

# Theoretical Models for Opinion Polarization via Local Edge Dynamics

---

Cameron Musco

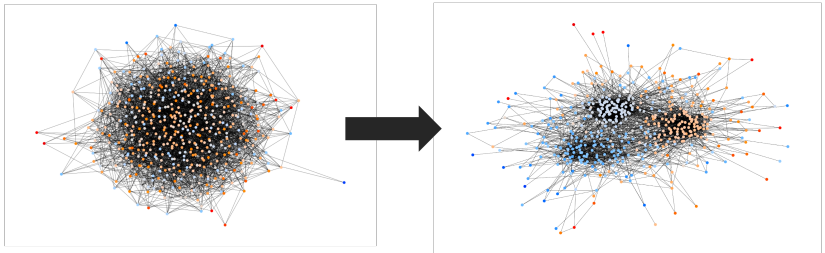
University of Massachusetts Amherst

Joint With: Adam Lechowicz and Nikita Bhalla

Integrity Workshop. WSDM 2023.

# Outline

- We propose a simple theoretical model for **network and opinion co-evolution**.
- How do node opinions drive the addition and deletion of edges in a social network, and in turn, how do these edge modifications drive opinion evolution?
- We use this model to study the role of phenomena such as **confirmation bias** and **recommender systems** in driving opinion polarization and filter bubble formation in social networks.



**Disclaimer:** Our model is a highly stylized and theoretical.

- Similar in spirit to DeGroot learning, Friedkin-Johnsen dynamics, preferential attachment models, stochastic block models, etc.
- We should not think of it as an accurate reflection of the real world, but as a theoretical tool for getting a handle on the driving forces behind opinion polarization.

## Related Work

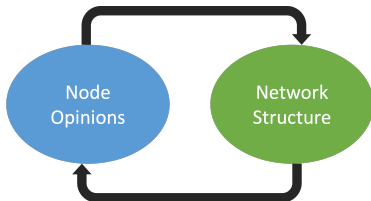
Our work fits into a broad literature in recent years studying opinion polarization and filter bubble formation.

- **Effects of edge rewiring on opinion polarization:** [Musco, Musco, Tsourakakis '18], [Abebe, et al. '18], [Uthsav, Musco '20], [Chen, Ráz '20], [Gaitonde, Kleinberg, Tardos '20].
- **Effects of recommendation systems on network structure:** [Daly, Geyer, Millen '10], [Su, Sharma, Goel '16].

## Related Work

Our work fits into a broad literature in recent years studying opinion polarization and filter bubble formation.

- **Effects of edge rewiring on opinion polarization:** [Musco, Musco, Tsourakakis '18], [Abebe, et al. '18], [Uthsav, Musco '20], [Chen, Ráz '20], [Gaitonde, Kleinberg, Tardos '20].
- **Effects of recommendation systems on network structure:** [Daly, Geyer, Millen '10], [Su, Sharma, Goel '16].
- **Opinion and network co-evolution:** [Holme, Newman '06], [Dandekar, Goel, Lee '13], [Sasahara et al. '20]



# Opinion and Network Co-Evolution Model

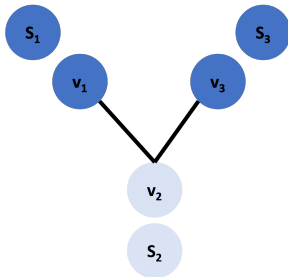
# Opinion Dynamics Model

## Underlying Opinion Dynamics Model: Friedkin-Johnsen

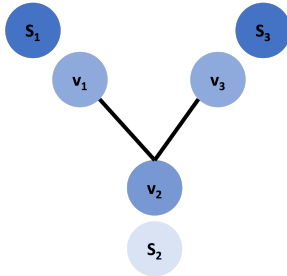
- $n$  individuals are represented as nodes in an undirected, unweighted social network graph.
- Each node has an **innate opinion**  $\mathbf{s}_i \in [-1, 1]$ .
- Each node also has an **expressed opinion**  $\mathbf{z}_i \in [-1, 1]$ . Initially  $\mathbf{z}_i = \mathbf{s}_i$ .
- At any time step, a node's **expressed opinion** is obtained by averaging its innate opinion with the expressed opinions of its neighbors.

$$\mathbf{z}_i := \frac{\mathbf{s}_i + \sum_{j \in \mathcal{N}(i)} \mathbf{z}_j}{|\mathcal{N}(i)| + 1}.$$

