

# Local Edge Dynamics and Opinion Polarization

Nikita Bhalla

University of Massachusetts Amherst

Amherst, MA, USA

[nbhalla@cs.umass.edu](mailto:nbhalla@cs.umass.edu)

Adam Lechowicz

University of Massachusetts Amherst

Amherst, MA, USA

[alechowicz@cs.umass.edu](mailto:alechowicz@cs.umass.edu)

Cameron Musco

University of Massachusetts Amherst

Amherst, MA, USA

[cmusco@cs.umass.edu](mailto: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 code and supplemental materials are available at <https://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

## ACM Reference Format:

Nikita Bhalla, Adam Lechowicz, and Cameron Musco. 2023. Local Edge Dynamics and Opinion Polarization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27–March 3, 2023, Singapore, Singapore*. ACM, New York, NY, USA, ?? pages. <https://doi.org/10.1145/3539597.3570442>

## 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 [HN12, LCKK14, Del20, BG08, GMJCK13]. 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 [Nic98]. 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 [MSFD19, TJ19, PS21].

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 [Par12, LCKK14, GMJCK13].

### 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, [FJ90], 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 [MMT18]: *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 [MMT18, XBZ21], 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 [SSG16, DGM10].

We also fix 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

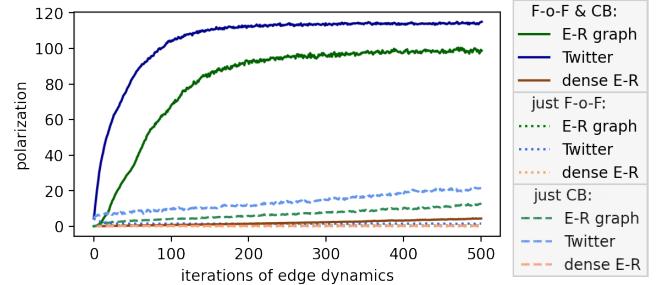
Our main findings are as follows:

**1. Confirmation bias and recommendations drive polarization.** We find 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 no fixed 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 just made randomly), 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. So, 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 find 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.

**2. Our model is tractable to theoretical analysis.** To complement our simulation results, we give theoretical bounds which



**Figure 1: Opinion polarization over time for an Erdős-Renyi (ER) graph with 1000 nodes and connection probability  $p = 0.05$ , a real-world Twitter social network [DBB<sup>+</sup>14], and a *dense* ER graph with 1000 nodes and connection probability  $p = 0.1$ . 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.**

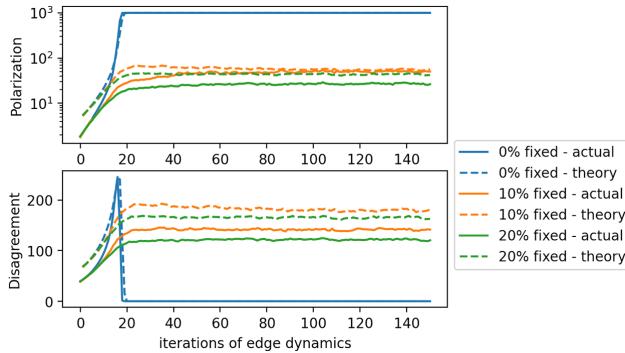
illustrate how edge removals and insertions can drive polarization and disagreement in the Friedkin-Johnsen model. We prove that swapping an edge with large expressed opinion disagreement across it for a new edge with small disagreement monotonically drives up the *sum of polarization and disagreement* in the graph, a combined quantity studied in prior work [MMT18], 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 [CM20], 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 splinters 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 of fixed 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.



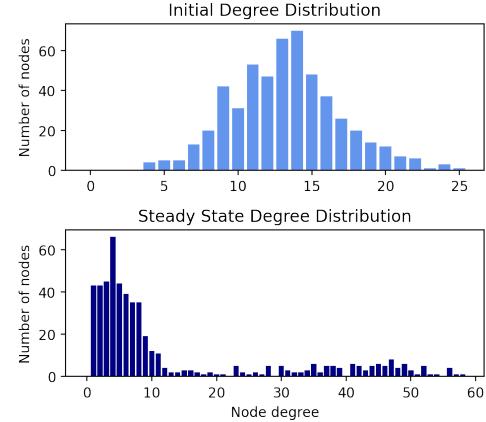
**Figure 2: Polarization and disagreement for an Erdős-Renyi graph with 1000 nodes, varying percentages of fixed edges, and average degree 25. Innate opinions are uniformly distributed in  $[-1, 1]$ . Estimates of polarization and disagreement based on our SBM approximation closely match observed values and reflect important patterns – e.g., that even a small percentage of fixed edges significantly 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 Erdős-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 Erdős-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 [MPP<sup>+</sup>13]. See Fig. 3. Such findings 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 significant research interest, particularly in the wake of Pariser’s 2012 ‘filter bubble’ hypothesis [Par12].

A number of papers consider the effect of link recommendations, including friend-of-friend recommendations, on network structure [DGM10, SSG16]. Others consider the effect of edge rewiring and innate opinion perturbation on polarization, including within the Friedkin-Johnsen model [MMT18, AKPT18, CR20, CM20, GKT20, RR22]. 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, often finding that relatively minor changes can have significant



**Figure 3: Initial and steady state degree histograms for an Erdős-Renyi graph with 531 nodes & average degree  $\approx 14$ , subject to our edge dynamics. The steady state distribution differs significantly from the initial distribution. It is closer to a power law distribution, reflecting a more realistic network structure.**

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 [HN06, CFL09, DGL13, BLSSS20, SCP<sup>+</sup>20]. A common finding, matching our results, is that confirmation bias (also called homophily), itself is not sufficient to drive significant polarization [DGL13, SCP<sup>+</sup>20]. Dandekar et al. [DGL13] 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 confirmation bias. Sasahara et al. [SCP<sup>+</sup>20] present a model in which a user’s expressed opinion only takes into account sufficiently 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 [SCP<sup>+</sup>20, EF18, SSG16]. Although specific validation methods vary, many works examine graph structures such as clustering, triangles, and degree distribution. Sasahara et al. [SCP<sup>+</sup>20] 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 reflect changes to the recommendation dynamics in a social network [SSG16], or correspond with parameters that change the behavior of their synthetic model [EF18].

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 effect of fixed 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 actually mitigate filter bubble effects [NHH<sup>+</sup>14, AGS20]. 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 [CKSV19, HMRU21, RR22, CM20]. Some of these works [RR22, CM20] 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 [FJ90]. 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:  $z(i) := \frac{s(i) + \sum_j w_{ij} z(j)}{1 + \sum_j w_{ij}}$ , 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 [MMT18].

### 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 so not subject to deletion. They can be thought of as modeling connections that are independent of opinions or recommendations, e.g., to family or co-workers.

**Edge Removals.** We set a percentage  $p$  of edges in the graph to be removed in each step. Typically,  $p = 10\%$ . In the Appendix, in Fig. 15, we show that the choice of  $p$  has little effect on the model’s behavior – a smaller or larger value of  $p$  simply scales 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 based 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  $k = \lfloor p \cdot e \rfloor$  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 [MMT18, AKPT18, BP20]. 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 [MMT18], 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

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

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

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

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 [MMT18] and defined below.

