Modeling Algorithmic Polarization in Social Media Networks Using Markov Models

Jack Betzing

University of Washington, Mathematics

Nicholas Mundy

University of Washington, Applied and Computational Mathematical Sciences

Abstract

Polarization on social media platforms has grown alongside increasing reliance on algorithmic content recommendation. Empirical studies show that online political discourse frequently fragments into ideologically aligned clusters, while algorithmic curation can amplify these divisions by selectively reinforcing user preferences. In this work, we develop a dynamical model to study how platform-driven recommendations shape opinion evolution in a connected population. Building on the Friedkin–Johnson opinion dynamics framework, we represent users as nodes in a weighted graph whose opinions update as a function of both prior beliefs and the opinions of their neighbors. We introduce a novel recommendation mechanism in which edge weights evolve according to a Beta-distributed process influenced jointly by social affinity and a platform-selected target opinion. This formulation allows us to examine how varying levels of algorithmic bias affect long-term convergence, polarization, and network connectivity. Simulation results reveal that moderate algorithmic influence reliably steers the population toward the platform's target opinion, while excessive influence collapses interaction entirely. These findings highlight a fundamental tradeoff in algorithmic environments where influence can guide collective behavior, but too much influence suppresses the very interactions required for meaningful opinion change.

Introduction

Understanding how social media platforms shape collective opinion has become a central problem in the study of online polarization. A large body of empirical work has shown that user–user interactions online are shaped not only by individual preferences, but also by the structural and algorithmic design of platforms themselves. Studies by Conover [5], Barberá [3], and others demonstrate that political communication on platforms such as Twitter often organizes into ideologically segregated communities, while Centola's experiments on complex contagions [4] show how network structure can amplify alignment within groups. At the same time, research on recommender systems which includes work by Bakshy et al. [2] and Bail [1], suggests that algorithmic curation can narrow users' content exposure, reinforcing homogeneity and contributing to diverging opinion trajectories.

Motivated by these findings, we model polarization as a dynamical process on a complete graph social network in which each user holds an opinion that evolves through weighted averaging of their neighbors' opinions. We adopt the classical Friedkin–Johnson framework [7] to capture inertia and interpersonal influence, and augment it with a platform-controlled recommendation mechanism that modulates interaction strengths. In our formulation, the platform's algorithmic bias strength determines how strongly it encourages interactions whose average opinion aligns with a chosen target value. This allows us to explore how different recommendation strategies shape long-term opinion trajectories and network connectivity.

We show that moderate algorithmic intervention can reliably guide the population toward the platform's desired target opinion, but excessive intervention collapses interaction altogether by suppressing edge formation. This results in stagnation rather than convergence, a nonlinear effect that suggests a fundamental tradeoff in real recommender-system environments. Our model provides a simple, flexible framework for analyzing how platform-level decisions interact with network structure to either amplify or dampen polarization.

Methods

Model Overview

We model the joint evolution of user opinions and interaction strengths as a time-homogeneous Markov chain defined over the state space

$$\Omega \;=\; [0,1]^N \times [0,1]^{\binom{N}{2}},$$

where the first component represents user opinions and the second represents edge weights along the complete graph. At each discrete time step t, the state consists of $z_i(t) \in [0,1]$: the opinion of user i at time t, $w_{ij}(t) \in [0,1]$: the interaction weight between users i and j at time t. Opinions are fully observed in figures and simulations, while edge weights are latent and not plotted. Two zealot nodes are fixed at opinions 0 and 1 and do not update.

Parameter Definitions

The model depends on the parameters listed in Table 1. All parameters were fixed during simulation except where otherwise noted.

Parameter	Value / Range
Number of users (N)	50
Time horizon (t)	150 steps
Zealot opinions	0 and 1
Self-weight (η)	0.5
Responsiveness rate (δ)	0.01
Beta concentration (v)	40
Algorithmic bias strength (ζ)	4 (Figs. 1–3), 0–50 (heatmap)
Target opinion (z')	0, 0.5, 1
Recommendation score ($r_{ij}(t)$)	$[0, 1]$
Social affinity	$[0, 1]$
Mean interaction level (μ)	$[0, 1]$, defined below

Table 1: Model parameters used in simulations.

