Algorithmic Filtering, Out-Group Stereotype, and Polarization on Social Media

Jean Springsteen
Washington University in St. Louis
Saint Louis, MO, United States
jmspringsteen@wustl.edu

William Yeoh
Washington University in St. Louis
Saint Louis, MO, United States
wyeoh@wustl.edu

Dino Christenson
Washington University in St. Louis
Saint Louis, MO, United States
dinopc@wustl.edu

ABSTRACT

The introduction of social media websites touted the idea of global communication — exposing users to a worldwide audience and a diverse range of experiences, opinions, and debates. Unfortunately, studies have shown that social networks have instead contributed to growing levels of polarization in society across a wide variety of issues. Social media websites employ algorithmic filtering strategies to drive engagement, which can lead to the formation of filter bubbles and increased levels of polarization. In this paper, we introduce features of affective polarization — feelings towards one's in-group and out-group — into an opinion dynamics model. Specifically, we show that incorporating a negative out-group stereotype into the opinion dynamics model (1) affects the level of polarization present among agents in the network; (2) changes the effectiveness of algorithmic filtering strategies; and (3) is exacerbated by the presence of extremists in the network. Hence, the inclusion of an affective group mechanism in opinion dynamics modeling provides novel insights into the effects of algorithmic filtering strategies on the extremity of opinions in social networks.

KEYWORDS

Social Media; Opinion Dynamics; Algorithmic Filtering; Polarization

ACM Reference Format:
Jean Springsteen, William Yeoh, and Dino Christenson. 2024. Algorithmic Filtering, Out-Group Stereotype, and Polarization on Social Media. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

1 INTRODUCTION

Communication of information has taken many forms over time, and recently, social media networks have become a common place to discuss ideas, opinions, and events. Theoretically, social media networks and their global audience encourage the spread of diverse ideas, but recent work suggests the opposite may be occurring. While social media networks have provided a platform for globalized communication, there have also been unintended consequences. For example, they have been associated with increased polarization in society across a wide range of issues, including politics [2, 4, 12], science [34], and, more recently, healthcare [24]. Social media has the ability to expose individuals to diverse opinions and perspectives, but individuals seemingly gravitate towards opinions and posts they already agree with, unwilling to consider different opinions.

Given the severity of a polarized society — eroding democratic values, making rational discourse impossible, and potentially causing partisan violence — it is of extreme importance to identify the cause of this phenomena. A recent study concluded that social networks are likely not the root causes of political polarization, but they do exacerbate it [6]. There are several reasons social media websites may contribute to increased levels of polarization: (1) Fake news travels faster than true stories on social networks [15] and significant misinformation and polarization arise in social networks even when only a small percentage (∼15%) of individuals believe fake news to be true [3]; (2) Social networks filter the opinions that individuals are exposed to and prioritize opinions that are well aligned with the opinions of the individuals [9]. This concept of *filter bubbles* [38] is based on the idea that individuals are more likely to engage with and be exposed to content that they agree with. Thus, social network companies are incentivized to increase user engagement and drive revenue by promoting content similar to what the user likes, forming filter bubbles around individuals, creating echo chambers and promoting polarization.

Recent work has studied the effects of filter bubbles in social networks, including how a network administrator can alter how likely individuals are exposed to the opinions of others in a social network and how to mitigate filter bubbles in an influence maximization setting [31]. However, one limitation of previous work [9, 31] is that these models assume that an individual in the network can still be exposed to the opinions of *all* of its neighbors in the social network. This assumption may not hold in some social networks, where an individual is exposed to the opinions of a *subset* of its neighbors only. A prominent example is Facebook, where only updates of some friends are shown in the news feed of users.

We address this limitation by investigating how several simple algorithmic filtering strategies for deciding whose opinions an individual is exposed to can impact the polarization of opinions in social networks. Further, our model allows for both *assimilation and boomerang effects* [8, 42, 43], where the former describes how the opinion of an individual will *converge closer* to the opinion of an individual who shares a "similar enough" opinion and the latter describes how the opinion of an individual will *diverge further away* from the opinion of an individual whose opinion is too different.

We make another novel contribution by accounting for groups in studying online opinion changes. By incorporating an individual's group, we are able to better connect studies of social media algorithmic filtering to the political science literature on polarization. Partisan identity serves as the fundamental grouping in studies of

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

both ideological and affective polarization.[1] Studies of ideological polarization test if partisans have increasingly diverged on their positions on various social, civil, and moral issues, both at the elite level [23] and among the masses [1, 19]. Studies of affective polarization test divergence between partisans in their feelings towards members of their in-group, or their own party, and their out-group, or the other party [27].[2] While groups are therefore essential to understanding ideological and affective polarization, we are not aware of any studies of social media opinion dynamics that have recognized the potential for differences in how individuals interact within and across groups.

Affective polarization suggests that the messages from in-group and out-group members are likely to be received differently. While partisans generally communicate with their own [36], when they are exposed to the out-group, it is frequently of the most politically engaged and extreme in the news [30] or social media [11, 25, 41]. Thus, while there is little doubt that out-group dislike and distrust have increased over time in the US [10, 26, 29], part of this may be because the public thinks of the out-group as narrowly portrayed in those contexts [17].[3] Indeed, recent work by Druckman et al. [16] finds that individuals rely on these negative out-group stereotypes.

Our experimental results show that incorporating a negative out-group stereotype into the opinion dynamics model (1) affects the level of polarization present among agents in the network; (2) changes the effectiveness of algorithmic filtering strategies; and (3) is exacerbated by the presence of extremists in the network. Hence, the inclusion of an affective mechanism in opinion dynamics modeling provides insights into the effects of algorithmic filtering strategies on the extremity of opinions in social networks.

2 BACKGROUND

Popular opinion dynamics models explain how interactions between two agents on a network lead to their opinions becoming more similar. The foundational DeGroot model of opinion dynamics [13] has been widely studied extended to include the concept of bounded confidence [22], the stubbornness of agents to stay committed to their initial opinions [21], multi-dimensional extensions [37, 39], and the inclusion of a network administrator that can make changes to the graph [9].

