Filter Bubbles

Indu Ramesh and R. Teal Witter

December 18, 2020

1 Introduction

Over the past decade, social media has transformed societal interaction. In theory, social media allows individuals and groups access to a diverse array of ideas. However, recent studies suggest that social networks are associated with societal polarization, separating individuals into groups unable to find agreement [9].

A popular explanation for this phenomenon is the emergence of *filter bubbles*, a term coined by internet activist Eli Pariser [13]. Filter bubbles reflect the idea that users' news feeds in social media networks are simply echo chambers that prevent individuals from accessing a variety of viewpoints. Since user feed content is constrained by metrics that aim to increase user engagement and ad revenue (i.e. friends' and followers' views, internet search history, user location, etc.), social media companies explicitly incentivize users to pay preferential attention to like-minded content. In this manner, users end up living in a 'filter bubble' of their own ideas. Filter bubbles have been blamed for the spread of misinformation in Brexit, the 2016 U.S. presidential election, and increased distrust in democracy [6].

In this project, we study *polarization* in social networks, aiming to advance a mathematical theory behind the formation of filter bubbles.

1.1 Related Work

Through studying various metrics and models of polarization, a mathematical theory has begun to emerge. We give a summary of interesting recent approaches and results.

In [4], Dandekar et al. studied the idea of *biased assimilation*: when given mixed evidence on a complex issue, individuals use that evidence to support their innate opinion, arriving at a more extreme version of their original opinion. They provided evidence that in simple models of social networks, DeGroot's model of opinion formation [5] results in convergence at a less diverse set of opinions, falling short of explaining polarization. They also analyzed the effects of three internet content recommendation algorithms on polarization, showing that if individuals start out sufficiently biased, these algorithms lead to increased polarization.

In [3], Chitra and Musco studied social media companies' roles in creating filter bubbles using a related model of opinion formation, the Friedkin-Johnsen model [7]. They explored the effect of adding an important outside actor to the picture: the *network administrator*. The network administrator's job is to minimize disagreement among users by modifying the edge weights of the graph such that users interact with more content from users with similar opinions. In their model, the network administrator's modifications are subject to certain constraints: the administrator cannot change the degree of any vertex, and can only modify edge weights by a small amount. Chitra and Musco ran experiments on the social networks Twitter and Reddit to simulate the effect of a network administrator on polarization. Their experimental results confirm filter bubble theory, and they show that there is a network administrator action that leads to increased polarization.

Other formulations and solutions to the polarization problem have been studied. In [11], Musco et al. asked: what is the *topological structure* of a network that minimizes both polarization and disagreement, given an opinion dynamics model? They posed this question as an optimization problem, and gave a polynomial-time algorithm that approximates the optimum for a fixed graph. In [10], Matakos et al. formulated polarization as an influence maximization problem, providing an approximation algorithm to break up filter bubbles.

Finally, consider that the work we mentioned has addressed the question: how sensitive are networks to polarization? Chitra and Musco's results in [3] suggest that many networks are in a state of *fragile consensus*. The Netflix documentary "The Great Hack" [12] discusses Cambridge Analytica's role in the 2016 presidential election, swaying undecided voters toward electing Trump through targeted intervention on Facebook. Could we exploit this sensitivity, and model some sort of attack by an outside actor on increasing polarization?

1.2 Overview of Results

In this project, we study polarization in the Friedkin-Johnson (FJ) dynamics model, formally defined below. In Section 3, we present empirical results on the impacts of several additions to the FJ opinion formation process. Notably, we introduce two natural network administrator actions and describe their effects on social networks. In Section 4, we extend the results of [3]. In particular, we show that the polarization of the equilibrium opinions after running FJ dynamics on the Stochastic Block Model (SBM) defined below converges with high probability for any mean-centered innate opinions (rather than the half 1, half -1 innate opinions considered before). When the graph is in the special case of an Erdős–Rényi our bounds are even stronger. Finally, we discuss our work and provide directions for future work in Section 5.

2 Background and Preliminaries

The formulation of the polarization problem we study focuses specifically on mathematical models of opinion formation in social networks.

Social networks arise randomly, with individuals developing connections based on social groups. Thus, the natural mathematical representation of a social network is a graph. Each node represents an individual and an edge represents a friendship or connection in the network. Edges have weights associated with the extent of their friends' influence on their opinions– in other words, the "closeness" of the friendship. In a real-life social network like Facebook, a higher edge weight between users corresponds to a increased interactions between those users (e.g. their respective stories pop up in their news feeds more often).

2.1 The Stochastic Block Model

In this project, we use a random graph model termed the Stochastic Block Mode (SBM) to construct a graph representing a social network. The SBM is a common generative model seen in numerous fields, including statistics, theoretical computer science, and machine learning [1]. We refer the interested reader to [1] for a comprehensive survey of applications of the SBM.

An SBM graph G has n vertices, divided into several communities. The probability that there is an edge between two nodes in the same community is p while the probability there is an edge between two nodes in different communities is q.

Often, we have that $p \geq q$, to model the idea that relationships within communities are more likely than relationships between communities. However, there are some real-world situations where $p \leq q$. Consider a patient-doctor network and the way opinions of e.g. a vaccine are formed. Connections representing conversations about a vaccine are most likely between a doctor and a patient and less likely between patient and patient or doctor and doctor. Such a model is particularly relevant to the acceptance of COVID-19 vaccines [8]. When $p = q$, the SBM is known as an Erdős-Rényi graph.

2.2 Opinion Formation

There are two main models of opinion formation in social networks. The input to both models is a social network given by a graph and an innate (starting) opinion for each node in the graph. That is, each node i has an innate opinion s_i and expressed opinion (that changes according to the respective model) z_i where $s_i, z_i \in [-1, 1]$.

