Local Edge Dynamics and Opinion Polarization

Nikita Bhalla  Adam Lechowicz  Cameron Musco
University of Massachusetts Amherst  University of Massachusetts Amherst  University of Massachusetts Amherst
Amherst, MA, USA  Amherst, MA, USA  Amherst, MA, USA
nbhalla@cs.umass.edu  alechowicz@cs.umass.edu  cmusco@cs.umass.edu

ABSTRACT
The proliferation of social media platforms, recommender systems, and their joint societal impacts have prompted significant interest in opinion formation and evolution within social networks. We study how local edge dynamics can drive opinion polarization. In particular, we introduce a variant of the classic Friedkin-Johnsen opinion dynamics, augmented with a simple time-evolving network model. Edges are iteratively added or deleted according to simple rules, modeling decisions based on individual preferences and network recommendations.

Via simulations on synthetic and real-world graphs, we find that the combined presence of two dynamics gives rise to high polarization: 1) confirmation bias — i.e., the preference for nodes to connect to other nodes with similar expressed opinions and 2) friend-of-friend link recommendations, which encourage new connections between closely connected nodes. We show that our model is tractable to theoretical analysis, which helps explain how these local dynamics erode connectivity across opinion groups, affecting polarization and a related measure of disagreement across edges. Finally, we validate our model against real-world data, showing that our edge dynamics drive the structure of arbitrary graphs, including random graphs, to more closely resemble real social networks.

Our appendix and code are available at github.com/adamlechowicz/opinion-polarization/.

CCS CONCEPTS
• Mathematics of computing → Random graphs; • Theory of computation → Random network models.

KEYWORDS
opinion polarization, Friedkin-Johnsen model, polarization and disagreement, recommender systems, random graph models

1 INTRODUCTION
Over the last twenty years, the rise of massive social media platforms has significantly increased information sharing and human interaction around the globe. While information availability and richer online interactions are positive effects of this social shift, there is increasing concern about the ability of these platforms to polarize and divide us [3, 14, 18, 20, 23]. This polarization seems to be driven by both established factors of human social dynamics, along with new dynamics, driven by the behavior of the social media platforms themselves.

An example of a human behavior driving polarization is confirmation bias, the tendency to avoid information that challenges our own views and to seek information that confirms them [30]. Confirmation bias is amplified on social media platforms due to the increased availability of opinion-confirming content. Indeed, it is thought to be a key driver of the online spread of conspiracy theories and fake news in recent years [26, 32, 38].

An example of the behavior of social media platforms driving polarization is the use of recommender systems to filter and deliver content that maximizes user engagement. Such recommendations can create filter bubbles, which further strengthen the power of confirmation bias and drive polarization [18, 23, 31].

1.1 Our Model
We seek to understand how local edge updates (i.e., insertions and deletions) driven by both human behavior and the behavior of social media platforms may cause opinion polarization. To do so we introduce a simple model of network and opinion coevolution, which 1) is based on established opinion dynamics models, 2) captures the effects of both confirmation bias and recommender systems, and 3) remains tractable to theoretical analysis and efficient simulation. This model is detailed in Section 2 and summarized below. Beyond our work, we hope that the model will serve as a useful platform for further investigation of network and opinion coevolution.

Opinion Dynamics and Polarization. We build on the classic Friedkin-Johnsen opinion model, [16], which models how individuals’ expressed opinions (represented as real numbers) are influenced by their innate opinions, along with the expressed opinions of their neighbors in a network. A node’s innate opinion is fixed at initialization, and its expressed opinion at any time is an average of this innate opinion with the expressed opinions of its neighbors.

We consider two metrics studied in prior work [28]: opinion polarization, which is the variance of the expressed opinions, and disagreement, which is the total squared difference of expressed opinions summed over all edges in the network. Under the Friedkin-Johnsen model, polarization and disagreement tend to counteract each other — with low disagreement, nodes tend to be connected to other nodes with similar opinions, and polarization tends to be high. With high disagreement, nodes are connected to other nodes with...
a diversity of opinions, limiting polarization. Both values can be efficiently computed in closed form [28, 39], making them tractable for both theoretical and empirical investigation.

**Edge Dynamics.** We consider a model where the network and expressed node opinions coevolve over time. At each time step, users within the network stochastically delete connections to some of their neighbors and add new connections. To model confirmation bias, edges that are more disagreeable are deleted with higher probability. To model recommender systems, new edges are added to random friend-of-friends – i.e., two-hop connections in the network. Friend-of-friend recommendations are a popular form of edge recommendations made by social media platforms [10, 37].

We also soft x a small percentage of edges at initialization, which are not considered for deletion at any time step. These edges can be thought of as modeling connections that are independent of opinions or recommendations, e.g., to family members or co-workers.

**Simulation and Theoretical Analysis.** To study how the above opinion and edge dynamics interact to drive opinion polarization, we employ both simulation and theoretical analysis. In our simulations, we start with an initial graph, either randomly generated from an established model, or taken from a snapshot of a real social network. We also start with innate opinions, which are randomly generated over the interval $[-1, 1]$. We then simulate our edge and opinion dynamics, recording how opinions, polarization, disagreement, and the graph structure evolve over time. Our theoretical results, which help explain many of the dynamics observed in simulation, leverage well-established formulas for polarization and disagreement in the Friedkin-Johnsen model, defined in Section 2.

### 1.2 Our Findings

Our main findings are as follows:

1. **Confirmation bias and recommendations drive polarization.** Well nd that, when both confirmation bias and friend-of-friend link recommendations are part of the edge dynamics model, the network becomes polarized – nodes sort into distinct clusters of similar and opposing opinions. When the network has n o f f edges, edge dynamics splinter it into clusters such that expressed opinions are very close to the innate opinions, and polarization nearly reaches its maximum value (see Secs. 4.2 and 4.3).