While these models provide conditions for the agents in the network to reach a consensus, it suffers from the constraint that weights of interpersonal influence must be non-negative. After extensions to the DeGroot model [20], individual agents are no longer *equally* susceptible to outside influence; however, when two agents interact, their opinions can still only become more similar — a limitation when comparing this to interactions on present-day social networks. Other opinion dynamics models [7, 18, 46] have modeled how trust and skepticism within a network influence opinion dynamics. However, these models are all limited by the assumption that trust is non-negative and that agents are interacting with the true opinions of their neighbors.

We address this limitation by incorporating methods from social judgement theory. Social judgement theory states the attitude change of an individual is a judgemental process, where external stimuli and influence are judged relative to an individual's own opinion [42, 43]. In social judgement theory, there are three zones within which individuals judge external influence or attitudes. If an outside opinion is close enough to an individual's own views, this opinion is "acceptable"; whereas an opinion sufficiently far from an individual's own opinion would be "unacceptable." If the perceived opinion is neither close enough or sufficiently different, it falls in a zone of noncommitment. Chau et al. [8] expands the model by Jager and Amblard [28] by incorporating these effects (also called the assimilation and boomerang effects) from social judgement theory. We follow this framework and allow agents to judge their neighbors' opinions relative to their own.

Opinion dynamics models have been used to study rising polarization on social networks. While most agree that polarization is increasing, understanding how to reduce polarization is an active area of research. In addition, most opinion dynamics models in the literature focus solely on the extremity of opinions in a single distribution, leaving out important group distinctions. We incorporate group identity — motivated by studies of partisan polarization in political science, though the model is general enough to account for any grouping — into our model of social media opinion dynamics. When opinion dynamics models do not take group identity into account, they fail to recognize an important mechanism in opinion formation. We address this by developing a model that allows people to (1) judge the distance between their opinion and the opinion of someone they are interacting with; and (2) know the group (e.g., party) of an individual, allowing for interactions driven by negative out-group stereotypes. Further, this enables us to study how the incorporation of these mechanisms drives the extremity of opinions and is impacted by the presence of filtering strategies.

3 OPINION DYNAMICS MODEL

We now describe the opinion dynamics model we use to study the impact of interactions on social media *polarization*, by which we mean here simply the extremity of opinions in the network at large. To incorporate both social identity theory and social judgment theory into the model, we extend the model by Tsang and Larson [46] to include a trust function that adapts to the magnitude of opinion difference between two agents, using the same conventions as Chau et al. [8]. Additionally, the trust function depends on whether two agents are of the same group, as individuals respond to content based on similarity of social identity.

We model a network where n agents are embedded in a weighted, undirected graph $G = \langle V, E \rangle$. The vertices, $V = \{v_1, ..., v_n\}$ correspond to the agents, while an edge, $e_{i,j} = (v_i, v_j) \in E$ indicates that agents v_i and v_j are neighbors on the social network. If two agents are neighbors, they are able to interact with content from the other.

Our focus is on the propagation of an opinion during a set of discrete time steps, $t \in \{1, ..., T\}$. Each agent's opinion is confined to the $[0, 1]$ interval, where 0 and 1 are referred to as "extreme"

[1] While much of the political science literature, particularly in the US, has been concerned with partisan polarization — i.e., political parties as the cleavage — the grouping could be drawn on other divisions, including geography, race, class or religion, among others [33]. Though motivated by partisan polarization, our approach generalizes to other groupings.

[2] "In-group" refers to the group that the "self" agent belongs to, while "out-groups" refer to the other groups that the agent does not belong to. The origins lie in social identity theory's efforts to understand group perceptions and intergroup behaviors by positing identities based in group membership [45].

[3] The failure to recognize heterogeneity in the out-group is a long-standing finding in the study of intergroup relations [see 40].

opinions, and 0.5 represents a moderate opinion. Each agent v_i is also assigned a group $p_i \in \{0, 1\}$ that corresponds loosely to their opinion. Specifically, $p_i \sim \text{Bernoulli}(x_i)$. This group does not change even while opinions shift. At each time step, agent v_i has an opinion, denoted x_i^t, and it shares that opinion with its neighbors, $N_i = \{v_j \in V | (v_i, v_j) \in E\}$. An agent's opinion at time t is updated based on the weighted opinion of their neighbors in the previous time step:

$$x_i^t = \frac{w_{i,i}^{t-1} x_i^{t-1} + \sum_{v_j \in N_i} w_{i,j}^{t-1} x_j^{t-1}}{w_{i,i}^{t-1} + \sum_{v_j \in N_i} w_{i,j}^{t-1}} \tag{1}$$

where $w_{i,j}^{t-1}$ indicates the weight agent v_i places on the opinion of agent v_j at time $t-1$. This value also evolves over time, according to Equation 2:

$$w_{i,j}^t = \alpha w_{i,j}^{t-1} + (1-\alpha) T(x_i^t, x_j^t) \tag{2}$$

where $\alpha \in [0, 1]$ is a parameter describing how set an agent is in their own opinion, and $T(x_i^t, x_j^t)$ defines the trust between two agents. To incorporate social judgement theory [8, 42, 43], the trust function has three components, where agents behave differently according to their absolute difference in opinion, $|x_i^t - x_j^t|$. The trust function is given by:

$$T(x_i^t, x_j^t) = \begin{cases} e^{\frac{(|x_i^t - x_j^t| - d_1)^2}{-(d_1/\ln(2))^2}} - 1 & \text{if } |x_i^t - x_j^t| < d_1 \\ 0 & \text{if } d_1 \leq |x_i^t - x_j^t| \leq d_2 \\ 1 - e^{\frac{(|x_i^t - x_j^t| - d_2)^2}{-((1-d_2)/\ln(2))^2}} & \text{if } |x_i^t - x_j^t| > d_2 \end{cases} \tag{3}$$

where d_1 and d_2 are threshold parameters ($0 \leq d_1 \leq d_2 \leq 1$). Trust values are confined to the interval $[-1, 1]$, where the highest trust value is assigned to neighbors with the exact same opinion. Therefore, agents are more likely to assign high trust values to their neighbors of the same group and similar opinion and the lowest trust values to agents of the opposite group with a large absolute difference of opinion.

Group identity also influences how two agents interact. As Druckman et al. [16] find, individuals not only mis-estimate the extremity of those in the out-group, they rely on these misconceptions when making their own decisions. The trust function in Equation 3 allows agents to judge the distance between their opinion and the opinion of another agent, but it does not allow for an agent to take into account the group of the other individual.

