Impact of Simple Algorithmic Filtering Strategies on Polarization in Social Networks due to Filter Bubbles: Preliminary Results

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

Yevgeniy Vorobeychik
Washington University in St. Louis
Saint Louis, MO, United States
yvorobeychik@wustl.edu

ABSTRACT

Many of us hoped that the introduction of online social networks and their widespread adoption would result in an increased dissemination of diverse ideas and a growing convergence on a set of universally accepted good ideas. Unfortunately, studies have shown that social networks have instead contributed to the growing degree of polarization in society at large across a wide range of issues. By employing algorithmic filtering algorithms to prioritize opinions that social network users agree with, and is thus more likely to engage with, over opinions that users disagree with, social networks encourage users to form filter bubbles around themselves, creating echo chambers, and promote polarization.

In this paper, we contribute to the growing body of work on studying the impact of such filter bubbles from the lens of empirical simulations. Specifically, we build upon an existing opinion dynamics model to incorporate both assimilation and boomerang effects, representing how opinions can either converge or diverge when two individuals interact, and empirically evaluate the polarization impact of several simple algorithmic filtering strategies through simulations.

Our results show that (1) when prioritizing similar opinions to expose to individuals, polarization increases as the number of individuals affected by boomerang effects increases; (2) when prioritizing popularity of individuals, extreme polarization seldom occur; and (3) when prioritizing the neutrality of opinions (i.e., preferring non-extreme opinions over extreme opinions), polarization is unsurprisingly minimized.

KEYWORDS

Polarization, Social Networks, Algorithmic Filtering, Filter Bubbles, Social Simulation

ACM Reference Format:
Jean Springsteen, William Yeoh, and Yevgeniy Vorobeychik. 2022. Impact of Simple Algorithmic Filtering Strategies on Polarization in Social Networks due to Filter Bubbles: Preliminary Results. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Auckland, New Zealand, May 9–13, 2022*, IFAAMAS, 7 pages.

1 INTRODUCTION

Introduced by Justice Oliver Wendell Holmes in the 1919 case of Abrams *v.* United States, the notion of a *marketplace of ideas*, where all ideas can compete in a free and unregulated market and the best ideas will ultimately emerge through widespread acceptance among the population, is a key rationale behind the need for the freedom of speech. Consequently, in the modern era of social networks, where an even more diverse set of ideas can be shared faster than ever before and with a global audience that is larger than ever before ever, one would expect that even better ideas should emerge from the discourse.

Unfortunately, instead of realizing our lofty dreams, social networks have brought forth unintended consequences of arguably nightmarish proportions. For example, they have been associated with increased polarization in society across a wide range of issues, including politics [2, 4, 10], science [21], and, more recently, healthcare [18]. Despite the exposure to diverse opinions and perspectives, individuals, unfortunately, gravitate toward clusters with diametric opposing views, unwilling to consider differing opinions. Given the severity of this phenomena – 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]. Researchers have identified several reasons why social networks exacerbate polarization: (1) Fake news travels faster than true stories on social networks [14] 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* is based on the idea that individuals are more likely to engage with content and opinions that they agree with over those that they do not. Therefore, as social network companies are incentivized to increase user engagement, as that is their primary revenue source, they form filter bubbles around individuals, creating echo chambers and promote polarization.

Researchers have recently studied the effect 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 [20] and how to mitigate filter bubbles in an influence maximization setting [20]. However, a limitation of the previous work by [9] is that it assumes that an individual 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 in this paper by investigating how several simple algorithmic

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Auckland, New Zealand.

filtering strategies for deciding whose opinions an individual is exposed to can impact the polarization of opinions in social networks. Further, our opinion dynamics model allows for both *assimilation and boomerang effects*, where the former describes how the opinion of an individual will *converge closer* to the opinion of the individual they are interacting with and the latter describes how the opinion of an individual will *diverge further away* from the opinion of the individual they are interacting with. Our experimental results show that (1) when prioritizing similar opinions to expose to individuals, polarization increases as the number of individuals affected by boomerang effects increases; (2) when prioritizing popularity of individuals, extreme polarization seldom occur; and (3) when prioritizing the neutrality of opinions (i.e., preferring non-extreme opinions over extreme opinions), polarization is unsurprisingly minimized.

2 BACKGROUND: OPINION DYNAMICS MODEL

We follow the opinion dynamics model by Tsang and Larson [25] in this paper, where agents are embedded in a weighted undirected graph $G = \langle V, E \rangle$, where the vertices $V = \{v_1, v_2, \ldots, v_n\}$ correspond to the agents and the edges (v_i, v_j) indicate that agents v_i and v_j are neighbors on the social network and can interact with each other.

