}^{2n \times (2n-1)}$ be an orthonormal basis for the span of $I - P$. We will apply Lemma 4.7 to the matrices $\tilde{U}^T\overline{L}\tilde{U}$ and $\tilde{U}^TL\tilde{U}$. Since $\mathbf{u}^{(1)}$ is an eigenvector of both $\overline{L}$ and L, the eigenvectors of $\tilde{U}^T\overline{L}\tilde{U}$ and $\tilde{U}^TL\tilde{U}$ are equal to the remaining $2n - 1$ eigenvectors of $\overline{L}$ and L left multiplied by $\tilde{U}^T$. The eigenvalues of $\tilde{U}^T\overline{L}\tilde{U}$ and $\tilde{U}^TL\tilde{U}$ are simply the non-zero eigenvalues of $\overline{L}$ and L.

Let $\mathbf{y}$ be the eigenvector of L associated with $\lambda_2(L)$. Theorem 4.5 implies that $\|\tilde{U}^T\overline{L}\tilde{U} - \tilde{U}^TL\tilde{U}\| \leq c_0\sqrt{pn \log n}$ and so by Lemma 4.7, we have:

$$(\mathbf{u}^{(2)T}\tilde{U}^T\tilde{U}\mathbf{y})^2 \geq 1 - \frac{c_0^2 pn \log n}{((p - q)n)^2}.$$

Since $p - q \geq p/2$, our assumption that $p = \Omega(\log^4 n/n)$ implies that $(\mathbf{u}^{(2)T}\tilde{U}^T\tilde{U}\mathbf{y})^2 \geq 1 - O(1/\log^3 n))$, which is $\geq 1/2$ for large enough n. Since $\mathbf{y}$ and $\mathbf{u}^{(2)}$ are eigenvalues of L and $\overline{L}$ respectively, both are orthogonal to $\mathbf{u}^{(1)}$. So $(\mathbf{u}^{(2)T}\tilde{U}^T\tilde{U}\mathbf{y})^2 = (\mathbf{u}^{(2)T}\mathbf{y})^2$. We conclude:

$$1/2 \leq (\mathbf{y}^T\mathbf{u}^{(2)})^2 \leq 1. \tag{15}$$

In other words, L's second eigenvector $\mathbf{y}$ has a large inner product with $\overline{L}$'s second eigenvector $\mathbf{u}^{(2)}$.

Since L and $(L + I)^{-2}$ have the same eigenvectors, we can bound $\mathcal{P}_{\mathbf{z}^*} = \mathbf{s}^T(L + I)^{-2}\mathbf{s} = 2n \cdot \mathbf{u}^{(2)T}(L + I)^{-2}\mathbf{u}^{(2)}$ as follows:

$$\begin{aligned}(\mathbf{y}^T\mathbf{u}^{(2)})^2(\lambda_2(L) + 1)^{-2} &\leq \frac{1}{2n}\mathcal{P}_{\mathbf{z}^*} \\ &\leq (\mathbf{y}^T\mathbf{u}^{(2)})^2(\lambda_2(L) + 1)^{-2} + (1 - (\mathbf{y}^T\mathbf{u}^{(2)})^2)\|(L + I)^{-2}R\|_2\end{aligned}$$

where $R = I - \mathbf{y}\mathbf{y}^T - \mathbf{u}^{(1)}\mathbf{u}^{(1)T}$ is a projection matrix onto $(L + I)$'s largest $n - 2$ eigenvectors. From the same argument used for Case 1, all of these eigenvectors have corresponding eigenvalues $\geq \frac{1}{2}pn + 1$, and thus $\|(L + I)^{-2}R\|_2 \leq \frac{1}{(\frac{1}{2}pn+1)^2} \leq \frac{1}{(qn+1)^2}$. Applying (15) and Lemma 4.6, we have:

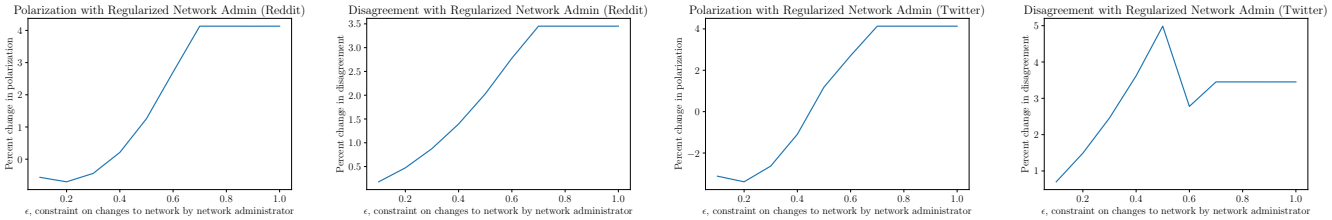
$$\frac{1}{2}\frac{2n}{(4nq + 1)^2} \leq \mathcal{P}_{\mathbf{z}^*} \leq \frac{3n}{(\frac{1}{2}nq + 1)^2},$$

which establishes the theorem. □

5 A SIMPLE REMEDY

Throughout this paper, our results have largely been pessimistic. The introduction of a network administrator who filters user content causes polarization to rises and echo chambers form, along the lines of Pariser's filter bubble theory [55]. Our analysis in the SBM further evidences that social networks can easily be in a state of "fragile consensus", which leaves them vulnerable to extreme polarization, even when only a small number of edges are modified.