To address this, when interacting with agents of the other group, an agent does not judge their *true opinion*. Instead, they use an *estimated opinion*. Since individuals rely on more extreme stereotypes when estimating the opinion of people in the out-group, we model this opinion as the average of the most extreme 10% of the agents of the group. This quantity can evolve over time; an agent re-estimates the extremity of an agent of the opposite group at each iteration. While Equation 3 still calculates the trust value for two agents of the same group, Equation 4 models how trust changes after interacting with agents of the other group:

$$T_{\text{opp}}(x_i^t, x_j^t) = T(x_i^t, \hat{x}_j^t) \tag{4}$$

where the difference is that it uses an estimated opinion $\hat{x}_j^t$ instead of the true opinion x_j^t.

Note that while we proposed a specific way of accounting for groups in the trust function, where the actual opinions of out-group members are unknown and are estimated to be extreme, there exists other ways to account for groups in trust functions as well. For example, actual opinions of in-group members may be assumed to be unknown as well and must be estimated. We leave the study of these variants to future work.

4 ALGORITHMIC FILTERING STRATEGIES

To test the impact algorithmic filtering strategies can have on the distribution of opinions on a network, we implement several simple filtering strategies. At each time step, the filtering strategy selects k neighbors for each agent v_i to interact with, where k is a user-defined parameter that is the same for each agent and at each time step. We use $S_i \subseteq N_i$ to denote the subset of neighbors that agent v_i interacts with, formally defined as:

$$S_i \equiv \{\hat{v_1}, \hat{v_2}, ...\hat{v_k} | \hat{v_j} \in N_i \wedge \mathcal{P}\} \tag{5}$$

where $\mathcal{P}$ corresponds to the constraints of the filtering strategies.

We propose five filtering strategies, where the first two strategies take into account the agents' attributes — requiring an agent to interact with neighbors of the same group (Section 4.1) or neighbors with similar opinions (Section 4.2). The next three strategies are common baselines used in the literature — requiring an agent to interact with neighbors with least extreme opinions (Section 4.3), neighbors who are the most popular (Section 4.4), or neighbors who are randomly chosen (Section 4.5).

4.1 In-Group Neighbors

In an attempt to reduce polarization within the network, the first filtering strategy prioritizes interactions between agents of the same group. On social media networks, cross-party ties exist. However, as Facebook reported in 2015, their algorithms lead individuals to experience slightly less cross-cutting content [5]. The same report notes that people are more likely to interact with and consume information with which they already agree. To drive engagement and prevent individuals from believing they are interacting with extremists of the other group, this strategy forces interactions between agents of the same group and prevents agents from different groups from interacting. If an agent did not have neighbors of the same group, they interacted with random neighbors.

$$\mathcal{P} \equiv \forall v \in N_i \setminus \{\hat{v}_1, \hat{v}_2, ..., \hat{v}_k\} : p_i = p_j \tag{6}$$

4.2 Most Similar Neighbors

Another intuitive approach to reducing opinion polarization on a network is to allow individuals to interact with agents similar in opinion, reducing the number of interactions with less similar agents. While being in the same group is one measure of similarity, this filtering strategy focuses on neighbors with similar opinions, meaning an agent can interact with an agent from the other group. This leads to more interactions in the assimilation zone and fewer in the boomerang zone. More formally,

$$\mathcal{P} \equiv \forall x \in X_i \setminus \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_k\} : |x - x_i| \geq |\hat{x}_j - x_i| \tag{7}$$

where $\hat{x}_j$ is the opinion of agent v_j and X_i is the set of opinions corresponding to neighboring agents $v \in N_i$.

4.3 Least Polar Neighbors

In trying to reduce polarization of opinions on the network, an obvious filtering strategy is to only allow individuals to interact with their least extreme, or most moderate, neighbors, where polarity is measured here as the absolute distance from 0.5. This is formalized as the constraint:

$$\mathcal{P} \equiv \forall x \in X_i \setminus \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_k\} : |x - 0.5| \geq |\hat{x}_j - 0.5| \quad (8)$$

where neighboring agent v_j has opinion $\hat{x}_j$, and X_i is the set of opinions corresponding to the neighboring agents, N_i.

4.4 Most Popular Neighbors

Social networks have provided a way for individuals to follow and interact with influential people or organizations, often with a higher number of followers than a typical individual. For this filtering strategy, an agent is shown their k neighbors with the highest degree. Letting $\deg(v)$ denote the degree of agent v in the graph, we model this constraint as:

$$\mathcal{P} \equiv \forall v \in N_i \setminus \{\hat{v}_1, \hat{v}_2, ..., \hat{v}_k\} : \deg(v) \leq \deg(\hat{v}_j) \quad (9)$$

4.5 Random Neighbors

Finally, as a baseline strategy, k neighbors are chosen randomly for an agent v_i. For this baseline strategy, there are no additional constraints beyond being in the neighbor set.

$$\mathcal{P} \equiv true \quad (10)$$

Clearly, these are relatively simple filtering strategies, especially in comparison to the filtering and ranking algorithms used by actual social media companies. However, as simple as they are, their impact on polarization is pronounced. Therefore, if algorithms do not explicitly address the impact they have on polarization, it is likely they are further exacerbating it.

5 EXPERIMENTAL EVALUATION

To determine the effects that algorithmic filtering strategies can have on polarization, we run a set of simulation experiments [44], varying whether agents take the group of another agent into account and varying the extremity of agents in the network. For each set of parameters, we run 25 trials, averaging quantities of interest over the trials. In each run, the experiment terminates when all opinions have changed by no more than a small value $\delta = 0.001$ or the experiment reaches the maximum number of iterations $i_{\max} = 500$, though $i_{\max}$ was rarely reached.

For each experiment, we first generate a graph G with $n = 200$ agents (nodes), where each agent v_i has opinion x_i and group p_i. We follow the literature [46] and use the following empirical setup:

- We define G to be a Barabási-Albert random graph with homophily, which allows one to model the tendency of individuals to self-select similar neighbors on the network.
- In each iteration, individual agents are allowed to interact with only a subset of its neighbors defined by the filtering strategy.
- We consider "extremists," which are agents that have fixed opinions at one extreme (0 or 1) and do not update their opinions as a result of interacting with other agents. In half of the experiments, there are no extremists present, and in the other half of experiments, 10% of agents are randomly assigned to be 0-extremists and 10% are assigned to be 1-extremists.