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

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

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  $(I + L)^{-2}$  are at most 1. Thus, using Fact 1,  $P(L, s) = s^T(I + L)^{-2}s \leq \|s\|_2^2$ .  $\square$

In our setting, the innate opinions  $s$  are fixed at initialization. Thus, we typically write the polarization, disagreement and polarization + disagreement at time  $t$  simply as  $P(L_t)$ ,  $D(L_t)$ ,  $PD(L_t)$ .

### 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  $(u_t, v_t)$  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  $(u_t, v_t)$ . That is,  $E_t(u_t, u_t) = E_t(v_t, v_t) = 1$  and  $E_t(u_t, v_t) = E_t(v_t, u_t) = -1$ . We can also write  $E_t = \chi_{u_t, v_t} \chi_{u_t, v_t}^T$  where  $\chi_{u_t, v_t} \in \mathbb{R}^n$  is the edge indicator vector with a 1 at position  $u_t$  and a -1 at position  $v_t$ . 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 [SM50].

**LEMMA 5 (P+D – EDGE ADDITION/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  $(u, v)$ . Let  $\delta = z(u) - z(v)$  and  $r_{u,v} = \chi_{u,v}^T(I + L)^{-1}\chi_{u,v}$ .

$$PD(L + E) = PD(L) - \delta^2 / (1 + r_{u,v}).$$

$$PD(L - E) = PD(L) + \delta^2 / (1 - r_{u,v}).$$

We defer the proof of Lem. 5 to the appendix. Note that  $r_{u,v} = \chi_{u,v}^T(I + L)^{-1}\chi_{u,v}$  is the *effective resistance* between  $(u, v)$  in the graph given by  $L$  plus a small copy of the complete graph [Spi19]. 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  $r_{u,v} \leq \|\chi_{u,v}\|_2^2 \leq 2$ . If  $(u, v)$  is already in the graph (as in a deletion),  $r_{u,v} \in (0, 1]$ . This gives:

**COROLLARY 6 (P+D – EDGE ADDITION/DELETION BOUNDS).** 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  $(k, \ell)$ , then removing edge  $(i, j)$  and adding  $(k, \ell)$  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 – EDGE SWAP).** 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  $r_1 = \chi_1^T(I + L)^{-1}\chi_1$ ,  $r_2 = \chi_2^T(I + L)^{-1}\chi_2$ , and  $r_{2,1} = \chi_2^T(L + I + E_1)^{-1}\chi_2$ .

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

$$\text{where } \alpha = \frac{\chi_1^T(L+I)^{-1}\chi_2}{1+r_1} \text{ and } |\alpha| \leq \frac{\sqrt{r_1 \cdot r_2}}{1+r_1}.$$

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  $\frac{\delta_2^2}{1-r_{2,1}} - \frac{\delta_1^2}{1+r_1} \geq \delta_2^2 - \delta_1^2$ . 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  $|\delta_2| > \frac{3\sqrt{3}}{4} \cdot |\delta_1|$  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 [CM20], we can compute closed form expressions for the polarization and disagreement for this SBM graph and innate opinion vector. We find that these approximations closely 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  $A, B$ , we let  $E(A, B)$  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  $q = \frac{E(V_+, V_-)}{|V_+| \cdot |V_-| / 2}$  and  $p = \frac{E(V_+, V_+) + E(V_-, V_-)}{(|V_+|^2 + |V_-|^2) / 2}$  be the fraction of out-group and in-group edges, respectively.

Let  $\bar{L}$ , be expected SBM graph with out-group and in-group connection probabilities  $q, p$ . In particular,  $\bar{L} = \frac{(p+q)n}{2}I - \bar{A}$  where  $\bar{A}(i, j) = \bar{A}(j, i) = p$  if  $i, j \in [1, n/2]$  or  $i, j \in [n/2+1, n]$ .  $\bar{A}(i, j) = \bar{A}(j, i) = q$  if  $i \in [1, n/2]$  and  $j \in [n/2, n]$ . Let  $\bar{s}$  have  $\bar{s}(i) = 1$  for  $i \in [1, n/2]$  and  $\bar{s}(i) = -1$  for  $i \in [n/2+1, n]$ .

Observe that opinion groups in the expected SBM graph  $\bar{L}$  correspond to the first and last  $n/2$  nodes. If  $s$  is chosen randomly

according to 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  $2 \times 2$  block matrix, with its top-left and bottom-right  $n/2 \times n/2$  blocks filled with  $p$ 's and its top-right and bottom-left blocks filled 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,  $P(\bar{L}, \bar{s}) = \bar{s}^T(I + \bar{L})^{-2}\bar{s}$  and  $D(\bar{L}, \bar{s}) = \bar{s}^T(I + \bar{L})^{-1}\bar{L}(I + \bar{L})^{-1}\bar{s}$ . Since  $\bar{s}$  is an eigenvector of  $\bar{L}$  with eigenvalue  $qn$ , it is an eigenvector of both  $(I + \bar{L})^{-2}$  and  $(I + \bar{L})^{-1}\bar{L}(I + \bar{L})^{-1}$  with eigenvalues  $\frac{1}{(qn+1)^2}$  and  $\frac{qn}{(qn+1)^2}$ , respectively. The fact then follows from observing that  $\|\bar{s}\|_2^2 = n$ .  $\square$

Via Fact 10, we can approximate polarization and disagreement of a graph  $L$  and innate opinion vector  $s$  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  $\frac{\|s\|_2^2}{n}$  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

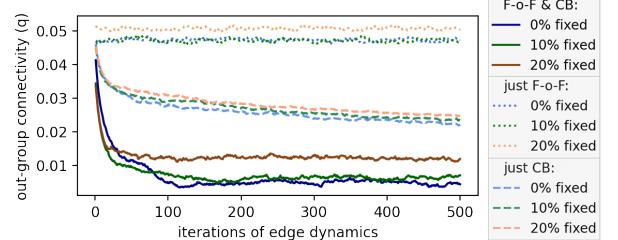
Interestingly, our disagreement approximation  $D(\bar{L}, \bar{s}) = \frac{qn^2}{(qn+1)^2}$  is not a monotonic function of  $q$ .  $D(\bar{L}, \bar{s})$  is decreasing in  $q$  when  $q \geq 1/n$ , but increasing when  $q \leq 1/n$ . This explains an interesting phenomena seen in our model with no fixed 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  $\gamma$  fraction of fixed edges,  $q$  is effectively lower bounded by  $\gamma$ . As long as  $\gamma \geq 1/n$ , we never see disagreement fall – it rises jointly with polarization. See Fig. 2. Relatedly, when  $q$  is not lower bounded by  $\gamma$ ,  $P(\bar{L}, \bar{s})$  is able to reach its maximum value of  $n$ , giving a predicted polarization of  $\|s\|_2^2$ , which matches the maximum polarization derived in Prop. 4.

## 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-Renyi (ER) and Barabási-Albert (BA)



**Figure 4: Out-group connectivity  $q$  over time for Erdős-Renyi graphs with 1000 nodes and connection probability  $p = 0.05$ , with varying percentages of fixed 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.**

distributions. ER graphs have each pair of nodes is connected independently with probability  $p$ . We modulate this probability in our experiments to study the effects of network density. BA graphs are scale-free networks generated via a preferential attachment process. For these graphs, we modulate the primary parameter  $m$ , which determines the number of edges to attach from a new node to existing nodes as the graph is generated.