   If either one of these dynamics is removed (i.e., either edge additions or removals are random), polarization remains low. This finding is generally robust to the initial graph structure and innate opinion distribution, with the exception of dense random graphs, where friend-of-friend recommendations lose their power – they behave similarly to random recommendations. See Fig. 1 for an illustration and Sec.4.2.1 for more detailed discussion.

   While perhaps intuitive, the above finding is far from obvious. In the Friedkin-Johnsen model, a node’s expressed opinion is heavily influenced by those of its neighbors. Initially, a node’s expressed opinion will depend very little on its innate opinion. Thus, it is not clear that removing disagreeable edges (where disagreement is with respect to the expressed opinions) and adding friend-of-friends will lead to polarization. We also nd it surprising that both friend-of-friend recommendations and confirmation bias are needed to drive polarization, initially conjecturing that just confirmation bias would suffice. Recommender systems seem to act as a key catalyst in amplifying human behavior that favors polarization.

![Figure 1: Opinion polarization over time for an Erdös-Renyi (ER) graph with 1000 nodes and connection probability $p = 0.05$, a real-world Twitter social network [12], and a dense ER graph with 1000 nodes and connection probability $p = 0.1$](Image)

Each network has 5% fixed edges. Innate opinions are uniformly distributed in $[-1, 1]$. Except in the dense ER graph, with edge dynamics influenced by both confirmation bias (CB) and friend-of-friend recommendations (F-o-F), polarization rises significantly before asymptoting. When either factor is removed, polarization remains low.

2. **Our model is tractable to theoretical analysis.** To complement our simulation results, we give theoretical bounds which illustrate how edge removals and insertions can drive polarization and disagreement in the Friedkin-Johnsen model. We prove that swapping an edge with large expressed opinion disagreement across it for a new edge with small disagreement monotonically drives up the sum of polarization and disagreement in the graph, a combined quantity studied in prior work [28], whose rise seems closely correlated with a rise in polarization itself.

This finding helps explain how both confirmation bias and friend-of-friend recommendations drive polarization. In each iteration of our edge dynamics, we remove disagreeable edges, and replace them with friend-of-friend connections. Initially, these friend-of-friend connections are somewhat random (i.e., uncorrelated with the expressed node opinions), but still more agreeable on average than the removed connections. Eventually, as opinion groups separate and the removed edges become less disagreeable, so do the friend-of-friend connections. Thus, polarization continues rising.

We also give an understanding of our model based on the Stochastic Block Model (SBM). Using a simple method to split nodes into two opinion groups (i.e., one group with innate opinions $< 0$ and one group with innate opinions $> 0$), we coarsely approximate the underlying network as an SBM graph with two blocks, drawn from a distribution with in-group connection probability $p$ and out-group connection probability $q$. We set $p$ and $q$ to match the in-group and out-group connection densities in the true graph.

Following prior work [9], we employ closed form expressions for the polarization and disagreement in the expected SBM graph. We demonstrate empirically that these expressions yield good approximations to the true polarization and disagreement over time – see Fig. 2. This indicates that the evolution of polarization and disagreement in our model can be understood largely in terms of in-group and out-group connection densities. With confirmation bias and friend-of-friend recommendations, the out-group connection density is eroded over time. If either dynamic is replaced with
the random control, we observe that this connection density does not decrease to the level necessary for high polarization.

Our SBM-based approximations also predict and explain interesting phenomena in our model. For instance, in networks with no fixed edges, we see that the graph eventually splits into many connected components with very similar innate opinions, causing polarization to reach a maximum and disagreement to drop to near zero. However, in the presence of a small fraction offi xed edges, while the network becomes more polarized, it remains connected. Surprisingly, in this setting, disagreement, like polarization, tends to increase over time, before asymptoting – see Fig. 2.

Figure 2: Polarization and disagreement for an Erdös-Renyi graph with 1000 nodes, varying percentages offi xed edges, and average degree 25. Innate opinions are uniformly distributed in [−1, 1]. Estimates of polarization and disagreement based on our SBM approximation closely match observed values and reflect important patterns – e.g., even a small percentage offi xed edges signifi cantly limits polarization and prevents disagreement from dropping over time.

3. Our model creates “natural-looking” networks. Finally, we give evidence that our model gives rise to graph structures that resemble real-world networks. Even when the initial network is an Erdös-Renyi graph, measures such as the degree distribution and triangle density shift to resemble real social networks due to our edge dynamics. E.g., while the degree distribution of the Erdös-Renyi graph is initially binomial (approximately normal), our edge dynamics drive the network to a steady state degree distribution that appears closer to a power-law, as expected in real social networks [27]. See Fig. 3. Such fi ndings help to validate our model’s realism, as a mechanism for opinion and network coevolution.

1.3 Related Work

Polarization and its connection to recommender systems and opinion dynamics has seen signifi cant research interest, particularly in the wake of Pariser’s 2012 “fi lter bubble” hypothesis [31].

A number of papers consider the effect of link recommendations, including friend-of-friend recommendations, on network structure [10, 37]. Others consider the effect of edge rewiring and innate opinion perturbation on polarization, including within the Friedkin-Johnsen model [1, 8, 9, 17, 28, 33]. Edge rewiring can model a recommender system, an adversary that seeks to maximize polarization, or a benevolent administrator that seeks to minimize polarization. Generally, these works focus on how a single intervention can effect polarization, oft en nding that relatively minor changes can have signifi cant impact. Unlike our work, they do not consider how opinions and the underlying network coevolve and drive polarization over time.

