Biased assimilation, homophily, and the dynamics of polarization

Pranav Dandekar[a,1], Ashish Goel[b], and David T. Lee[c]

Departments of [a]Management Science and Engineering, [b]Management Science and Engineering and (by courtesy) Computer Science, and [c]Electrical Engineering, Stanford University, Stanford CA 94305

Edited by Jon Kleinberg, Cornell University, Ithaca, NY, and approved February 28, 2013 (received for review October 3, 2012)

We study the issue of polarization in society through a model of opinion formation. We say an opinion formation process is polarizing if it results in increased divergence of opinions. Empirical studies have shown that homophily, i.e., greater interaction between like-minded individuals, results in polarization. However, we show that DeGroot's well-known model of opinion formation based on repeated averaging can never be polarizing, even if individuals are arbitrarily homophilous. We generalize DeGroot's model to account for a phenomenon well known in social psychology as biased assimilation: When presented with mixed or inconclusive evidence on a complex issue, individuals draw undue support for their initial position, thereby arriving at a more extreme opinion. We show that in a simple model of homophilous networks, our biased opinion formation process results in polarization if individuals are sufficiently biased. In other words, homophily alone, without biased assimilation, is not sufficient to polarize society. Quite interestingly, biased assimilation also provides a framework to analyze the polarizing effect of Internet-based recommender systems that show us personalized content.

The issue of polarization in society has been extensively studied and vigorously debated in the academic literature as well as the popular press over the last few decades. In particular, are we as a society getting more polarized? If so, why, and how can we fix it? Different empirical studies arrive at different answers to this question depending on the context and the metric used to measure polarization.

Evidence of polarization in politics has been found in the increasingly partisan voting patterns of the members of Congress (1, 2) and in the extreme policies adopted by candidates for political office (3). McCarty et al. (4) claim via rigorous analysis that America is polarized in terms of political attitudes and beliefs. Phenomena such as segregation in urban residential neighborhoods (5–7), the rising popularity of overtly partisan television news networks (8, 9), and the readership and linking patterns of blogs along partisan lines (10–13) can all be viewed as further evidence of polarization. On the other hand, it has also been argued on the basis of detailed surveys of public opinion that society as a whole is not polarized, even though the media and the politicians make it seem so (14, 15). We adopt the view that polarization is not a property of a state of society; instead it is a property of the dynamics through which individuals form opinions. We say that opinion formation dynamics are polarizing if they result in an increased divergence of opinions.

It has been argued using empirical studies that homophily, i.e., greater interaction between like-minded individuals, results in polarization (12, 16, 17). This argument has been used to claim that the rise of cable news, talk radio, and the Internet has contributed to polarization: the increased diversity of information sources coupled with the increased ability to narrowly tailor them to one's specific tastes (either manually or algorithmically through, for example, recommender systems) has an echo-chamber effect that ultimately results in increased polarization.

A rich body of work attempts to explain polarization through variants of a well-known mathematical model of opinion formation proposed by DeGroot (18). In DeGroot's model, individuals are connected to each other in a social network. The edges of the network have associated weights representing the extent to which neighbors influence each other's opinions. Individuals update their opinion as a weighted average of their current opinion and that of their neighbors. Variants of this model (e.g., refs. 19–22) explain the empirically observed persistent disagreement on many issues by, for example, introducing stubborn individuals (i.e., individuals with unchanging opinions) into the original model. However, we show that repeated averaging of opinions, which underlies these models, always results in opinions that are less divergent compared with the initial opinions, even if individuals are arbitrarily homophilous. As a result, this entire body of work appears to fall short of explaining polarization which is generally perceived to mean an increased divergence of opinions, not just persistent disagreement. In this paper, we seek a more satisfactory model of opinion formation that (*a*) is informed by a theory of how individuals actually form opinions and (*b*) produces an increased divergence of opinions under intuitive conditions.

We base our model on a well-known phenomenon in social psychology called biased assimilation, according to which individuals process new information in a biased manner whereby they readily accept confirming evidence while critically examining disconfirming evidence. Suppose that individuals with opposing views on an issue are shown mixed or inconclusive evidence. Intuitively, exposure to such evidence would engender greater agreement, or at least a moderation of views. However, in a seminal paper, Lord et al. (23) showed that biased assimilation causes individuals to arrive at more extreme opinions after being exposed to identical, inconclusive evidence. This finding has been reproduced in many different settings over the years (e.g., refs. 24–26). We use biased assimilation as the basis of our model of opinion formation and show that in our model homophily alone, without biased assimilation, is not sufficient to polarize society.

It has been argued (27) that biased assimilation can be countered by surprising validators: Individuals are more likely to carefully consider disconfirming evidence if it is presented by a source that is otherwise similar to them. Centola (28) empirically showed that individuals are much more likely to adopt health behaviors when they are a part of more homophilous networks. We show that a stylized model of surprising validators does indeed reduce polarization as we define it in this paper.

Finally, we analyze the polarizing effects of recommender systems that are widely used on the Internet to make personalized recommendations (e.g., search results, news articles, products) to individuals. We analyze three recommender algorithms—SimpleSALSA, SimplePPR, and SimpleICF—that are similar in spirit to commonly used recommender algorithms. For a simple, natural model of the underlying user-item graph, and under reasonable assumptions, we show that SimplePPR, which recommends the item that is most relevant to a user based on a PageRank-like (29) score, is always polarizing. On the other hand, SimpleSALSA and SimpleICF, which first choose a random item liked by the user

Author contributions: P.D., A.G., and D.T.L. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ppd@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1217220110/-/DCSupplemental.

and recommend an item similar to that item, are polarizing only if individuals are biased. Thus, biased assimilation also provides a useful framework to understand whether recommender systems contribute to polarization.