The dynamics model models the propagation of an opinion on an issue during a discrete set of time steps $t \in \{0, 1, \ldots, T\}$. Further, it assumes that the opinion has exactly two poles, which an agent's opinion encoded by a continuous real value in $[0, 1]$, where 0 and 1 represent the most extreme opinions in the spectrum and 0.5 represents a neutral opinion.

The opinion of agent v_i at time step t is denoted by x_i^t and it can share it's opinion with all its neighbors $N_i = \{v_j \in V \mid (v_i, v_j) \in E\}$. This opinion is updated according to:

$$x_i^t = \frac{w_{ii}^{t-1} x_i^{t-1} + \sum_{v_j \in N_i} w_{ij}^{t-1} x_j^{t-1}}{w_{ii}^{t-1} + \sum_{v_j \in N_i} w_{ij}^{t-1}} \tag{1}$$

where w_{ij}^{t-1} corresponds to a *trust value* maintained by agent v_i indicating the weight it gives to the opinions of v_j at time step $t-1$. This trust value also evolves over time according to:

$$w_{ij}^t = \frac{w_{ij}^{t-1} + rT(x_i^t, x_j^t)}{1 + r} \tag{2}$$

where $T(x_i^t, x_j^t)$ is a *trust function* and r is a learning rate of the population. The trust function is defined by:

$$T(x_i^t, x_j^t) = \exp\left(-\frac{(x_i^t - x_j^t)^2}{h}\right) \tag{3}$$

and depends on the difference between the two opinions $x_i^t - x_j^t$ and an empathy parameter h via the Gaussian kernel. Intuitively, the larger the difference of opinions, the more unwilling the agent is to be persuaded by the other opinion. In contrast, the larger the empathy parameter, the more willing the population is to be persuaded by other opinions.

Finally, the model also includes a subset of agents who are *extremists* whose opinions are either 0 or 1 (called 0-extremists and 1-extremists, respectively) and their opinions will remain unchanged regardless of who they are interacting with.

3 OPINION DYNAMICS WITH ASSIMILATION AND BOOMERANG EFFECTS

One of the key properties of the opinion dynamics model described in the previous section is that the trust value is non-decreasing over time as the range of the trust function is greater that or equal to one (see Equations 2 and 3). Consequently, the opinion x_i^t of an agent v_i will always be drawn towards the opinion x_j^t of agent v_j that it is interacting with. This property is based on the rationale that when two people interact and learn about each other's opinions, they revise their own opinions to become more similar [1, 24].

However, there is also evidence that the opposite happens, that is, that the opinion x_i^t of an agent v_i moves further away in the opposite direction of the opinion x_j^t of agent v_j that it is interacting with [5]. One possible explanation for this effect is through the theory of cognitive dissonance from psychology [15, 19]. According to this theory, individuals seek to have all of their beliefs and opinions to be consistent with one another. Therefore, when individuals are in a state of cognitive dissonance (i.e., they have to rationalize two conflicting opinions), they take steps to reduce the extent of that dissonance [7]. For example, they may associate the other opinion as "fake news" and will even more firmly defend their opinion.

With this motivation in mind, we extend the opinion dynamics model described in the previous section to better account for these two opposite effects. Specifically, we modify the trust function and parameterize it with two threshold parameters d_1 and d_2 (with $0 \le d_1 \le d_2 \le 1$). There are the following three cases based on the difference between the two opinions $|x_i^t - x_j^t|$:

- $|x_i^t - x_j^t| < d_1$: An *assimilation effect* occurs, that is, the trust value w_{ij}^t of agent v_i on the opinion of agent v_j increases. Consequently, the opinion x_i^t of agent v_i moves *closer* to the opinion x_j^t of agent v_j that it was exposed to.
- $|x_i^t - x_j^t| > d_2$: A *boomerang effect* occurs, that is, the trust value w_{ij}^t of agent v_i on the opinion of agent v_j decreases. Consequently, the opinion x_i^t of agent v_i moves *further away* to the opinion x_j^t of agent v_j that it was exposed to.
- $d_1 \le |x_i^t - x_j^t| \le d_2$: It is in a *neutral zone*, that is, the trust value w_{ij}^t of agent v_i on the opinion of agent v_j remains unchanged. Consequently, the opinion x_i^t of agent v_i remains unaffected by the opinion x_j^t of agent v_j that it was exposed to.

Intuitively, the larger the threshold d_1, the more empathetic an agent is to an opinion that is different from theirs, resulting in them more likely to align their opinion in response. Similarly, the smaller the threshold d_2, the more adamant an agent is on their own opinion, resulting in them more likely to oppose a differing opinion.