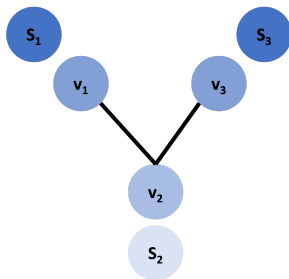
# Opinion Dynamics Model



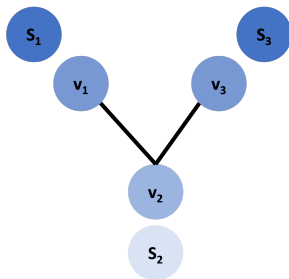
# Opinion Dynamics Model



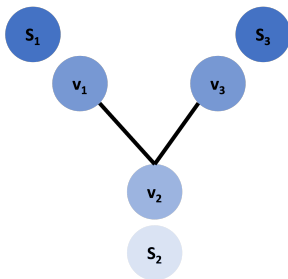
# Opinion Dynamics Model



# Opinion Dynamics Model



# Opinion Dynamics Model



At equilibrium,  $\mathbf{z} = (\mathbf{L} + \mathbf{I})^{-1}\mathbf{s}$ , where  $\mathbf{L} \in \mathbb{R}^{n \times n}$  is the Laplacian of the network graph and  $\mathbf{z}, \mathbf{s} \in \mathbb{R}^n$  are the vectors of expressed and innate opinions respectively.

## Measures of Polarization

[Musco, Musco, Tsourakakis '18] demonstrate that several natural measures of polarization have simple linear algebraic forms under the Friedkin-Johnsen model at equilibrium.

# Measures of Polarization

[Musco, Musco, Tsourakakis '18] demonstrate that several natural measures of polarization have simple linear algebraic forms under the Friedkin-Johnsen model at equilibrium.

- The **polarization**, which is the variance of expressed opinions is given by  $\|\mathbf{z}\|_2^2 = \mathbf{z}^T \mathbf{z} = \mathbf{s}^T (\mathbf{L} + \mathbf{I})^{-2} \mathbf{s}$ .
- The **disagreement**, which is the sum of squared differences of expressed opinions across edges in the graph is given by  $\sum_{(i,j) \in E} (\mathbf{z}_i - \mathbf{z}_j)^2 = \mathbf{z}^T \mathbf{L} \mathbf{z} = \mathbf{s}^T (\mathbf{L} + \mathbf{I})^{-1} \mathbf{L} (\mathbf{L} + \mathbf{I})^{-1} \mathbf{s}$ .
- The **polarization + disagreement** is given by  $\mathbf{s}^T (\mathbf{L} + \mathbf{I})^{-1} \mathbf{s}$ .



## Edge Dynamics Model

Starting from a graph at a Friedkin-Johnsen opinion equilibrium, edges are modified via two local update rules.

## Edge Dynamics Model

Starting from a graph at a Friedkin-Johnsen opinion equilibrium, edges are modified via two local update rules.

**Edge Deletions:** Remove edge  $(i, j)$  with probability proportional to the disagreement  $(z_i - z_j)^2$ . Models **confirmation bias** – the human tendency to seek out information that confirms existing beliefs.

# Edge Dynamics Model

Starting from a graph at a Friedkin-Johnsen opinion equilibrium, edges are modified via two local update rules.

**Edge Deletions:** Remove edge  $(i, j)$  with probability proportional to the disagreement  $(z_i - z_j)^2$ . Models **confirmation bias** – the human tendency to seek out information that confirms existing beliefs.

**Edge Additions:** Add a random friend-of-friend connection. I.e., a random edge  $(i, j)$  such that  $i$  and  $j$  share a neighbor  $k$ . Models a very simple **recommendation system**.

# Edge Dynamics Model

Starting from a graph at a Friedkin-Johnsen opinion equilibrium, edges are modified via two local update rules.

**Edge Deletions:** Remove edge  $(i, j)$  with probability proportional to the disagreement  $(z_i - z_j)^2$ . Models **confirmation bias** – the human tendency to seek out information that confirms existing beliefs.

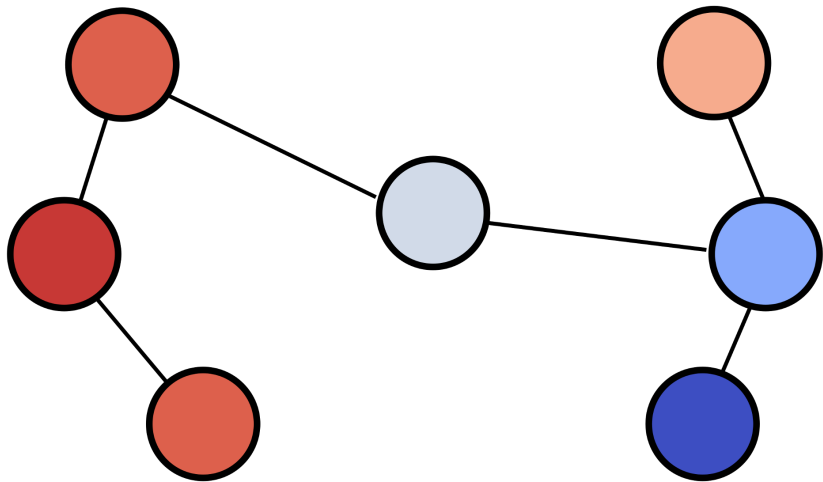
**Edge Additions:** Add a random friend-of-friend connection. I.e., a random edge  $(i, j)$  such that  $i$  and  $j$  share a neighbor  $k$ . Models a very simple **recommendation system**.