**Real-world Networks.** We also simulate our model on several real-world networks. We preprocess all networks by computing their 2-core, which is a maximal subgraph such that each node has degree at least two. This eliminates the large number of single-edge nodes present in some of these datasets, which would distort the results. Basic graph parameters (after preprocessing) are listed below, with more details on the datasets given in Appendix B.

- **Reddit** [DBB<sup>+</sup>14]  $n = 546$  nodes,  $e = 8962$  edges
- **Twitter** [DBB<sup>+</sup>14]  $n = 531$  nodes,  $e = 3621$  edges
- **Facebook** [LM12]  $n = 3964$  nodes,  $e = 88159$  edges

**Innate Opinions.** In our primary experiments, innate opinions are sampled uniformly at random from the interval  $[-1, 1]$ . In the appendix we also give results for a bimodal truncated Gaussian innate opinion distribution, observing similar behavior.

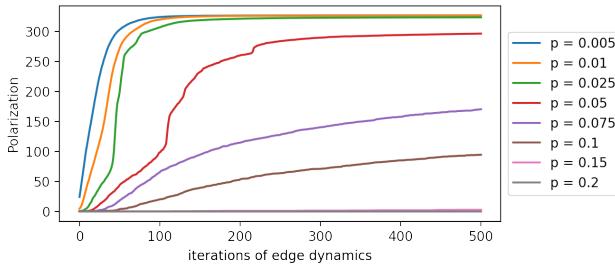
**Trials.** For all experiments, we report the average behavior over five independent trials.

### 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 fixed edges, this finding is robust to the initial graph size and structure, and to the innate opinion distribution. See Figs. 11 and 14 in the appendix, which show similar behavior for large random graphs and real social network graphs, respectively. See Fig. ?? for an illustration with a bimodal innate opinion distribution. When fixed edges are present, they limit and slow the increase in polarization – this effect is amplified for larger fractions of fixed 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 [DBB<sup>+</sup>14], and a dense ER random graph with 1000 nodes. Polarization is essentially constant with just friend-of-friend recommendations, and it rises at a very slow rate 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 finding below.



**Figure 5: Polarization for ER random graphs with 1000 nodes, no fixed edges, and varied connection probabilities. Dynamics include both friend-of-friend recommendations and confirmation bias. We see polarization start to decrease once the density passes a high enough threshold, near  $1/\sqrt{n} \approx 0.03$ .**

**4.2.1 Effect of Density.** For particularly dense graphs, the asymptotic behavior of our model with both friend-of-friend recommendations and confirmation bias changes, as observed in Fig. 5. We see that denser graphs exhibit lower polarization when the simulation is complete, and that polarization also increases at a slower rate.

For ER graphs, a change in behavior seems to occur roughly when the average degree  $d_{avg} = pn$  is large enough such that the friends-of-friends set with size roughly  $d_{avg}^2$  encompasses nearly all nodes. I.e., when  $d_{avg}^2 = p^2 n^2 \gtrsim n$ , and so  $p \gtrsim 1/\sqrt{n}$ . In this case, friend-of-friend recommendations at initialization are essentially uniformly random, mirroring our control setting. A more precise theoretical understanding of this behavior would be very interesting and would help clarify the importance of friend-of-friend recommendations in amplifying the polarizing effects of confirmation bias.

### 4.3 Evolution of Polarization & Disagreement

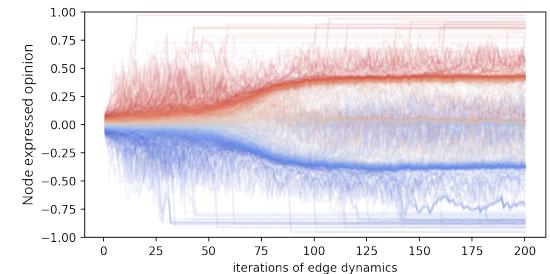
For both synthetic graphs and real-world networks, using friend-of-friend recommendations and confirmation bias, we find that expressed opinions evolve through roughly three distinct states.

**Initial state:** Opinions converge to be very close to the mean of the innate opinions – which is  $\approx 0$  in our setting. Polarization and disagreement are low. The network’s connectivity within similar opinions and between differing opinions is roughly equal.

**Bimodal polarization:** Expressed opinions bifurcate, “pulling apart” into a few distinct clusters on either side of the mean. The network’s connectivity is strengthening between similar opinions, and eroding between differing opinions (see Fig. 4). Both polarization and disagreement rise, with disagreement reaching a steady state if the network has fixed edges and a peak if it does not, as predicted by the SBM analysis of Sec. 3.3.

**Maximal polarization:** If fixed edges are not present in the network, the edge dynamics finally “splinter” the network into many components. Out-group connectivity  $q$  is near zero. Polarization is near maximal as defined in Prop. 4 and disagreement is near zero,

as predicted by the SBM analysis of Sec 3.3. Nodes cluster into a structure which keeps expressed opinions very close to innate opinions – in the appendix, Fig. 13 illustrates this by plotting the mean squared error between expressed opinions and innate opinions over time for a few experiment settings.



**Figure 6: Evolution of opinions on a Barabási-Albert graph with 1000 nodes, 5% fixed edges, and uniform random innate opinions. Each line represents the evolving expressed opinion of one node. The color of each line represents that node’s innate opinion, ranging on a gradient from -1 (blue) to 1 (red). We see opinions initially grouped near 0 (Initial state). We then see a bifurcation of the major opinion clusters, indicated by the red and blue “branches” (Bimodal polarization). The fixed edges in the network prevent it from fully splintering, so nodes are “stuck” in this bimodal state.**

In Figure 6, we show this progression of opinion states for a Barabási-Albert graph with 5% fixed edges. In Fig. ?? in the appendix, we show a similar plot for a network with no fixed edges, which continues onto the maximal polarization stage. Appendix Figure 12 also shows histograms of node opinions which provide another visual dimension for this state evolution.

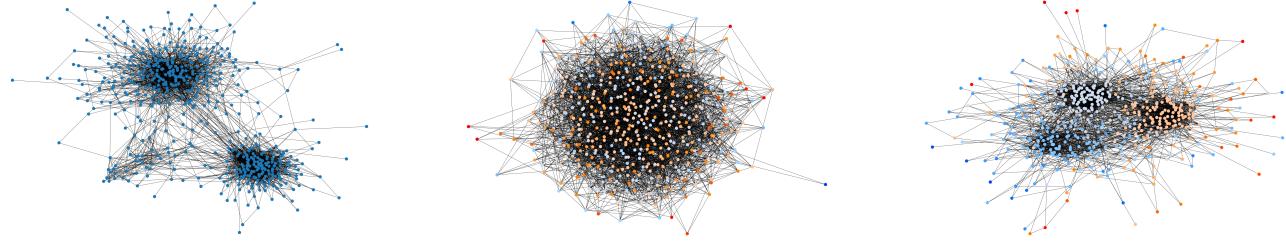
### 4.4 Validation Against Real-World Data

Following prior work [SCP<sup>+</sup>20, EF18], in this section, we use social network datasets as a “benchmark” to ascertain whether our synthetic model can recreate realistic social network structures. We show validation against the Twitter data set, with results on the Facebook data set appearing 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. We fix 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.