More formally, we define the trust function as:

$$T(x_i^t, x_j^t) = \begin{cases} e^{\frac{(|x_i^t - x_j^t| - d_1)^2}{-(d_1/\ln(2))^2}} - 1 & \text{if } |x_i^t - x_j^t| < d_1 \\ 0 & \text{if } d_1 \leq |x_i^t - x_j^t| \leq d_2 \\ 1 - e^{\frac{(|x_i^t - x_j^t| - d_2)^2}{-((1-d_2)/\ln(2))^2}} & \text{if } |x_i^t - x_j^t| > d_2 \end{cases} \quad (4)$$

Additionally, unlike the trust function from the earlier model, we ensure that the range of the new function is within -1 and 1 and modify the trust value update according to:

$$w_{i,j}^t \leftarrow \alpha w_{i,j}^{t-1} + (1-\alpha) T(x_i^t, x_j^t) \quad (5)$$

where $\alpha \in [0, 1]$ is a parameter akin to the learning rate that tunes how fast the weights change over time.

Finally, we use the exact same opinion update rule (see Equation 1) as in the previous model and allow for the possibility of *extremist* agents as in the previous model.

4 ALGORITHMIC FILTERING STRATEGIES

We now describe several simple algorithmic filtering strategies for selecting at each time step, which k neighbors an agent is interacting with, where k is a user-defined input parameter. We will use $S_i \subseteq N_i$ to denote the subset of neighbors of agent v_i that the agent is interacting with. More formally:

$$S_i = \{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k \mid \hat{v}_j \in N_i \wedge \mathcal{P}\} \quad (6)$$

where $\mathcal{P}$ corresponds to the constraints of different algorithmic filtering strategies described below.

- **Random Neighbors:** As a baseline strategy, for each agent v_i, we select k neighbors randomly. Therefore, for this strategy:

$$\mathcal{P} \equiv \text{true} \quad (7)$$

  as no additional constraint is imposed.
- **Least Polar Neighbors:** In this strategy, for each agent v_i, we select the k neighbors whose opinions are the least polar, that is, the k neighbors whose opinions are closest to 0.5. More formally,

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

  where $\hat{x}_j$ is the opinion of neighboring agent $\hat{v}_j$ and X_i is the set of opinions of all neighboring agents $v \in N_i$. The intuition behind this strategy is that one might want to combat polarization through exposure to neutral neighbors only.
- **Most Popular Neighbors:** In this strategy, for each agent v_i, we select the k most connected neighbors. This is motivated by the idea that social networks may prefer to expose individuals to neighbors who are "influencers" or popular (i.e., neighbors with many neighbors) under the assumption that the opinions of influencers carry more weight than opinions of non-influencers. More formally,

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

  where $\deg(v)$ is the degree of agent v in the graph.
- **Most Similar Neighbors:** In this strategy, for each agent v_i, we select the k neighbors whose opinions are closest to that of the agent. This is motivated by homophily – the notion that individuals are likely to be attracted to others who are similar to them. More formally,

$$\mathcal{P} \equiv \forall x \in X_i \setminus \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k\} : |x - x_i| \geq |\hat{x}_j - x_i| \quad (10)$$

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

5 RELATED WORK

5.1 Opinion Dynamics Models

While we extend the Tsang and Larson model of opinion dynamics [25], there are other models worth noting. The DeGroot model [12] is one such foundational model. Similar to our model, it too models social networks as a graph of agents that interact with neighbors in each time step. However, each agent v_i updates its opinion x_i^t at time step t according to:

$$x_i^t = \sum_{v_j \in N_i} T_{ij} x_j^t \quad (11)$$

where $T_{ij} \geq 0$ is the weight that agent i has on agent j's opinion. One benefit of this model is that it provides clear conditions for determining whether it is possible for the agents in the network to reach a consensus. Then, if we let X^k denote the vector of opinions at time step t, such that $X^t = [x_1^t, x_2^t, \ldots, x_n^t]^T$, and define a matrix T whose i, j-th element is T_{ij}, Equation 11 can be written as:

$$X^t = TX^{t-1} \quad (12)$$

A popular extension of the DeGroot model is the Friedkin-Johnsen model [16], which introduces a positive diagonal matrix Λ that quantifies the extent to which each agent is open to outside influence. This model is then defined by:

$$X^t = \Lambda T X^{t-1} + (\mathrm{I} - \Lambda) X^0 \quad (13)$$

where I is the identity matrix. Then, if $\Lambda = \mathrm{I}$, the Friedkin-Johnsen model reduces to the original DeGroot model. With the incorporation of Λ, this model allows each agent to hold true to their own opinion and not only be influenced by their neighbors. However, one shortcoming of both models is that they assume that when two agents interact, their opinions will become more similar. In other words, they only model assimilation effects.