- We always add and remove the same number of edges to keep the total number of edges constant.
- For efficiency in simulation we often add and delete edges in batches, consisting of say 5% of the total edges. We find that the behavior of the model is fairly constant across different batch sizes.

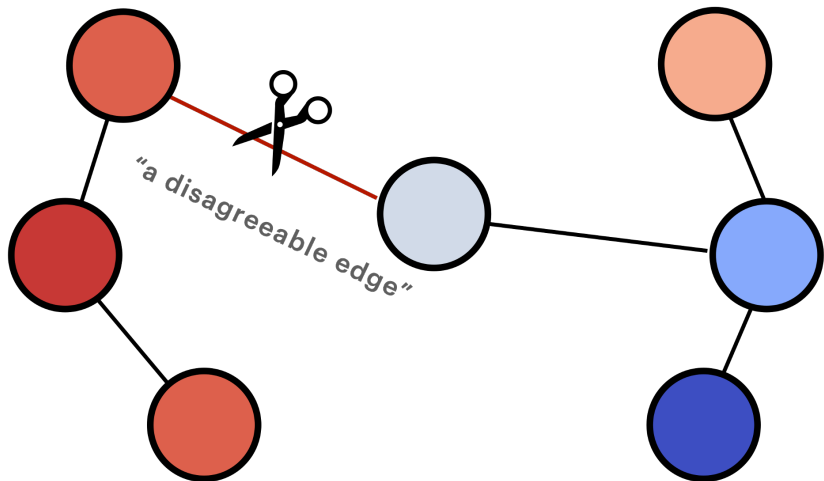
# Co-Evolution Model



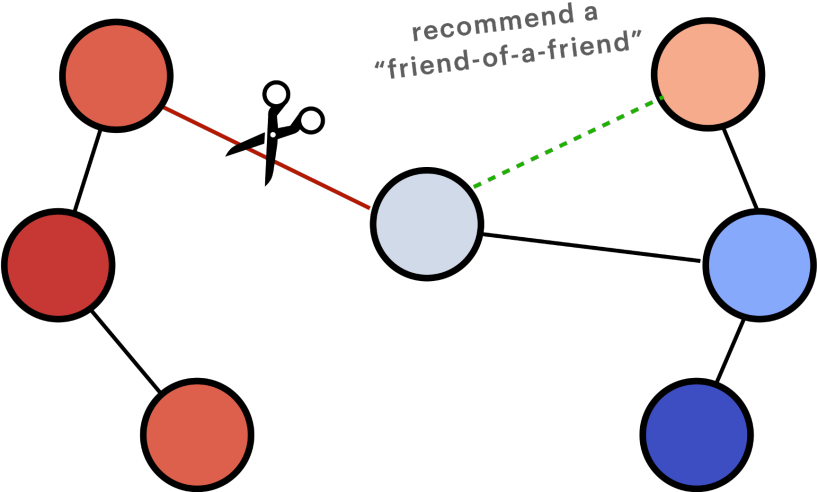
## Graph Evolution Example



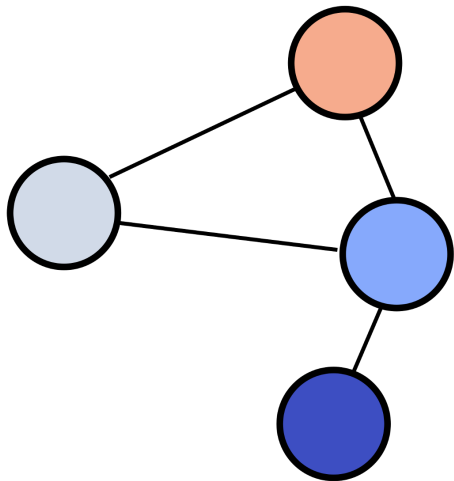
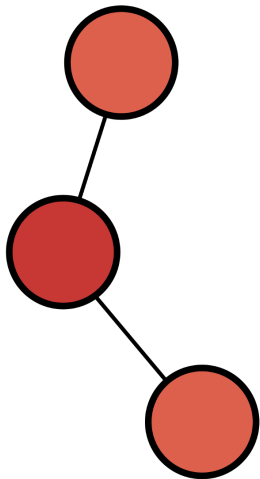
# Graph Evolution Example