Opinion Dynamics

User opinions update according to a modified Friedkin–Johnson rule:

$$z_i(t+1) = \frac{\eta z_i(t) + \sum_j w_{ij}(t)\, z_j(t)}{\eta + \sum_j w_{ij}(t)}. \tag{1}$$

The self-weight parameter $\eta > 0$ captures stubbornness where larger η slows opinion change. Only the deterministic update in (1) was used in simulation experiments.

Alternative stochastic update (not used in simulations)

For completeness, we considered a probabilistic variant:

$$\hat{z}_i(t+1) = \frac{z_i(t) + \sum_j w_{ij}(t) z_j(t)}{1 + \sum_j w_{ij}(t)}, \tag{2}$$

$$p_i(t) = \exp(-\eta\, |\hat{z}_i(t+1) - z_i(t)|)\,. \tag{3}$$

The parameter η here acts as an inverse temperature as in the Boltzmann distribution, distinct from its role as self-weight in (1). Because the deterministic model produced smoother and more stable behavior, all results from the simulation use only (1).

Edge-Weight Update

To understand how users interact in our model, it is helpful to interpret what the edge weights represent. Each weight $w_{i,j}(t)$ measures the level of interaction between users i and j at time t. A large weight corresponds to frequent or influential interactions, while a small weight reflects limited contact. If $w_{ij}(t) = 0$, we treat the pair as effectively disconnected meaning no information is exchanged between them. This interpretation also applies to interactions with the zealots, whose fixed opinions allow them to act as stable anchors within the network.

Edges update independently at each time step via a Beta distribution:

$$w_{ij}(t+1) \sim \text{Beta}(\alpha_{ij}(t), \beta_{ij}(t)).$$

The update depends only on the current opinions, not on $w_{ij}(t)$; this choice was intentional to minimize memory effects and isolate the role of algorithmic recommendations.

We Define:

$$\text{userOpinion}_{i,j}(t) = |z_i(t) - z_j(t)|,$$

$$\text{socialAffinity}_{i,j}(t) = 1 - \delta\, \text{userOpinion}_{i,j}(t),$$

$$r_{ij}(t) = \exp\left(-\zeta \left|\frac{z_i(t) + z_j(t)}{2} - z'\right|\right).$$

The expected interaction level is then

$$\mu_{ij}(t) = \delta \cdot r_{ij}(t) \cdot \text{socialAffinity}_{i,j}(t),$$

ensuring $\mu_{ij}(t) \in [0, 1]$. The Beta parameters are given by its mean and concentration as parametrized by Kruschke (2011)[6]:

$$\alpha_{ij}(t) = v\, \mu_{ij}(t), \qquad \beta_{ij}(t) = v\,(1 - \mu_{ij}(t)).$$

Higher concentration v produces more stable weights; the choice $v = 40$ provided balance between volatility and stability in simulations.

Simulation Settings

All experiments used the parameter values in Table 1. Figures 1–3 show single simulation runs for $\zeta = 4$ and $z' \in \{0, 0.5, 1\}$. For the heatmap, each point corresponds to one of 50 independent simulations, one for each value of algorithmic bias strength ζ spanning 0 to 50.

Results

Opinion Convergence Under Moderate Influence

Figures 1–3 illustrate the behavior of the model under moderate platform influence ($\zeta = 4$). In each case, the population converges toward the target opinion:

- $z' = 0$: contraction toward the left extreme,
- $z' = 0.5$: convergence toward the center,
- $z' = 1$: contraction toward the right extreme.

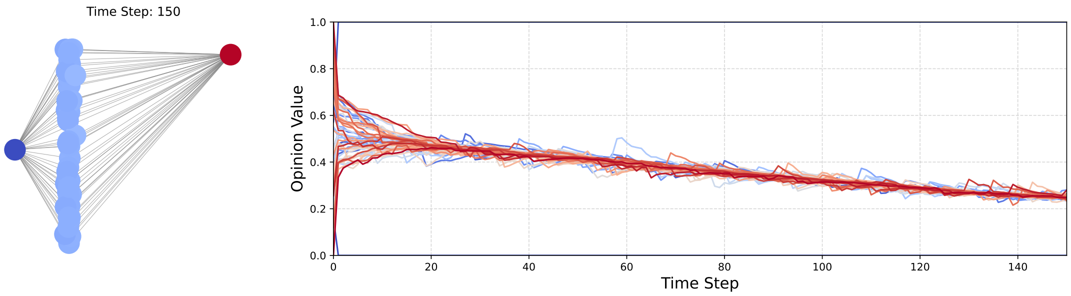

Figure 1: **Opinion trajectories targeting minimal extreme** ($\zeta = 4$, $z' = 0$, $\delta = .01$). User opinions contract toward 0, demonstrating effective directional steering.