We vary the d_1 and d_2 parameters used in the trust functions (Equations 3 and 4) to analyze how different assimilation and boomerang zones affect opinion dynamics. For each experiment, we use three values for d_1 and d_2, where $d_1 \in \{0.3, 0.5, 0.7\}$ and $d_2 \in \{0.5, 0.7, 0.9\}$, creating 8 combinations (since $d_1 \leq d_2$).

For each of the five filtering strategies, we compare four results — whether or not agents reacted to a neighbor's group or used their true opinion and whether or not the network contains extremists. The polarity results are presented without the opinions of agents who were designated as extremists. We use average distance from 0.5 as a measure of polarization, following the convention by Tsang and Larson [46], but we also investigate the variance of opinions in the final opinion distribution, incorporating another measure of polarity used in the literature [9, 32, 35].

5.1 In-Group Neighbors

Individuals seek out information they already agree with and often interact with other individuals and content similar to their own beliefs. First, we assume similarity is based on group. To understand how opinions on the network change over time, Figure 1 shows the distribution on the network at five steps, as a fraction of the total number of iterations required for convergence, t^*, for the in-group neighbors filtering strategy. Since each trial does not take the same number of iterations, we combine trials by the percentage of the way to convergence. The first column of Figure 1 shows the distribution of *initial* opinions and the final column shows the distribution of *converged* opinions. While the number of iterations across trials is not the same, this provides a way to understand how opinion distributions change over the course of an experiment.

The first row of Figure 1 shows the distribution of opinions over time for the in-group filtering strategy when agents use estimated out-group members' opinions, and the second row shows the distribution of opinions when agents use the true opinions of out-group members, both without extremists present in the network. The stacked histograms show the distribution by agent group, where agents with $p_i = 0$ shown in blue and agents with $p_i = 1$ shown in purple. While both experiments begin with a uniform distribution of opinions, by time $0.25t^*$, we already see a difference in opinion distribution between the two experiments. In the top row, the case where agents are estimating opinion, there are already a significant number of agents at the extremes; whereas, in the second row, opinions have become more moderate overall. *Even though the majority of interactions are taking place among agents in the same group, the use of estimated opinions still creates more extreme agents.*

This trend continues as the experiments progress; by the time the experiments have reached the halfway mark, the use of estimated opinions leads to more extreme opinions, while the second row displays more movement towards the moderate opinion. The movement to the extremes is largely along group lines. By the time these experiments converge, the impact of group dynamics is clear. First, note that while this filtering strategy prioritizes interactions with in-group neighbors, it is still possible for an agent to interact with out-group neighbors, specifically when there is an insufficient

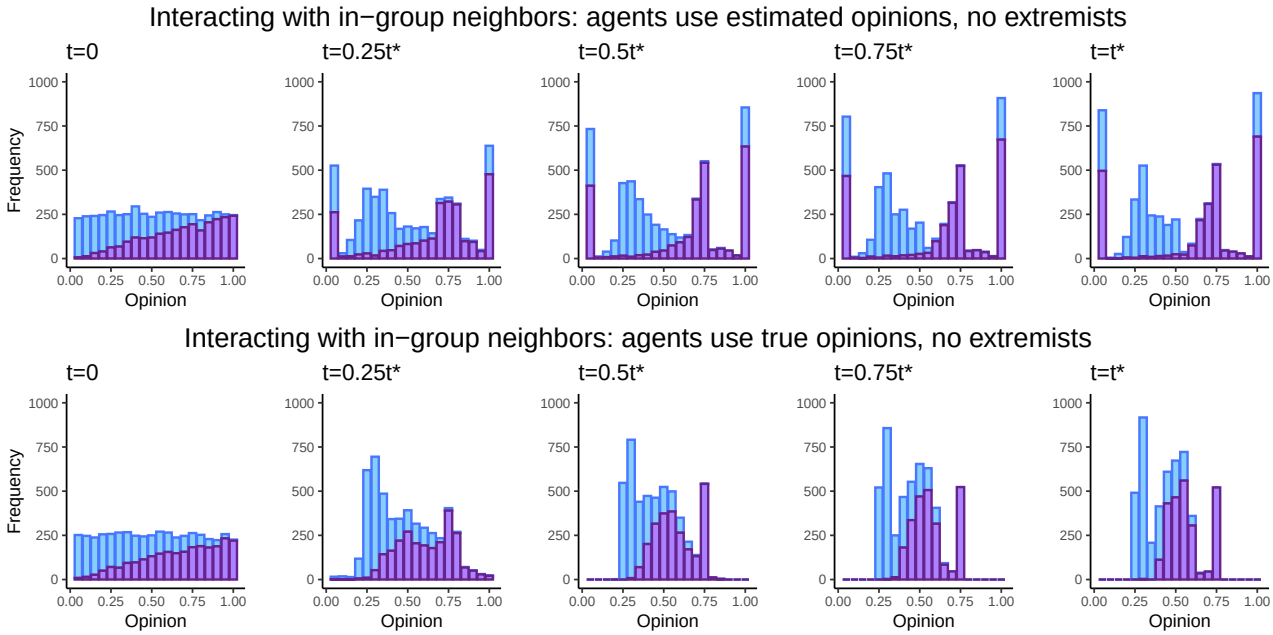

Figure 1: Distribution of opinions over time, where t^* denotes the convergence time, of moderates with the *in-group neighbors* filtering strategy without the presence of extremists; agents use estimated opinions of out-group members (top row) and use true opinion (bottom row). The stacked histograms show agents with $p_i = 0$ in blue and agents with $p_i = 1$ in purple.