Several works do consider opinion and network coevolution [4, 6, 11, 22, 34]. A common fi nding, matching our results, is that confi rmation bias (also called homophily), itself is not suffi cient to drive signifi cant polarization [11, 34]. Dandekar et al. [11] introduce another psychological factor, biased assimilation, in which nodes that are presented a mixture of opinions give undue support to their initial opinion. They show that this dynamic drives polarization in combination with confi rmation bias. Sasahara et al. [34] present a model in which a user’s expressed opinion only takes into account suffi ciently agreeable neighbors. They show that this behavior drives the network into a bimodal opinion distribution. They also show that direct recommendations of agreeable edges accelerate polarization.

Like our work, several related works study the validity of their synthetic opinion network models compared to real-world data [15, 34, 37]. Although specifi c validation methods vary, many works examine graph structures such as clustering, triangles, and degree distribution. Sasahara et al. [34] use such measures to show that their model can produce a snapshot which has similar features to real-world social network data. Others show that these graph structures can refl ect changes to the recommendation dynamics in a social network [37], or correspond with parameters that change the behavior of their synthetic model [15].

The above works are complementary to ours. They use custom opinion dynamics models, while we build on the standard Friedkin-Johnsen model. This ties our work to well-established studies of opinion dynamics and allows us to leverage theoretical tools from prior work. Additionally, prior work does not consider the eff ect offi xed edges or friend-of-friend link recommendations, focusing instead on other important dynamics that can drive polarization.

Contrasting our work, some works suggest that recommendation systems can mitigate fi lter bubble effects [2, 29]. These works focus on diversity of content consumption, showing that it can be increased by recommendations based on natural collaborative
filtering. Other works propose remedies for ‘fixing’ the polarizing effect of recommender systems, or augmenting networks to reduce polarization [7, 9, 19, 33]. Some of these [9, 33] also work with the Friedkin-Johnsen model. It would be interesting to understand the effects of these remedies in our opinion and edge dynamics model.

2 OPINION AND EDGE DYNAMICS MODEL

We start by defining preliminaries and detailing our model of opinion formation under local edge dynamics.

2.1 Opinion Dynamics

We work with the Friedkin-Johnsen opinion model [16]. There are $n$ individuals, connected by an undirected graph $G$ with Laplacian matrix $L \in \mathbb{R}^{n \times n}$. There is an innate opinion vector $s \in \mathbb{R}^n$, whose entries represent each individual’s opinion without influence from neighbors. Node opinions are numerically coded along the interval $[-1, 1]$, and we assume they are drawn from a distribution with mean 0. An expressed opinion vector $z \in \mathbb{R}^n$ models the individuals’ opinions under influence from their neighbors. $z$ is obtained by repeatedly applying the opinion averaging update:

$$z(i) = \frac{s(i) + \sum_{j \neq i} w_{ij} z(j)}{1 + \sum_j w_{ij}},$$

where $w_{ij}$ is the weight of the edge between node $i$ and node $j$. For simplicity, we only consider unweighted graphs, where $w_{ij} = 0$ if there is no edge and $w_{ij} = 1$ if there is an edge. This update converges to an equilibrium, with $z = (I + L)^{-1}s$. Note that since the innate opinion vector $s$ is mean-centered, $z$ will also be mean-centered, as shown in Proposition 2 from [28].

2.2 Edge Dynamics

In our model, the network and the expressed opinions coevolve over time. The innate opinions are fixed at initialization. Let $L_t \in \mathbb{R}^{n \times n}$ denote the network Laplacian at time $t$, and $z_t = (I + L_t)^{-1}s$ denote the expressed opinions. At each time step, we compute a set of edges to be removed and a set of edges to be added to $G$. These sets typically depend on the expressed opinions $z_t$. After the removals and additions, $L_t$ is updated to $L_{t+1}$, and the process continues.

At initialization, we select a small percentage (typically 5%) of random edges to be fixed, and set the number of time steps necessary for the edge dynamics to converge.

We then select the set to be removed according to a probability distribution. In the control, this distribution is uniform over non-fixed edges. To model confirmation bias, the distribution is biased on expressed opinion disagreement. A non-fixed edge $(i,j)$ is removed with probability proportional to $|z_t(i) - z_t(j)|$. In this way, nodes tend to remove connections to other nodes that express conflicting opinions, while keeping edges that confirm their own opinion.

We sample $k = [p \cdot e]$ edges to be removed, without replacement, where $e$ is the number of edges in the graph, excluding any fixed edges. We then iterate over the sampled edges, removing each from the graph in turn.

Edge Additions. Given the number of edges $r$ that were removed at the current step, we select $r$ edges to be inserted. For the control, we simply sample $r$ edges that are not currently in the graph uniformly at random. To model friend-of-friend recommendations, we select $r$ edges iteratively. We select a random node and compute its friend-of-friends set – i.e., its two-hop neighbors. We then select one of these friend-of-friends uniformly at random and add an edge between it and the source node to an ‘addition set’. The process continues until there are $r$ edges in the addition set, at which point all edges in the set are added to the graph.

2.3 Polarization and Disagreement

The primary quantities that we measure as the network $L_t$ and expressed opinions $z_t$ coevolve over time are polarization and disagreement. The polarization is the variance of the expressed opinions, a common choice in the literature [1, 5, 28]. The disagreement measures the variance of expressed opinions just across edges currently in the graph. It is high if nodes tend to be connected to other nodes with very different opinions, and low if nodes tend to be connected to other nodes with similar opinions.

As shown in [28], in the Friedkin-Johnsen model, polarization and disagreement can be written as quadratic forms involving the network Laplacian and the innate opinion vector. For simplicity, we assume throughout that our innate opinions have mean 0.