Chau et al. [8] addresses this limitation by introducing boomerang effects in their opinion dynamics model. Similar to previous models, they too model social networks as a graph of agents that interact with neighbors at each time step. However, unlike previous models where an agent interacts with all their neighbors in each time step, in this model, only a pair of neighboring agents v_i and v_j interact with each other in each time step. When they interact, both agents update their respective opinions x_i^t and x_j^t at time step t simultaneously. If $|x_i^{t-1} - x_j^{t-1}| < d_1$, then they are in an assimilation zone, and the opinions are updated according to:

$$x_i^t = x_i^{t-1} + \mu(x_j^{t-1} - x_i^{t-1}) \quad (14)$$

$$x_j^t = x_j^{t-1} + \mu(x_i^{t-1} - x_j^{t-1}) \quad (15)$$

where $\mu \in (0, 0.5]$ is a convergence parameter. If $|x_i^{t-1} - x_j^{t-1}| > d_2$, then they are in a boomerang zone, and the opinions are updated according to:

$$x_i^t = \mathcal{N}(x_i^{t-1} - \lambda(x_j^{t-1} - x_i^{t-1})) \quad (16)$$

$$x_j^t = \mathcal{N}(x_j^{t-1} - \lambda(x_i^{t-1} - x_j^{t-1})) \tag{17}$$

where $\lambda > 0$ is a divergence parameter and $\mathcal{N}(\cdot)$ is a normalization function to keep the opinions between 0 and 1. If $d_1 \leq |x_i^{t-1} - x_j^{t-1}| \leq d_2$, then they are in a neutral zone and the opinions remain unchanged. We draw inspirations from this model by also incorporating the three regions of interactions in our model. However, aside from the difference in the number of agents that interact and change their opinions in each time step, another key difference is that both agents that interact with each other change their opinions simultaneously here. In contrast, only one agent changes its opinion in our model (as well as in the other models described above). This assumption is better reflects one-way interactions in some social networks, such as Facebook and Twitter, where an individual who posted some information on the network (e.g., a tweet on Twitter) does not change its opinion based on the other individuals who accesses that information (e.g., reads the tweet). In contrast, the individuals who access the information may have their opinions changed.

5.2 Polarity Measures

There are a number of approaches to measure and quantify polarity and there is often no consensus on which is the right measure. In this paper, we follow the same measure as Tsang and Larson by defining it as the absolute difference in opinion from 0.5 [25] and report that measure in our experimental results. In contrast, Musco et al. [23] present "a class of group-based polarization measures that capture the extent to which opinions are clustered into distinct groups."

One statistical definition of polarization is Sarle's Bimodality Coefficient, which has been used to measure opinion polarization, for example in the work by DiMaggio et al. [13]. Let X denote a vector of opinions and let γ and κ denote the skewness and kurtosis of $\bar{X}$, respectively, where $\bar{X} = X - \text{mean}(X)$. Then, Equation 18 defines a bimodality value $\beta(\bar{X})$:

$$\beta(\bar{X}) = \frac{\gamma^2 + 1}{\kappa} \tag{18}$$

This measure returns a value in the range $[0, 1]$ with 0 indicating no polarization and 1 indicating maximum polarization.

Another polarity measure is one that is localized and is based on an "average local agreement" that accounts for the local structure in the social graph [23]. Let X denote a vector of opinions and $S = \text{sign}(X - \text{mean}(X) \cdot \vec{1})$, where $\vec{1}$ is a vector of ones. Then, Equation 19 defines the average local agreement $\mathcal{L}(X)$:

$$\mathcal{L}(X) = \frac{1}{|V|} \sum_{v_i \in V} \frac{1}{\deg(v_i)} \sum_{v_j \in N_i} \mathbf{1}[S_i = S_j] \tag{19}$$

where $\deg(v_i)$ is the degree of agent v_i in the graph, S_i is the i-th element in the vector S, and $\mathbf{1}$ is the indicator function. This measure also returns a value in the range $[0, 1]$ with 0 indicating no polarization and 1 indicating maximum polarization. Additionally, with a randomly initialized vector of opinions X, $\mathcal{L}(X) = 0.5$, indicating that values between 0 and 0.5 constitute expected levels of polarization.