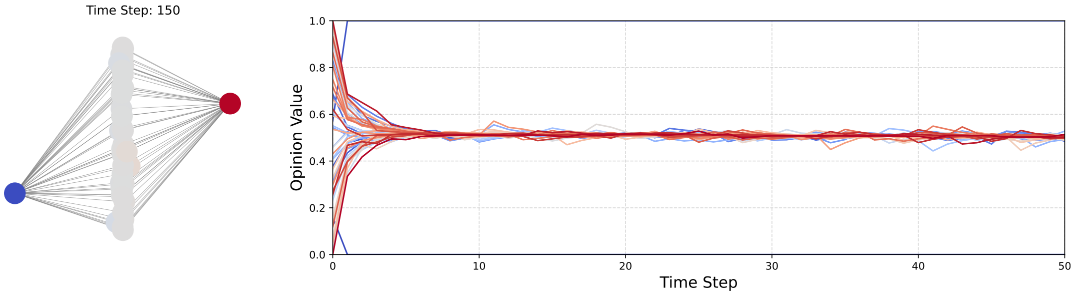

Figure 2: **Opinion evolution toward a moderate target** ($\zeta = 4$, $z' = 0.5$, $\delta = .01$). The population converges to the central target, yielding relatively balanced structure. (Plot cut at $t = 50$ as it remained constant)

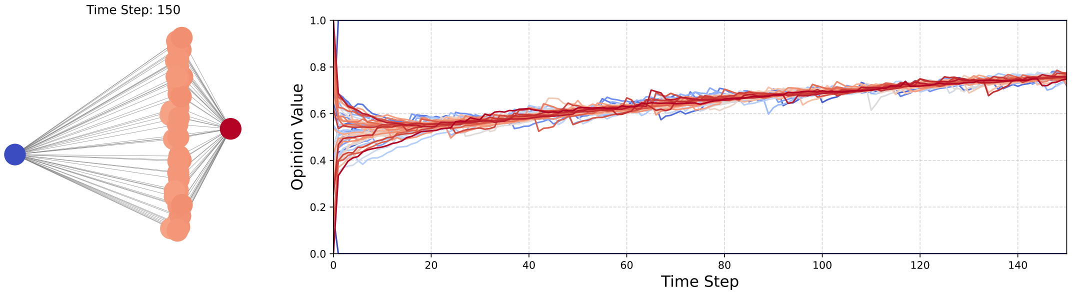

Figure 3: **Opinion trajectories under a maximal target** ($\zeta = 4$, $z' = 1$, $\delta = .01$). User opinions concentrate near the extreme, mirroring the $z' = 0$ case.

Conjecture 1: Convergence Toward the Target Opinion

Let $a(t) = \frac{1}{N}\sum_i z_i(t)$ denote the population mean opinion. From the edge-weight update,

$$\mathbb{E}[w_{ij}(t+1)] = \mu_{ij}(t) = \delta\, r_{ij}(t)\, \mathrm{socialAffinity}(i, j, t).$$

Edges between users whose mean opinion lies closer to the platform's target z' have larger expected weights, while edges far from z' weaken. Substituting these weights into the update

$$z_i(t+1) = \frac{\eta z_i(t) + \sum_j w_{ij}(t) z_j(t)}{\eta + \sum_j w_{ij}(t)},$$

users closer to z' exert proportionally greater influence. This induces a directional drift of the population mean toward z'.

Because two zealots remain fixed at 0 and 1, exact convergence cannot occur for small t, but the dynamics exhibit approach toward a limiting distribution centered at z':

$$\lim_{t \to \infty} a(t) = z',$$

provided the algorithmic bias strength ζ is sufficiently large to overcome initial dispersion but not so large as to collapse edge formation entirely.

Polarization Metrics

While Figures 2 and 3 illustrate how the platform can substantially steer user opinions, we also require a quantitative measure of how polarization manifests in the evolving network. Our initial metric was the spectral gap of the weighted graph Laplacian,

$$L = D - A,$$

where D is the diagonal matrix of weighted degrees and A contains the edge weights $w_{ij}(t)$. The spectral gap $\lambda_2(L)$ provides a measure of connectivity where larger gaps indicate more cohesive clusters, while smaller gaps correspond to fragmented or weakly integrated networks.