Model

Our opinion formation process unfolds over a social network represented by a connected weighted undirected graph $G = (V, E, w)$. The nodes in V represent individuals and the edges represent friendships or relationships between them. Let $|V| = n$. An edge $(i, j) \in E$ is associated with a weight $w_{ij} > 0$ representing the degree of influence i and j have on each other. Each individual $i \in V$ also has an associated weight $w_{ii} \geq 0$ representing the degree to which the individual weights his own opinions. We denote by $N(i)$ the set of neighbors of i; that is, $N(i) := \{j \in V: (i, j) \in E\}$.

An individual i has an opinion $x_i(t) \in [0, 1]$ at time step $t = 0, 1, 2 \ldots$. The extreme opinions 0 and 1 represent opposing points of view on an issue. So $x_i(t)$ can be interpreted as individual i's degree of support at time t for the position represented by 1 and $1 - x_i(t)$ as the degree of support for the position represented by 0. Let $\mathbf{x}(t) \in [0, 1]^n$ denote the vector of opinions at time t. An opinion formation process is a description of how individuals update their opinions; i.e., for each individual $i \in V$, it defines $x_i(t + 1)$ as a function of the vector of opinions, $\mathbf{x}(t)$, at time t.

Measuring Polarization. We view polarization as a property of an opinion formation process instead of a property of a state of the network. We characterize polarization as a verb as opposed to a noun; i.e., we say that an opinion formation process is polarizing if it results in an increased divergence of opinions. One could mathematically capture divergence of opinions in many different ways. We measure divergence in terms of the network disagreement index defined below.

Definition 1. Network Disagreement Index. *Given a graph $G = (V, E, w)$ and a vector of opinions $\mathbf{x} \in [0, 1]^n$ of individuals in V, the network disagreement index $\eta(G, \mathbf{x})$ is defined as*

$$\eta(G, \mathbf{x}) := \sum_{(i,j)\in E} w_{ij}\left(x_i - x_j\right)^2. \qquad [1]$$

Consider an opinion formation process over a network $G = (V, E, w)$ that transforms a set of initial opinions $\mathbf{x} \in [0, 1]^n$ into a set of opinions $\mathbf{x}' \in [0, 1]^n$. Then, we say the process is polarizing if $\eta(G, \mathbf{x}') > \eta(G, \mathbf{x})$, and vice versa.

The network disagreement index (NDI) is similar to the notion of social cost used by Bindel et al. (22). Each term $w_{ij}(x_i - x_j)^2$ can be viewed as the cost of disagreement imposed upon i and j. This view that the social cost depends on the magnitude of the difference of opinions along edges is consistent with theories in social psychology according to which attitude conflicts in relationships are a source of psychological stress or instability (30, 31). The NDI captures the phenomenon of issue radicalization, i.e., preexisting groups of individuals becoming progressively more extreme. Admittedly, it does not entirely capture an aspect of polarization called issue alignment (32) whereby individuals with diverse opinions organize into ideologically coherent, but opposing factions. However, there is significant empirical evidence (4, 27, 32) that issue radicalization is more prevalent compared with issue alignment, and hence NDI captures the most salient aspects of polarization. Many of our results hold for more general measures of divergence, which we discuss in a later section.

DeGroot's Repeated Averaging Process. In his seminal work on opinion formation, DeGroot (18) proposed a process where at each time step, individuals simultaneously update their opinion to the weighted average of their neighbors' and their own opinion at the previous time step.

Definition 2. DeGroot's Repeated Averaging Process. *The opinion of individual i at time $t + 1$, $x_i(t + 1)$, is given by*

$$x_i(t+1) = \frac{w_{ii}x_i(t) + s_i(t)}{w_{ii} + d_i}, \qquad [2]$$

where $s_i(t) := \sum_{j\in N(i)} w_{ij}x_j(t)$ is the weighted sum of the opinions of i's neighbors, and $d_i := \sum_{j\in N(i)} w_{ij}$ is i's weighted degree.

Recall that $x_j(t)$ and $1 - x_j(t)$ represent the degree of support for extremes 1 and 0, respectively. Then, opinion update under DeGroot's process is equivalent to taking a weighted average of the total support for 0 and that for 1. The weight that individual i places on 1 (and on 0) is computed by summing the degrees of support of i's neighbors weighted by the influence of each neighbor on i.

Biased Opinion Formation Process. We generalize DeGroot's process to account for biased assimilation. Biased assimilation is a well-known phenomenon in social psychology described by Lord et al. (23, p. 2098) in their seminal paper as follows:

> People who hold strong opinions on complex social issues are likely to examine relevant empirical evidence in a biased manner. They are apt to accept "confirming" evidence at face value while subjecting "disconfirming" evidence to critical evaluation, and as a result to draw undue support for their initial positions from mixed or random empirical findings.

Lord et al. (23) showed through experiments that biased assimilation of mixed or inconclusive evidence does indeed result in more extreme opinions.