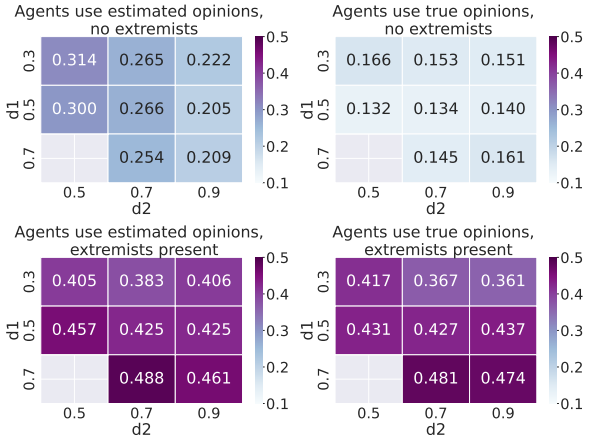

Figure 2: Average polarization of moderates with the *in-group neighbors* filtering strategy without the presence of extremists (top row) and with extremists (bottom row); agents use estimated opinions of out-group members (left column) and use true opinion (right column).

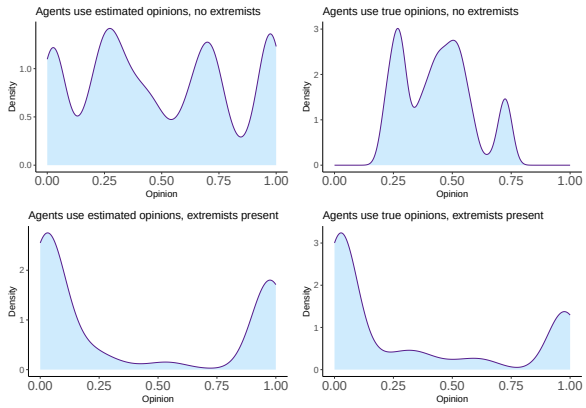

Figure 3: Distribution of final opinions with the *in-group neighbors* filtering strategy without the presence of extremists (top row) and with extremists (bottom row); agents use estimated opinions of out-group members (left column) and use true opinion (right column).

number of in-group neighbors. Even though such out-group interactions are relatively rare, they can lead to a high proportion of agents with extreme opinions when estimated opinions of out-group members are used instead of true opinions. The impact goes beyond extremists, as even the opinions of moderate agents are more polarized when agents use group identity. This is evidenced by a clear bimodal distribution in the top, right panel in Figure 1.

While there are still many agents with non-extreme opinions, they are still clustered into two distinct groups, with peaks around 0.25 and 0.75. When agents use the true opinion of their neighbors, not only are agents with extreme opinions absent, but there is no clear bimodal distribution, indicating a less polarized distribution. *Even in the case where the filtering strategy attempts to mitigate negative group dynamics, a small number of out-group interactions leads to an increase in polarity.* Due to space constraints, we do not present

analysis over time for the remaining filtering strategies; in general, for each filtering strategy, the distribution moves slowly from the uniform distribution to the final distributions presented.

We investigate this strategy further by looking at results across more parameters. Figure 2 shows average distance from 0.5 (polarity) across threshold parameters. The main difference in polarity is a result of the presence of extremists. This aligns with our expectations — with extremists present, an agent is likely to interact with in-group extremists, increasing polarity. This should intensify as the threshold parameter d_1 increases, and agents become more susceptible to in-group extremist influence. However, we also see a difference in the top two panels in Figure 2, where the difference in the experiment was whether agents estimate the extremity of out-group neighbors. This corresponds to the discussion of Figure 1, where the difference comes from the agents who did not have neighbors of the same group. If an agent did not have a sufficient number of in-group neighbors for filtering, they interacted randomly with neighbors of the out-group, causing higher levels of polarity, especially when the d_1 threshold was relatively low.

This filtering strategy was not effective in reducing polarization in the presence of extremists, as evidenced in the bottom row of Figure 2. The average distance to 0.5 was greater than 0.35 for all combinations of threshold parameters, indicating that under this strategy, most agents who began with a moderate opinion ended up at the extreme of their corresponding group. We explore this further by looking at the opinion distributions in Figure 3. In each instance of this filtering strategy, the final opinion distribution is characterized by two distinct peaks of opinions. The final opinions are extremely polarized in the case when extremists are present in the network and agents overestimate the extremity of out-group neighbors. In this instance, there are no remaining moderates — every agent in the network has moved toward an extreme.

These results show that including a mechanism for agents to mis-estimate the opinion of their neighbors of the other group affects the final distribution of opinions. In general, when agents hold negative stereotypes about their out-group neighbors, there are increased levels of polarity, and this is only exacerbated by the presence of extremists. While there are certainly other factors that influence the formation of an individual's opinion, we emphasize the importance of including a mechanism where an individual's *perceptions* of their neighbors influence their interactions.

5.2 Most Similar Neighbors

In this filtering strategy, we expect there to be little variation in the results across threshold parameters since the filtering strategy prioritizes neighbors whose opinion is already close to that of the agent. Since agents are more likely to be neighbors with agents who have similar opinions and are thus more likely to be of the same group, we do not expect a polarized distribution. Figure 4 shows the average polarization of agents after opinions converge under this strategy. As expected, there is no significant difference based on threshold parameters, the presence of extremists, or whether agents are estimating the opinion of neighbors of the other group.

In comparing the values in Figure 4 to the levels of polarization from other strategies, the values seem high, but not entirely extreme. Upon further investigation, the variance in these opinions

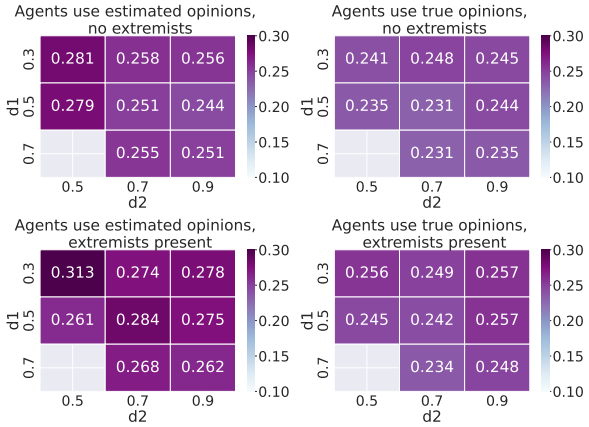

Figure 4: Average polarization of moderates with the *most similar neighbors* filtering strategy.