DeGroot's Model: Perhaps the most popular model of opinion dynamics studied, DeGroot's model updates a node's opinion with the weighted average of its expressed opinion and its friends opinions. Formally,

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

where $z_i^{(t)}$ is the opinion of node i at iteration t, $z_i^{(t-1)}$ is the opinion of node i in the previous iteration, and d_i is the degree of node i in the graph.

Friedkin-Johnsen Model: The Friedkin-Johnsen (FJ) opinion dynamics model is similar to DeGroot's Model but, instead of averaging with the node's own expressed opinion, averages with the innate opinion. Intuitively, FJ opinions are more 'stubborn.' Formally,

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

where $z_i^{(t)}$ is the opinion at iteration t, s_i is the innate opinion of node i, and d_i is the degree of node i in the graph.

We will use FJ dynamics because [4] have shown that polarization (formally defined in the next section) always converges in DeGroot's model.

2.3 Measuring Polarization

One measure of polarization with a natural mathematical interpretation is is the *network disagreement* index. In [4], a process is viewed as *polarizing* if it increases the network disagreement index. In a social network graph $G = V, E, w$, the network disagreement index is calculated as follows:

$$\eta(G, x) = \sum_{i,j \in E} w_{ij}(x_i - x_j)^2$$

where x_i and x_j are the opinions at nodes i and j.

In [3], Chitra and Musco add some structure to the measure of polarization. They define polarization as the *variance* of a set of opinions. For a vector of n opinions $z \in [-1, 1]$, let

$$\mathbb{E}[z] = \frac{1}{n} \sum_{j=1}^{n} x_j$$

where z_j is the mean opinion of z. Then, polarization, P_z can be defined as:

$$P_z = \sum_{i=1}^{n} (x_i - \mathbb{E}[z])^2$$

Another measure is *bimodality*. Instead of simply of measuring the difference from the mean, bimodality gives a sense for whether opinions are centered at two different means. Formally, the bimodality coefficient is given by

$$\beta = \frac{\gamma^2 + 1}{\kappa}$$

where γ is the *skewness* (third standardized moment) and κ is the *kurtosis* (fourth standardized moment). Measuring bimodality is particularly relevant to the modern political situation in the United States [2].

2.4 Linear Algebraic Interpretation of FJ Opinion Dynamics

To analyze polarization on an SBM graph, it is helpful to have a linear algebraic interpretation of FJ opinion dynamics. We follow the approach in [3].

Let $A \in R^{n \times n}$ be the adjacency matrix of G, where $A_{ij} = A_{ji}$ since the graph is undirected. Let D be a diagonal matrix where $D_{ii} = d_i$. That is, each diagonal entry is equal to the degree of node i, and the other entries are 0's. Finally, let $L = D - A$ be the *Laplacian* of G. Recall that s is the innate opinion vector of all nodes in G. It is not too hard to see that Equation (1) is equivalent to

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

where $z^{(t)} = \begin{bmatrix} z_1^{(t)} & z_2^{(t)} & \cdots & z_n^{(t)} \end{bmatrix}$.

Let $z^{(*)}$ be the final equilibrum opinion vector after t time steps. Then

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

3 Experiments

It is often challenging to mathematically analyze variations on the opinion formation process, so we simulate them instead. We analyze the effects of two network administrator actions on polarization and bimodality in SBM graphs and two real-world networks, Reddit and Twitter. The variations we consider include: network administrator actions, attack nodes, imbalanced innate opinions, and three (rather than two) blocks in the Stochastic Block Model (SBM).

Listing 1 presents the baseline pseudocode we use for our experiments.

Algorithm 1: Opinion formation.

```
input: adjacency matrix A_0 of dimension n × n,
       innate set of opinions s with s_i ∈ [−1, 1] for i ∈ [n],
       number of iterations T to run
output: a set of equilibrium opinions z^(T)
z^(0) ← s
for each t in {0,...,T} do
    D ← degree matrix of A_t
    z^(t) ← (D + I)^−1 (A_t z^(t−1) + s)
    A_t ← network administrator modifications of A_{t−1}
return z^(T)
```

3.1 Extreme Administrator Action

The first administrator update we consider is a simple, intuitive update rule: add weight to the neighbor with closest opinion and subtract the same weight from the neighbor with furthest opinion.

Listing 2 presents the pseudocode for the update. The overall weight of friendships is preserved in the graph, but many friendships disappear and some nodes even lose their connections altogether.

Algorithm 2: Extreme Administrator Update.

```
input: adjacency matrix A of dimension n × n,
       expressed opinions z^(t),
       maximum edge change ε per iteration
output: modified adjacency matrix A'
A' ← all-zero n × n matrix
foreach u in [n] with more than one neighbor do
    c ← neighbor with closest opinion to u
    f ← neighbor with furthest opinion to u
    Δ ← min{ε, A_{u,f}}
    A'_{u,f} ← A'_{f,u} ← A_{f,u} − Δ
    A'_{u,c} ← A'_{c,u} ← A_{c,u} + Δ
return A'
```

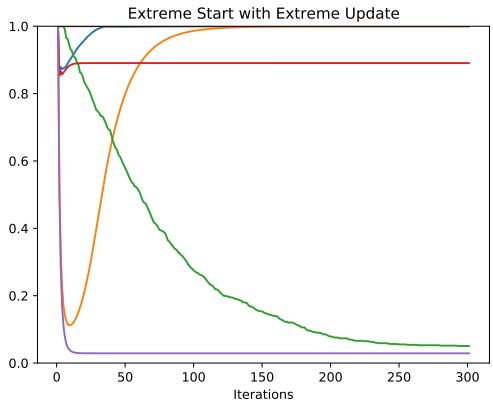
Figure 1 shows the ratio of remaining friendships, bimodality, and polarization in four different pairs of social networks and innate opinions. The SBM social network in both Figure 1 and Figure 2 is the same random network on $n = 100$ in order to preserve consistency. We use $p = 30/n$ and $q = 5/n$.