6 EXPERIMENTAL EVALUATION

We now describe our experimental evaluation of the five algorithmic filtering algorithms described in the previous section applied on a simulated social network with agents whose opinions change based on the opinion dynamics model described in Section 3.

6.1 Experimental Design

For our experiments, similar to others in the literature [8, 25], we simulated our social network with Barabasi-Albert random graphs, where we set the number of agents $V = 200$ and the attachment parameter of the graphs $m = 4$. The attachment parameter controls the number of nodes each new node in the graph connects to during the graph generation process. Each data point reported is an average of 25 runs, each on a different problem instance.

In each problem instance, 20% of agents are randomly selected, where half of them are 0-extremists and the other half are 1-extremists. The opinions of these agents are initialized to 0 and 1, respectively, and do not change over time even when they are interacting with other agents. The opinions of all other non-extremist agents are uniformly sampled from the range $[0, 1]$. The trust values of all agents are initialized to 1, which corresponds to the uniform trust model in the literature [25].

In each experiment, we varied the threshold parameters d_1 and d_2 that are used by the trust function (see Equation 4), varying each parameter between 0 and 1 in increments of 0.1. The simulation for a single run terminates when each agent's opinion changed no more than a small value $\epsilon = 10^{-3}$ or when the maximum number of iterations $t_{\max} = 500$ is reached. Finally, the number of neighbors chosen by the algorithmic filtering strategies is $k = 2$.

We ran our experiments on a MacBook Pro machine comprising of an Intel Core i7 2.6GHz processor with 16GB of memory. Our implementation of the simulation is written in Python and is extended from the source code of Tsang and Larson for their simulations [25].

6.2 Experimental Results: Polarization

Figures 1 to 4 show the average polarization of agents for the different d_1 and d_2 values and different algorithmic filtering strategies. In these heatmaps, the polarization of agent v_i is defined as $|x_i - 0.5|$, that is, the absolute difference between the agent's opinion and the neutral opinion of 0.5 [25]. We now discuss the results for each of those strategies separately.

Random Neighbors: Figure 1 show the results for the random neighbors filtering strategy. We make the following observations:

- In the top left quadrant (where $d_1 \leq 0.5$ and $d_2 \leq 0.5$), opinions are very polarized as the polarization values are close to the maximum of 0.5. The reason for this behavior is that when $d_2 \leq 0.5$, every agent experiences a boomerang effect from their filtered neighbors with extreme opinions. Specifically, agents whose opinions are in the range of $[0, 0.5]$ will experience boomerang effects from filtered neighbors with extreme opinions close to 1 and, consequently, move their opinions closer to 0. Similarly, agents whose opinions are in the range of $[0.5, 1]$ will experience boomerang effects from filtered neighbors with extreme opinions close to 0 and move their opinions closer to 1.

  As 20% of agents in the problem are extremists, their neighbors that are 1-hop away will first experience the boomerang effects

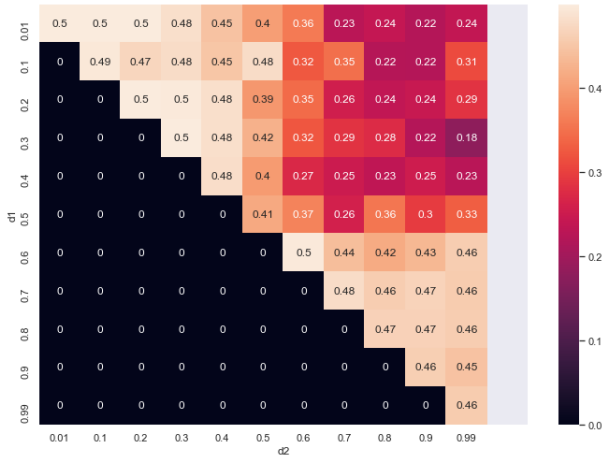

Fig. 1: Average Polarization for the Random Neighbors Filtering Strategy

Fig. 2: Average Polarization for the Least Polar Neighbors Filtering Strategy

described above. However, as their opinions move closer to the extremes, the agents that are 2-hops away will also experience the boomerang effects. This process continues until a large majority of agents have extreme opinions.