Figure 4 shows λ_2 across values of the target opinion z'. Although this metric captures coarse structural changes, its interpretability is limited because the network includes two zealots fixed at opinions 0 and 1. These nodes necessarily weaken connectivity near the extremes, making it hard to distinguish genuine polarization from trivial boundary effects.

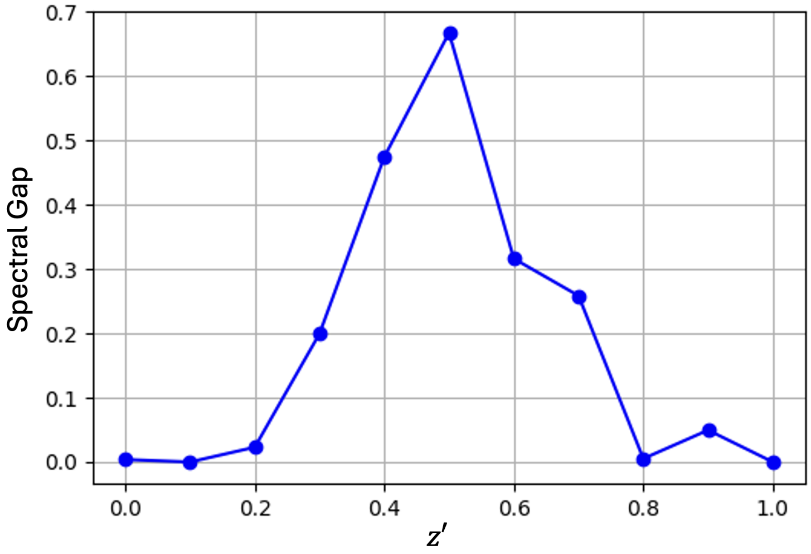

Figure 4: **Spectral gap of the weighted Laplacian** across target opinions z'. Larger gaps correspond to stronger connectivity. The influence of zealots complicates interpretation near $z' = 0$ and $z' = 1$.

To better capture how opinions align with the evolving network structure, we introduce a second metric that compares interactions among similar opinions to interactions among opposing ones. Specifically, we compute

$$\sum_{i,j} \left(z_i(t) - \frac{1}{2} \right) \left(z_j(t) - \frac{1}{2} \right) w_{ij}(t),$$

which measures how strongly edge weights concentrate around opinions above or below the midpoint $\frac{1}{2}$. Large positive values indicate strong clustering of like-minded users, while values near zero correspond to mixed or weakly aligned interactions.

Figure 5 displays this metric across the two key parameters: the target opinion z' (horizontal axis) and the algorithmic bias strength ζ (vertical axis). Darker regions indicate strong alignment between user opinions and interaction weights.