In this section, however, we conclude with a positive result. We find that, with a slight modification to the network administrator dynamics, the filter bubble effect is vastly mitigated. Even more surprisingly, disagreement also barely increases, showing that it



(a) Change in polarization vs ϵ for Reddit network. **(b) Change in disagreement vs ϵ for Reddit network.** **(c) Change in polarization vs ϵ for Twitter network.** **(d) Change in disagreement vs ϵ for Twitter network.**

Figure 4: Applying *regularized* network administrator dynamics to real-world social networks, $\gamma = 0.2$. Details in Section 5.

is possible for the network administrator to reduce polarization in the network while not hurting its own objective.

5.1 Regularized Dynamics

We modify the role of the network administrator by adding an L^2 regularization term to its objective function.

Regularized Network Administrator Dynamics.
Given initial graph $G^{(0)} = G$ and initial opinions $z^{(0)} = s$, in each round $r = 1, 2, 3, \ldots$

- First, the users adopt new expressed opinions $z^{(r)}$. These opinions are the equilibrium opinions (Equation 3) of the FJ dynamics model applied to $G^{(r-1)}$:
$$z^{(r)} = (L^{(r-1)} + I)^{-1} s. \tag{16}$$
Here $L^{(r-1)}$ is the Laplacian of $G^{(r-1)}$.
- Then, given user opinions $z^{(r)}$, the network administrator minimizes disagreement by modifying the graph, subject to certain restrictions:
$$G^{(r)} = \arg\min_{G \in S} \mathcal{D}_{G, z^{(r)}} + +\gamma \|W\|_F^2 \tag{17}$$
S is the constrained set of graphs the network administrator is allowed to change to, W is the adjacency matrix of G, and $\gamma > 0$ is a fixed constant.

$\gamma > 0$ is a fixed constant that controls the strength of regularization. We use L^2 regularization because $\left|\arg\min_{x:\|x\|_1=1} \|x\|_2\right| = \mathbf{1}_n/n$ for $x \in \mathbb{R}^n$. So intuitively, since the network administrator must keep the total edge weight of the graph constant, the addition of the regularization term encourages the network administrator to make modifications to many edges in the graph, instead of making large, concentrated changes to a small number of edges.

5.2 Results

Figure 4 shows the results of the *regularized* network administrator dynamics on the Reddit and Twitter networks, with $\gamma = 0.2$. Polarization increases by at most 4%, no matter the value of ϵ. This is a drastic difference from the non-regularized network administrator dynamics, where polarization increased by over 4000%. Disagreement, which the network administrator is incentivized to decrease, increases by at most 5%.

6 CONCLUSION AND FUTURE DIRECTIONS

Despite enabling users access to a diversity of information, social media has been linked to increased societal polarization [31]. One proposed explanation for this counterintuitive phenomenon is the *filter bubble theory*, which posits that, by automatically recommend content that a user is likely to agree, content filtering algorithms on social networks create polarized "echo chambers" of users [55].

In this work, we provide experimental and theoretical support that the filter bubble theory holds. Specifically, we propose an extension to the Friedkin-Johnsen opinion dynamics model that explicitly models recommendation systems in social networks. Using this model, we experimentally show the emergence of filter bubbles in real-world networks, and provide theoretical justification for why social networks are vulnerable to outside actors.

Our work poses many follow-up questions. For example, as discussed earlier, variants of the Bounded Confidence Model (BCM) have been used to argue that polarization is caused by "biased assimilation" of content by users [22, 27, 35]. In this work, we use the Friedkin-Johnsen model because of its linear algebraic interpretation, which allows us to establish concrete theoretical results. It could be interesting incorporate our network administrator dynamics into the more complex BCM variants used by [35] and [27].

Another interesting direction is modeling the interference of other outside actors, as our theoretical analysis is not limited to recommendation systems. Can we develop a similar framework for modeling the effects of cyber warfare (see e.g. [56]) on societal polarization? And perhaps more importantly, can we also develop methods to mitigate the effects of cyber warfare on polarization?

