Filter Bubbles

Indu Ramesh and R. Teal Witter

December 17, 2020

1 Introduction

In theory, social media offers 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 [6].

A popular explanation for this phenomenon is the emergence of *filter bubbles*, a term coined by internet activist Eli Pariser [10]. 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 [4].

In this project, we study emphpolarization in social networks, aiming to provide 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 now give a summary of interesting recent approaches and results.

In [2], Dandekar et al. studied the idea of *biased assimilation*: when given mixed evidence on a complex issue, individuals use that evidence to support their intrinsic opinion, arriving at a more extreme opinion [2]. They provide empirical evidence that in a simple models of social networks, DeGroot's model of opinion formation [3] results in polarization. They also analyze the effects of internet content recommendation algorithms on biased assimilation. They show that biased assimilation leads to polarization in society, if individuals start out sufficiently biased.

In [1], Chitra and Musco studied social media companies' roles in creating filter bubbles. In particular, Chitra and Musco studied a slightly different model of opinion formation, the Friedkin-Johnsen model [5], and added 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. The network administrator's modifications are subject to certain constraints; for instance, they cannot change the degree of any vertex, and they 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. That is, a social network modeled by the stochastic block model is, with high probability, already in a state of *fragile consensus*.

Other formulations and solutions to the polarization problem have been studied. In [8], Musco et al. ask: what is the *topological structure* of a network that minimizes both polarization and disagreement, given an opinion dynamics model? They pose this question as an optimization problem, and for a fixed graph, give a polynomial-time algorithm that approximates the optimum. When generalized as an influence maximization problem, [7] gives an approximation algorithm to break filter bubbles.

Another measure related to disagreement in a network is *biomdality.* While [1] and [2] both find polarization increases in social networks, there might be another force at work driving the formation of filter bubbles:

bimodality. We hypothesize that opinion dynamics cause the distribution of opinions in the network to become more *bimodal* as time passes; in other words, opinions converge in clusters around two centers, rather settling far away from the mean. One measure of bimodality, derived by Warren Sarle, is the *bimodality coefficient*, β:

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

where γ is the *skewness* (third standardized moment) and κ is the *kurtosis* (fourth standardized moment).

The work we mentioned has all addressed the question: how sensitive are networks to polarization? [1] suggests that many networks are in a state of *fragile consensus.* The Netflix documentary "The Great Hack" [9] 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 and DeGroot settings, which we will more formally define in the "Background" section below. First, we give experimental results on polarization and bimodality in these models, as affected by two network administrator actions. We also introduce "attack" nodes into the social network, with strong opinions, and analyze their influence.

In our project, we also provide theoretical results, studying polarization as a linear algebraic problem, modeled after the approach in [1]. We generalize Chitra and Musco's results on the relationship of polarization to the innate opinions of the users in a social network.

2 Background

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 individual is a node, and is connected by undirected edges to their friends 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 (i.e. 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 model* to construct a graph representing a social network. The stochastic block model is a common generative model seen in numerous fields, including statistics, theoretical computer science, and machine learning.

The Graph G that the stochastic block model generates follows the general layout of a social network, where each individual corresponds to a node, with weighted edges corresponding to social relationships.

We now describe the mathematical particulars of the model. Graph G has $2n$ vertices, divided into two communities, S and T. The set of vertices $v_1, v_2, ..., v_n \in S$, and $v_{n+1}, v_{n+2}, ..., v_{2n} \in T$. The edge set of G is generated as follows:

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

In other words, the probability of two nodes being connected is p when the nodes are in the same community, and q when the nodes are in different communities in different communities.

2.2 Mathematical models of opinion formation

Prior work has attempted to explain polarization via the two mathematical models of opinion formation. Opinion dynamics are studied, in these models, by adding an extra parameter to the graph corresponding to the social network: opinions. Each individual node is augmented with a certain opinion, z_i, modeled as a real number in the interval $[-1, 1]$.