The ratio of remaining friendships with the extreme update (green) drops quickly across all four figures until it stabilizes around .1. When the innate opinion is extremely polarized (half 1, half -1) in Figure 1a the final ratio of remaining friendships is even smaller.

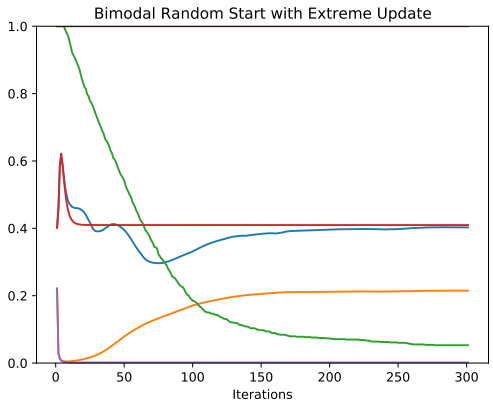
Bimodality is highest with the extremely polarized innate opinions (orange). Otherwise, it stabilizes in the middle of the range. Interestingly, the SBM with two normal innate opinions centered at .5 and -.5, respectively looks similar to the Reddit and Twitter social networks. This suggests that the bimodal SBM is a good model for real-world social networks. Polarization only increases to 1 when innate opinions are extremely polarized. Otherwise, it slowly increases over time and stabilizes near 0.

The extreme administrator update is a naive approach to the natural mechanism social media platforms use to boost engagement: increase exposure to content similar in nature to user preferences and simultaneously decrease exposure to content dissimilar to user preferences.

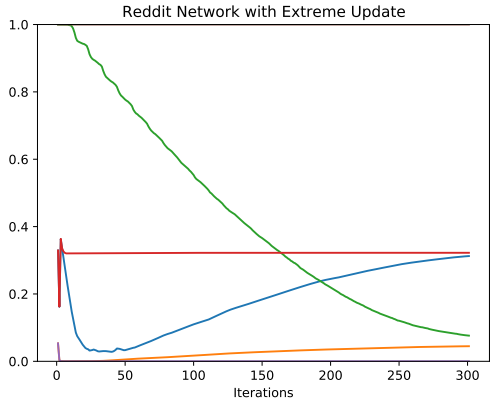
While the overall weight (interpreted as time by [3]) stays constant, the individual level weight can move from one node to another. This combined with the elimination of friendships suggest that the extreme update is not a reasonable administrator action. Nonetheless, analyzing it provides insight into an intuitive idea and the extremes of a potential administrator action.



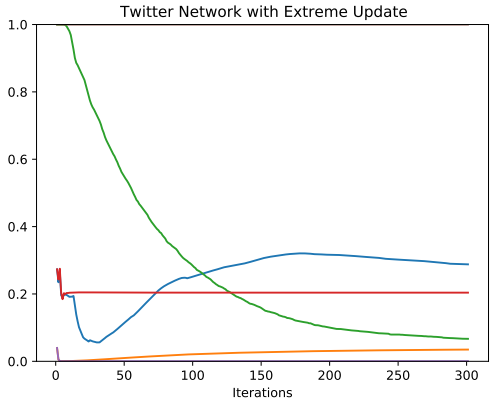
(a) SBM with half 1, -1 innate opinions.



(b) SBM with two scaled normal opinions.



(c) Posts on r/politics and other subreddits.



(d) 2013 Delhi election opinions on Twitter.

Figure 1: Difference between baseline and extreme administrator update on four different social network and opinion pairs. Brown corresponds to the ratio of remaining friendships in the baseline opinion formation process, green to friendships in the extreme update, red to bimodality in baseline, blue to bimodality in extreme, purple to polarization in baseline, and orange to polarization in extreme.

3.2 Scaled Administrator Action

The other administrator action we consider is a less extreme variation. Instead of subtracting weights from edges and risking eliminating connections, the scaled administrator only adds weights and then normalizes to preserve the weights of each row. The unfortunate result is that the matrix is no longer symmetric.

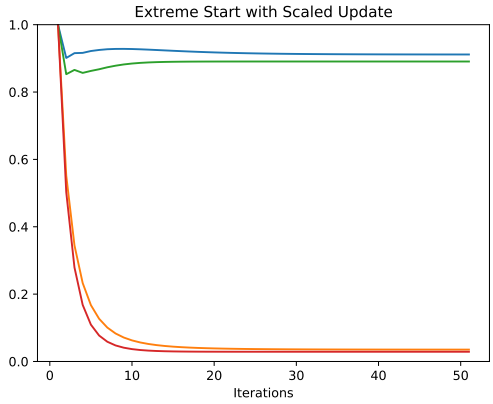
Listing 3 presents the pseudocode for the scaled update. The idea is that we want to add more weight to edges close in opinion and less to edges far in opinion. We do this by taking 2 (the most extreme possible difference between opinions) minus the absolute difference in opinion for each node and multiplying by ϵ. (Empirically, varying $0 \leq \epsilon \leq 1$ has a limited impact on the resulting figures.) If two nodes have the same opinion, we add 2ϵ; if their opinions are as far apart as possible, we do not add any weight.

Algorithm 3: Scaled Administrator Update.