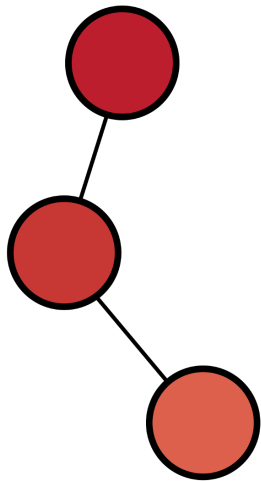
# Graph Evolution Example



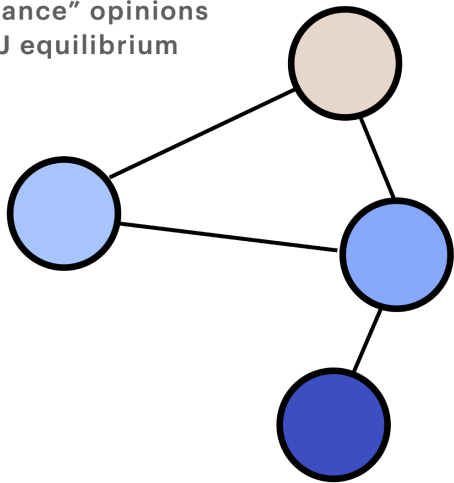
## Graph Evolution Example



## Graph Evolution Example

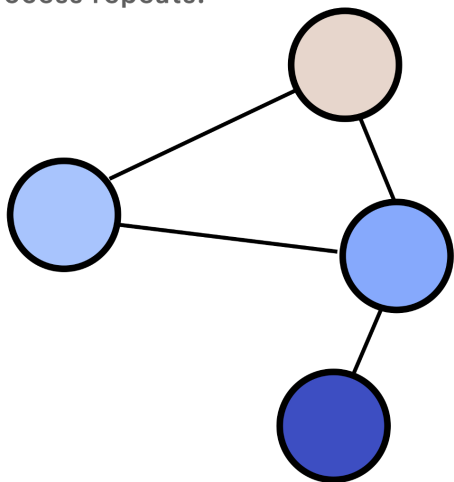
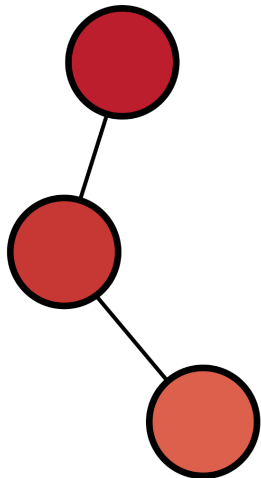


“rebalance” opinions  
to FJ equilibrium

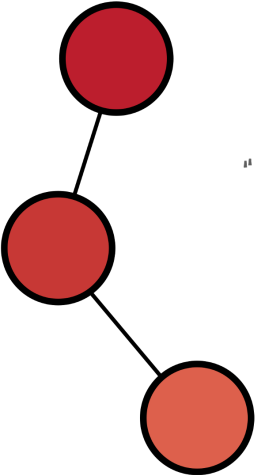


# Graph Evolution Example

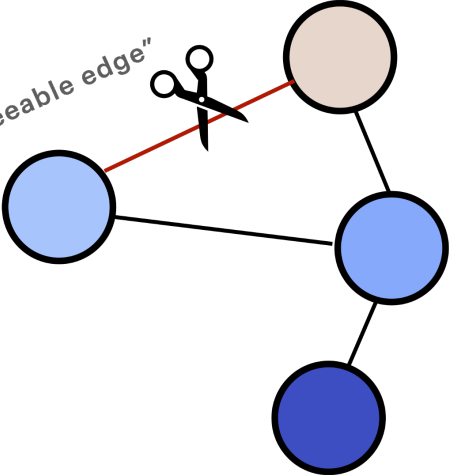
the process repeats.



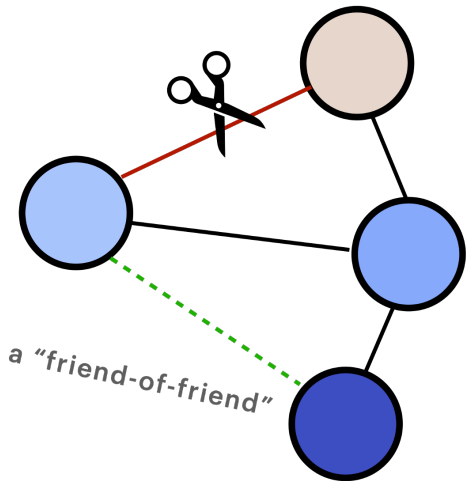
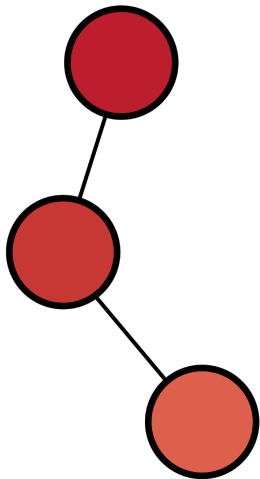
# Graph Evolution Example