To account for biased assimilation, we propose a biased opinion formation process. Recall that $x_i(t)$ can be viewed as the degree of support for the position represented by 1. Individuals weight confirming evidence more heavily relative to disconfirming evidence by updating their opinions as follows: Individual i weights each neighbor j's support for 1 [i.e., $x_j(t)$] by an additional factor $(x_i(t))^{b_i}$, where $b_i \geq 0$ is a bias parameter. Therefore, $x_i(t+1) \propto (x_i(t))^{b_i} w_{ij}x_j(t)$. Similarly, i weights j's support for 0 [i.e., $1 - x_j(t)$] by $(1 - x_i(t))^{b_i}$, and so $(1 - x_i(t+1)) \propto (1 - x_i(t))^{b_i} w_{ij}(1 - x_j(t))$. Informally, b_i represents the bias with which i assimilates his neighbors' opinions.

Illustrative Example. Consider a graph with two nodes, i and j, connected by an edge with a weight w_{ij}. Then, according to the biased opinion formation process, i's opinion at time $t + 1$, $x_i(t + 1)$, is given by

$$x_i(t+1) = \frac{w_{ii}x_i(t) + (x_i(t))^{b_i} w_{ij}x_j(t)}{w_{ii} + (x_i(t))^{b_i} w_{ij}x_j(t) + (1 - x_i(t))^{b_i} w_{ij}\left(1 - x_j(t)\right)}.$$

More generally, the opinion update of individual i in the biased opinion formation process is defined as shown below.

Definition 3. Biased Opinion Formation Process. *Under the biased opinion formation process, the opinion of individual i at time $t + 1$, $x_i(t + 1)$, is given by*

$$x_i(t+1) = \frac{w_{ii}x_i(t) + (x_i(t))^{b_i}\, s_i(t)}{w_{ii} + (x_i(t))^{b_i}\, s_i(t) + (1 - x_i(t))^{b_i}(d_i - s_i(t))}, \qquad [3]$$

where, as before, $s_i(t) := \sum_{j\in N(i)} w_{ij}x_j(t)$ is the weighted sum of the opinions of i's neighbors, and $d_i := \sum_{j\in N(i)} w_{ij}$ is i's weighted degree. Observe that when $b_i = 0$, [3] is identical to [2]; i.e., DeGroot's process is a special case of our process and corresponds to unbiased assimilation. More generally, biased assimilation can be modeled by making i's opinion update proportional to $\beta_i(x_i(t))\, s_i(t)$, where the bias function β_i: $[0, 1] \to [0, 1]$ is nondecreasing.

Connection with Urn Models. Urn models are an elegant abstraction and have been used to analyze the properties of a wide

variety of probabilistic processes. DeGroot's process and our biased opinion formation process have the following analogous urn dynamics. Let $x_i(t)$ denote the fraction of RED balls in individual i's urn at time t and $1 - x_i(t)$ denote the corresponding fraction of BLUE balls:

- Step 1 (common): At each time step, i chooses a neighbor j with probability proportional to w_{ij} and inspects a ball chosen uniformly at random from j's urn. Note that i does not remove the ball from j's urn.
- Step 2 (DeGroot's process): i adds a ball of the same color as the inspected ball to his urn and discards a ball chosen uniformly at random from his urn.
- Step 2 (Biased opinion formation process with $b_i = 1$): i also inspects a ball chosen uniformly at random from his own urn. If the colors of the two inspected balls match, i adds a ball of the same color to his urn and discards a ball chosen uniformly at random from his urn.

Biased Assimilation by a Single Agent in a Fixed Environment. Here we demonstrate that our model of biased assimilation mathematically reproduces the empirical findings of Lord et al. (23). We analyze the change in opinion of a single individual as a function of his bias parameter when he is exposed to opinions from a fixed environment. The fixed environment represents sources of information that influence the individual's opinion, but can be assumed to remain unaffected by the individual's opinion, such as the news media, the Internet, the organizations that the individual is a part of, etc.

For this section, we denote by $x(t) \in [0, 1]$ the individual's opinion at time t and by $b \geq 0$ the individual's bias parameter. Let the individual's weight on his own opinion be $w_{ii} = w$. Let $s \in (0, 1)$ denote the (time-invariant) weighted average of the opinions of all sources in the individual's environment. Then, from [**3**], the individual's opinion at time $t + 1$ is given by

$$x(t+1) = \frac{wx(t) + (x(t))^b s}{w + (x(t))^b s + (1-x(t))^b (1-s)}. \quad [4]$$

Given $s \in (0, 1)$, and $b \neq 1$, we define

$$\hat{x}(s,b) := \frac{s^{1/(1-b)}}{s^{1/(1-b)} + (1-s)^{1/(1-b)}} \quad [5]$$

as the polarization threshold for the individual. We show that when the individual is sufficiently biased (i.e., $b > 1$), the polarization threshold $\hat{x}$ is an unstable equilibrium; i.e., in equilibrium the individual's opinion goes to 1 or 0 depending on whether the initial opinion was greater than or less than $\hat{x}$. On the other hand, when $b < 1$, $\hat{x}$ is a stable equilibrium.

Theorem 1.
Fix $t \geq 0$. Let $x(t) \in (0, 1)$.