REFERENCES

[1] Emmanuel Abbe. 2018. Community Detection and Stochastic Block Models. *Foundations and Trends in Communications and Information Theory* 14, 1-2 (2018).
[2] 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 and Data Mining (KDD)*. 1089–1098.
[3] Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*. 36–43.
[4] Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer Publishing Company, Incorporated.
[5] Cigdem Aslay, Antonis Matakos, Esther Galbrun, and Aristides Gionis. 2018. Maximizing the diversity of exposure in a social network. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 863–868.
[6] Drake Baer. 2016. The 'Filter Bubble' Explains Why Trump Won and You Didn't See It Coming. *Science of Us* (2016).
[7] Pablo Barberá. 2014. How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. *Job Market Paper, NYU* (2014).

[8] Luca Becchetti, Andrea Clementi, Emanuele Natale, Francesco Pasquale, and Luca Trevisan. 2017. Find Your Place: Simple Distributed Algorithms for Community Detection. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 940–959.

[9] A. Beck. 2015. On the Convergence of Alternating Minimization for Convex Programming with Applications to Iteratively Reweighted Least Squares and Decomposition Schemes. *SIAM Journal on Optimization* 25, 1 (2015), 185–209.

[10] D.P. Bertsekas. 1999. *Nonlinear Programming*. Athena Scientific.

[11] David Bindel, Jon Kleinberg, and Sigal Oren. 2015. How bad is forming your own opinion? *Games and Economic Behavior* 92 (2015), 248 – 265.

[12] Sarah Binder. 2014. Polarized we govern?

[13] Paul T. Boggs and Jon W. Tolle. 1995. Sequential Quadratic Programming. *Acta Numerica* 4 (1995), 1–51.

[14] Paul Boutin. 2011. Your Results May Vary: Will the information superhighway turn into a cul-de-sac because of automated filters? *The Wall Street Journal* (2011).

[15] Jennifer Brundidge. 2010. Encountering "Difference" in the Contemporary Public Sphere: The Contribution of the Internet to the Heterogeneity of Political Discussion Networks. *Journal of Communication* 60, 4 (11 2010), 680–700.

[16] T. Carletti, D. Fanelli, S. Grolli, and A. Guarino. 2006. How to make an efficient propaganda. *Europhysics Letters* 74, 2 (2006), 222.

[17] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 224–232.

[18] Xi Chen, Jefrey Lijffijt, and Tijl De Bie. 2018. The Normalized Friedkin-Johnsen Model (A Work-in-progress Report). In *ECML PKDD 2018-PhD Forum*.

[19] Xi Chen, Jefrey Lijffijt, and Tijl De Bie. 2018. Quantifying and Minimizing Risk of Conflict in Social Networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 1197–1205.

[20] Uthsav Chitra and Christopher Musco. 2019. Understanding Filter Bubbles and Polarization in Social Networks. *arXiv:1906.08772* (2019).

[21] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

[22] Pranav Dandekar, Ashish Goel, and David T Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5791–5796.

[23] Abhimanyu Das, Sreenivas Gollapudi, and Kamesh Munagala. 2014. Modeling opinion dynamics in social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. ACM, 403–412.

[24] C. Davis and W. Kahan. 1970. The Rotation of Eigenvectors by a Perturbation. III. *SIAM J. Numer. Anal.* 7, 1 (1970), 1–46.

[25] Abir De, Sourangshu Bhattacharya, Parantapa Bhattacharya, Niloy Ganguly, and Soumen Chakrabarti. 2014. Learning a Linear Influence Model from Transient Opinion Dynamics. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 401–410.

[26] Morris H DeGroot. 1974. Reaching a consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121.

[27] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2017. Modeling confirmation bias and polarization. *Scientific reports* 7 (2017).

[28] Lisa Eadicicco. 2019. Apple CEO Tim Cook urges college grads to 'push back' against algorithms that promote the 'things you already know, believe, or like'. *Business Insider* (2019).

[29] Markos Epitropou, Dimitris Fotakis, Martin Hoefer, and Stratis Skoulakis. 2017. Opinion Formation Games with Aggregation and Negative Influence. In *Algorithmic Game Theory - 10th International Symposium, SAGT 2017, L'Aquila, Italy, September 12-14, 2017, Proceedings*. 173–185.

[30] Benedict Evans. 2018. The death of the newsfeed. https://www.ben-evans.com/benedictevans/2018/4/2/the-death-of-the-newsfeed.

[31] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80, S1 (2016), 298–320.

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

[33] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. 2017. Balancing information exposure in social networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 4663–4671.

[34] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying Controversy on Social Media. *ACM Transactions on Social Computing* 1, 1 (2018), 3:1–3:27.