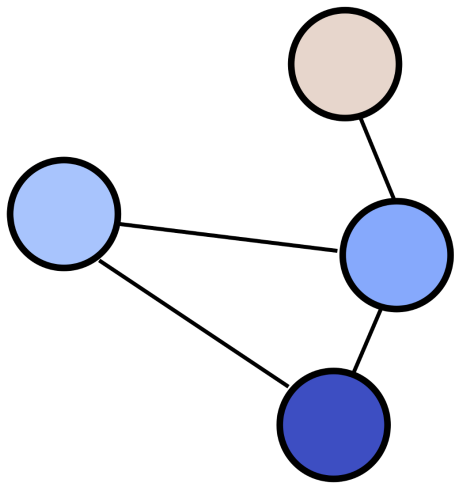
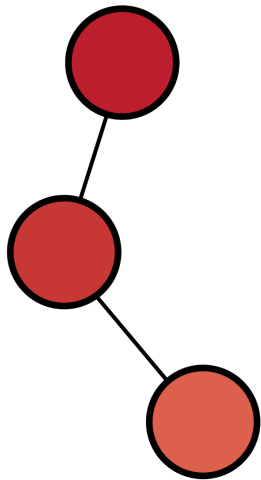
"a disagreeable edge"



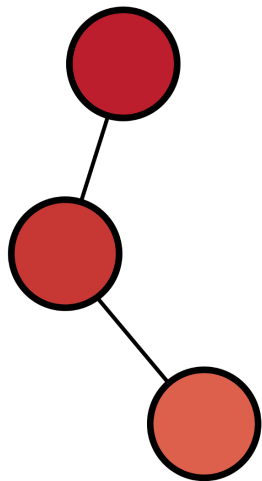
# Graph Evolution Example



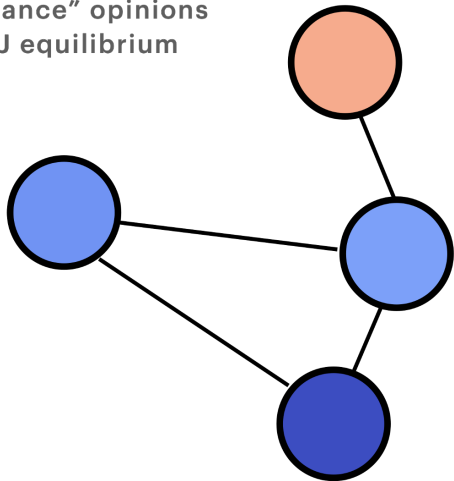
## Graph Evolution Example



## Graph Evolution Example



“rebalance” opinions  
to FJ equilibrium

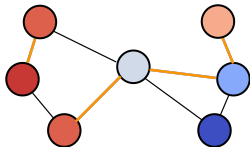


## Other Parameters: Initialization

- We typically initialize the network to be a real-world social network or a random Erdős-Renyi or Barabási-Albert graph.
- Innate opinions are typically chosen uniformly at random in  $[-1, 1]$ , independent of the network structure.
- Generally, we find that the network evolution and final network equilibrium are fairly uncorrelated with the initial conditions.

## Other Parameters: Fixed Edges

- We typically choose a small percentage (e.g., 5%) of the initial edges to be **fixed** – i.e., not subject to edge deletions.



- These edges can be thought of as modeling edges which are not subject to deletion based on disagreement – e.g., between family members or co-workers.
- The small percentage of fixed edges leads to more realistic network evolution.
- Without them, the graph tends to splinter into disconnected components corresponding to innate opinion clusters. I.e., polarization is maximized and disagreement is minimized.

## Our Findings

# Main Results

1. In conjunction, confirmation bias and friend-of-friend recommendations drive the network to a highly polarized state. If either is replaced with a random control, the network does not polarize significantly.

# Main Results

1. In conjunction, confirmation bias and friend-of-friend recommendations drive the network to a highly polarized state. If either is replaced with a random control, the network does not polarize significantly.
2. Give initial theoretical results explaining convergence in our model. Analyze the effect of local 'edge swaps' on polarization + disagreement and a introduce simple surrogate mode based on in/out group connectivity.

# Main Results

1. In conjunction, confirmation bias and friend-of-friend recommendations drive the network to a highly polarized state. If either is replaced with a random control, the network does not polarize significantly.
2. Give initial theoretical results explaining convergence in our model. Analyze the effect of local 'edge swaps' on polarization + disagreement and a introduce simple surrogate mode based on in/out group connectivity.
3. Find that at equilibrium, the networks produced by our model look 'realistic' in some aspects.