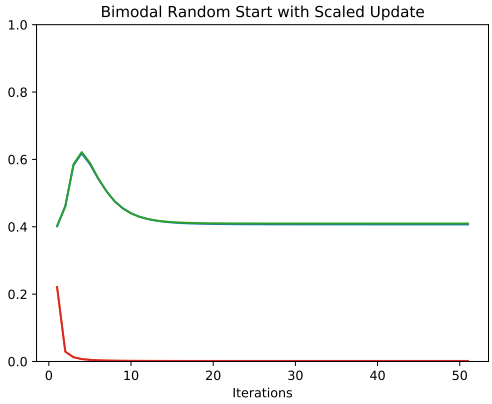
input: adjacency matrix A of dimension $n \times n$,
  expressed opinions $z^{(t)}$,
  maximum edge change ϵ per iteration
output: modified adjacency matrix A'
$A' \leftarrow$ copy of A
foreach u **in** $[n]$ **with more than one neighbor do**
  $\Delta \leftarrow \epsilon(A_{u,:} > 0) \times (2 - |z^{(t)} - z_u^{(t)}|)$ # resulting $n \times 1$ vector
  norm $\leftarrow |\text{neighbors}|/(\sum_{v \in [n]} A_{u,v} + \Delta_v)$
  $A'_{u,:} \leftarrow (A_{u,:} + \Delta)/\text{norm}$
return A'

Figure 2 shows bimodality and polarization as a function of iterations for the baseline and scaled administrator update. Notably, the time until the trends converge is substantially smaller than the times for the extreme update.

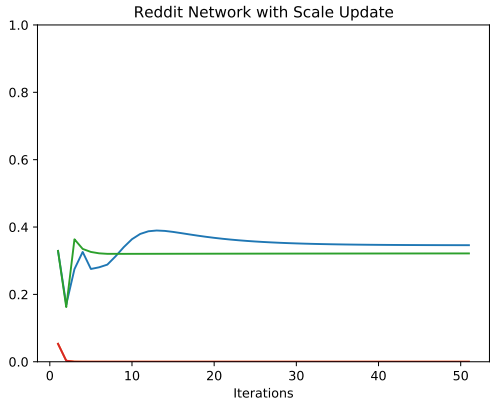
Across all four figures, polarization reaches 0 fairly quickly. The difference between the baseline and scaled update is that the scaled bimodality is higher most clearly in Figure 2d but also in Figure 2a and Figure 2c. Curiously, the bimodal SBM shows no discernible difference between the baseline and the scaled update.



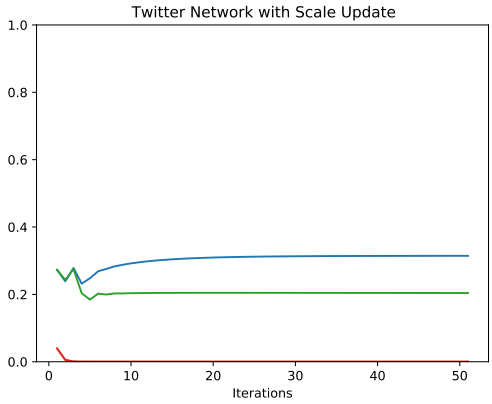
(a) SBM with half 1, -1 innate opinions.



(b) SBM with two normal opinions.



(c) Posts on r/politics and other subreddits.



(d) 2013 Delhi election opinions on Twitter.

Figure 2: Difference between baseline and scaled administrator update on four different social network and opinion pairs. Green corresponds to bimodality in the baseline update, blue to bimodality in the scaled update, red to polarization in the baseline, and orange to polarization in the scaled update.

The scaled update administrator action takes a more subtle approach. While the action boosts like-minded content, the increase in bimodality is marginal and the increase in polarization is negligible.

3.3 Miscellaneous Variations

In this section, we analyze the following variations and their effects on bimodality: attacker nodes with an extreme opinion that do not update their own opinion, three stochastic blocks rather than two, and an imbalanced opinion (1/3-2/3 vs 1/2-1/2).

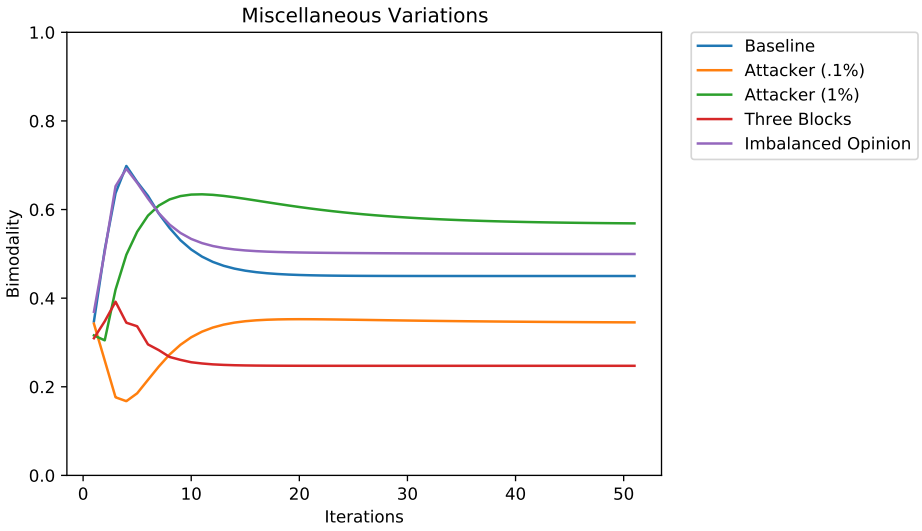


Figure 3: SBM with half 1, -1 innate opinions.