[35] Daniel Geschke, Jan Lorenz, and Peter Holtz. 2019. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology* 58, 1 (2019), 129–149.

[36] Aristides Gionis, Evimaria Terzi, and Panayiotis Tsaparas. 2013. Opinion Maximization in Social Networks. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA*. 387–395.

[37] Christopher Hare and Keith T Poole. 2014. The polarization of contemporary American politics. *Polity* 46, 3 (2014), 411–429.

[38] Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation* 5 (2002), 1–24.

[39] Rainer Hegselmann, Ulrich Krause, et al. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation* 5, 3 (2002).

[40] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social networks* 5, 2 (1983), 109–137.

[41] Harald Holone. 2016. The filter bubble and its effect on online personal health information. *Croatian medical journal* 57, 3 (2016), 298.

[42] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. 2014. Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation. *Management Science* 60, 4 (2014), 805–823.

[43] Jasper Jackson. 2017. Eli Pariser: activist whose filter bubble warnings presaged Trump and Brexit: Upworthy chief warned about dangers of the internet's echo chambers five years before 2016's votes. *The Guardian* (2017).

[44] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[45] Yonghwan Kim. 2011. The contribution of social network sites to exposure to political difference: The relationships among SNSs, online political messaging, and exposure to cross-cutting perspectives. *Computers in Human Behavior* 27, 2 (2011), 971–977.

[46] Geoffrey C. Layman, Thomas M. Carsey, and Juliana Menasce Horowitz. 2006. Party Polartization in American Politics: Characteristics, Causes, and Consequences. *Annual Review of Political Science* 9, 1 (2006), 83–110.

[47] Jae Kook Lee, Jihyang Choi, Cheonsoo Kim, and Yonghwan Kim. 2014. Social media, network heterogeneity, and opinion polarization. *Journal of communication* 64, 4 (2014), 702–722.

[48] Charles G. Lord, Lee Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology* 37, 11 (1979), 2098.

[49] Frederik Mallmann-Trenn, Cameron Musco, and Christopher Musco. 2018. Eigenvector Computation and Community Detection in Asynchronous Gossip Models. In *Proceedings of the 45th International Colloquium on Automata, Languages and Programming (ICALP)*. 159:1–159:14.

[50] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. 2017. Measuring and Moderating Opinion Polarization in Social Networks. *Data Min. Knowl. Discov.* 31, 5 (2017), 1480–1505.

[51] Aaron M McCright and Riley E Dunlap. 2011. The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly* 52, 2 (2011), 155–194.

[52] Frank McSherry. 2001. Spectral Partitioning of Random Graphs. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[53] Cameron Musco, Christopher Musco, and Charalampos Tsourakakis. 2018. Minimizing Controversy and Disagreement in Social Networks. In *Proceedings of the 27th International World Wide Web Conference (WWW)*.

[54] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW 2014*. ACM, 677–686.

[55] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin.

[56] Michael Riley and Jordan Robertson. 2017. Russian Cyber Hacks on US Electoral System Far Wider Than Previously Known. *Bloomberg, June* 13 (2017).

[57] Elisa Shearer and Katerina Eva Matsa. 2018. News Use Across Social Media Platforms: Most Americans continue to get news on social media, even though many have concerns about its accuracy. *Pew Research Center Report* (2018). https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

[58] Xinyue Shen, Steven Diamond, Madeleine Udell, Yuantao Gu, and Stephen Boyd. 2016. Disciplined Multi-Convex Programming. *arXiv:1609.03285* (2016).

[59] Aaron Smith and Monica Andersen. 2018. Social Media Use in 2018. *Pew Research Center Report* (2018).

[60] Kirsten P Smith and Nicholas A Christakis. 2008. Social networks and health. *Annu. Rev. Sociol* 34 (2008), 405–429.

[61] Daniel Spielman. 2015. Lecture notes on Spectral Partitioning in a Stochastic Block Model. http://www.cs.yale.edu/homes/spielman/561/lect21-15.pdf..

[62] Cristian Vaccari, Augusto Valeriani, Pablo Barberá, John T. Jost, Jonathan Nagler, and Joshua A. Tucker. 2016. Of Echo Chambers and Contrarian Clubs: Exposure to Political Disagreement Among German and Italian Users of Twitter. *Social Media + Society* 2, 3 (2016).

[63] Van H. Vu. 2007. Spectral norm of random matrices. *Combinatorica* 27, 6 (2007).

[64] Martin J. Wainwright. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

[65] Frederik Zuiderveen Borgesius, Damian Trilling, Judith Moeller, Balázs Bodó, Claes H. de Vreese, and Natali Helberger. 2016. Should We Worry About Filter Bubbles? *Internet Policy Review, Journal on Internet Regulation* 5, 1 (2016).