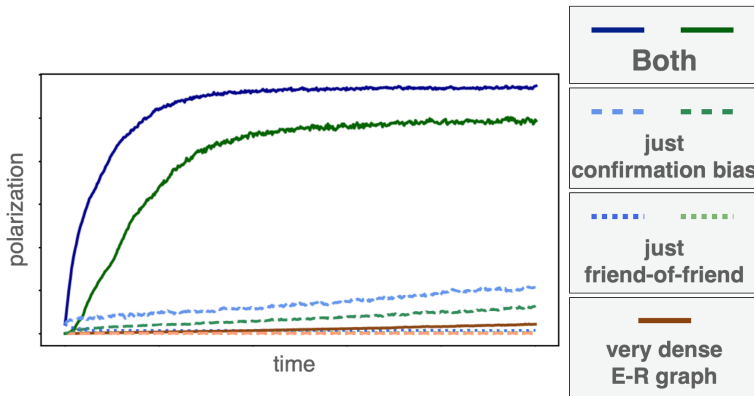
Finding 1: Confirmation Bias +  
Recommendations Drive Polarization

# How Does Polarization Arise

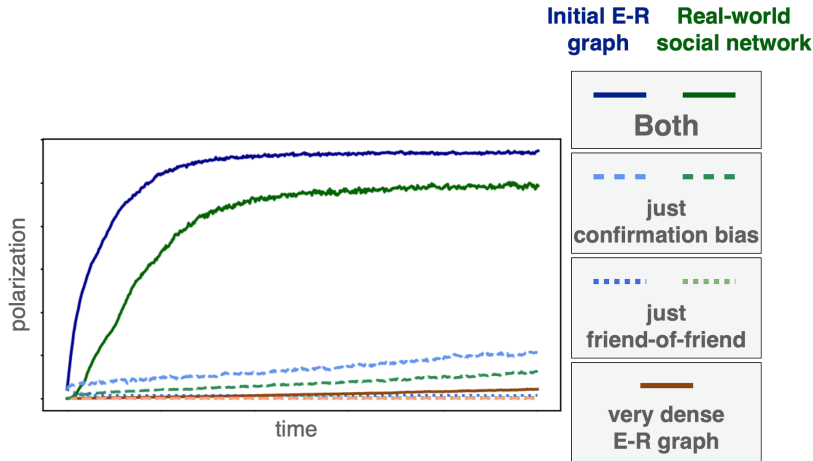
- In our model of confirmation bias, edges are removed with probability proportional to the **expressed opinion disagreement**  $(z_i - z_j)^2$ .
- If opinions are assigned at random, the expressed opinions are initially only very weakly correlated with the innate opinions.
- So a priori, it is not clear that these edge removals will lead to nodes with similar innate opinions being connected and in turn drive polarization.
- Intuitively friend-of-friend recommendations should not alone drive polarization. However they may enforce it.

# Confirmation Bias + Recommendations Drive Polarization

Initial E-R graph    Real-world social network



# Confirmation Bias + Recommendations Drive Polarization

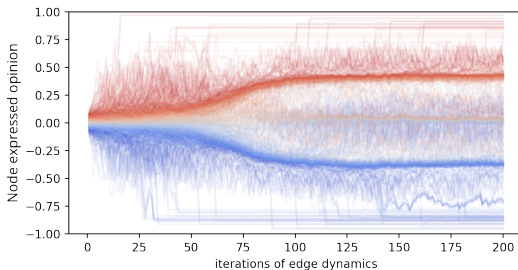
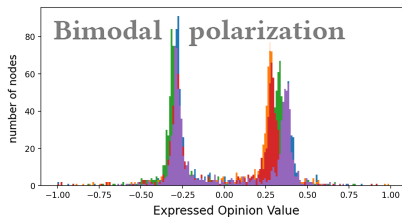
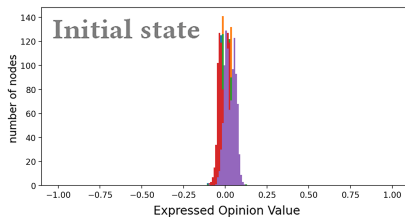


In very dense graphs, friend-of-friend recommendations are similar to random edge additions. Significant polarization does not arise.

**Open Question:** Can we explain our findings theoretically?

- E.g., show that initializing with a sparse enough E-R graph and random initial opinions, confirmation bias alone requires exponential time to reach a highly polarized state, while confirmation bias + friend-of-friend recommendations only requires polynomial time?