Figure 3 shows the bimodality of each variation on social networks with $n = 1000$ nodes. The imbalanced opinion is closest to the baseline, maintaining essentially the same bimodality until the 10th iteration when the bimodality of the imbalanced opinion increases. The attacker node that constitutes .1% of the population ironically seems to reduce bimodality. The attacker nodes that constitute 1% of the population markedly increase bimodality as we would expect. Finally, the three block model has the lowest bimodality. This makes sense given that the three blocks are centered at .5, 0, -.5 respectively rather than the more polarized .5 and -.5.

4 Theory

We present our theoretical results in this section. We use the work of [3] as a foundation for our results.

We consider the Stochastic Block Model (SBM) with two blocks each of size n where the probability of an edge between nodes in the same block is p and the probability of an edge between nodes in different blocks is q.

4.1 Equilibrium in Expectation

We first analyze the expected SBM. In this subsection, we do not require any bounds on p and q besides $0 \leq p, q \leq 1$ to ensure both are valid probabilities. Our first result is that we can express the equilibrium vector, z^* as a linear combination of *any* mean-centered opinion vector, rather than simply the innate opinion vector s (as in [3]).

Theorem 1. *Let $\bar{G}$ be the expected SBM graph with $2n$ nodes and adjacency matrix $\bar{A}$ where*

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

Let s be any mean-centered innate opinion vector and let z^* be the resulting equilibrium opinion vector according to FJ dynamics. Recall $z^* = (\bar{L}+I)^{-1}s$ where $\bar{L}$ is the Laplacian of $\bar{G}$. Let $(\bar{L}+I) = \bar{U}\bar{S}\bar{U}$ where $\bar{U}$ and $\bar{S}$ the eigendecomposition defined below. Then

$$z^* = \frac{c_2}{2nq+1}u^{(2)} + \frac{c_3}{nq+np+1}u^{(3)} + \cdots + \frac{c_3}{nq+np+1}u^{(3)}$$

where $u^{(i)}$ is an eigenvector of $(\bar{L}+I)$ for $i \in \{2, \dots, 2n\}$.

Proof. Let $U' = \begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix}$ where $u^{(1)}$ is the normalized all 1's vector and $u^{(2)}$ is the normalized vector with n 1's followed by n -1's. A back-of-the-envelope calculation shows that $\bar{A} + pI = U'S'U'^T$ where $S' = \text{diag}(np+nq, np-nq)$. Now, define $\bar{U} = \begin{bmatrix} u^{(1)} & u^{(2)} & Z \end{bmatrix}$, where $Z \in \mathbb{R}^{2n \times (2n-2)}$ is a matrix with orthonormal columns satisfying $Z^T u^{(1)} = 0 = Z^T u^{(2)}$ built by extending $u^{(1)}$ and $u^{(2)}$ to an orthonormal basis.

With the observation that $\bar{U}\bar{U}^T = \bar{U}^T\bar{U} = I$, we have:

$$\begin{aligned} \bar{L} + I &= \bar{D} + I - \bar{A} = (np+nq)I + I - (\bar{A} + pI) \\ &= (np+nq+1)I - U'S'U'^T = \bar{U}\bar{S}\bar{U}^T \end{aligned} \tag{4}$$

where $\bar{S} = \text{diag}(1, 2nq+1, np+nq+1, \dots, np+nq+1)$. Since $\bar{U}$ is an orthonormal basis, it follows that $(\bar{L}+I)^{-1} = \bar{U}\bar{S}^{-1}\bar{U}$. Additionally, because $\bar{U}$ is an orthonormal basis, we can write:

$$s = c_1 u^{(1)} + c_2 u^{(2)} + \cdots + c_{2n} u^{(2n)}$$

for scalars $c_i \in \mathbb{R}$ and columns $u^{(i)}$ of $\bar{U}$ where $i \in [2n]$. Notice that $c_1 = 0$ since s is mean-centered. Then we have that:

$$\begin{aligned} z^* = (\bar{L} + I)^{-1} s &= \bar{U}\bar{S}\bar{U}^T(c_2 u^{(2)} + c_3 u^{(3)} + \cdots + c_{2n} u^{(2n)}) \\ &= \bar{U}\bar{S}\begin{bmatrix} 0 & c_2 & c_3 & \cdots & c_{2n} \end{bmatrix}^T \\ &= \frac{c_2}{2nq+1} u^{(2)} + \frac{c_3}{nq+np+1} u^{(3)} + \cdots + \frac{c_3}{nq+np+1} u^{(3)} \end{aligned}$$

Theorem 1 immediately follows. □

4.2 Polarization Bounds

We next present bounds for the polarization of the equilibrium opinion resulting from FJ dynamics on any mean-centered innate opinion vector. We extend the results of [3] in the following ways:

- Instead of requiring innate opinion vectors where the first block has opinion all 1 and the second block has opinion all -1, we allow for any mean-centered innate opinion vector.
- We consider the additional case of $q \geq p$. This case has real-world applications: for example, when dealing with opinion formation in a network of doctors and patients, opinions may be more likely to develop between doctors and patients than doctors and doctors or patients and patients.
- We explicitly describe the factors on the bound of polarization, and explain how they tighten as n increases.
- We describe the special case $p = q$, equivalent to the Erdős-Rényi graph, and tightly characterize the bounds in this case.

Unfortunately, we trade the extension to any mean-centered innate opinion to a limited result. Instead of allowing $q \geq 1/n$, as in [3], we require $q \geq p/2$.