**FACT 1 (POLARIZATION).** Consider a graph with Laplacian matrix $L \in \mathbb{R}^{n \times n}$, along with a mean 0 innate opinion vector $s \in \mathbb{R}^n$. Let $z = (I + L)^{-1}s$ be the equilibrium expressed opinion vector in the Friedkin-Johnsen model. The polarization is given by

$$P(L,s) = \|s\|_2^2 = s^T(I + L)^{-2}s.$$  

**FACT 2 (DISAGREEMENT).** Consider the setting of Fact 1. The disagreement is given by

$$D(L,s) \overset{\text{def}}{=} \sum_{ij} w_{ij} \cdot (z(i) - z(j))^2 = z^T L z = s^T(I + L)^{-1} L(I + L)^{-1} s.$$  

Due to its simple formulation as a quadratic form over $(I + L)^{-1}$, in some of our theoretical bounds, we work with the polarization + disagreement, introduced in [28] and defined below.

**FACT 3 (POLARIZATION + DISAGREEMENT).** Consider the setting of Facts 1 and 2. The polarization + disagreement is given by

$$PD(L,s) \overset{\text{def}}{=} P(L,s) + D(L,s) = s^T(1 + L)^{-1}.$$  

From Fact 1, we can derive a simple upper bound on polarization, which is useful in interpreting our simulation results.

**PROPOSITION 4.** Consider any Laplacian matrix $L \in \mathbb{R}^{n \times n}$, along with a mean 0 innate opinion vector $s \in \mathbb{R}^n$. We have $P(L,s) \leq \|s\|_2^2$; i.e., polarization is bounded by the innate opinion variance.

**PROOF.** Since $L$ is positive semidefinite, all eigenvalues of $I + L$ are at least 1. Thus, all eigenvalues of $(I + L)^{-1}$ are at most 1. Thus, using Fact 1, $P(L,s) = s^T(I + L)^{-2}s \leq \|s\|_2^2$. □
3 THEORETICAL RESULTS

We now present our theoretical results, which help explain how polarization and disagreement evolve under local edge dynamics.

3.1 Single Edge Updates

Using the expressions for polarization and disagreement in Sec. 2.3, we can understand how these quantities evolve as edges are added and removed from the graph. We consider a simplified setting in which just a single edge \((u, v)\) is added or removed at a time. This can be represented by the update \(L_{t+1} = L_t \pm E_t\) where \(E_t \in \mathbb{R}^{n \times n}\) is a rank-1 edge Laplacian for \((u, v)\). That is, \(E_t(u, v) = E_t(v, u) = 1\) and \(E_t(u, w) = E_t(v, w) = -1\) for all \(w \neq u, v\). We can also write \(E_t = \chi_{u,v} \chi_{v,u}^T\) where \(\chi_{u,v} \in \mathbb{R}^n\) is the edge indicator vector with a 1 at position \(u\) and a -1 at position \(v\). We only allow adding an edge not currently in the graph and removing one in the graph. This ensures that \(L_t\) remains a valid unweighted graph Laplacian.

We compute the update in polarization + disagreement under edge additions/deletions via the Sherman-Morrison formula [35].

**Lemma 5 (P+D – EdgeAddition/Delete).** Consider any unweighted graph Laplacian \(L \in \mathbb{R}^{n \times n}\) and let \(s, z \in \mathbb{R}^n\) be the innate and expressed opinion vectors in the Friedkin-Johnson model. Let \(E \in \mathbb{R}^{n \times n}\) be the edge Laplacian for edge \((u, v)\). Then \(L = \chi_{u,v} \chi_{v,u}^T\) is the effective resistance between \((u, v)\) in the graph given by \(L\) plus a small copy of the complete graph \([36, 37]\). It will be small if \((u, v)\) are algebraically well-connected. For any \(L\) and \((u, v)\) we have \(r_{u,v} \geq 0\) and \(r_{u,v} \leq \|\chi_{u,v}\|^2 \leq 2\). If \((u, v)\) is already in the graph (as in a deletion), \(r_{u,v} \in (0, 1]\). This gives:

**Corollary 6 (P+D – EdgeAddition/Delete).** Consider the setting of Lemma 5. We have:

\[
PD(L) - \delta^2 \leq PD(L + E) \leq PD(L) - \delta^2 / 3.
\]

\[
PD(L - E) \geq PD(L) + \delta^2.
\]

From Cor. 6, we see that adding an edge decreases the polarization + disagreement. Subtracting an edge increases it. In both cases, the magnitude of change is roughly linear in the squared disagreement across the edge. We highlight that, since the disagreement is in terms of the expressed opinions, which may differ substantially from the innate opinions, this finding is non-obvious. It is surprising that the change in polarization + disagreement only depends on the innate opinions \(s\) through the expressed opinions \(z\).

3.2 Edge Swaps

Building on the above, we next consider how \(PD(L)\) changes when a disagreeable edge is swapped out for a more agreeable one. Our main result is Cor. 8: if there is a sufficient gap in disagreement across pair \((i,j)\) and pair \((k,t)\), then removing edge \((i,j)\) and adding \((k,t)\) will strictly increase the polarization + disagreement. Thus, as long as there remain edges in the graph with higher disagreement than non-edges, swapping out disagreeable connections for agreeable ones will drive up \(PD(L)\). This finding helps explain why a combination of confirmation bias and friend-of-friend recommendations leads to a significant increase in polarization.