- In the bottom right quadrant (where $d_1 \geq 0.5$ and $d_2 \geq 0.5$), opinions are also very polarized. The reason for this behavior is similar to the one described above, except that it is due to assimilation effects instead of boomerang effects. Specifically, all agents experience assimilation effects from their filtered neighbors with extreme opinions. Agents whose opinions are in the range of $[0, 0.5]$ will experience assimilation effects from filtered neighbors with extreme opinions close to 0 and, consequently, move their opinions closer to 0. The opposite happens for agents whose opinions are in the range of $[0.5, 1]$.
- In the top right quadrant (where $d_1 \leq 0.5$ and $d_2 \geq 0.5$), opinions are less polarized with a general trend of decreasing polarization with decreasing d_1 and increasing d_2. The reason for this trend is twofold:
  - Agents whose opinions are in the range of $[0, d_1]$ and $[1-d_1, 1]$ will move their opinions closer to 0 and 1, respectively, due to assimilation effects. As d_1 decreases, the size of the ranges decrease and, thus, fewer agents experience the assimilation effects.
  - Agents whose opinions are in the range of $[0, 1 - d_2]$ and $[d_2, 1]$ will move their opinions closer to 0 and 1, respectively, due to boomerang effects. Similar to the previous case, as d_2 increases, the size of the ranges decrease and fewer agents experience the boomerang effects.

Least Polar Neighbors: Figure 2 show the results for the least polar neighbors filtering strategy, where agents are exposed to neighbors whose opinions are closest to 0.5. We make the following observations:

- Unsurprisingly, agents are significantly less polarized with this filtering strategy compared to the random neighbors filtering strategy since they are exposed to least polarizing neighbors.
- In the bottom right quadrant (where $d_1 \geq 0.5$ and $d_2 \geq 0.5$), opinions are very unpolarized as the polarization values are close to the minimum of 0. The reason for this behavior is that every agent experiences assimilation effect from their filtered neighbors with neutral opinions and, consequently, move their opinions closer to 0.5.
- There is a general trend of decreasing polarization with increasing d_1 and d_2. We describe the reason for this trend below, where we use x_{mid} (= ~0.5) to refer to the opinion of the least polar neighbor:
  - Agents whose opinions are in the range of $[x_{\text{mid}} - d_1, x_{\text{mid}} + d_1]$ will move their opinions closer to x_{mid} due to assimilation effects. As d_1 increases, the size of the range increases and, thus, more agents experience the assimilation effect and have less extreme opinions.
  - Agents whose opinions are in the range of $[0, x_{\text{mid}} - d_2]$ and $[x_{\text{mid}} + d_2, 1]$ will move their opinions closer to 0 and 1, respectively, due to boomerang effects. As d_2 increases, the size of the ranges decrease and fewer agents experience the boomerang effects to have more extreme opinions.

Most Similar Neighbors: Figure 3 show the results for the most similar neighbors filtering strategy, where agents are exposed to neighbors whose opinions are closest to their opinions. We make the following observations:

- Overall, agents are more polarized with this filtering strategy compared to the least polar neighbors filtering strategy since they are exposed to more polarizing neighbors. However, they are less polarized than when a random neighbors filtering strategy is used, reflecting that they are less often influenced by neighbors with extreme opinions.
- There is a general trend of decreasing polarization with increasing d_2. As this filtering strategy selects neighbors with similar opinions as that of the agent, it is likely that the differences in opinion will be small. Further, as d_2 increases, there will be less of a boomerang effect from the neighbors with differing opinions and the agents will be predominantly affected by assimilation