We find that the percentage of fixed 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 of fixed edges in our synthetic graph. E.g., for the Twitter network, the global clustering coefficient is  $\approx 0.227$ . The percentage of fixed edges leading to the closest steady-state clustering coefficient for both ER and BA networks is 25%.

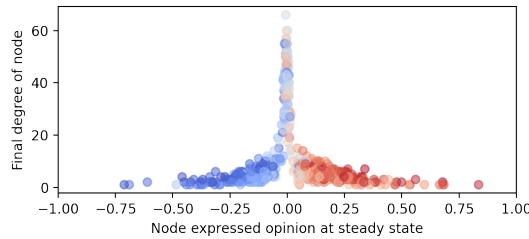


**Figure 7:** (left): Twitter [DBB<sup>+</sup>14] real-world snapshot. (middle): Initial ER graph with 25% fixed edges. (right): Steady state ER graph with 25% fixed edges, after edge dynamics simulation.

**Degree Distribution.** As in many social networks, Twitter has a power law degree distribution [MPP<sup>+</sup>13]. See Fig. ?? 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 loose connections. This behavior may drive the observed degree distribution shift. An improved theoretical understanding would be very interesting.

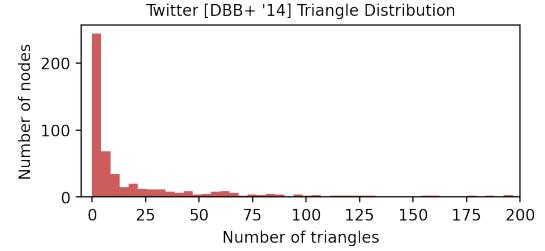
We show similar results for a BA network with 25% fixed edges in the appendix, Fig. ???. Note that the BA graph initially has a power-law degree distribution, so these results serve to validate that our model *preserves* a realistic degree distribution.



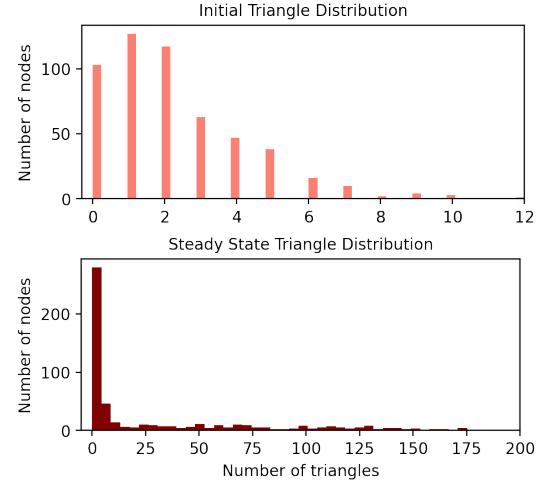
**Figure 8:** Scatter plot comparing expressed opinions and degree for each node in the steady state ER graph with 25% fixed edges. The color of each circle corresponds to the node’s innate opinion, ranging on a gradient from -1 (blue) to 1 (red).

**Triangle Distribution.** Similar to the degree distribution, in the Twitter network, the triangle distribution is power law like – see Fig. 9. In Fig. 10, we show that our edge dynamics drive the triangle distribution of an initial ER graph to be much closer to this power law distribution. We show results for the BA generated network with 25% fixed edges in the Appendix, in Fig. ???. Again, in this case, our dynamics preserve the already relatively realistic triangle distribution of the initial BA graph.

**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  $[-1, 1]$ .



**Figure 9:** Triangle distribution for Twitter data set [DBB<sup>+</sup>14].



**Figure 10:** Initial and steady state triangle distributions for an ER network, subject to our edge dynamics with 25% fixed edges. Notably, the steady state distribution appears closer to the Twitter network’s triangle distribution (Fig. 9).

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. ?? & ??, 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 [DeG74, HK02, EF18, SCP<sup>+</sup>20], 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

- [AGS20] Guy Aridor, Duarte Goncalves, and Shan Sikdar. Deconstructing the filter bubble: User decision-making and recommender systems. *Fourteenth ACM Conference on Recommender Systems*, 2020.
- [AKPT18] Rediet Abebe, Jon Kleinberg, David Parkes, and Charalampos E Tsourakakis. Opinion dynamics with varying susceptibility to persuasion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1089–1098, 2018.
- [BG08] Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.
- [BLSSS20] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4):048301, 2020.
- [BP20] Heather Z Brooks and Mason A Porter. A model for the influence of media on the ideology of content in online social networks. *Physical Review Research*, 2(2):023041, 2020.
- [CFL09] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591, 2009.
- [CKSV19] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nishanth Vishnoi. Controlling polarization in personalization. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.
- [CM20] Uthsav Chitra and Christopher Musco. Analyzing the impact of filter bubbles on social network polarization. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020.
- [CR20] Mayee Chen and Miklós Rácz. Network disruption: maximizing disagreement and polarization in social networks, 2020.
- [DBB<sup>+</sup>14] Abir De, Sourangshu Bhattacharya, Parantapa Bhattacharya, Niloy Ganguly, and Soumen Chakrabarti. Learning a linear influence model from transient opinion dynamics. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 401–410, New York, NY, USA, 2014. Association for Computing Machinery.
- [DeG74] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121, 1974.
- [Del20] Daniel DellaPosta. Pluralistic collapse: The “oil spill” model of mass opinion polarization. *American Sociological Review*, 85(3):507–536, 2020.
- [DGL13] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.
- [DGM10] Elizabeth M Daly, Werner Geyer, and David R Millen. The network effects of recommending social connections. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, 2010.
- [EF18] Tucker Evans and Feng Fu. Opinion formation on dynamic networks: identifying conditions for the emergence of partisan echo chambers. *Royal Society Open Science*, 5:181122, 2018.
- [FJ90] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4):193–206, 1990.
- [GKT20] Jason Gaitonde, Jon Kleinberg, and Eva Tardos. Adversarial perturbations of opinion dynamics in networks. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 471–472, 2020.
- [GMJCK13] P.H. Guerra, Wagner Meira Jr, Claire Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. *Proceedings of the 7th International Conference on Weblogs and Social Media*, ICWSM 2013, pages 215–224, 2013.
- [HK02] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.
- [HMRU21] Shahrzad Haddadan, Cristina Menghini, Matteo Riondato, and Eli Upfal. Republik: Reducing polarized bubble radius with link insertions. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021.
- [HN06] Petter Holme and M. E. J. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74, 2006.
- [HN12] P. Sol Hart and Erik C. Nisbet. Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, 39(6):701–723, 2012.
- [LCKK14] Jae Kook Lee, Jihyang Choi, Cheonsoo Kim, and Yonghwan Kim. Social Media, Network Heterogeneity, and Opinion Polarization. *Journal of Communication*, 64(4):702–722, 2014.
- [LM12] Jure Leskovec and Julian McAuley. Learning to discover social circles in ego networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [MMT18] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 27th International World Wide Web Conference (WWW)*, 2018.
- [MPP<sup>+</sup>13] Lev Muchnik, Sen Pei, Lucas C. Parra, Saulo D. S. Reis, José S. Andrade Jr, Shlomo Havlin, and Hernán A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific Reports*, 3(1):1783, May 2013.
- [MSFD19] Corine S. Meppelink, Edith G. Smit, Marieke L. Fransen, and Nicola Di Viani. “I was right about vaccination”: Confirmation bias and health literacy in online health information seeking. *Journal of Health Communication*, 24(2):129–140, 2019. PMID: 30895889.
- [NHH<sup>+</sup>14] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd International World Wide Web Conference (WWW)*, 2014.
- [Nic98] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220, 1998.
- [Par12] Eli Pariser. *The filter bubble: what the Internet is hiding from you*. Viking, 2012.
- [PS21] Eric M. Patashnik and Wendy J. Schiller. The larger forces behind the January 6, 2021 insurrection | Watson Institute, 2021.
- [RR22] Miklós Rácz and Daniel Rigobon. Towards Consensus: Reducing Polarization by Perturbing Social Networks, 2022.
- [SCP<sup>+</sup>20] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 2020.
- [SM50] Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.
- [Spi19] Daniel Spielman. *Spectral and Algebraic Graph Theory*. <http://cs-www.cs.yale.edu/homes/spielman/sagt/>, 2019.
- [SSG16] Jessica Su, Aneesh Sharma, and Sharad Goel. The effect of recommendations on network structure. In *Proceedings of the 25th International World Wide Web Conference (WWW)*, pages 1157–1167, 2016.
- [TJ19] Edson C. Tandoc Jr. The facts of fake news: A research review. *Sociology Compass*, 13(9):e12724, 2019. e12724 SOCO-1481.R1.
- [XBZ21] Wanyue Xu, Qi Bao, and Zhongzhi Zhang. Fast evaluation for relevant quantities of opinion dynamics. In *Proceedings of the 30th International World Wide Web Conference (WWW)*, 2021.