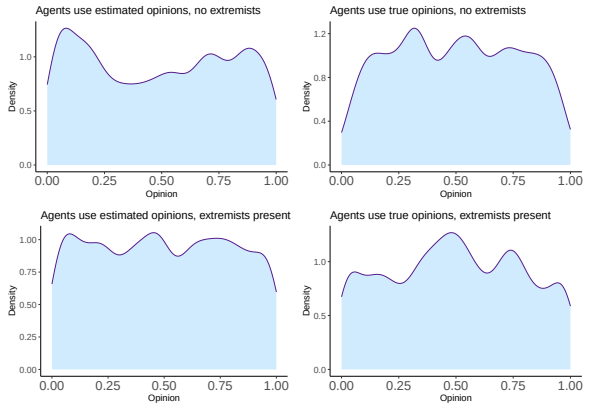

Figure 5: Distribution of final opinions with the *most similar neighbors* filtering strategy.

was nearly always between 0.25 and 0.3, indicating that agents did not reach a consensus. Figure 5 shows the final distribution of opinions for each of the four experiments, across all 25 trials for parameters $d_1 = 0.5$ and $d_2 = 0.5$. The average distance from the moderate opinion is in this range because the opinion distribution remains relatively uniform. There is no significant movement towards the extremes, even in the presence of extremists. According to the definition of polarization by DiMaggio et al. [14], a bimodal distribution with relatively high levels of variance indicates a polarized opinion distribution. However, this is not the case with the distributions in Figure 5, even though the average distance from the moderate opinion may indicate polarization. Therefore, when prioritizing *similarity* in algorithmic filtering, it is important to make the distinction between opinion similarity and group similarity, specifically in the presence of extremists.

5.3 Least Polar Neighbors

Unsurprisingly, prioritizing neighbors with the most moderate opinions results in the lowest levels of polarity of agents in the network,

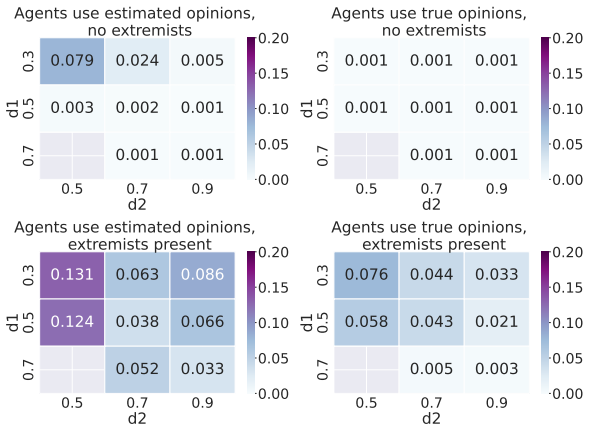

Figure 6: Average polarization of moderates with the *least polar neighbors* filtering strategy.

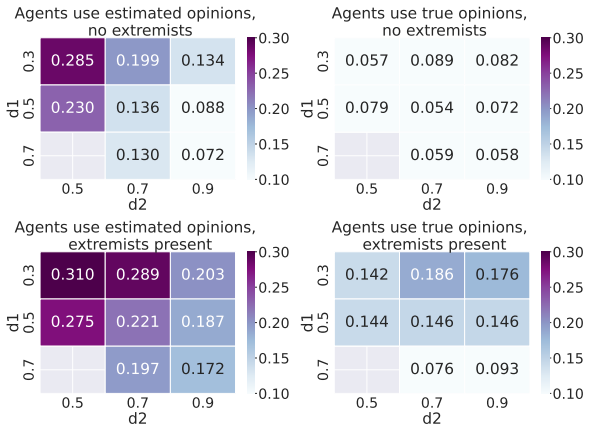

Figure 7: Average polarization of moderates with the *most popular neighbors* filtering strategy.

as shown in Figure 6. The highest levels of polarity come from the experiments with the smallest threshold parameter corresponding to assimilation zones. When d_1 is small, agents may still boomerang away from moderate opinions. However, agents still overwhelmingly assimilate when interacting with their moderate neighbors.

Even in the case where agents estimate the opinion of moderates to be extreme because they are of the other group, average polarization remains relatively low. In the bottom, left panel in Figure 6, we see the highest levels of polarization for this strategy, corresponding to a network with extremists where agents estimate the opinion of the other group. However, both the average distance to 0.5 and the variance in the agents' opinions (~0.08) are relatively small when compared to other filtering strategies. Additionally, in these experiments, originally moderate agents (the 80% who were not extremists) typically reach a consensus, almost exactly at 0.5. Therefore, in our model, it is possible for agents to reach a consensus, even when over-estimating the extremism of agents of the other group, but it requires prioritizing the least polar opinions.

We omit the distribution of final opinions for this strategy because, in each case, all agents (except the initial extremists) end the experiment with opinions close to 0.5. There is no difference in the final distribution when varying the presence of group dynamics.

5.4 Most Popular Neighbors

This filtering strategy is unique because an agent will interact with the same neighbors at every iteration. Since the network structure does not change over the course of the experiment, an agent's most popular neighbors will not change. Therefore, we expect the presence of extremists and the threshold parameters to influence the polarity of the agents in the network. Additionally, the nodes with the highest degree in the network may be the most extreme, so the exact network structure plays a significant role in this filtering strategy. The results of this experiment are presented in Figure 7.

As expected, polarity varies based on the threshold parameters and the presence of extremists. When agents use the true opinions of their neighbors and there are no extremists in the network, there is very little opinion polarization. In addition, the variance

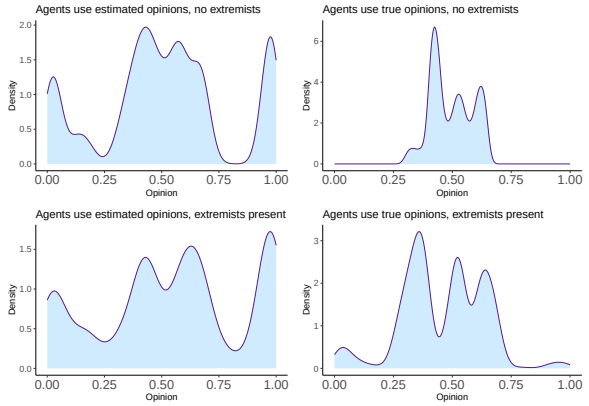

Figure 8: Distribution of final opinions with the *most popular neighbors* filtering strategy.

of opinions in these experiments is quite small (< 0.008). In this instance, agents tend to reach a consensus at the opinion of the agent with the highest degree. This can be at any value on the spectrum, but over the course of 25 trials, the average highest-degree node has opinion 0.5. In general, this filtering strategy leads to consensus at the opinion of the node with the highest degree.