1. *If $b > 1$,*
   (a) *if $x(t) > \hat{x}$, then $x(t + 1) > x(t)$, and $x(t) \to 1$ as $t \to \infty$;*
   (b) *if $x(t) < \hat{x}$, then $x(t + 1) < x(t)$, and $x(t) \to 0$ as $t \to \infty$;*
   (c) *if $x(t) = \hat{x}$, then for all $t' > t$, $x(t') = \hat{x}$.*
2. *If $b < 1$,*
   (a) *if $x(t) > \hat{x}$, then $x(t + 1) < x(t)$;*
   (b) *if $x(t) < \hat{x}$, then $x(t + 1) > x(t)$;*
   (c) *$x(t) \to \hat{x}$ as $t \to \infty$.*

The opinion $x(t)$ can be interpreted as the individual's degree of support for the extreme represented by 1. So, the above theorem shows that when the individual is sufficiently biased (i.e., $b > 1$), exposure to the environment pushes him away from the threshold $\hat{x}$ (unless $x(0) = \hat{x}$), and he holds an extreme opinion ($x(t) = 0$ or $x(t) = 1$) in equilibrium. Thus, $\hat{x}$ is an unstable equilibrium. This mathematically captures the biased assimilation behavior observed empirically. On the other hand, if the individual has low bias (i.e., $b < 1$), then he gravitates toward the polarization threshold $\hat{x}$ over time. Thus, $\hat{x}$ is a stable equilibrium in this case. The behavior of the individual when $b = 1$ is a limiting case of the two cases proved in the theorem; as $b \to 1$, $\lim_{t\to\infty} x(t) = \hat{x}$, but $\hat{x}$ goes to 1, $\frac{1}{2}$, or 0 depending on whether s is greater than, equal to, or less than $\frac{1}{2}$. When the individual is connected to other individuals in a social network, we show below that the biased opinion formation process produces polarization even when $b = 1$.

DeGroot's Process Is Not Polarizing

It is easy to see that if DeGroot's process was asynchronous, i.e., individuals update their opinions one at a time, each opinion update can only lower the NDI. However, here we show that each opinion update can only lower the NDI even when individuals update opinions simultaneously. As a result, the repeated averaging process is depolarizing.

Theorem 2. *Consider an arbitrary connected, weighted, undirected graph $G = (V, E, w)$. Let $\mathbf{x} \in [0, 1]^n$ be an arbitrary vector of opinions of nodes in G at time $t \geq 0$. Assume that for all $i \in V$, $b_i = 0$. Then, $\eta(G, \mathbf{x}(t + 1)) \leq \eta(G, \mathbf{x}(t))$; i.e., the network disagreement index at time $t + 1$ is no more than that at time t.*

Our result holds for arbitrary weights w_{ij} and an arbitrary vector of opinions $\mathbf{x} \in [0, 1]^n$, i.e., when the underlying network is arbitrarily homophilous. Moreover, it holds for a number of variants of DeGroot's model that have been proposed to explain the empirically observed lack of consensus on many issues. We defer that discussion to a later section of the paper.

Polarization Due to Biased Assimilation

In this section we illustrate using a simple model of networks with homophily that the biased opinion formation process may result in polarization, persistent disagreement, or consensus depending on how biased the individuals are. We model homophilous networks using a deterministic variant of multitype random networks (33). Multitype random networks are a generalization of Erdös–Rényi random graphs. Nodes in V are partitioned into types, say, $\tau_1, \tau_2, \ldots, \tau_k$. The network is parameterized by a vector $(n_1, \ldots, n_k)$ where n_i is the number of nodes of type τ_i, and a symmetric matrix $P \in [0, 1]^{k\times k}$, where P_{ij} is the probability that there exists an undirected edge between a node of type τ_i and another of type τ_j. The class of multitype random networks where $P_{ii} > P_{ij}$ for all i, j is called the islands model and is used to model homophily (because an individual is more likely to be connected with individuals of the same type). We analyze the biased opinion formation process over a deterministic variant of the islands model, which we call a two-island network.

Definition 4. *Given integers $n_1, n_2 \geq 0$ and real numbers $p_s, p_d \in (0, 1)$, a (n_1, n_2, p_s, p_d)-two-island network is a weighted undirected graph $G = (V_1, V_2, E, w)$, where*

- *$|V_1| = n_1$, $|V_2| = n_2$, and $V_1 \cap V_2 = \emptyset$.*
- *Each node $i \in V_1$ has $n_1 p_s$ neighbors in V_1 and $n_2 p_d$ neighbors in V_2.*
- *Each node $i \in V_2$ has $n_2 p_s$ neighbors in V_2 and $n_1 p_d$ neighbors in V_1.*[†]
- *$p_s > p_d$.*

We define the degree of homophily as follows.

Definition 5. *Let $G = (V_1, V_2, E, w)$ be a (n_1, n_2, p_s, p_d)-two-island network. Then the degree of homophily in G, h_G, is defined to be the ratio p_s/p_d.*

[†]For clarity of exposition, we assume that the quantities $n_1 p_s, n_2 p_s, n_1 p_d$ and $n_2 p_d$ are all integers.

Informally, a high value of h_G implies that nodes in V are much more likely to form edges to other nodes of their own type, thereby exhibiting a high degree of homophily.