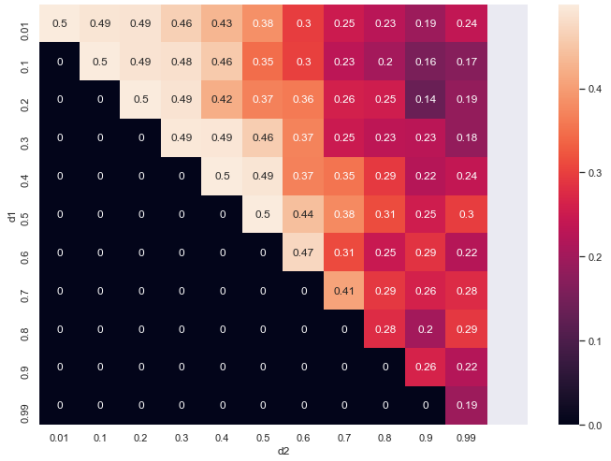

Fig. 3: Average Polarization for the Most Similar Neighbors Filtering Strategy

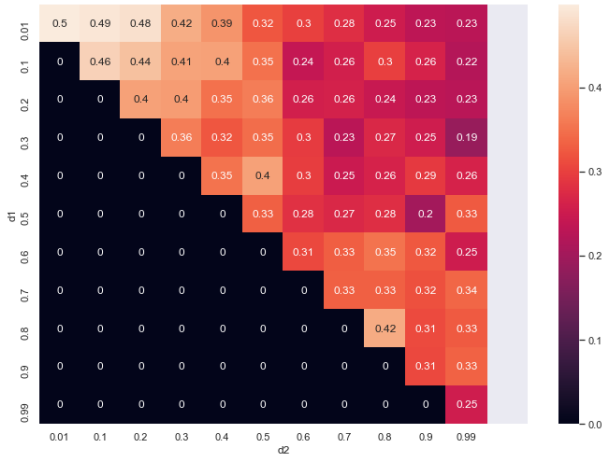

Fig. 4: Average Polarization for the Most Popular Neighbors Filtering Strategy

effects from neighbors with similar opinions. Therefore, with the exception of agents whose initial opinions are close to the extremes, all other agents will mostly have similar opinions as their initial opinions.

Most Popular Neighbors: Figure 4 show the results for the most popular neighbors filtering strategy, where agents are exposed to neighbors that have the largest number of neighboring agents. We observe that, unsurprisingly, agents are more polarized with this strategy compared to the least polar neighbors strategy. However, they are less polarized than the random neighbors and most similar neighbors strategies. Further, there is no observable trend for the different values of d_1 and d_2. The reason for these observations is that unlike other strategies where the choice of selected neighboring agents depend on their opinions, in this strategy, the choice is independent of opinions and is time-invariant. Consequently, the same set of neighboring agents will be chosen in each iteration of the simulation. Therefore, the resulting polarity of opinions is largely dependent on whether extremist agents have a small or large number of neighbors, and since the extremist agents are selected randomly, there are no observable trends.

6.3 Experimental Results: Distribution of Final Opinions

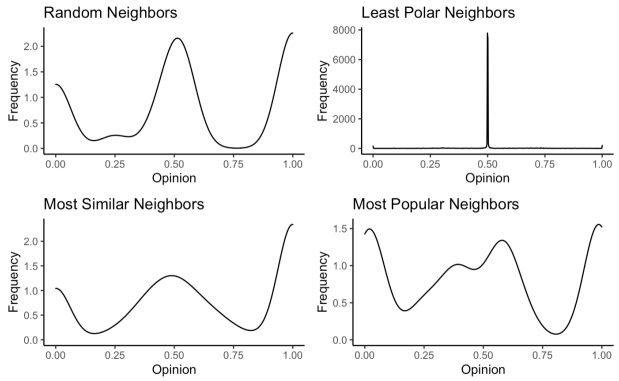

Fig. 5: Distribution of Final Opinions of Non-Extremist Agents for the Various Filtering Strategies with $d_1 = 0.4$ and $d_2 = 0.7$

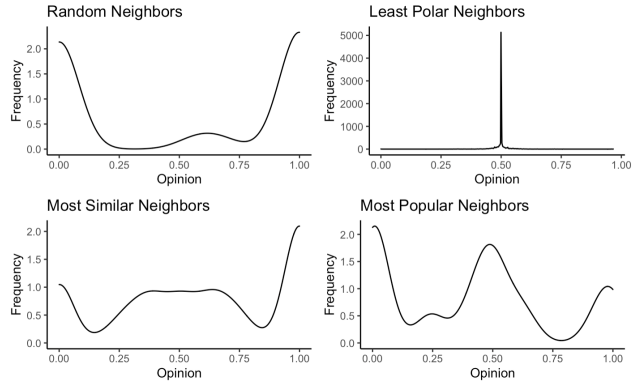

Fig. 6: Distribution of Final Opinions of Non-Extremist Agents for the Various Filtering Strategies with $d_1 = 0.7$ and $d_2 = 0.75$

Figures 5 and 6 show the distribution of opinions of non-extremist agents in the final iteration of the simulations for two d_1 and d_2 configuration pairs. Unsurprisingly, the least polar neighbors filtering strategy results in almost all agents having an opinion of 0.5 in both figures.

With the random neighbors filtering strategy, the agents are more polarized when $d_1 = 0.7$ than when $d_1 = 0.4$, consistent with our observation in the heatmaps. However, it is interesting to note that the variance in the opinions of the neutral cluster is not very large, reflecting a convergence to the neutral opinion of 0.5.

With the most similar neighbors filtering strategy, the same observation that the polarization of agents is not noticeably different as there is a relatively large neutral cluster. However, the variance

of the cluster is larger when $d_1 = 0.7$ compared to $d_1 = 0.4$. The reason is that when d_1 is small, only a subset of neighboring agents selected by the filtering strategy will influence the agent. Therefore, the opinion of the agent will sway in the direction of the more similar neighbor. On the other hand, when d_1 is larger, the agent is influenced by more of its neighbors and its opinion may not change as much if it is influenced in both directions – towards 0 and 1.