In the top, left panel of Figure 7, we see the threshold parameters influence the effectiveness of this strategy. When d_1 and d_2 are small, there is only a small zone of acceptable opinions for each agent. If the agents with the highest degree have opinions that fall outside of this zone, an agent will not assimilate towards that opinion, and since d_2 is also small, an agent is more likely to boomerang away from the opinion of the most popular agents. This results in higher levels of polarization than in experiments with larger d_1 thresholds. Additionally, we see higher levels of polarization with this strategy when agents overestimate the extremity of their neighbors. This strategy is most affected by overestimation of opinion because of the repeated interaction with the same neighbors. When the most

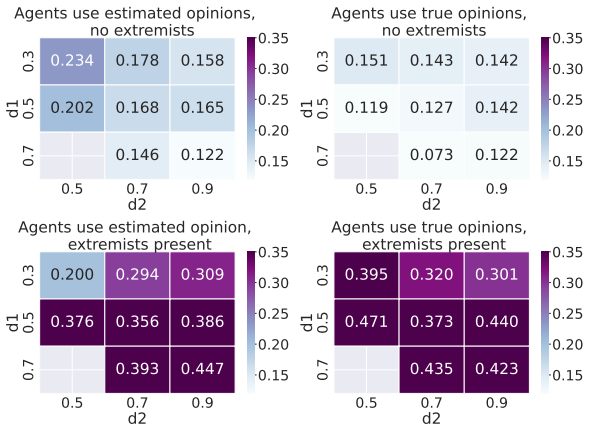

Figure 9: Average polarization of moderates with the *random neighbors* filtering strategy.

popular neighbors are of the other group, agents repeatedly interact with them, causing an overestimation of their extremity.

Figure 8 shows the distribution of opinions at the end of each experiment for all 25 trials, for $d_1 = 0.5$ and $d_2 = 0.5$ (corresponding to the results in the center square of Figure 7). The main difference in these results stems from how an agent interacts with out-group neighbors. In the two panels in the first column of Figure 8, agents estimate the opinion of out-group neighbors, and in these two experiments, there is a high proportion of agents with extreme opinions, even in the top, right panel, where there are no initial extremists. This is a result of popular neighbors not necessarily belonging to the same group as the agent. If an agent is constantly interacting with and over-estimating the extremity of out-group neighbors, they move towards the extreme of their own group. This strategy emphasizes the impact group dynamics can have on opinion dynamics; if agents constantly interact with out-group neighbors and over-estimate their extremity, the final distribution will be more polarized than if agents disregard group stereotypes.

5.5 Random Neighbors

The final strategy we implement allows agents to interact with their neighbors randomly. By construction of the network, it is still more likely that an agent will interact with someone similar, both in terms of opinion and group, due to homophily. Without the presence of a filtering strategy, we cannot rule out consistent interaction with extremists or agents of the other group. Figure 9 shows the polarization on the network after the experiment has terminated, for each of the d_1 and d_2 combinations. The top, left panel shows the polarization of moderates on a network without extremists where agents estimate the opinion of neighbors of the other group. In general, the agents on these networks reached a consensus, as there was very little variance in the 200 opinions. While the average opinion over the 25 trials was roughly 0.5 for each combination of threshold parameters, closer inspection of the results indicates that opinions rarely settled around 0.5. Opinions more frequently converged to slightly polarized values, such as 0.35 and 0.65, meaning opinion formation was largely dependent on initial graph structure.

The top, right panel of Figure 9 shows the average polarity from networks free of extremists where agents use the true opinion of their neighbors to form their opinions. In this case, agents always reached a consensus, and there was little variance (< 0.006) in the opinions for all trials in all combinations of threshold parameters. The average polarization reflects the fact that consensus rarely occurred at the moderate opinion, 0.5, and was more likely to occur at a slightly polarized opinion. The random interactions that did occur significantly influenced the final consensus — meaning the filtering strategies should greatly influence the final opinion distribution. The bottom row of Figure 9 shows the polarization of the agents in the same experiments when there are extremists present in the network. Unsurprisingly, the average polarization on the network increases as a result. There is little difference between agents interacting with the true opinion of their neighbors as opposed to an estimated opinion, likely due to the fact that agents interact with extremists over the course of the experiment. The final opinion distributions of these experiments provide little insight beyond what Figure 9 shows. The main difference is the presence of extremists, which leads to a bimodal distribution at the extremes, whereas when there are no extremists, the two poles are closer to the moderate opinion. These facts are reflected in Figure 9.

6 CONCLUSIONS

Understanding the formation of opinions on a network is a complicated question. Previous opinion dynamics models have focused on how individuals change their opinion after interacting with others, but each model relies on an individual interacting with their neighbor's actual opinion. We address this by allowing individuals to overestimate the extremity of out-group neighbors, a phenomenon with support in the study of partisan polarization [e.g., 16]. We find that this leads to generally higher levels of polarity, especially in networks with extremists. Further, the effectiveness of filtering strategies is impacted by the presence of the group mechanism.

It is important to emphasize the role group identity plays in opinion formation. If individuals do not interact with someone's true opinion, and they assume they are interacting with an out-group member of more extreme opinions, filtering strategies can contribute to increasing levels of polarity, and the presence of extremists in the network only contributes to agents' negative stereotypes about the other group. Further, we cannot accurately study opinion dynamics if we do not incorporate the ways in which people actually interact. Given the evidence that individuals primarily interact with people of their in-group, and use negative stereotypes when interacting with people of the out-group, opinion dynamics models should take this into account, and we have shown that the presence of such a mechanism may lead to higher levels of polarization.

ACKNOWLEDGMENTS

This work was supported in part by a seed grant from the Transdisciplinary Institute in Applied Data Sciences (TRIADS) at Washington University in St. Louis. We would like to also thank Jenna Pedersen and Yevgeniy Vorobeychik for helpful discussions.

REFERENCES

[1] Alan I Abramowitz and Kyle L Saunders. 2008. Is polarization a myth? *The Journal of Politics* 70, 2 (2008), 542–555.