Theorem 3. *Let $G = (V_1, V_2, E, w)$ be a (n, n, p_s, p_d)-two-island network. For all $i \in V = V_1 \cup V_2$, let $w_{ii} = 0$. For all $(i, j) \in E$, let $w_{ij} = 1$. Assume for all $i \in V_1$, $x_i(0) = x_0$, where $\frac{1}{2} < x_0 < 1$. Assume for all $i \in V_2$, $x_i(0) = 1 - x_0$. Assume for all $i \in V$, the bias parameter $b_i = b > 0$. Then,*

1. (*Polarization*) *If $b \geq 1$, $\forall i \in V_1$, $\lim_{t\to\infty} x_i(t) = 1$, and $\forall i \in V_2$, $\lim_{t\to\infty} x_i(t) = 0$.*
2. (*Persistent disagreement*) *If $1 > b \geq \frac{2}{h_G+1}$, then there exists a unique $\hat{x} \in \left(\frac{1}{2}, 1\right)$ such that $\forall i \in V_1$, $\lim_{t\to\infty} x_i(t) = \hat{x}$, and $\forall i \in V_2$, $\lim_{t\to\infty} x_i(t) = 1 - \hat{x}$.*
3. (*Consensus*) *If $b < \frac{2}{h_G+1}$, then for all $i \in V$, $\lim_{t\to\infty} x_i(t) = \frac{1}{2}$.*

Let us analyze the implications of this theorem. Let η_∞ be the NDI at equilibrium; i.e., $\eta(G, \mathbf{x}(t)) \to \eta_\infty$ as $t \to \infty$. Then, the above result implies that when $b \geq 1$, $\eta_\infty > \eta(G, \mathbf{x}(0))$; i.e., the biased opinion formation process is polarizing. On the other hand, when individuals are moderately biased [i.e., $1 > b \geq 2/(h_G + 1)$], $\eta_\infty > \eta(G, \mathbf{x}(0))$ if and only if $\mathbf{x}_0 < \hat{\mathbf{x}}$; so the opinion formation process may not be polarizing, but it does not produce consensus either. Finally, when individuals have low bias [i.e., $b < 2/(h_G + 1)$], $\eta_\infty = 0$. So, the opinion formation process is depolarizing. This illustrates the importance of the bias parameter in causing polarization. Also, observe that $b = 1$ corresponds to the urn dynamic described earlier; hence the above result shows that that the urn dynamic causes polarization for an arbitrarily small degree of homophily.

Nonhomogeneous Opinions. Theorem 3 holds in the restrictive setting where initial opinions in each island are homogeneous. However, the biased opinion formation process produces polarization even when initial opinions in each island are not homogeneous. If $b \geq 1$, and the initial opinions of individuals in the two islands are sufficiently far apart relative to the degree of homophily h_G, the equilibrium opinions of individuals in V_1 go to 1 and those in V_2 go to 0 (*SI Appendix*, Theorem 4.1). Admittedly, in this case, the NDI in equilibrium might be lower than the initial NDI depending on the initial distribution of opinions. However, let us consider another natural measure of opinion divergence, namely, the global disagreement index (GDI).

Definition 6. Global Disagreement Index. *Given a vector of opinions $\mathbf{x} \in [0, 1]^n$ of individuals in V, the global disagreement index $\gamma(\mathbf{x})$ is defined as*

$$\gamma(\mathbf{x}) := \sum_{i<j} (x_i - x_j)^2. \qquad [6]$$

The GDI is maximized when half the individuals have opinion 0 and other half 1. So, regardless of the initial distribution of opinions, the biased opinion formation process produces polarization even in this case, if opinion divergence is measured using GDI.

Variants of DeGroot's Process

Our result about DeGroot's process (Theorem 2) in fact holds for a number of variants that are all based on repeated averaging of opinions. Here we discuss some of the variants.

Stubborn Individuals. One variant (34) of DeGroot's model attempts to explain the observed lack of consensus on many issues by allowing some nodes to have an infinite self-weight w_ii. Such nodes are called stubborn individuals. Because our result holds for arbitrary weights, this variant is also depolarizing according to our definition.

Surprising Validators. It has been argued (27) that biased assimilation can be countered by surprising validators: Individuals are more likely to carefully consider disconfirming evidence if it is presented by a source that is otherwise similar to them. An opinion formation process with surprising validators can be viewed as individual i adopting an opinion held by j, if i finds j to be similar to him. This process can be interpreted as the following natural urn dynamic: At each time step, i chooses a neighbor j with probability proportional to w_{ij} and inspects a ball chosen uniformly at random from j's urn. i also inspects a ball uniformly at random from his own urn. If the colors of the two inspected balls match, i inspects another ball chosen uniformly at random from j's urn, adds a ball of the same color to his urn, and discards a ball chosen uniformly at random from his urn.

Observe that conditioned on the colors of the two inspected balls matching, the probability that i adds a RED ball to his urn is $x_j(t)$, which is identical to the corresponding (unconditional) probability in DeGroot's process. In other words, this process is a conditional version of DeGroot's process. Mathematically, the opinion update in this process is given by

$$x_i(t+1) := \frac{w_{ii}x_i(t) + \sum_{j\in N(i)} w_{ij}p_{ij}(t)\,x_j(t)}{w_{ii} + \sum_{j\in N(i)} w_{ij}p_{ij}(t)}, \qquad [7]$$

where the additional term $p_{ij}(t) := (x_i(t))^{b_i}x_j(t) + (1 - x_i(t))^{b_i}(1 - x_j(t))$ corresponds to the probability in the urn model that i finds j to be similar to him. Observe that if we define $w'_{ij}(t) := w_{ij}p_{ij}(t)$ [7] is identical to [2], except that the weights may vary with time. Therefore, like DeGroot's process, each update in the opinion formation process with surprising validators can only lower the NDI, regardless of the value of the bias parameter b_i. This stylized model validates the claim (27, 28) that biased assimilation can be countered with surprising validators.