DeGroot's Model: Perhaps the most popular model of opinion dynamics studied, DeGroot's updates an individual's opinion using the following update rule: the individual node's current opinion is added with that of its neighbors' opinions, weighted by the edge weight connecting the individual to each neighbor, and averaged over the node's degree plus one. More formally, here is the update rule:

$$z_i^{(t)} = \frac{z_i^{(t-1)} + \sum_{j \neq i} w_{ij} z_j^{(t-1)}}{d_i + 1} \tag{2}$$

where $z_i^{(t)}$ is the opinion of node i at iteration t, $z_i^{(t-1)}$ 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 opinion dynamics model is similar to DeGroot's Model, but follows a slightly different update rule: the individual node's *innate* opinion is added with that of its neighbors' opinions, weighted by the edge weight connecting the individual to each neighbor, and averaged over the node's degree plus one. In other words, in this model, individuals can be seen as more "stubborn" than in DeGroot's model. Mathematically, we can write the update rule as follows:

$$z_i^{(t)} = \frac{s_i + \sum_{j \neq i} w_{ij} z_j^{(t-1)}}{d_i + 1} \tag{3}$$

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

2.3 Measuring Polarization

One measure of polarization with a natural mathematical interpretation is is the *network disagreement* index. In [2], 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:

$$\eta(G, x) = \sum_{i,j \in E} w_{ij} (x_i - x_j)^2 \tag{4}$$

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

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

$$mean(\mathbf{z}) = \frac{1}{n} \sum_{j=1}^{n} x_j \tag{5}$$

where z_j is the mean opinion of $\mathbf{z}$. Then, polarization, P_z can be defined as:

$$P_z = \sum_{i=1}^{n} (x_i - mean(\mathbf{z}))^2 \tag{6}$$

2.4 Linear Algebraic Interpretation of Polarization

In [1], Chitra and Musco provide simple linear algebraic interpretations of polarization on a graph generated by the stochastic block model.

First, we need a clear formulation Friedkin-Johnsen dynamics in our social network graph, G. We describe the approach given in [1].

Let $A \in R^{nxn}$ be the adjacency matrix of G, where $A_{ij} = A_{ji} = w_{ij}$. Recall that the adjacency matrix is symmetric. Let D be a diagonal matrix where $D_{ii} = d_i$; in other words, 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 $\mathbf{s}$ is the innate opinion vector of all nodes in G. It is not too hard to see that 3 is equivalent to:

$$\mathbf{z}^{(t)} = (D + I)^{-1}(A\mathbf{z}^{(t-1)} + \mathbf{s} \tag{7}$$

where $\mathbf{z}^{(t)} = [z_1^{(t)}, z_2^{(t)}, ..., z_n^{(t)}]$.

Let $\mathbf{z}^{(*)}$ be the final opinion vector after t time steps. We have that:

$$\mathbf{z}^{(*)} = (L + I)^{-1}\mathbf{s} \tag{8}$$

Finally, Note that in the theoretical work in our project, we will strengthen 8.

Equation 14

2.5 Network Administrator

Additionally, they show that there is a network administrator action that leads to high polarization. They define the network adminsitrator dynamics as follows:

3 Experiments

It is often challenging to mathematically analyze variations on the opinion formation process so we simulate them instead. 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).

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

Algorithm 1: Opinion formation.

```
input: adjacency matrix $A_0$ of dimension $n \times n$,
       innate set of opinions $s$ with $s_i \in [-1, 1]$ for $i \in [n]$,
       number of iterations $T$ to run
output: a set of equalibrium opinions $z^{(T)}$
$z^{(0)} \leftarrow s$
foreach $t$ in $\{0, \ldots, T\}$ do
    $D \leftarrow$ degree matrix of $A_t$
    $z^{(t)} \leftarrow (D + I)^{-1}(A_t z^{t-1} + s)$
    $A_t \leftarrow$ network administrator modifications of $A_{t-1}$
return $z^{(T)}$
```