Finally, with the most popular neighbors filtering strategy, similar to the heatmap for this strategy, there is no observable trend in the distributions as well for the same reason as described previously.

7 CONCLUSIONS AND FUTURE WORK

Social networks have been accused of exacerbating polarization due to the algorithmic filtering strategies that they employ. By prioritizing opinions that social network users agree with, and is thus more likely to engage with, over opinions that users disagree with, users form filter bubbles around themselves, creating echo chambers, and promote polarization.

In this paper, we build upon the opinion dynamics model of Tsang and Larson [25], which models the *assimilation effect* of interactions (i.e., the opinion of an agent will converge closer to the opinion that it is exposed to). We extend the model to also include *boomerang effects* (i.e., the opinion of an agent will diverge further away from the opinion that it is exposed to). Using this extended model, we empirically evaluated several simple algorithmic filtering strategies for choosing which opinions to expose to individuals. Our results show that (1) when prioritizing similar opinions to expose to individuals, polarization increases as the number of individuals affected by boomerang effects increases; (2) when prioritizing popularity of individuals, extreme polarization seldom occur; and (3) when prioritizing the neutrality of opinions (i.e., preferring non-extreme opinions over extreme opinions), polarization is unsurprisingly minimized.

As part of future work, we plan to investigate if these observations carry over to other opinion dynamics models, such as Friedkin-Johnsen model [16], as well as other forms of graphs (e.g., Stochastic Block Model [17]) and real-world datasets (e.g., Twitter and Reddit datasets [11, 22]). We also plan to investigate more complex algorithmic filtering strategies, such as ones that better model likelihood of engagements into account in social networks. Finally, we plan to theoretically characterize convergence properties of this approach under the different algorithmic filtering strategies.

REFERENCES

[1] Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. 2011. Bayesian Learning in Social Networks. *The Review of Economic Studies* 78, 4 (2011), 1201–1236.

[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. 2018. *Social Media Networks, Fake News, and Polarization.* Working Paper 24462. National Bureau of Economic Research.

[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] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to Opposing Views on Social Media Can Increase Political Polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[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] Hitesh Bhasin. 2021. Boomerang Effect - Definition, Theory and Examples. https://www.marketing91.com/boomerang-effect/

[8] H. F. Chau, C. Y. Wong, F. K. Chow, and C. H. F. Fung. 2013. Social Judgment Theory Based Model On Opinion Formation, Polarization And Evolution. *arXiv preprint arXiv:1308.2042v2* (2013).

[9] Uthsav Chitra and Christopher Musco. 2020. *Analyzing the Impact of Filter Bubbles on Social Network Polarization.* Association for Computing Machinery, 115–123.

[10] 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.

[11] Abir De, Sourangshu Bhattacharya, Parantapa Bhattacharya, Niloy Ganguly, and Soumen Chakrabarti. 2014. Learning a Linear Influence Model From Transient Opinion Dynamics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 401–410.

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

[13] Paul DiMaggio, John Evans, and Bethany Bryson. 1996. Have American's Social Attitudes Become More Polarized? *Amer. J. Sociology* 102, 3 (1996), 690–755.

[14] 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

[15] Leon Festinger. 1957. *A Theory of Cognitive Dissonance.* Vol. 2. Stanford University Press.

[16] Noah E Friedkin and Eugene C Johnsen. 1999. Influence Networks and Opinion Change. *Advances in Group Processes* 16, 1 (1999), 1–29.

[17] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic Blockmodels: First Steps. *Social Networks* 5, 2 (1983), 109–137.

[18] Harald Holone. 2016. The Filter Bubble and its Effect on Online Personal Health Information. *Croatian Medical Journal* 57, 3 (2016), 298.

[19] Myeongki Jeong, Hangjung Zo, Chul Ho Lee, and Yasin Ceran. 2019. Feeling Displeasure from Online Social Media Postings: A Study Using Cognitive Dissonance Theory. *Computers in Human Behavior* 97 (2019), 231–240.

[20] 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* (2020).

[21] 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.

[22] 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.

[23] Christopher Musco, Indu Ramesh, Johan Ugander, and R Teal Witter. 2021. How to Quantify Polarization in Models of Opinion Dynamics. *arXiv preprint arXiv:2110.11981* (2021).

[24] David G Myers and George D Bishop. 1971. Enhancement of Dominant Attitudes in Group Discussion. *Journal of Personality and Social Psychology* 20, 3 (1971), 386.

[25] Alan Tsang and Kate Larson. 2014. Opinion Dynamics of Skeptical Agents. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*. 277–284.