Flocking Model. The flocking model is a well-known model for decentralized consensus (35) based on repeated averaging. Under this model, at each time step $t \geq 0$, an arbitrary set $S(t) \subseteq V$ of individuals simultaneously updates their opinions to be closer to the average opinion of the set.

Definition 7. Flocking Process. *Let $\epsilon \in [0, 1]$. For $t \geq 0$, let $S(t) \subseteq V$ be an arbitrary set of individuals such that $|S(t)| \geq 2$. Let $s(t) := \frac{1}{|S(t)|}\sum_{i\in S(t)} x_i(t)$ be the average opinions of individuals in $S(t)$. Under the flocking process, the opinion of individual $i \in V$ at time $t + 1$, $x_i(t + 1)$, is given by*

$$x_i(t+1) = \begin{cases} (1-\epsilon)\,x_i(t) + \epsilon s(t), & \text{if } i \in S(t) \\ x_i(t), & \text{otherwise.} \end{cases} \qquad [8]$$

Observe that in the flocking process, there is no notion of an underlying network. Therefore, the GDI (Definition 6) is a natural measure of opinion divergence under this process. Next we show that each opinion update in the flocking process can only lower the GDI.

Theorem 4. *Let $\mathbf{x} \in [0, 1]^n$ be an arbitrary vector of opinions of nodes in V at time $t \geq 0$. Let $\mathbf{x}(t + 1) \in [0, 1]^n$ be the vector of opinions at time $t + 1$ after one step of the flocking process. Then, $\gamma(\mathbf{x}(t + 1)) \leq \gamma(\mathbf{x}(t))$; i.e., the GDI at time $t + 1$ is no more than that at time t.*

A generalization of the GDI is the following: $\sum_{i<j} h(|x_i - x_j|)$, where h is an arbitrary convex function. The flocking process has the property that the vector $\mathbf{x}(t + 1)$ is majorized by $\mathbf{x}(t)$. Therefore, as noted in the proof of Theorem 4, each opinion update of the flocking process is depolarizing under this definition or, more generally, when divergence is defined by any symmetric convex function of $\mathbf{x}$.

Observe that it is possible to assign weights w_{ij} such that a single opinion update in DeGroot's process increases the GDI (or any symmetric convex function of $\mathbf{x}$) because the latter is independent of the weights. However, DeGroot's process converges to

consensus under fairly general conditions (18). Thus, under those conditions, DeGroot's process is depolarizing in equilibrium.

Recommender Systems and Polarization

Recommender systems are widely used on the Internet to present personalized information (e.g., search results, news articles, products) to individuals. This personalization is typically done by algorithms that use an individual's past behavior (e.g., history of browsing and purchases) and that of other individuals that are similar in some way to that individual, to discover items of possible interest to the user. It has been argued (17) that this personalization has an echo-chamber effect where individuals are exposed only to information they agree with, and this ultimately leads to increased polarization. Here we investigate this question: Do recommender systems have a polarizing effect?

We consider the following simple model: Let $G = (V_1, V_2, E)$ be an unweighted undirected bipartite graph. Nodes in V_1 represent individuals. Nodes in V_2 represent items. The items could be books, webpages, news articles, products, etc. For concreteness, we refer to nodes in V_2 as books. For a node $i \in V_1$ and a node $j \in V_2$, an edge $(i, j) \in E$ represents ownership, i.e., individual i owns book j. For our purpose, we define a recommender algorithm as below.

Definition 8. *A recommender algorithm takes as input a bipartite graph $G = (V_1, V_2, E)$ and a node $i \in V_1$ and outputs a node $j \in V_2$.*

Thus, given a graph representing which users own which books and a specific user i, a recommender algorithm outputs a single book j to be recommended to i. We analyze three simple recommender algorithms—SimpleSALSA (Algorithm 1), SimpleICF (Algorithm 2), and SimplePPR (Algorithm 3)—that are similar in spirit to three well-known recommender algorithms from the literature: SALSA (36), Personalized PageRank (29), and item-based collaborative filtering (37), respectively. All three algorithms are based on performing random walks on the graph G. Informally speaking, SimpleSALSA and SimpleICF first choose a random item liked by user i and recommend an item similar to that item, whereas SimplePPR recommends the item that is most relevant to user i on the basis of a PageRank-like score.

We assume that i can buy a book only if it is recommended to him. However, he may choose to reject a recommendation, i.e., to not buy a recommended book. Therefore, i buying a book j requires two steps: The recommender algorithm must recommend j to i, and then i must accept the recommendation.

Because we are interested in analyzing the polarizing effects of recommender systems, we assume that each book in V_2 is labeled either "RED" or "BLUE". These labels are purely for the purpose of analysis; the algorithms we study are agnostic to these labels. For each individual $i \in V_1$, let $x_i \in [0, 1]$ be the fraction of RED books owned by i and $1 - x_i$ be that of BLUE books. Individuals may be biased or unbiased, as we define below.