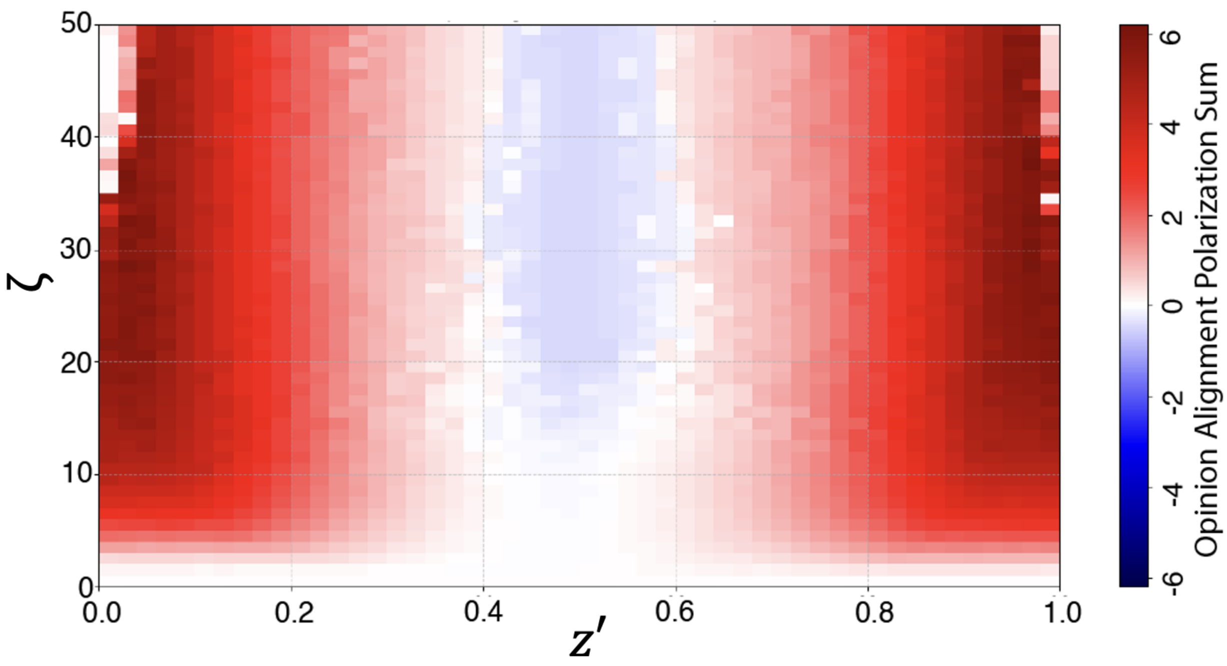

Figure 5: **Opinion–weight alignment heat map**. Each cell corresponds to a single simulation (50 total across ζ), showing how interaction clustering varies with target opinion z' and algorithmic bias strength ζ. Dark regions indicate strong polarization between similar opinions.

A clear pattern emerges. When z' lies near either extreme (0 or 1), interactions concentrate tightly among like-minded users, producing strong polarization. When the target is moderate ($z' = 0.5$), users remain more evenly mixed, leading to a weaker clustering signal. Interestingly, the dark polarized regions begin to fade as ζ increases beyond roughly 30–35, indicating that excessive algorithmic bias suppresses edge formation and disrupts coherent clustering. In this regime, the platform attempts to bias interactions so strongly that too few edges form to sustain opinion change.

Collapse Under Excessive Influence

This breakdown is visible in Figure 6, which plots the final average opinion as a function of ζ. For $0 < \zeta < 20$, stronger bias reliably pulls opinions toward the platform's target. However, once ζ exceeds approximately 20, the system undergoes a sharp transition: opinion change collapses and the population remains close to its initial distribution.