**Lemma 7 (P+D – EdgeSwap).** Consider any unweighted graph Laplacian \(L \in \mathbb{R}^{n \times n}\), and let \(s, z \in \mathbb{R}^n\) be the innate and expressed opinion vectors in the Friedkin-Johnson model. Let \(E_1\) be the edge Laplacian for edge \((u_1, v_1)\) and \(\delta_1 = z(u_1) - z(v_1)\). Let \(E_2\) be the edge Laplacian for edge \((u_2, v_2)\) and \(\delta_2 = z(u_2) - z(v_2)\). Assume that the edge \((u_2, v_2)\) is in the graph corresponding to \(L\). Let \(r_1 = \chi_{u_1,v_1}^T(I + L)^{-1} \chi_{u_1,v_1}\) and \(r_2 = \chi_{u_2,v_2}^T(I + L)^{-1} \chi_{u_2,v_2}\).

\[
PD(L + E_2 - E_1) \geq PD(L) - \frac{\delta_1^2}{1 + r_1} + \frac{\delta_2 - \alpha \cdot \delta_1^2}{1 - r_2},
\]

where \(\alpha = \frac{\chi_{u_1,v_1}^T(I + L)^{-1} \chi_{u_1,v_1}}{r_1}\) and \(\|\alpha\| \leq \sqrt{r_1^{-1}r_2^{-1}}\).

We defer the proof of Lem. 7 to the appendix. It follows by applying the two formulas of Lem. 5 in sequence, and simplifying via a third application of the Sherman-Morrison formula.

In a well-connected graph, \(r_1, r_2 \ll 1\). Thus, the increase in \(PD(L)\) will be roughly \(\frac{\delta_2^2}{1} - \frac{\delta_1^2}{1} \geq \delta_2^2 - \delta_1^2\). I.e., adding a more agreeable edge and removing a more disagreeable one will increase PD. Formally, in the appendix we prove:

**Corollary 8 (PD Increase with Swap).** Consider the setting of Lemma 7. If \(|\delta_2| > \frac{3\sqrt{3}}{\sqrt{2}}, |\delta_1|\) then \(PD(L + E_1 - E_2) > PD(L)\).

3.3 Stochastic Block Model Analysis

The results of Secs. 3.1 and 3.2 help explain how polarization and disagreement evolve incrementally as edges are added and removed. As discussed, we find that the evolution of these quantities at larger time scales can be well-approximated by looking simply at connectivity across opinion groups. In particular, we approximate \(L_t\) with an expected Stochastic Block Model (SBM) graph with the same in-group and out-group connectivity. We approximate the innate opinion vector \(s\) with a discretization of that vector that is constant on each opinion group. Following work of [9, 25], we can compute closed form expressions for the polarization and disagreement for this SBM graph and innate opinion vector. We find that these approximations match empirical observations – see Fig. 2.

We start by formally defining our SBM approximation. For simplicity, we assume that the number of nodes \(n\) is even – odd \(n\) is easily handled via rounding. For two vertex sets \(A, B\), let \(E(AB)\) denote the number of edges between these sets in the graph.

**Definition 9 (SBM Approximation).** Consider a graph Laplacian \(L \in \mathbb{R}^{n \times n}\) and innate opinion vector \(s \in [-1, 1]^n\). Let \(V_+\) and \(V_-\) be the vertex sets corresponding to nodes with positive and negative opinions, respectively. Let \(q = \frac{E(V_+V_+) + E(V_-V_-)}{2|E(V_+V_-)|}\) be the fraction of out-group and in-group edges, respectively. Let \(L\), be the expected SBM graph with out-group and in-group connection probabilities \(q, p\). In particular, \(L = \frac{|p + q|}{2} I - \Delta\), where \(\Delta(i,j) = \Delta(j,i) = p\) if \(i,j \in [1, n/2]\) or \(i,j \in [n/2 + 1, n]\), \(\Delta(i,j) = q\) if \(i \in [1, n/2]\) and \(j \in [n/2, n]\). Let \(s\) have \(s(i) = 1\) for \(i \in [1, n/2]\) and \(s(i) = -1\) for \(i \in [n/2 + 1, n]\)

Observe that opinion groups in the expected SBM graph \(L\) correspond to the fiftieth and last \(n/2\) nodes. If \(s\) is chosen randomly
When either dynamic is removed, we next discuss our experimental results simulating our opinion dynamics bias and friend-of-friend recommendations. We simulate our model on several random networks, respectively. See Fig. 17 (in Appendix) for an illustration with a bimodal innate opinion distribution. When removing either dynamic, we find that replacing either with a random connection drive a large increase in polarization. 

#### 4 EXPERIMENTAL RESULTS

We next discuss our experimental results simulating our opinion and edge dynamics model on synthetic and real world data.

##### 4.1 Experimental Set Up

**Synthetic Networks.** We simulate our model on several random graphs, drawn from the Erdős-Rényi (ER) and Barabási-Albert (BA) distributions. In an ER graph, each pair of nodes is connected independently with probability \( p \). We modulate this probability in a symmetric distribution, \( |V_+| \approx |V_-| \approx n/2 \). Thus, it is reasonable to approximate both groups as having size exactly \( n/2 \). Also observe that \( \bar{A} \) is simply a 2x2 block matrix, with its top-left and bottom-right \( n/2 \times n/2 \) blocks ill-defined with \( p \)'s and its top-right and bottom-left blocks ill-defined with \( q \)'s. We can check that \( \bar{s} \) is an eigenvector of \( \bar{A} \) with eigenvalue \( \frac{(p-q)n}{2} \). Thus, \( \bar{s} \) is an eigenvector of \( \bar{L} = \frac{(p-q)n}{2} I - \bar{A} \) with eigenvalue \( qn \). Using this, we derive:

**Fact 10 (SBM Polarization and Disagreement).** Let \( \bar{L} \in \mathbb{R}^{n \times n} \) and \( \bar{s} \in \mathbb{R}^n \) be as defined in Def. 9. We have:

\[
P(\bar{L}, \bar{s}) = \frac{n}{(qn + 1)^2} \quad \text{and} \quad D(\bar{L}, \bar{s}) = \frac{qn^2}{(qn + 1)^2}.
\]

**Proof.** Using Facts 1 and 2, \( P(\bar{L}, \bar{s}) = \bar{s}^T ((\bar{I} + \bar{L})^{-2} \bar{s} \) and \( D(\bar{L}, \bar{s}) = \bar{s}^T (\bar{I} + \bar{L}^{-1}) \bar{L} (\bar{I} + \bar{L}^{-1})^{-1} \bar{s} \). Since \( \bar{s} \) is an eigenvector of \( \bar{L} \) with eigenvalue \( qn \), it is an eigenvector of both \((\bar{I} + \bar{L})^{-2} \) and \((\bar{I} + \bar{L}^{-1})^{-1} \bar{L} (\bar{I} + \bar{L}^{-1})^{-1} \) with eigenvalues \( \frac{1}{(qn + 1)^2} \) and \( \frac{qn^2}{(qn + 1)^2} \), respectively. The fact then follows from observing that \( ||\bar{s}||_2^2 = n \).

Via Fact 10, we can approximate to polarization and disagreement of a graph \( \bar{L} \) and innate opinion vector \( \bar{s} \) with a simple formula that depends just on the out-group connection density \( q \). Surprisingly, the formula has no dependence on the in-group density \( p \). Since we always have \( ||\bar{s}||_2^2 = n \), we scale the quantities in Fact 10 by a factor of \( \frac{1}{qn} \) to adjust for the innate opinion variance. The resulting approximations are quite accurate in predicting polarization and disagreement evolution in our model – see Fig. 2.

**Interpretation.** The accuracy of our SBM-based approximation indicates that the evolution of polarization and disagreement in our edge dynamics model is largely governed by the evolution of out-group connection probabilities. We observe that a combination of confirmation bias and friend-of-friend recommendations tends to drive down \( q \) over time and hence drive up polarization.

When either dynamic is removed, \( q \) remains relatively high, and polarization does not rise significantly. See Fig. 4.

Interestingly, our approximation \( D(\bar{L}, \bar{s}) = \frac{(q^2n^2)}{(qn + 1)^2} \) is not a monotonic function of \( q \). \( D(\bar{L}, \bar{s}) \) is decreasing in \( q \) when \( q \geq 1/n \), but increasing when \( q < 1/n \). This explains an interesting phenomena seen in our model with no fof edges: the disagreement initially ‘spikes’ as out-group connections are removed, before falling to near 0, as the opinion groups become fully disconnected (i.e., as \( q \) becomes very small). When we fix a y fof off er edges, \( q \) is effectively lower bounded by \( y \). As long as \( y \geq 1/n \), we never see disagreement fall – it rises jointly with polarization. See Fig. 2. Relatively, when \( q \) is not lower bounded by \( y \), \( P(\bar{L}, \bar{s}) \) is able to reach its maximum value of \( n \), giving a predicted polarization of \( \frac{nn}{2} \), which matches the maximum polarization derived in Prop. 4.

---

**Figure 4: Out-group connectivity \( q \) over time for ER graphs with 1000 nodes and connection probability \( p = 0.05 \), with varying percentages fof ed edges. Innate opinions are uniformly distributed in \([-1, 1]\). With edge dynamics influenced by both confirmation bias (CB) and friend-of-friend recommendations (F-o-F), out-group connectivity drops significantly, mirroring a rise in polarization, as predicted by Fact 10 and observed in Figs. 1 and 2.**

---

The accuracy of our SBM-based approximation indicates that the evolution of polarization and disagreement in our edge dynamics model is largely governed by the evolution of out-group connection probabilities. We observe that a combination of confirmation bias and friend-of-friend recommendations tends to drive down \( q \) over time and hence drive up polarization.

When either dynamic is removed, \( q \) remains relatively high, and polarization does not rise significantly. See Fig. 4.

Interestingly, our approximation \( D(\bar{L}, \bar{s}) = \frac{(q^2n^2)}{(qn + 1)^2} \) is not a monotonic function of \( q \). \( D(\bar{L}, \bar{s}) \) is decreasing in \( q \) when \( q \geq 1/n \), but increasing when \( q < 1/n \). This explains an interesting phenomena seen in our model with no fof edges: the disagreement initially ‘spikes’ as out-group connections are removed, before falling to near 0, as the opinion groups become fully disconnected (i.e., as \( q \) becomes very small). When we fix a y fof off er edges, \( q \) is effectively lower bounded by \( y \). As long as \( y \geq 1/n \), we never see disagreement fall – it rises jointly with polarization. See Fig. 2. Relatively, when \( q \) is not lower bounded by \( y \), \( P(\bar{L}, \bar{s}) \) is able to reach its maximum value of \( n \), giving a predicted polarization of \( \frac{nn}{2} \), which matches the maximum polarization derived in Prop. 4.

---

### 4.2 Drivers of Polarization

As discussed, our edge dynamics with both confirmation bias and friend-of-friend recommendations drive a large increase in polarization. Without fof edges, this fof ed driving is robust to the initial graph size, structure, and innate opinion distribution. Figs. 11 and 14 (in Appendix) show similar behavior for large random graphs and real social network graphs, respectively. See Fig. 17 (in Appendix) for an illustration with a bimodal innate opinion distribution. Whenever fof edges are present, they limit and slow the increase in polarization – this effect is amplified for larger fractions fof ed edges.