Theorem 2. *Let G be a graph generated by the Stochastic Block Model with $p/2 \leq q \leq p$ and $p \geq C \log^4 n/n$ or with $1/n \leq p \leq q$ and $q \geq C \log^4 n/n$ for some universal constant C. Let s be any mean-centered innate opinion vector on $2n$ nodes and let z^* be the equilibrium opinion vector according to FJ dynamics. Then for sufficiently large n,*

$$C' \frac{||s||_2^2}{(2nq+1)^2} \leq \mathcal{P}_{z^*} \leq C'' \frac{||s||_2^2}{(2nq+1)^2}$$

with probability 97/100 where $C' \geq 1/6$ and approaches $1/2$ as n grows while $C'' \leq 16$ and approaches 4 as n grows.

Proof. At a high level, we prove Theorem 2 by writing $\mathcal{P}_{z^*} = s^T(L+I)^{-2}s$ and bounding the eigenvalues of $(L+I)^{-2}$ through clever comparisons to $\bar{L}$, where L is the Laplacian of G and $\bar{L}$ is the Laplacian of the expected SBM $\bar{G}$.

First, notice that the normalized all 1's vector $u^{(1)}$ is an eigenvector of $(\bar{L}+I)$ by Equation (4); $u^{(1)}$ is also an eigenvector of $(L+I)$, since the columns and rows of L must sum to 0 (by the definition of a graph Laplacian). Unfortunately, $u^{(1)}$ has eigenvalue 1 for both $(L+I)^{-2}$ and $(\bar{L}+I)^{-2}$, which poses a problem for the bound we want to achieve. We deal with this by observing that s is mean-centered and so does not have any scaling of $u^{(1)}$. It follows that

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

where P is the projection $u^{(1)}{u^{(1)}}^T$. The point of $(I-P)$ is to remove the eigenvector $u^{(1)}$ and corresponding eigenvalue 1. Therefore, to bound $\mathcal{P}_{z^*}$, it is sufficient to bound the eigenvalues of $(I-P)(L+I)(I-P)$, since s is some linear combination of the eigenvectors of $(I-P)(L+I)(I-P)$. To bound the eigenvalues of $(I-P)(L+I)(I-P)$, we use Lemma 3, which tells us that the eigenvalues of L and $\bar{L}$ are within $.5n\max\{\mathrm{p},\mathrm{q}\}$ of each other for sufficiently large n. We leave the proof of Lemma 3 to appendix A.1 for brevity.

Lemma 3 (Extension of Lemma 4.5 in [3]). *Let L be the Laplacian of graph G drawn from the SBM and let $\bar{L} = \mathbb{E}[L]$. For fixed constant C', with probability 98/100,*

$$||L - \bar{L}||_2 \leq C'\sqrt{n \log n \max\{\mathrm{p}, \mathrm{q}\}}.$$

Note that when $\max\{\mathrm{p}, \mathrm{q}\} \geq C \log^4 n/n$, $C'\sqrt{n \log n \max\{\mathrm{p}, \mathrm{q}\}} \leq \frac{C'}{\sqrt{C}\log^{1.5} n} n \max\{\mathrm{p}, \mathrm{q}\}$. So for $n \geq \exp(\frac{C'}{\sqrt{C}\epsilon})^{2/3}$, $||L - \bar{L}||_2 \leq \epsilon n \max\{\mathrm{p}, \mathrm{q}\}$. Choose $\epsilon = 1/2$.

From Weyl's Inequality and Lemma 3, we now have that

$$|\lambda_i - \bar{\lambda}_i| \leq ||L - \bar{L}||_2 \leq \frac{1}{2} n \max\{\mathrm{p}, \mathrm{q}\} \tag{5}$$

for sufficiently large n, where λ_i is the ith eigenvalue of L and $\bar{\lambda}_i$ is the ith eigenvalue of $\bar{L}$. We now bound the largest $2n-1$ eigenvalues of $L + I$ or, equivalently, the eigenvalues of $(I - P)(L + I)(I - P)$. The eigenvalues of $(I - P)(\bar{L} + I)(I - P)$ are $2nq + 1$ and $np + nq + 1$.

Case 1: $q \geq p$ The smallest eigenvalue of $(I - P)(\bar{L} + 1)(I - P)$ is $nq + np + 1 \geq nq + 1$ while the largest eigenvalue is $2nq + 1$. Then, Equation (5) implies that:

$$(.5nq + 1)(I - P) \preceq (I - P)(L + I)(I - P) \preceq (2.5nq + 1)(I - P).$$

We square and invert each term, which yields:

$$(2.5nq + 1)^{-2}(I - P) \preceq (I - P)(L + I)^{-2}(I - P) \preceq (.5nq + 1)^{-2}(I - P).$$

Now, we can bound

$$\frac{1}{2}\frac{||s||_2^2}{(2nq+1)^2} \leq \frac{||s||_2^2}{(2.5nq+1)^2} \leq s^T(L + I)^{-2}s \leq \frac{||s||_2^2}{(.5nq+1)^2} \leq \frac{16||s||_2^2}{(2nq+1)^2}.$$

Case 2: $p \geq q \geq p/2$ The smallest eigenvalue of $(I - P)(L + I)(I - P)$ is $2nq + 1 \geq np + 1$ while the largest eigenvalue is $nq + np + 1 \leq 2np + 1$. Similar analysis to Case 1 yields

$$\frac{1}{6}\frac{||s||_2^2}{(2nq+1)^2} \leq \frac{||s||_2^2}{(2.5np+1)^2} \leq s^T(L + I)^{-2}s \leq \frac{||s||_2^2}{(.5np+1)^2} \leq \frac{16||s||_2^2}{(2nq+1)^2}.$$

We now justify the upper and lower bound on C' and C'', respectively. Consider a small ϵ such that Lemma 3 gives a small additive error. Then, without loss of generality, the largest eigenvalue between Cases 1 and 2 is $(2 + \epsilon)np + 1$. It follows that C' gets arbitrarily close to $1/2$ as n grows. The smallest eigenvalue between Cases 1 and 2 $(1 - \epsilon)np + 1$ without loss of generality. It follows that C'' gets arbitrarily close 4 as n grows. □

For $p = q$, the SBM is simply an Erdős-Rényi random graph. Therefore, a special case follows immediately from Theorem 2.