# Opinion Evolution Over Time



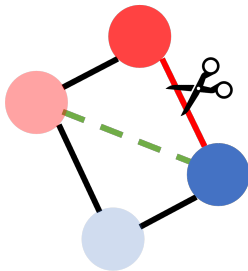
## Finding 2: Initial Theoretical Results

# Effect of Edge Swaps

**Theorem:** Consider removing edge  $e_1 = (u_1, v_1)$  and adding edge  $e_2 = (u_2, v_2)$ . Then the **polarization + disagreement**  $\mathbf{s}^T(\mathbf{L} + \mathbf{I})^{-1}\mathbf{s}$  will monotonically increase as long as

$$|\mathbf{z}(u_1) - \mathbf{z}(v_1)| \geq \alpha \cdot |\mathbf{z}(u_2) - \mathbf{z}(v_2)|$$

where  $\alpha = \frac{\sqrt{r_1}}{1+r_1} + \frac{1}{\sqrt{1+r_1}} \leq \frac{3\sqrt{3}}{4}$ .  $\alpha \approx 1$  in well-connected graphs



## Effect of Edge Swaps

**Theorem:** Consider removing edge  $e_1 = (u_1, v_1)$  and adding edge  $e_2 = (u_2, v_2)$ . Then the polarization + disagreement  $\mathbf{s}^T(\mathbf{L} + \mathbf{I})^{-1}\mathbf{s}$  will monotonically increase as long as

$$|\mathbf{z}(u_1) - \mathbf{z}(v_1)| \geq \alpha \cdot |\mathbf{z}(u_2) - \mathbf{z}(v_2)|$$

where  $\alpha = \frac{\sqrt{r_1}}{1+r_1} + \frac{1}{\sqrt{1+r_1}} \leq \frac{3\sqrt{3}}{4}$ .  $\alpha \approx 1$  in well-connected graphs

- Explains why removing edges with expressed opinion disagreement can drive up polarization.
- Proven via the linear algebraic formula for polarization + disagreement in combination with the Sherman-Morrison formula for rank-1 updates of the matrix inverse.

## Limitations of This Result

- Does not explain the importance of friend-of-friend recommendations.
- Both F-o-F and random edge additions should lead to swaps that increase polarization + disagreement. But why do F-o-F recommendations drive up polarization so much faster?

# Limitations of This Result

- Does not explain the importance of friend-of-friend recommendations.
- Both F-o-F and random edge additions should lead to swaps that increase polarization + disagreement. But why do F-o-F recommendations drive up polarization so much faster?
- Applies to polarization + disagreement but not polarization itself.

$$P + D = \mathbf{s}^T(\mathbf{L} + \mathbf{I})^{-1}\mathbf{s} \quad \text{vs.} \quad P = \mathbf{s}^T(\mathbf{L} + \mathbf{I})^{-2}\mathbf{s}$$

## Surrogate Model

Our second set of theoretical results attempts to understand the evolution of polarization and disagreement over larger time scales.

# Surrogate Model

Our second set of theoretical results attempts to understand the evolution of polarization and disagreement over larger time scales.

- For a given graph and mean 0 expressed opinion vector, we divide the nodes into two groups based on the sign of their opinions:  $V_+$  and  $V_-$ . We round each opinion to  $-1$  or  $1$ .

# Surrogate Model

Our second set of theoretical results attempts to understand the evolution of polarization and disagreement over larger time scales.

- For a given graph and mean 0 expressed opinion vector, we divide the nodes into two groups based on the sign of their opinions:  $V_+$  and  $V_-$ . We round each opinion to  $-1$  or  $1$ .
- We let  $p$  be the fraction of in-group edges that are present and  $q$  be the fraction of out group edges.

# Surrogate Model

Our second set of theoretical results attempts to understand the evolution of polarization and disagreement over larger time scales.

- For a given graph and mean 0 expressed opinion vector, we divide the nodes into two groups based on the sign of their opinions:  $V_+$  and  $V_-$ . We round each opinion to  $-1$  or  $1$ .
- We let  $p$  be the fraction of in-group edges that are present and  $q$  be the fraction of out group edges.
- We approximate the graph via an **expected stochastic block model graph**, where in-group edges are added with probability  $p$  and out-group edges are added with probability  $q$ .

# Surrogate Model

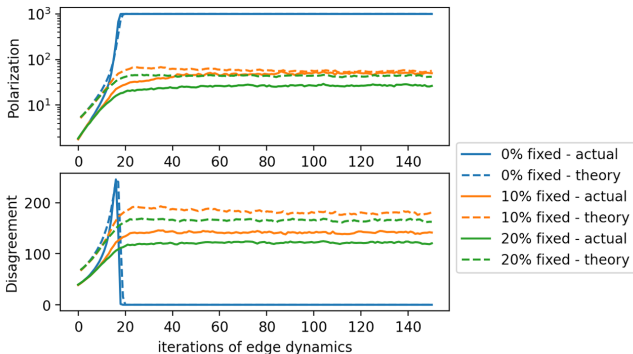
Our second set of theoretical results attempts to understand the evolution of polarization and disagreement over larger time scales.