Isolating the effects of friend-of-friend recommendations and confirmation bias, we find that replacing either with a random control drastically changes the behavior of polarization. See Fig. 1 for an illustration on an ER random graph with 1000 nodes, the Twitter real-world data set [12], and a dense ER random graph with 1000 nodes. Polarization is essentially constant with just friend-of-friend recommendations, and it rises very slowly with just confirmation bias. Note that polarization in the dense ER graph does not rise significantly, even with both friend-of-friend recommendations and confirmation bias. We discuss this fof ed next.
expressed opinions and innate opinions over time. Fig. 13 (in Appendix) illustrates this by plotting the mean structure which keeps expressed opinions very close to innate opinions – Fig. 12 (in Appendix) also shows histograms of node opinions which provide another visual dimension for this state evolution.

Figure 6: Opinion evolution on a BA graph with 1000 nodes, 5% fixed edges, and uniform random innate opinions. Each line plots the evolving expressed opinion of one node. Line color represents that node’s innate opinion, ranging from -1 (blue) to 1 (red). We see opinions initially grouped near 0 (Initial state). We then see a bifurcation of the cluster, indicated by the distinct “branches” (Bimodal polarization). Fixed edges in this network prevent it from splintering.

4.4 Validation Against Real-World Data

Following prior work [15, 34], in this section, we use real data as a “benchmark” to ascertain whether our model can recreate realistic social network structures. We show validation for the Twitter data set, with results on the Facebook data set in the Appendix.

Set up. We start with a synthetic graph (ER or BA) with the same number of nodes as the real-world network, and roughly the same number of edges. Wep x a certain percentage of edges (e.g., 15%, 25%, and 35%), which are not subject to our edge dynamics.

We then simulate our edge dynamics with friend-of-friend recommendations and confirmation bias, running for 1500 iterations, until the network reaches a steady state. We measure the evolution of various structural properties of the network over time, and compare them to those of the real-world network. Wep nd that the percentage offi xed edges in a graph correlates with the steady state global clustering coefficient, which is defined as \( \frac{3 \times \text{number of triangles in } G}{\text{number of open & closed triads in } G} \). Using this quantity, we tune the percentage offi xed edges in our synthetic graph. E.g., for the Twitter network, the global clustering coefficient is \( \approx 0.227 \). The percentage offi xed edges leading to the closest steady-state clustering coefficient for both ER and BA networks is 25%.

Degree Distribution. As in many social networks, Twitter has a power law degree distribution [27]. See Fig. 19 in the appendix. For an ER network with 25% fixed edges, we show the initial degree distribution, along with the steady state degree distribution after running edge dynamics in Fig. 3. The initial ER graph has a bell-shaped binomial degree distribution. Surprisingly, our edge dynamics change this distribution significantly, and the steady state distribution appears closer to a power law distribution.

In Fig. 8, we compare each node’s equilibrium degree against its expressed opinion in the steady state network. It seems that nodes with near-mean expressed opinions gain connections under our edge dynamics model, while peripheral nodes lose connections.
This behavior may drive the observed degree distribution shift. An improved theoretical understanding would be very interesting.

Fig. 20 (in Appendix) shows similar results for a BA network with 25% fixed edges. Note that the BA graph initially has a power-law degree distribution, so these results serve to validate that our model preserves a realistic degree distribution.

Visual Similarity. While not rigorous as a method of comparison, in Fig. 7 we visualize the Twitter network, an initial ER graph, and the steady-state ER graph after undergoing our edge dynamics. The color of each node is set according to the node’s Friedkin-Johnsen innate opinion, which is uniformly sampled from the interval $[-1, 1]$. While the ER network’s steady state connectivity between clusters seems to be more dense than the Twitter network, the overall structure appears similar, particularly around the periphery nodes and in-group clusters. In the Appendix, in Figs. 24 & 25, we show a similar visual evolution for a BA network.

5 CONCLUSION

We present a simple extension of the Friedkin-Johnsen opinion model, in which the opinions and the underlying network coevolve under the influence of confirmation bias and recommendations. Via simulation, we find that both confirmation bias and friend-of-friend recommendations are required to noticeably increase polarization. We show theoretical results that explain how polarization and disagreement are increased via swaps of more agreeable edges for more disagreeable ones. We also show how polarization and disagreement can be accurately approximated as functions of connectivity between opinion groups, based on a stochastic block model analysis. Finally, we validate that our opinion dynamics tend to create relatively realistic looking networks in terms of structure.

Our findings leave open several questions. Theoretically, better explaining the role of density in our model, and the mechanisms behind degree and triangle distribution shift would be very interesting. One also might consider whether our findings generalize to variations of the Friedkin-Johnsen model or other opinion models [13, 15, 21, 34], multidimensional opinions, or network constraints beyond fixed edges. It would be very interesting to further validate our model’s realism, by extending our analysis to other graph measures, data sets with ground truth opinion data, and ideally to temporal data sets showing the coevolution of network structure and opinions in the real world.
REFERENCES

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

[2] Guy Aridor, Duarte Gonçalves, and Shan Sikdar. 2020. Deconstructing the Filter Bubble: User Decision-Making and Recommender Systems. Fourteenth ACM Conference on Recommender Systems (2020). https://doi.org/10.1145/3383313.3412442

[3] Delia Baldassarri and Andrew Gelman. 2008. Partisans without Constraint: Political Polarization and Trends in American Public Opinion. Amer. J. Sociology 114, 2 (2008), 408–446. https://doi.org/10.1086/590649

[4] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. 2020. Modeling echo chambers and polarization dynamics in social networks. Physical Review Letters 124, 4 (2020), 048301.

[5] Heather Z. Brooks and Mason A. Porter. 2020. A model for the influence of media on the ideology of content in online social networks. Physical Review Research 2, 2 (2020), 023041.