## A DEFERRED PROOFS

We now prove Lemmas 5 and 7 and Corollary 8, which bound the change in polarization+disagreement when adding or deleting an edge, and when swapping two edges respectively.

**LEMMA 5.** Consider any unweighted graph Laplacian  $L \in \mathbb{R}^{n \times n}$ , and let  $s, z \in \mathbb{R}^n$  be an innate and corresponding expressed opinion vectors in the Friedkin-Johnsen model. Let  $E \in \mathbb{R}^{n \times n}$  be the edge Laplacian for edge  $(u, v)$ . Let  $\delta = z(u) - z(v)$  and  $r_{u,v} = \chi_{u,v}^T(I + L)^{-1}\chi_{u,v}$ .

$$\begin{aligned} PD(L + E) &= PD(L) - \frac{\delta^2}{1 + r_{u,v}}. \\ PD(L - E) &= PD(L) - \frac{\delta^2}{1 - r_{u,v}}. \end{aligned}$$

**PROOF.** By the Sherman-Morrison Formula:

$$\begin{aligned} s^T(I + L + E)^{-1}s &= s^T(I + L)^{-1}s - \frac{s^T(I + L)^{-1}E(I + L)^{-1}s}{1 + \chi_{u,v}^T(I + L)^{-1}\chi_{u,v}} \\ &= PD(L) - \frac{z^T Ez}{1 + \chi_{u,v}^T(I + L)^{-1}\chi_{u,v}}. \\ s^T(I + L - E)^{-1}s &= s^T(I + L)^{-1}s + \frac{s^T(I + L)^{-1}E(I + L)^{-1}s}{1 - \chi_{u,v}^T(I + L)^{-1}\chi_{u,v}} \\ &= PD(L) + \frac{z^T Ez}{1 - \chi_{u,v}^T(I + L)^{-1}\chi_{u,v}}. \end{aligned}$$

Finally, note that  $z^T Ez = [z(u) - z(v)]^2 = \delta^2$ .  $\square$

**LEMMA 7.** Consider any unweighted graph Laplacian  $L \in \mathbb{R}^{n \times n}$ , and let  $s, z \in \mathbb{R}^n$  be the innate and corresponding expressed opinion vectors in the Friedkin-Johnsen 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  $r_1 = \chi_1^T(I + L)^{-1}\chi_1$ ,  $r_2 = \chi_2^T(I + L)^{-1}\chi_2$ , and  $r_{2,1} = \chi_2^T(L + I + E_1)^{-1}\chi_2$ .

**PROOF.** We apply Lemma 5 in sequence to give:

$$PD(L + E_1 - E_2) = PD(L) - \frac{\delta_1^2}{1 + r_1} + \frac{s^T(L + I + E_1)^{-1}E_2(L + I + E_1)^{-1}s}{1 - \chi_2^T(L + I + E_1)^{-1}\chi_2}.$$

We then expand out the third term using Sherman-Morrison again:

$$\begin{aligned} PD(L + E_1 - E_2) &= PD(L) - \frac{\delta_1^2}{1 + r_1} + \frac{\delta_2^2}{1 - r_{2,1}} \\ &\quad + \frac{s^T(L + I)^{-1}E_1(L + I)^{-1}E_2(L + I)^{-1}E_1(L + I)^{-1}s}{(1 - r_{2,1}) \cdot (1 + r_1)^2} \\ &\quad - \frac{2s^T(L + I)^{-1}E_1(L + I)^{-1}E_2(L + I)^{-1}s}{(1 - r_{2,1}) \cdot (1 + r_1)}. \end{aligned}$$

Letting  $\alpha = \frac{\chi_1^T(L+I)^{-1}\chi_2}{1+r_1}$ , we can simplify the above to:

$$\begin{aligned} PD(L + E_1 - E_2) &= PD(L) - \frac{\delta_1^2}{1 + r_1} + \frac{\delta_2^2}{1 - r_{2,1}} + \frac{\alpha^2 \cdot \delta_1^2}{1 - r_{2,1}} - \frac{2\alpha\delta_1\delta_2}{1 - r_{2,1}} \\ &= PD(L) - \frac{\delta_1^2}{1 + r_1} + \frac{(\delta_2 - \alpha\delta_1)^2}{1 - r_{2,1}}, \end{aligned}$$

completing the bound. Finally, note that since  $L + I$  is positive definite, we can bound  $|\alpha|$  using Cauchy-Schwarz by:

$$|\alpha| \leq \frac{\sqrt{\chi_1^T(L + I)^{-1}\chi_1 \cdot \chi_2^T(L + I)^{-1}\chi_2}}{1 + r_1} = \frac{\sqrt{r_1 \cdot r_2}}{1 + r_1}.$$

$\square$

**COROLLARY 8.** Consider the setting of Lemma 7. If  $|\delta_2| > \frac{3\sqrt{3}}{4} \cdot |\delta_1|$  then  $PD(L + E_1 - E_2) > PD(L)$ .