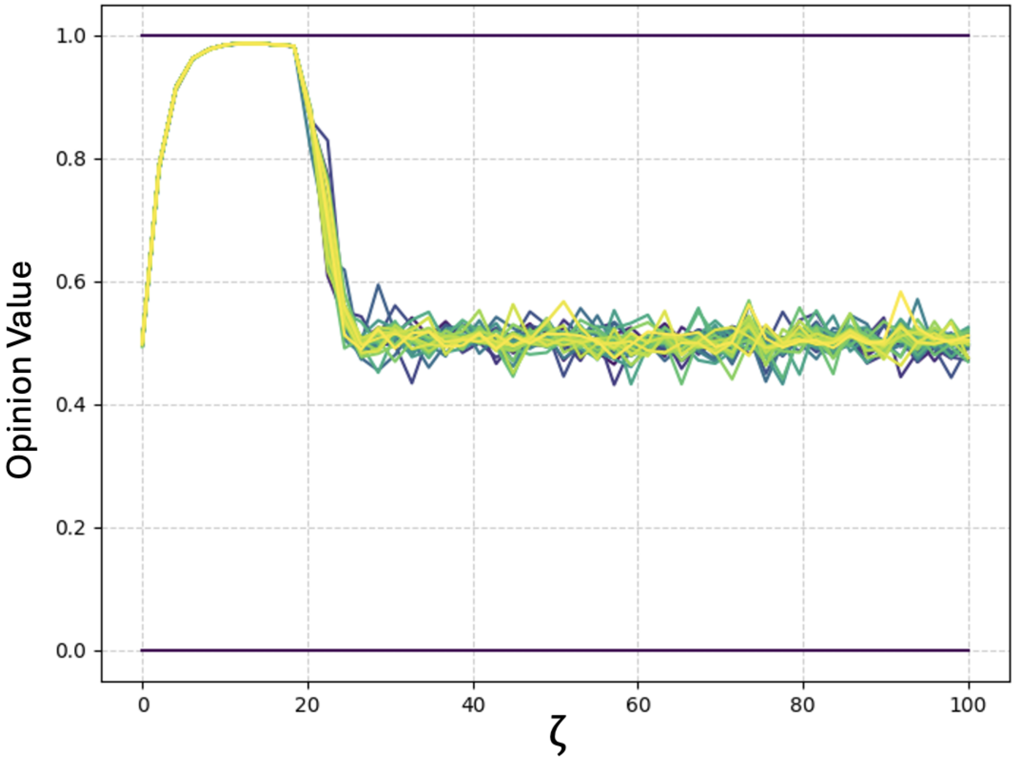

Figure 6: **Final average user opinions as a function of** ζ. A rapid falloff occurs near $\zeta \approx 20$, indicating a loss of meaningful interaction when the platform exerts excessive influence.

This phenomenon reflects the interaction between $r_{ij}(t)$ and the Beta update. As ζ becomes large, the recommendation term $r_{ij}(t)$ tends to zero for all but a narrow band of opinions extremely close to z'. Consequently, μ becomes very small and edge formation becomes negligible. With almost no interaction, the weighted averages in the Friedkin–Johnson update cease to change, and the system stagnates.

From a platform-design perspective, this effect is notable: while moderate bias can guide user behavior, excessive bias causes the platform to lose influence entirely because it suppresses the very interactions required to shift opinions.

Conjecture 2: Bound on ζ

These observations suggest the existence of a critical value ζ_0 such that beyond this threshold, the expected interaction weight satisfies

$$\mu < \varepsilon,$$

for some small $\varepsilon > 0$. In this regime, edge formation is too weak for opinions to evolve meaningfully. Thus we conjecture that

$$\zeta > \zeta_0 \quad \Longrightarrow \quad \text{opinion dynamics collapse},$$

where empirically ζ_0 lies between 20 and 35 depending on initial conditions and the self-weight parameter η. This conjecture emphasizes that algorithmic influence is effective only within a bounded range and that too little influence fails to steer the population, while too much eliminates the interactions required for any steering to occur.

Limitations

While our model captures important mechanisms of algorithmic influence, it simplifies many aspects of real social media environments. First, users are homogeneous aside from their initial opinions and two fixed zealots; real-world users vary widely in susceptibility to influence, posting activity, and content consumption. Second, edge weights evolve according to a Beta distribution controlled by platform recommendations, but the model does not incorporate endogenous user behavior such as selective exposure, content creation, or feedback between opinion change and network structure. Third, the platform's "target opinion" is represented as a single scalar value, whereas real recommender systems optimize engagement objectives rather than explicit ideological endpoints. Finally, the collapse of interaction under high bias strength arises from the specific mathematical form of our update rules; additional work is needed to determine whether similar thresholds occur under alternative modeling assumptions or real-world data.

Future Work

Future work may incorporate multiple competing platforms, heterogeneous user types, or dynamic user-generated content. Allowing several recommendation systems with different targets may reveal how users gravitate toward information sources. Introducing variable stubbornness or probabilistic content exposure could bring the model closer to empirical social media behavior.

Conclusion

Our results highlight the dual role of algorithmic recommendation systems in shaping opinion dynamics on social media. When platform intervention is moderate, the evolving interaction network reliably channels users toward the platform's chosen target opinion. This aligns with empirical findings that curated content streams can nudge collective behavior and reinforce desired ideological directions.

However, our results also reveal a sharp nonlinear effect where excessive algorithmic bias strength suppresses edge formation entirely, leaving users weakly connected and slowing opinion change to a halt. In this regime, the platform loses the ability to steer the population because it has undermined the very interactions needed to propagate influence. This collapse of connectivity provides a theoretical explanation for why overly aggressive content filtering may fail to produce the intended effects.

Taken together, our findings underscore a fundamental tension in the design of algorithmic recommendation systems. Influence can guide collective behavior, but too much influence destabilizes the network structure required for meaningful opinion change. Understanding this balance may help inform the design of healthier algorithmic environments and mitigate emergent polarization.

Use of Artificial Intelligence

AI tools were used to assist with visualization, debugging, and summarization of related literature. All modeling, derivations, and simulation design were developed by the authors.

References

[1] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, Matthew BF Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[3] Pablo Barberá. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542, 2015.

[4] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[5] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 89–96, 2011.

[6] John Kruschke. Tutorial: Doing bayesian data analysis with r and bugs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[7] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 world wide web conference*, pages 369–378, 2018.