[6] Claudia Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. Reviews of Modern Physics 81, 2 (2009), 591.

[7] L. Elisa Celis, Sayah Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling Polarization in Personalization. Proceedings of the Conference on Fairness, Accountability, and Transparency (2019). https://doi.org/10.1145/3285609.3287601

[8] Maye Chen and Miklos Racz. 2020. Network disruption: maximizing disagreement and polarization in social networks. https://arxiv.org/pdf/2005.08377.pdf

[9] Uthsav Chitra and Christopher Musco. 2020. Analyzing the Impact of Filter Bubbles on Social Network Polarization. Proceedings of the 14th International Conference on Web Search and Data Mining (2020). https://doi.org/10.1145/3361691.3371825

[10] Elizabeth M. Daly, Werner Geyer, and David R Millen. 2010. The network effects of recommending social connections. In Proceedings of the Fourth ACM Conference on Recommender Systems.

[11] Pranav Dandekar, Ashish Goel, and David Parkes. 2020. Deconstructing the Filter Bubble: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization. The 14th ACM International Conference on Web Search and Data Mining (2021). https://doi.org/10.1145/3437963.3441825

[12] Eric M. Patashnik and Wendy J. Schiller. 2021. The Larger Forces Behind the January 6, 2021 Insurrection | Watson Institute. https://watson.brown.edu/research/2021/larger-forces-behind-january-6-2021-insurrection

[13] 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 Proceedings of the 23rd International World Wide Web Conference (WWW).

[14] Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology 2 (1998), 175–220. https://doi.org/10.1037/1081-0730.2.2.175

[15] Eric P. K ripke. 2012. Th e filter bubble: what the Internet is hiding from you. Viking.

[16] D. M. Ashby. 2019. “I was Right about Vaccination”: Confirmation Bias and Health Literacy in Online Health Information Seeking. Journal of Communication 24, 2 (2019), 129–140. https://doi.org/10.1080/10810730.2019.1583701 PMID: 30895889.

[17] Jason Gaitonde, Jon Kleinberg, and Eva Tardos. 2020. Adversarial perturbations in online social networks. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/7af41d60c325499f16808986feedeb-Paper.pdf

[18] F. McIver. 2001. Spectral partitioning of random graphs. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science. 529–537. https://doi.org/10.1109/SFCS.2001.959929

[19] Corinne S. Meppelink, Edith G. Smit, Marielle L. Fransen, and Nicola Diviani. 2019. “I was Right about Vaccination”: Confirmation Bias and Health Literacy in Online Health Information Seeking. Journal of Communication 24, 2 (2019), 129–140. https://doi.org/10.1080/10810730.2019.1583701 PMID: 30895889.

[20] Magnús R. Ragnó. 2019. Opinion dynamics and bounded rationality. Reviews of Modern Physics 91, 4 (2019), 045003. https://doi.org/10.1103/revmodphys.91.045003

[21] Petter Holme and M. E. J. Newman. 2006. Nonequilibrium phase transition in the coevolution of networks and opinions. Physical Review E 74 (2006). https://doi.org/10.1103/physreve.74.056108

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

[23] R. Hegselmann, R. Winter, J. Los, C. D. Schmitt, and C. D. Schmitt. 2007. Chaotic group dynamics. In The 14th ACM International Conference on Conference on Recommender Systems.

[24] jack Sherman and Winifred J Morrison. 1950. Adjustment of an inverse matrix in the least-square sense. The Annals of Mathematical Statistics 21, 1 (1950), 124–127.

[25] Jack Sherman and Winifred J Morrison. 1950. Adjustment of an inverse matrix in the least-square sense. The Annals of Mathematical Statistics 21, 1 (1950), 124–127.

[26] Attila Török and Miklós Rácz. 2013. Towards Consensus: Reducing Polarization by Perturbing Social Networks. arXiv:2206.08996 (2022)

[27] Miklós Rácz, Christopher Musco, and Charalampos E. Toufarakakis. 2018. Minimizing polarization and disagreement in social networks. In Proceedings of the 27th International World Wide Web Conference (WWW).

[28] Eric M. Patashnik and Wendy J. Schiller. 2021. The Larger Forces Behind the January 6, 2021 Insurrection | Watson Institute. https://watson.brown.edu/research/2021/larger-forces-behind-january-6-2021-insurrection

[29] Jack Sherman and Winifred J Morrison. 1950. Adjustment of an inverse matrix in the least-square sense. The Annals of Mathematical Statistics 21, 1 (1950), 124–127.

[30] Attila Török and Miklós Rácz. 2013. Towards Consensus: Reducing Polarization by Perturbing Social Networks. arXiv:2206.08996 (2022)

[31] Eric M. Patashnik and Wendy J. Schiller. 2021. The Larger Forces Behind the January 6, 2021 Insurrection | Watson Institute. https://watson.brown.edu/research/2021/larger-forces-behind-january-6-2021-insurrection

[32] Miklós Rácz, Christopher Musco, and Charalampos E. Toufarakakis. 2018. Minimizing polarization and disagreement in social networks. In Proceedings of the 27th International World Wide Web Conference (WWW).

[33] Eric M. Patashnik and Wendy J. Schiller. 2021. The Larger Forces Behind the January 6, 2021 Insurrection | Watson Institute. https://watson.brown.edu/research/2021/larger-forces-behind-january-6-2021-insurrection

About Climate Mitigation Policies. Communication Research 39, 6 (2012), 701–723. https://doi.org/10.1177/00936502114116646

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

[35] R. Hegselmann, R. Winter, J. Los, C. D. Schmitt, and C. D. Schmitt. 2007. Chaotic group dynamics. In The 14th ACM International Conference on Conference on Recommender Systems.