3.1 Extreme Administrator Action

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

Algorithm 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 $n \times n$,
       expressed opinions $z^{(t)}$,
       maximum edge change $\epsilon$ per iteration
output: modified adjacency matrix $A'$
$A' \leftarrow$ all-zero $n \times n$ matrix
foreach $u$ in $[n]$ with more than one neighbor do
    c $\leftarrow$ neighbor with closest opinion to $u$
    f $\leftarrow$ neighbor with furthest opinion to $u$
    $\Delta \leftarrow \min\{\epsilon, A_{u,f}\}$
    $A'_{u,f} \leftarrow A'_{f,u} \leftarrow A_{f,u} - \Delta$
    $A'_{u,c} \leftarrow A'_{c,u} \leftarrow A_{c,u} + \Delta$
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. Otherwise, it stabilizes between .2 and .6. 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 overtime and stabilizes near 0.

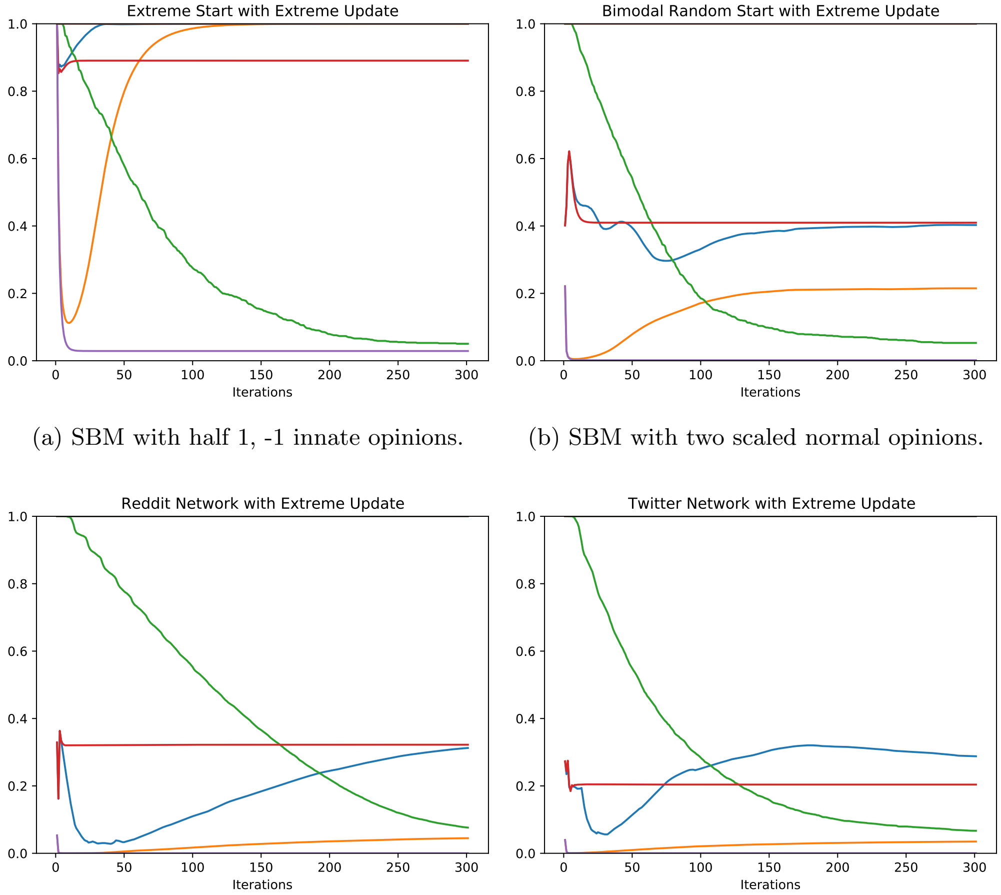
The extreme administrator update is a naive approach to the natural mechanism social media platforms use to boost engagement: increase exposure to content similar in nature to user preferences and simultaneously decrease exposure to content dissimilar to user preferences.

While the overall weight (interpreted as e.g. time by [1]) 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.

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.

Algorithm 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 then we add 2ϵ whereas if their opinions are as far apart as possible, we do not add any weight.