[2] Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the International Workshop on Link Discovery*. 36–43.

[3] Marina Azzimonti and Marcos Fernandes. 2023. Social media networks, fake news, and polarization. *European Journal of Political Economy* 76 (2023), 102256.

[4] Drake Baer. 2016. The 'filter bubble' explains why Trump won and you didn't see it coming. http://nymag.com/scienceofus/2016/11/how-facebook-and-the-filter-bubble-pushed-trump-to-victory.html

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

[6] Paul Barrett, Justin Hendrix, and Grant Sims. 2021. How tech platforms fuel U.S. political polarization and what government can do about it. https://www.brookings.edu/blog/techtank/2021/09/27/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/

[7] Arthur Carvalho and Kate Larson. 2012. A consensual linear opinion pool. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 2518–2524.

[8] HF Chau, CY Wong, FK Chow, and Chi-Hang Fred Fung. 2014. Social judgment theory based model on opinion formation, polarization and evolution. *Physica A: Statistical Mechanics and its Applications* 415 (2014), 133–140.

[9] Uthsav Chitra and Christopher Musco. 2020. *Analyzing the impact of filter bubbles on social network polarization*. Association for Computing Machinery, 115–123.

[10] Dino P Christenson and Herbert F Weisberg. 2019. Bad characters or just more polarization? The rise of extremely negative feelings for presidential candidates. *Electoral Studies* 61 (2019), 102032.

[11] Nate Cohn and Kevin Quealy. 2019. The Democratic electorate on Twitter is not the actual Democratic electorate. *The New York Times* (2019).

[12] Pranav Dandekar, Ashish Goel, and David T Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5791–5796.

[13] Morris H DeGroot. 1974. Reaching a consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121.

[14] Paul DiMaggio, John Evans, and Bethany Bryson. 1996. Have American's social attitudes become more polarized? *Amer. J. Sociology* 102, 3 (1996), 690–755.

[15] Peter Dizikes. 2018. Study: On Twitter, false news travels faster than true stories. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[16] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2022. (Mis) estimating affective polarization. *The Journal of Politics* 84, 2 (2022), 1106–1117.

[17] James N Druckman and Matthew S Levendusky. 2019. What do we measure when we measure affective polarization? *Public Opinion Quarterly* 83, 1 (2019), 114–122.

[18] Hui Fang, Jie Zhang, and Nadia M Thalmann. 2013. A trust model stemmed from the diffusion theory for opinion evaluation. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*. 805–812.

[19] Morris P Fiorina, Samuel J Abrams, and Jeremy C Pope. 2005. *Culture war? The myth of polarized America*. Longman.

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

[21] Noah E Friedkin and Eugene C Johnsen. 1999. Influence networks and opinion change. *Advances in Group Processes* 16, 1 (1999), 1–29.

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

[23] Marc J Hetherington and Suzanne Globetti. 2002. Political trust and racial policy preferences. *American Journal of Political Science* (2002), 253–275.

[24] Harald Holone. 2016. The filter bubble and its effect on online personal health information. *Croatian Medical Journal* 57, 3 (2016), 298–301.

[25] Adam Hughes. 2019. A small group of prolific users account for a majority of political tweets sent by U.S. adults. https://pewrsr.ch/3a31iDK

[26] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* 22 (2019), 129–146.

[27] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. 2012. Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly* 76, 3 (2012), 405–431.

[28] Wander Jager and Frédéric Amblard. 2005. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory* 10 (2005), 295–303.

[29] Samara Klar, Yanna Krupnikov, and John Barry Ryan. 2018. Affective polarization or partisan disdain? Untangling a dislike for the opposing party from a dislike of partisanship. *Public Opinion Quarterly* 82, 2 (2018), 379–390.

[30] Matthew Levendusky and Neil Malhotra. 2016. Does media coverage of partisan polarization affect political attitudes? *Political Communication* 33, 2 (2016), 283–301.

[31] Antonis Matakos, Cigdem Aslay, Esther Galbrun, and Aristides Gionis. 2020. Maximizing the diversity of exposure in a social network. *IEEE Transactions on Knowledge and Data Engineering* 34, 9 (2020), 4357–4370.

[32] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. 2017. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery* 31 (2017), 1480–1505.

[33] Jennifer McCoy and Murat Somer. 2019. Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The Annals of the American Academy of Political and Social Science* 681, 1 (2019), 234–271.

[34] Aaron M McCright and Riley E Dunlap. 2011. The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly* 52, 2 (2011), 155–194.

[35] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. 2018. Minimizing polarization and disagreement in social networks. In *Proceedings of the International World Wide Web Conference*. 369–378.

[36] Diana C Mutz. 2006. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

[37] Angelia Nedić and Behrouz Touri. 2012. Multi-dimensional Hegselmann-Krause dynamics. In *IEEE Conference on Decision and Control*. 68–73.

[38] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[39] Sergey E Parsegov, Anton V Proskurnikov, Roberto Tempo, and Noah E Friedkin. 2016. Novel multidimensional models of opinion dynamics in social networks. *IEEE Trans. Automat. Control* 62, 5 (2016), 2270–2285.

[40] George A Quattrone. 1986. On the perception of a group's variability. *Psychology of Intergroup Relations* 2 (1986), 25–48.

[41] Jaime E Settle. 2018. *Frenemies: How social media polarizes America*. Cambridge University Press.

[42] Carolyn W Sherif, Muzafer Sherif, and Roger Ellis Nebergall. 1965. *Attitude and attitude change: The social judgment-involvement approach*. Saunders Philadelphia.

[43] Muzafer Sherif and Carl I Hovland. 1961. *Social judgment: Assimilation and contrast effects in communication and attitude change*. Yale University Press.

[44] Jean Springsteen, William Yeoh, and Dino Christenson. 2024. Supplementary material: Algorithmic filtering, out-group stereotype, and polarization on social media. https://github.com/YODA-Lab/Algorithmic-Filtering-Out-Group-Stereotype-and-Polarization-on-Social-Media

[45] Henri Tajfel and John Turner. 1979. An integrative theory of intergroup conflict. *Organizational Identity: A Reader* (1979), 56–65.

[46] Alan Tsang and Kate Larson. 2014. Opinion dynamics of skeptical agents. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*. 277–284.