Theorem 4 (Erdős-Rényi Special Case). *Let G be a Erdős-Rényi graph where the probability of an edge between any pair of nodes is $p \geq C \log^4 n/n$ for some universal constant C. Let s be any mean-centered innate opinion vector on $2n$ nodes and let z^* be the equilibrium opinion vector according to FJ dynamics. Then for sufficiently large n,*

$$C'\frac{||s||_2^2}{(2np+1)^2} \leq \mathcal{P}_{\mathrm{z}^*} \leq C''\frac{||s||_2^2}{(2np+1)^2}$$

with probability 97/100 where $C' \geq 1/2$ and $C'' \leq 2$. Both C' and C'' approach 1 as n increases.

Proof. The theorem immediately follows as a special case of Theorem 2, except for the improved constants C' and C''. To obtain the improved bounds, observe that the only eigenvalue of $(I-P)(L+I)(I-P)$ is $2np+1$. Then, following the earlier analysis, we have that:

$$\frac{1}{2}\frac{||s||_2^2}{(2np+1)^2} \leq \frac{||s||_2^2}{(2.5np+1)^2} \leq s^T(L+I)^{-2}s \leq \frac{||s||_2^2}{(1.5np+1)^2} \leq \frac{2||s||_2^2}{(2np+1)^2}.$$

Setting ϵ to a small positive value yields an upper bound of $(2+\epsilon)np+1$ on the eigenvalues and a lower bound of $(2-\epsilon)np+1$. It is easy to see that the factors C' and C'' get arbitrarily close to 1 as ϵ decreases. □

We remark that extending our results to the case of $1/n \leq q \leq p/2$ would require bounding the inner product of the largest $n-2$ eigenvectors, even if the innate opinion vector s did not have any scaling of the all 1's vector or the second smallest eigenvector.

5 Conclusion

Filter bubbles are an important ethical issue. Improving the mathematical theory behind their formation presents an interesting and applicable area of research. In this work, we empirically analyzed several variations on the standard FJ dynamics opinion formation process including the introduction of two natural administrator actions. Our theoretical contribution is a series of extensions to the convergence bounds given by [3]. Notably, we apply the bounds to any mean-centered innate opinion vector and tightly characterize the constraints when the Stochastic Block Model is in the special case of the Erdős–Rényi graph.

5.1 Future Work

While answering some questions, our work points to several avenues for future research. The power of the theoretical bounds come from the natural linear-algebraic description of FJ dynamics. It would be interesting to write a network administrator action in terms of a matrix multiplication. Our main obstacle in this pursuit was balancing the 'macro' scale of any linear-algebraic action with the 'micro' scale employed in practice by social media platforms. Bimodality is a natural and interesting alternative measure of polarization. However, the bimodality coefficient involves the third and fourth moments of expectation. While polarization can easily be represented as a quadratic measure, cubic and quartic functions are more difficult to represent in terms of vector operations. An avenue for future research is either finding a different measure or writing bimodality in a neat expression.

In terms of theory, we envision a way to give an explicit expression for the bounds on the polarization convergence in terms of the size of the network n. We did not have sufficient time to write this up before the deadline but hope to extend it later.

5.2 Reflection and Responsibilities

We enjoyed working on a problem at the intersection of theoretical computer science and ethical issues. It was satisfying to apply mathematical tools to a problem facing the modern world. Our greatest challenge was thinking of linear algebraic formulations of network administrator actions and bimodality; we ended up focusing on experimental results for these ideas, and have left them as interesting future research directions. However, we feel that we extended existing theoretical results and introduced interesting variations to this problem, which we studied empirically. It is our hope that we can publish this work or a future iteration of it. Finally, we thank Professor Chris Musco and Raphael Meyer for discussing this project with us and providing input on generating theoretical results.

Responsibilities for the project and report: Indu focused on related work, background, and processing the data we used while Teal coded the experiments and came up with the theoretical extensions. We collaborated on the abstract ideas of each section and worked together to move through challenging parts in the theoretical analysis.

References

[1] Emmanuel Abbe. Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, 2018.

[2] Daniel W Bromley. A tale of two americas: Why is that a surprise? *Choices*, 32(2):1–4, 2017.

[3] Uthsav Chitra and Christopher Musco. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 115–123, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Pranav Dandekar, Ashish Goel, and David Lee. Biased assimilation, homophily and the dynamics of polarization. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 03 2013.

[5] Morris H. Degroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[6] Dominic Difranzo and Kristine Gloria-Garcia. Filter bubbles and fake news. *XRDS: Crossroads, The ACM Magazine for Students*, 23:32–35, 04 2017.

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

[8] Timothy Hoff. Patient-doctor trust and the COVID-19 Vaccine, December 2020.

[9] Cameron Brick Lee de wit, Sander Van der Linden. Are social media driving political polarization? 2019.

[10] Antonis Matakos, Cigdem Aslay, Esther Galbrun, and Aristides Gionis. Maximizing the diversity of exposure in a social network. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 11 2020.

[11] Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 369–378, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[12] Jehane Noujaim and Karim Amer. The great hack. 2019.

[13] Eli Pariser. The filter bubble: what the internet is hiding from you. 2011.