(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.

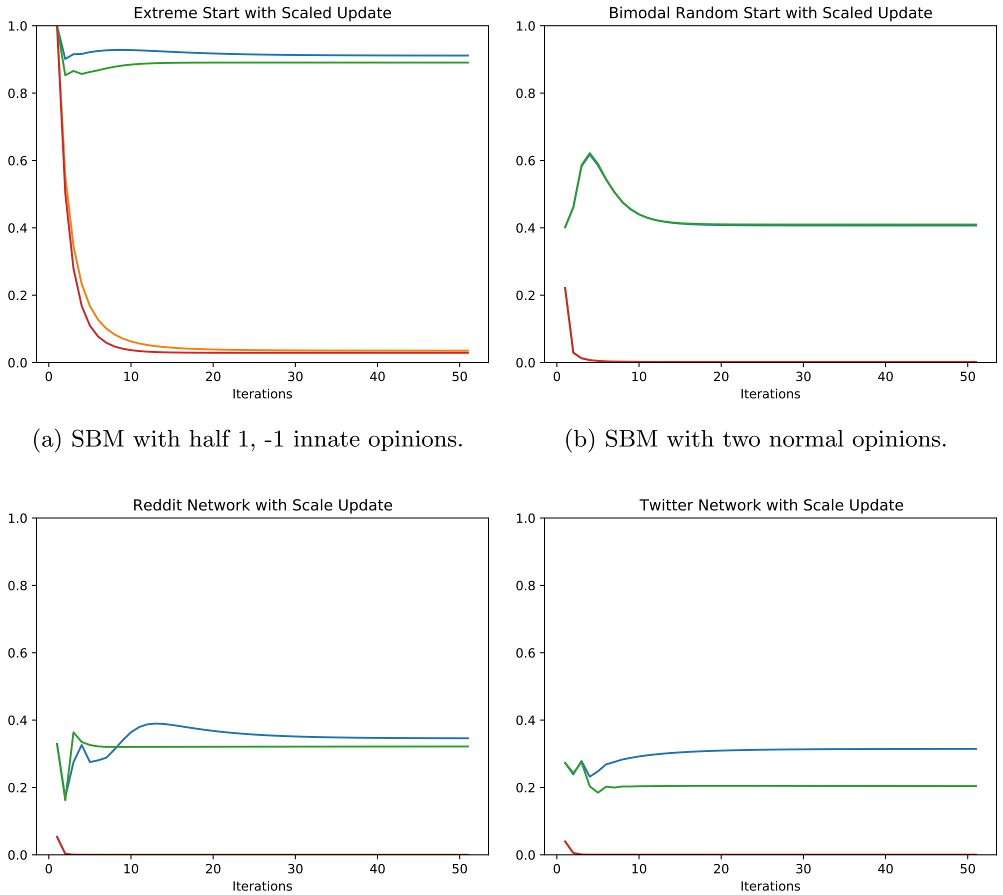
Algorithm 3: Scaled Administrator Update.

```
input: adjacency matrix $A$ of dimension $n \times n$,
       expressed opinions $z^{(t)}$,
       maximum edge change $\epsilon$ 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 $\leftarrow |\text{neighbors}|/(\sum_{v\in[n]} A_{u,v} + \Delta_v)$
    $A'_{u,:} \leftarrow (A_{u,:} + \Delta)/$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 discernable 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: 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).

4 Theory

We present our theoretical results in this section. We use the work of [1] 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.

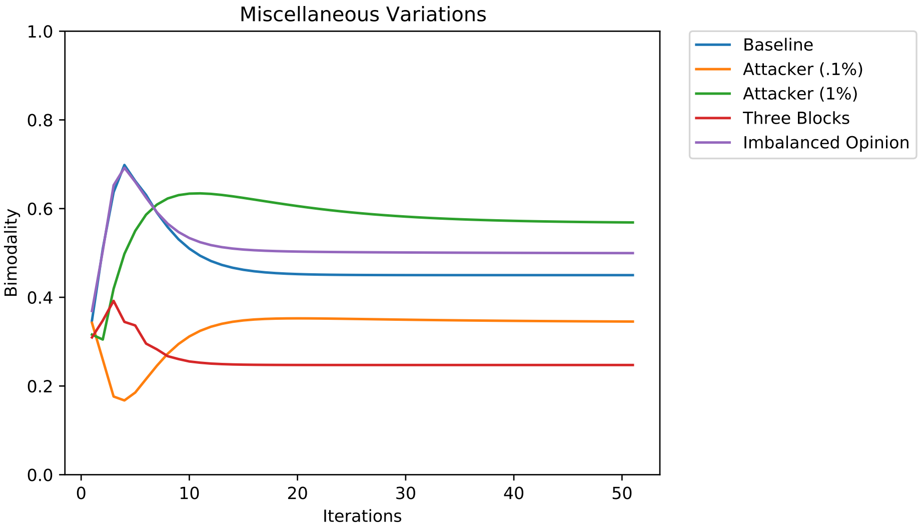


Figure 3: SBM with half 1, -1 innate opinions.

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.

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 $(L+I)$ for $i \in \{2, \dots, 2n\}$.

Proof. Let $U' = \begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix}$ where $u^{(1)}$ is the normalized all 1s vector and $u^{(2)}$ is the normalized vector with n 1s followed by n -1s. 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

$$\begin{aligned}\bar{L} + I &= \bar{D} + I - \bar{A} = (np + nq)I + I - (\bar{A} + pI) \\ &= (np + nq + 1)I - U'S'U'^T = \bar{U}\bar{S}\bar{U}^T \end{aligned} \tag{9}$$

where $\bar{S} = \text{diag}(1, 2nq + 1, np + nq + 1, \ldots, np + nq + 1)$. Since $\bar{U}$ is an orthonormal basis, it follows that $(\bar{L} + I)^{-1} = \bar{U}\bar{S}^{-1}\bar{U}$. Also 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

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

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. Observe that we extend the results of [1] in the following ways:

- Instead of the 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 $q \geq p$ which has real-world applications to e.g. doctor-patient opinion formation.
- We explicitly describe the factors on the bound of polarization and explain how they tighten as n increases.
- We describe the special case $p = q$ which is 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 when $p \geq q$. That is, instead of allowing $q \geq 1/n$ as in [1] 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 observe that the normalized all 1s vector $u^{(1)}$ is an eigenvector of $(\bar{L} + I)$ by Equation (9) and that $u^{(1)}$ is also an eigenvector of $(L + I)$ since the columns and rows of L must sum to 0 by the definition of a Laplacian. Unfortunately, $u^{(1)}$ is 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 $(I-P)(L+I)(I-P)$. In order 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 [1])**.** *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

$$|\lambda_i - \bar{\lambda}_i| \leq ||L-\bar{L}||_2 \leq \frac{1}{2} n \max\{\mathrm{p},\mathrm{q}\} \tag{10}$$

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: $\boldsymbol{q \geq p}$ 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 (10) implies that

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

We square then invert each term which yields

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

Then 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: $\boldsymbol{p \geq q \geq p/2}$ 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 the largest eigenvalue between Cases 1 and 2 is $(2+\epsilon)np+1$ without loss of generality. 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. □

Theorem 2 applies to the SBM but observe for $p = q$ that the SBM is simply an Erdős-Rényi 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,*

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

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 get the better bounds, observe that the only eigenvalue of $(I-P)(L+I)(I-P)$ is $2np+1$. Then following the earlier analysis

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

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 the second smallest eigenvector.

References

[1] 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.

[2] 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.

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

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

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

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

[7] 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.

[8] 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.

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

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

[11] 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 [1] except that we generalize it to the case that $q \geq p$.

Lemma 5 (Theorem 1.4 in [11])**.** *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 [1])**.** *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 [1]). 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 $X_1, \dots, X_m$ be independent random variables with variances $\sigma_1^2, \dots, \sigma_m^2$ 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 $C' = C + 1$. □