**PROOF.** First note that since we assume edge  $(u_2, v_2)$  is in the graph,  $r_{2,1} \leq r_2 < 1$ . Thus by Lemma 7 we have:

$$PD(L + E_1 - E_2) > PD(L) - \frac{\delta_1^2}{1 + r_1} + \left(|\delta_2| - \frac{\sqrt{r_1}}{1 + r_1} \cdot |\delta_1|\right)^2.$$

Writing  $|\delta_2| = \gamma \cdot |\delta_1|$  and solving for  $PD(L + E_1 - E_2) = PD(L)$ ,

$$\frac{\delta_1^2}{1 + r_1} = \left(\gamma - \frac{\sqrt{r_1}}{1 + r_1}\right)^2 \cdot \delta_1^2.$$

Solving for  $\gamma$  under the constraint that  $\gamma > 0$  gives  $\gamma = \frac{\sqrt{r_1}}{1+r_1} + \frac{1}{\sqrt{1+r_1}}$ .  $\gamma$  is maximized for all  $r_1 \geq 0$  at  $r_1 = 1/3$  and  $\gamma = \frac{3\sqrt{3}}{4}$ , which completes the proof.  $\square$

## B DATASET DETAILS

- **Reddit** [DBB<sup>+</sup>14]

$n = 546$  nodes,  $e = 8962$  edges

In this dataset, nodes (users) have an edge between them if there exist two subreddits in which both users posted during a given time period. Files for this data set were obtained from previous work that cites the original source [MMT18].

- **Twitter** [DBB<sup>+</sup>14]

$n = 531$  nodes,  $e = 3621$  edges

In this dataset aimed at analyzing discourse around the Delhi legislative elections of 2013, edges represent user interactions on the Twitter platform, discerned with the use of topical hashtags. Files for this data set were obtained from previous work that cites the original source [MMT18].

- **Facebook Egograph** [LM12]

$n = 3964$  nodes,  $e = 88159$  edges

Consists of ten anonymized ego networks, which are social circles of Facebook users – the ten overlapping networks are combined into a single connected component.

## C ADDITIONAL PLOTS

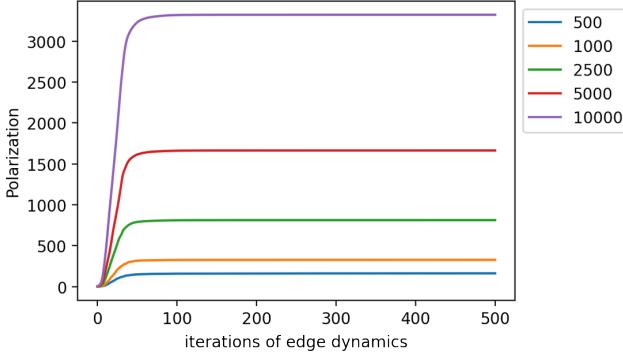


Figure 11: Polarization over time for varying sizes  $n$  of random ER graphs with average degree 25, yielding  $25(n - 1)/2$  expected edges, and no fixed edges. Dynamics include both friend-of-friend recommendations and confirmation bias. Innate opinions are distributed uniformly on the interval  $[-1, 1]$ . While the asymptotic value of polarization changes proportionally to the number of nodes in the graph, the polarizing behavior of edge dynamics generalizes across different graph sizes.

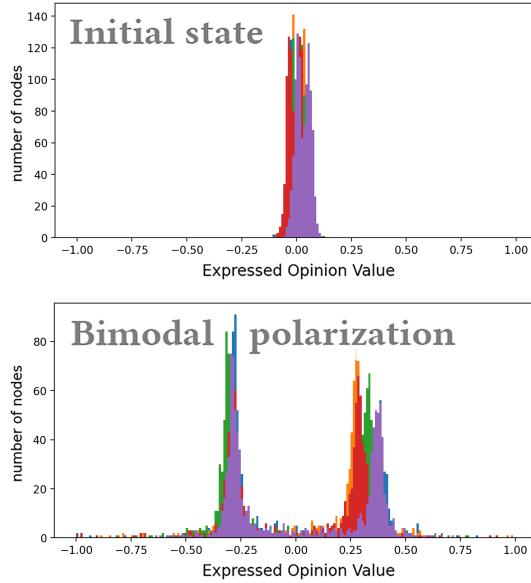


Figure 12: Two states of opinion polarization for an Erdős-Renyi graph with  $n = 1000$  nodes, uniform innate opinions on the interval  $[-1, 1]$ , and 5% of edges fixed. Plots are histograms of the expressed opinions at time steps  $t = 5$  (top) and  $t = 120$  (bottom). Each colors represent one of 5 independent trials. In the *initial state*, the opinions center around 0, due to opinion averaging. The second plot shows the *bimodal polarization* state, where opinions cluster into groups on either side of the mean. Since the presence of fixed edges prevents the network from splintering further, this bimodal state is the steady state in the network.

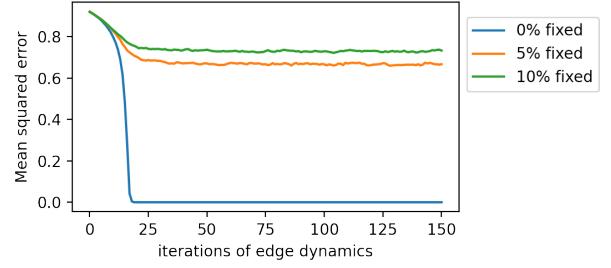


Figure 13: Mean squared error over time between expressed opinions and innate opinions for an Erdős-Renyi graph, with varying percentages of *fixed edges*, as described in Fig. 2. Innate opinions are uniformly assigned from  $\{-1, 1\}$ . When fixed edges are not present, the network can splinter and sort nodes such that expressed opinions are indistinguishable from innate opinions. Otherwise, the mean squared error between expressed and innate opinions is reduced proportionally to the rise in polarization.

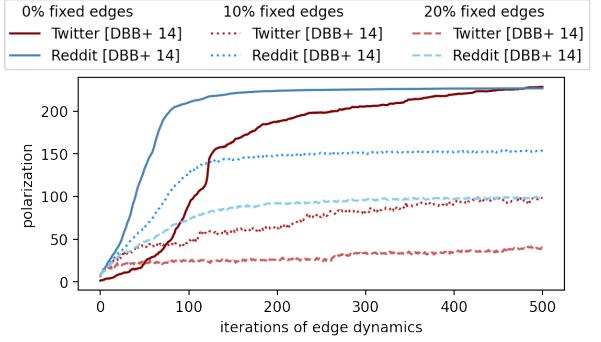


Figure 14: Varying percentages of *fixed edges*, on Twitter and Reddit real-world networks [DBB+14]. Edge dynamics include both friend-of-friend recommendations and confirmation bias.

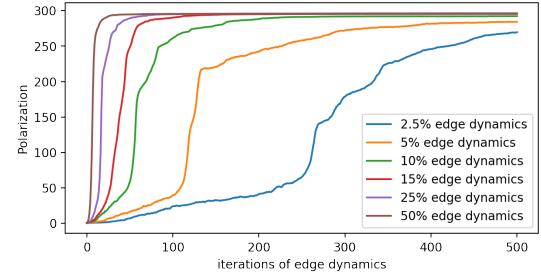
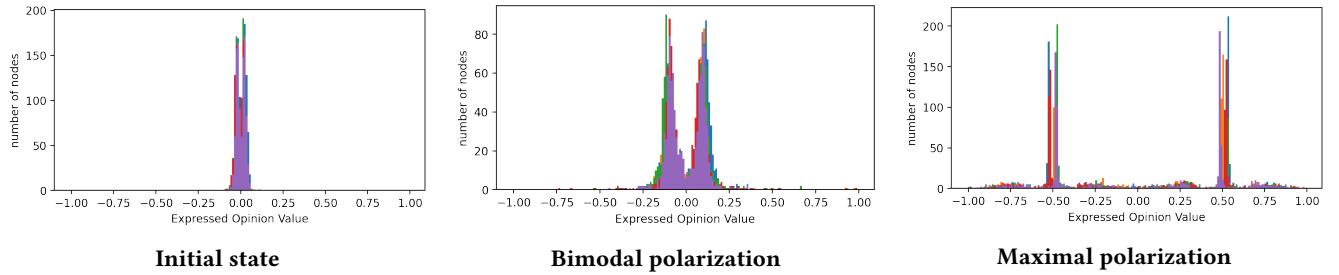
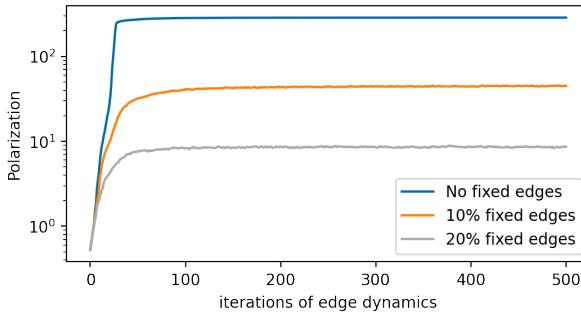