- For a given graph and mean 0 expressed opinion vector, we divide the nodes into two groups based on the sign of their opinions:  $V_+$  and  $V_-$ . We round each opinion to  $-1$  or  $1$ .
- We let  $p$  be the fraction of in-group edges that are present and  $q$  be the fraction of out group edges.
- We approximate the graph via an **expected stochastic block model graph**, where in-group edges are added with probability  $p$  and out-group edges are added with probability  $q$ .
- Polarization and disagreement can be computed in a simple closed form for these graphs [Uthsav, Musco '20].

$$P = \frac{n}{(qn + 1)^2} \quad D = \frac{qn^2}{(qn + 1)^2}.$$

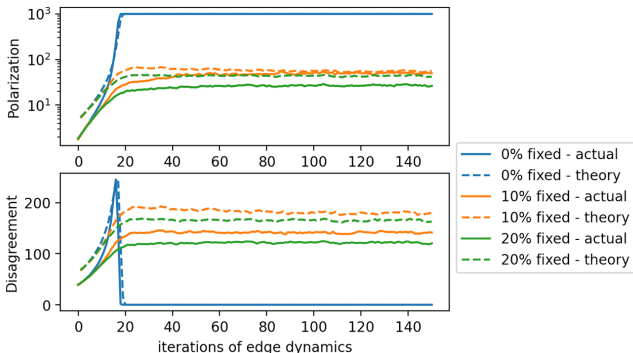
# Example Application: Understanding the effect of fixed edges

We find that this very simple approximation relatively accurately reflects the evolution of polarization and disagreement in our model, and is able to explain the effects of several different parameters.



## Example Application: Understanding the effect of fixed edges

We find that this very simple approximation relatively accurately reflects the evolution of polarization and disagreement in our model, and is able to explain the effects of several different parameters.



Demonstrates that in-group vs. out-group connection density is a key driver of polarization in our model.

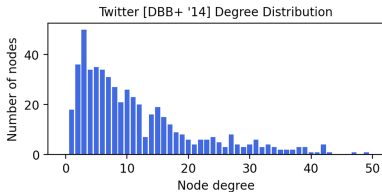
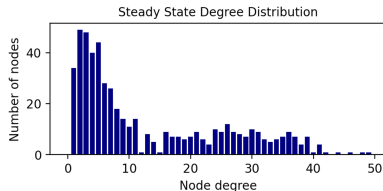
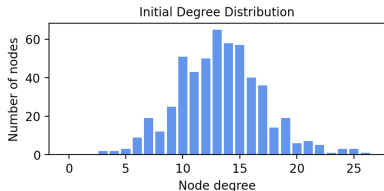
## Finding 3: Our Model Creates 'Realistic' Looking Networks

## Equilibrium Graph Structure

Generally, we find that at equilibrium, regardless of starting conditions, the networks produced by our dynamics display some hallmark features of real-world social networks.

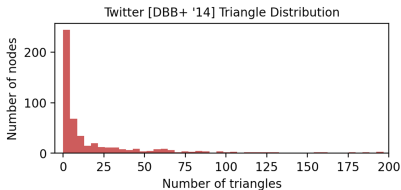
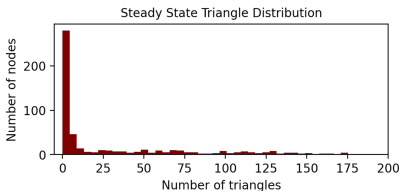
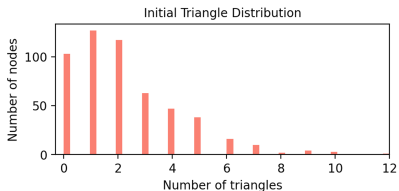
# Equilibrium Graph Structure

Generally, we find that at equilibrium, regardless of starting conditions, the networks produced by our dynamics display some hallmark features of real-world social networks.



# Equilibrium Graph Structure

Generally, we find that at equilibrium, regardless of starting conditions, the networks produced by our dynamics display some hallmark features of real-world social networks.



# Summary

- We propose a model of opinion and edge dynamics, building on the established Friedkin-Johnsen model.
- In this model, we find that both confirmation bias and friend-of-friend recommendations are required to drive polarization.
- Thanks to our model's relative simplicity, it is tractable for theoretical analysis, and we obtain preliminary theoretical results.
- We find that our model tends to create networks with realistic “social-network-like” structures.

# Open Questions

- Theoretically, being able to more fully explain aspects of our model, e.g. the role of friend-of-friend recommendations, density thresholds for rapid polarization, the mechanisms behind degree distribution shift, etc. would be very interesting.
- Beyond the Friedkin-Johnsen model, there are several other opinion dynamics models which may be more realistic. Are our findings robust to swapping out the underlying opinion model?
- It would be very interesting to further evaluate our model's realism, e.g. via comparison to real-world data sets with time-series opinion data. I.e., records of graph and opinion evolution over time.