[14] V. H. Vu. Spectral norm of random matrices. In *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '05, page 423–430, New York, NY, USA, 2005. Association for Computing Machinery.

A Appendix

A.1 Proof of Lemma 3

We introduce several lemmas that we will use to prove Lemma 3. Observe that this analysis closely follows that of [3] except that we generalize it to the case when $q \geq p$.

Lemma 5 (Theorem 1.4 in [14])**.** *There are constants C and C' such that the following hold. Let $M_{i,j}$ be an independent random variable with mean 0, variance at most σ^2 for $1 \leq i \leq j \leq m$, and absolute value bounded by 1 where $\sigma \geq C'\sqrt{m}\log^2 m$. Then almost surely*

$$\lambda(M) \leq 2\sigma\sqrt{m} + C\sqrt{\sigma}m^{1/4}\log m$$

where $\lambda(M)$ denotes the spectral norm (i.e. maximum singular value) of M.

Lemma 6 (Extension of Lemma 4.5 in [3]). *Let A be the adjacency matrix of a graph drawn from the SBM with intra-block probability p and inter-block probability q. Define $\bar{A} = \mathbb{E}[A]$. There exists a universal constant C such that if $p \geq C \log^4 n/n$ then with probability 99/100,*

$$||A - \bar{A}||_2 \leq 3\sqrt{n \max\{\mathrm{p}, \mathrm{q}\}}.$$

Proof (Omitted in [3]). Consider the matrix $M = A - \bar{A}$ with $m = 2n$. Each entry $M_{i,j}$ for $i \neq j$ is a binary variable with probability p or q of firing. Recall that $M_{i,i} = 0$. Then $\mathbb{E}[M_{i,j}] = \mathbb{E}[A_{i,j}] - \mathbb{E}[A]_{i,j} = 0$ and the variance σ^2 is less than $\max\{\mathrm{p}, \mathrm{q}\}$. The absolute value of each entry is bounded by 1, since every entry of A is either 0 or 1, and every entry of $\bar{A}$ is between 0 and 1. We use that $1 \geq \sqrt{\max\{\mathrm{p}, \mathrm{q}\}} > \sigma$ and $\sqrt{\sigma} \geq C'(2n)^{1/4} \log 2n$. Then Lemma 5 yields

$$\begin{aligned}\lambda(A - \bar{A}) &\leq 2\sqrt{\sigma}\sqrt{2n} + C\sqrt{\sigma}(2n)^{1/4} \log 2n \\ &\leq 2\sqrt{\max\{\mathrm{p}, \mathrm{q}\}\, 2n} + \frac{C}{C'}\sqrt{\sigma}\sqrt{\sigma} \leq 3\sqrt{n \max\{\mathrm{p}, \mathrm{q}\}}\end{aligned}$$

for sufficiently large n. Since $A - \bar{A}$ is a square matrix, $||A - \bar{A}|| \leq \lambda(A - \bar{A})$, which completes the proof. □

Lemma 7 (Bernstein Inequality). *Let $X_1, \ldots, X_m$ be independent random variables with variances $\sigma_1^2, \ldots, \sigma_m^2$ and $|X_i| \leq 1$ almost surely for $i \in [m]$. Let $X = \sum_{i \in [m]} X_i$, $\mu = \mathbb{E}[X]$, and $\sigma^2 = \sum_{i \in [m]} \sigma_i^2$. Then the following holds:*

$$\Pr(|X - \mu| > \epsilon) \leq \exp\left(\frac{e^2}{2\sigma^2 + \epsilon/3}\right)$$

With Lemma 6 and Lemma 7, we can now prove Lemma 3.

Proof of Lemma 3. Let D be the degree matrix of G and define $\mathbb{E}[D] = \bar{D}$. By the triangle inequality, $||L - \bar{L}||_2 \leq ||D - \bar{D}||_2 + ||A - \bar{A}||_2$. By Lemma 6, $||A - \bar{A}||_2 \leq 3\sqrt{n \max\{\mathrm{p}, \mathrm{q}\}}$. Additionally, $||D - \bar{D}||_2$ is bounded by $\max_{i \in [2n]} |D_{i,i} - \bar{D}_{i,i}|$. $D_{i,i}$ is a sum of Bernoulli random variables with total variance σ^2 upper bounded by $2n \max\{\mathrm{p}, \mathrm{q}\}$. It follows from Lemma 7 and our assumption $p = \Omega(1/n)$ that $|D_{i,i} - \bar{D}_{i,i}| \leq C\sqrt{n \log n \max\{\mathrm{p}, \mathrm{q}\}}$ with probability $1 - 1/200n$ for fixed universal constant C. By a union bound, we have that $\max_i |D_{i,i} - \bar{D}_{i,i}| \leq \sqrt{n \log n \max\{\mathrm{p}, \mathrm{q}\}}$ with probability 99/100. A second union bound with the event that $||A - \bar{A}||_2 \leq 3\sqrt{n \max\{\mathrm{p}, \mathrm{q}\}}$ yields the lemma with $C' = C + 1$. □

B Supplemental Material

The code, data, and LaTeX files are all publicly available at github.com/rtealwitter/Filter-Bubbles.

The two data sets we use are described below:

- Twitter is a graph with 548 nodes and 3638 edges. Nodes correspond to users who posted tweets about the Delhi legislative assembly elections of 2013, and edges represent user interactions debating that election.
- Reddit is a graph with 556 nodes and 8969 edges. Nodes represent users who posted in the r/politics subreddit. Edges correspond to users' posts in different subreddits. In particular, two users are connected by an edge if they have both posted in two subreddits (other than r/politics) during the given time period.