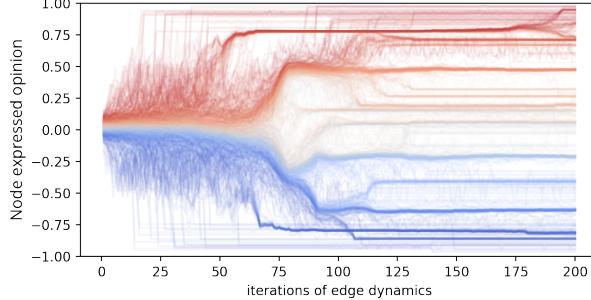
Figure 15: Polarization over time for an ER graph with  $n = 1000$  nodes and fixed degree 25, yielding  $25(n - 1)/2$  expected edges. Dynamics include both friend-of-friend recommendations and confirmation bias. We vary the percentage of edges which are added and removed at each time step, showing that the choice of this percentage simply scales the amount of time steps necessary for convergence. Innate opinions are distributed uniformly on the interval  $[-1, 1]$ .



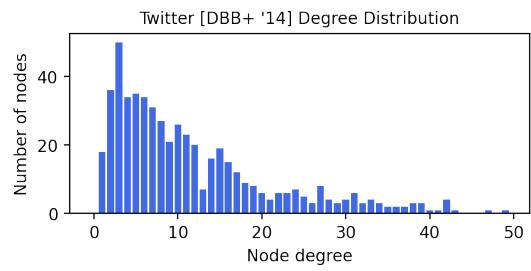
**Figure 16: States of opinion polarization for an Erdős-Renyi graph with no fixed edges, innate opinions sampled from mixture of two Gaussians with  $\mu = \pm 0.5$  and  $\sigma = 0.2$ , on interval  $[-1, 1]$ . Plots are histograms of the expressed opinions at  $t = 1$ ,  $t = 20$ , and  $t = 300$  time steps, from left to right. Colors represent 5 independent trials.**



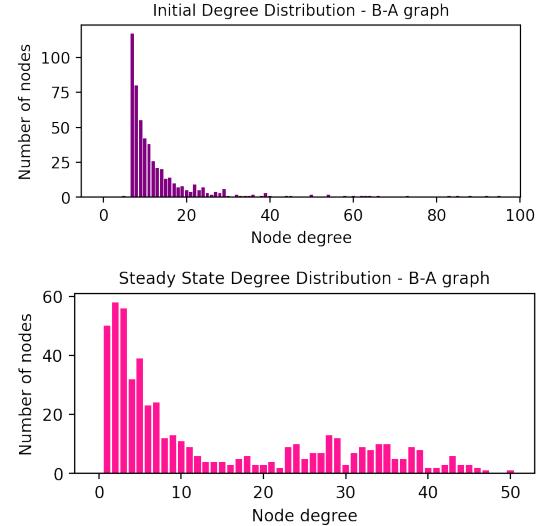
**Figure 17: Polarization over time for an Erdős-Renyi graph with 1000 nodes, varying percentages of fixed edges, and average degree 25, yielding  $25(n - 1)/2$  expected edges. Dynamics include both friend-of-friend recommendations and confirmation bias. Innate opinions are sampled from mixture of two Gaussians with  $\mu = \pm 0.5$  and  $\sigma = 0.2$ , on interval  $[-1, 1]$ . In this different setting for the innate opinions, polarization reaches a steady state slightly faster, but the behavior of our model is similar.**



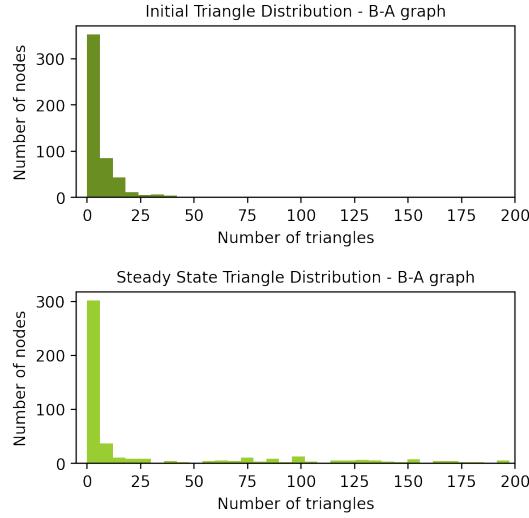
**Figure 18: Evolution of individual opinions on a Barabási-Albert graph with 1000 nodes,  $m = 10$ , and no fixed edges, with uniform random innate opinions. Each line represents the evolving expressed opinion of one node. The color of each line represents that node's innate opinion, ranging on a gradient from -1 (blue) to 1 (red). We see opinions initially grouped near 0 (Initial state). We then see a bifurcation of the major opinion clusters, denoted by the darker & more saturated lines (Bimodal polarization), and finally, a spreading of opinions into clusters based on their innate opinions, representing (Maximal polarization).**



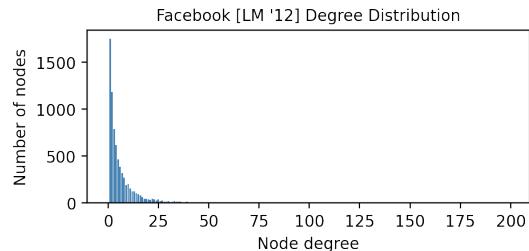
**Figure 19: Degree distribution histogram for Twitter data set [DBB<sup>+</sup>14]. The x-axis denotes the degree, and the height of each bar is the number of nodes with that degree.**



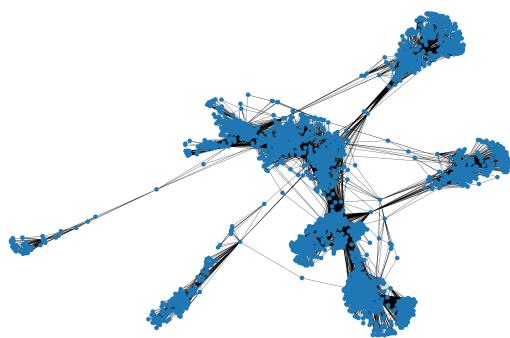
**Figure 20: Initial and steady state degree distribution histograms for BA generated network, validating against Twitter data set [DBB<sup>+</sup>14]. The x-axis denotes the degree, and the height of each bar is the number of nodes with that degree. The steady state histogram preserves a power law degree distribution, reflecting what can be considered a realistic social network structure.**