Definition 9. *Consider a book recommended to an individual $i \in V_1$. We say that i is unbiased if i accepts the recommendation with the same probability independent of whether the book is RED or BLUE. We say that i is biased if*

1. *i accepts the recommendation of a RED book with probability x_i and rejects it with probability $1 - x_i$ and*
2. *i accepts the recommendation of a BLUE book with probability $1 - x_i$ and rejects it with probability x_i.*

Observe that the above definition of an individual i being biased corresponds to the urn dynamic described earlier with $b_i = 1$. For an individual i, the fraction of RED books i owns, x_i, can be viewed as i's opinion in the interval [0, 1], and so a recommender algorithm can be viewed as an opinion formation process. The opinion x_i remains unchanged if i rejects a recommendation. However, if i accepts a recommendation, x_i increases or decreases depending on whether the recommended book was RED or BLUE. Thus, we are interested in the probability that a recommendation was for a RED (or BLUE) book given that i accepted the recommendation. The above probability determines whether a recommender algorithm is polarizing or not.

Algorithm 1. SimpleSALSA

Input: $G = (V_1, V_2, E)$, node $i \in V_1$.
1: Perform a three-step random walk on G starting at i.
2: Let the random walk end at node $j \in V_2$.
Output: j.

Algorithm 2. SimpleICF

Input: $G = (V_1, V_2, E)$, node $i \in V_1$.
Parameter: A large positive integer T.
1: Choose a neighbor k of i uniformly at random.
2: Perform T two-step random walks on G starting at k.
3: For each node $j \in V_2$, let count (j) be the number of random walks that end at node j.
4: Let j^*:= arg $\max_j$ count (j).
Output: j^*.

Definition 10. *Consider a recommender algorithm and an individual $i \in V_1$ that accepts the algorithm's recommendation. The algorithm is polarizing with respect to i if*

1. *when $x_i > \frac{1}{2}$, the probability that the recommended book was RED is greater than x_i, and*
2. *when $x_i < \frac{1}{2}$, the probability that the recommended book was RED is less than x_i.*

Informally speaking, a recommender algorithm is polarizing if it makes a "RED individual" more RED and a "BLUE individual" more BLUE. To analyze the recommender algorithms, we assume a generative model for G, which we describe next.

Generative Model for *G*. Let the number of individuals, $|V_1| = m > 0$. Let the number of books, $|V_2| = 2n$, with $n > 0$ books of each color. We assume that $m = f(n)$; and $\lim_{n\to\infty} f(n) = \infty$. For individual $i \in V_1$, we draw x_i independently from a distribution over [0, 1] with a probability density function $g(\cdot)$. We assume that g is symmetric about $\frac{1}{2}$; i.e., for all $y \in [0, 1]$, $g(y) = g(1 - y)$. This implies that for all $i \in V_1$, $\mathbb{E}[x_i] = \frac{1}{2}$. We assume that the variance of the distribution is strictly positive; i.e., $\mathrm{Var}(x_i) > 0$. For an individual i and a RED book j, there exists an edge $(i, j) \in E$ independently with probability $\frac{x_i k}{n}$, where $0 < k < n$. For an individual i and a BLUE book j, there exists an edge $(i, j) \in E$ independently with probability $\frac{(1-x_i)k}{n}$. So, in expectation, each individual i owns k books, and x_i fraction of them are RED.

For two books $j, j' \in V_2$, let $M_{jj'} := |N(j) \cap N(j')|$ be the number of individuals in V_1 that are neighbors of both j and j' in G. For any two nodes $i, j \in V$, let $\mathbb{P}[i \xrightarrow{\ell} j]$ be the probability that a ℓ-step random walk over G starting at i ends at j. For a node $i \in V_1$ and a node $j \in V_2$, let Z_{ij} be the indicator variable for edge (i, j); i.e., $Z_{ij} = 1$ if $(i, j) \in E$, and $Z_{ij} = 0$ otherwise.

Analysis. Next we prove our results about the polarizing effects of each of the three algorithms. Our results rely on a technical lemma, stated in *SI Appendix*, Lemma 6.1, which invokes the Strong law of large numbers to show that random quantities such as the number of neighbors of a user i or of a book j in the graph G all take their expected values with probability 1 as $n \to \infty$. First we show that SimplePPR (Algorithm 3) is polarizing with respect to i even if i is unbiased.

Theorem 5. *Fix a user $i \in V_1$. In the limit as $n \to \infty$ and as $T \to \infty$, SimplePPR is polarizing with respect to i.*

Next we show that SimpleSALSA (Algorithm 1) and SimpleICF (Algorithm 2) are polarizing only if i is biased.

Theorem 6. *Fix a user $i \in V_1$. In the limit as $n \to \infty$,*

Algorithm 3. SimplePPR

Input: $G = (V_1, V_2, E)$, node $i \in V_1$.
Parameter: A large positive integer T.
1: Perform T three-step random walks on G starting at i.
2: For each node $j \in V_2$, let count (j) be the number of random walks that end at node j.
3: Let j^*:= arg $\max_j$ count (j).
Output: j^*.