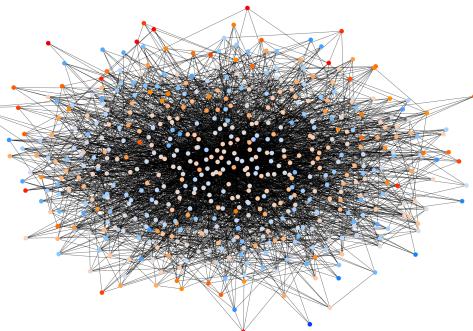
**Figure 21: Initial and steady state triangle distribution histograms for BA generated network, validating against Twitter data set [DBB<sup>+</sup>14].** The  $x$ -axis denotes the number of triangles, and the height of each bar is the number of nodes with that triangle count. The steady state histogram shows that edge dynamics preserve the distribution of triangles in the initial graph, which is similar to the distribution in the Twitter data set.



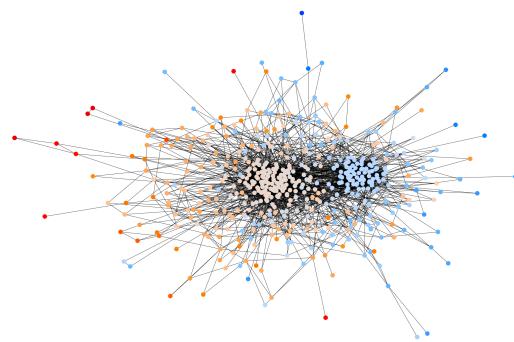
**Figure 22: Degree distribution histogram for Facebook data set [LM12].** The  $x$ -axis denotes the degree, and the height of each bar is the number of nodes with that degree.



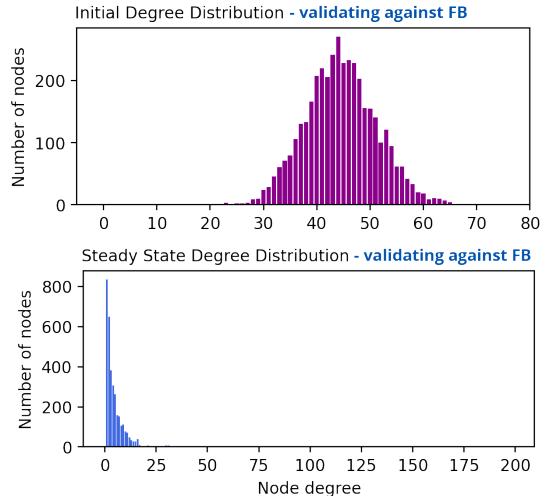
**Figure 23: Facebook Egograph [LM12] real-world snapshot.**



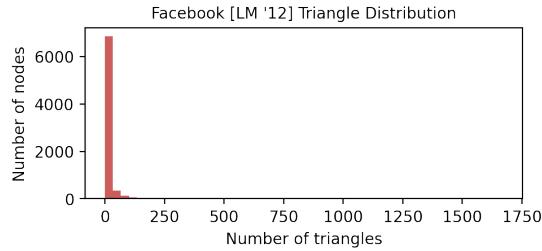
**Figure 24: Initial Barabási-Albert graph with 25% fixed edges, before edge dynamics simulation.**



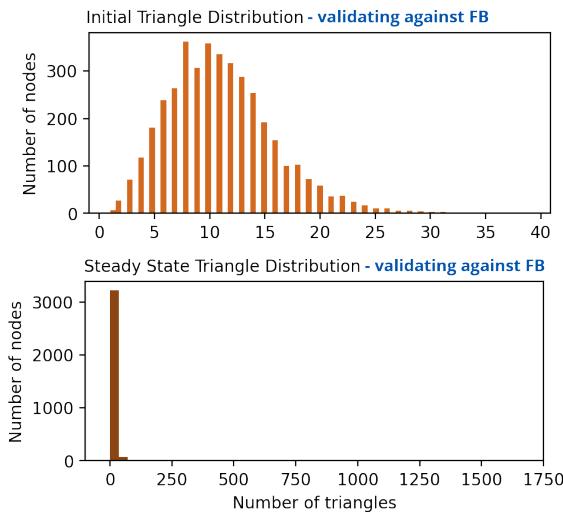
**Figure 25: Steady state Barabási-Albert graph with 25% fixed edges, after edge dynamics simulation.**



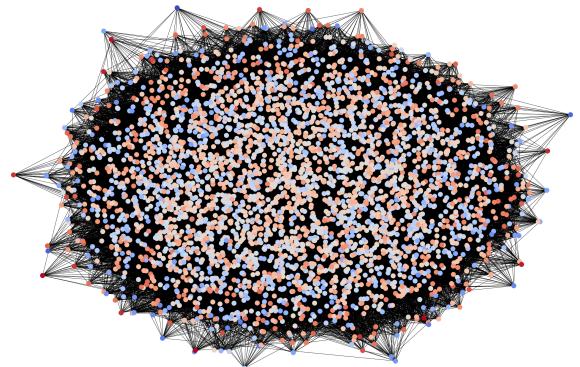
**Figure 26: Initial and steady state degree distribution histograms for ER generated network, validating against Facebook data set [LM12].** The  $x$ -axis denotes the degree, and the height of each bar is the number of nodes with that degree. The steady state distribution differs significantly from the initial distribution. It is closer to a power law distribution, reflecting a more realistic network structure.



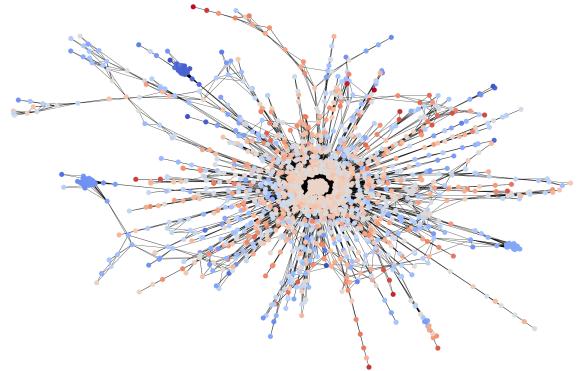
**Figure 27:** Triangle distribution histogram for Facebook data set [DBB<sup>+</sup>14]. The  $x$ -axis denotes the number of triangles incident on a node, and the height of each bar is the number of nodes with that number of triangles.



**Figure 28:** Initial and steady state triangle distribution histograms for BA generated network, validating against Twitter data set [DBB<sup>+</sup>14]. The  $x$ -axis denotes the number of triangles, and the height of each bar is the number of nodes with that triangle count. The steady state histogram shows that edge dynamics preserve the distribution of triangles in the initial graph, which is similar to the distribution in the Twitter data set.



**Figure 29:** Initial Erdős-Renyi graph with 3964 nodes, 87872 edges, and 5% fixed edges, validating against Facebook data set; visualized before edge dynamics simulation.



**Figure 30:** Steady state Erdős-Renyi graph with 3964 nodes, 87872 edges, and 5% fixed edges, validating against Facebook data set; visualized after edge dynamics simulation.