1. *SimpleSALSA is polarizing with respect to i if and only if i is biased.*
2. *In the limit as $T \to \infty$, SimpleICF is polarizing with respect to i if and only if i is biased.*

Both SimpleICF and SimpleSALSA first choose a random book owned by i: They choose a RED book with probability x_i and a BLUE book with probability $1 - x_i$. This initial random choice ensures that recommendations are sufficiently diverse; i.e., the book eventually recommended by these algorithms is RED with probability at most x_i when $x_i > \frac{1}{2}$. Recall that our definition of a biased individual in this section corresponds to $b = 1$. However, as we point out in the proof of Theorem 6, both algorithms are polarizing for all $b \geq 1$.

By contrast, the reason why SimplePPR is always polarizing is because the large number of three-step random walks serves to amplify user i's initial preference: If $x_i > \frac{1}{2}$, SimplePPR recommends a RED book is probability 1, and vice versa. Consequently, as we point out in the proof, Theorem 5 holds for all $b \geq 0$.

Analyzing the polarizing effect of recommender algorithms under a setting where the graph G evolves over time is an interesting question that we leave for future work.

Concluding Remarks

In this paper we attempted to explain polarization in society through a model of opinion formation. We showed that DeGroot-like repeated averaging processes can never be polarizing, even if individuals are arbitrarily homophilous. We generalized DeGroot's repeated averaging model to account for biased assimilation. We showed that in a two-island network, our biased opinion formation process results in polarization when individuals are sufficiently biased. In other words, homophily alone, without biased assimilation, is not sufficient to polarize society. We also used biased assimilation to provide insight into the polarizing effects of three recommender algorithms. We showed that SimplePPR, which recommends the item that is most relevant to a user on the basis of a PageRank-like (29) score, is always polarizing. The other two algorithms, which first choose a random item liked by the user and recommend an item similar to that item, are polarizing only if individuals are biased.

Our analysis raises a number of questions that we view as promising directions for further research. For example, are recommender algorithms that produce more relevant recommendations necessarily more polarizing? For certain applications (e.g., ecommerce), polarizing effects may not be an overriding concern. On the other hand, for online social systems designed expressly to facilitate collective decision making regarding complex societal issues, polarization might be a dominant concern. In the case of recommender algorithms for news articles, blogs, etc., there may well be a trade-off between relevance and polarizing effects. An understanding of polarization, its causes, and associated trade-offs is important for designing Internet-based socioeconomic systems.

As a final note, complete proofs of all theorems are presented in *SI Appendix*.*

ACKNOWLEDGMENTS. This research was supported in part by National Science Foundation (NSF) Grants 0904325 and 0947670. D.T.L. was supported in part by a NSF Graduate Research Fellowship under Grant DGE-1147470.

*A preliminary version of this paper was presented at the Eighth Workshop on Internet and Network Economics held at the University of Liverpool (Liverpool, UK), December 9–12, 2012.

1. Poole KT, Rosenthal H (1984) The polarization of american politics. *J Polit* 46(4): 1061–1079.
2. Poole KT, Rosenthal H (1991) Patterns of congressional voting. *Am J Pol Sci* 35(1):228–278.
3. Hill S (2009) Divided we stand: The polarization of American Politics. *Natl Civ Rev* 94(4):3–14.
4. McCarty N, Poole KT, Rosenthal H (2006) *Polarized America: The Dance of Ideology and Unequal Riches* (MIT Press, Cambridge, MA).
5. Schelling TC (1971) Dynamic models of segregation. *J Math Sociol* 1:143–186.
6. Bruch EE, Mare RD (2006) Neighborhood choice and neighborhood change. *Am J Sociol* 112(3):667–709.
7. Brandt C, Immorlica N, Kamath G, Kleinberg R (2012) An analysis of one-dimensional Schelling segregation. *Proc 44th symposium on Theory of Computing, STOC '12*, eds Karloff HJ, Pitassi T (Association for Computing Machinery, New York), pp 789–804.
8. Carter B (April 27, 2009) With rivals ahead, doubts for CNN's middle road. *NY Times*, Business Section, p B1.
9. Carter B (March 29, 2010) CNN fails to stop fall in ratings. *NY Times*, Business Section, p B1.
10. Adamic L, Glance N (2005) The political blogosphere and the 2004 U.S. election: Divided they blog. *In LinkKDD 05: Proc Third International Workshop on Link Discovery*, eds Adibi J, Grobelnik M, Mladenic D, Pantel P (Association for Computing Machinery, New York), pp 36–43.
11. Hargittai E, Gallo J (2007) Cross-ideological discussions among conservative and liberal bloggers. *Public Choice* 134:67–86.
12. Gilbert E, Bergstrom T, Karahalios K (2009) . Blogs are echo chambers: Blogs are echo chambers.*Proceedings of the Hawaii International Conference on System Sciences*, ed Sprague RH, Jr. (IEEE Computer Society, Wasington DC), pp 1–10.
13. Lawrence E, Sides J, Farrell H (2010) Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. *Perspect Polit* 8(1):141–157.
14. Wolfe A (1999) *One Nation, After All: What Americans Really Think About God, Country, Family, Racism, Welfare, Immigration, Homosexuality, Work, The Right, The Left and Each Other* (Penguin Group, New York).
15. Fiorina MP, Abrams SJ, Pope JC (2005) *Culture War? The Myth of a Polarized America* (Pearson Education, New York).
16. Baron RS, et al. (1996) Social corroboration and opinion extremity. *J Exp Soc Psychol* 32(6):537–560.
17. Sunstein CR (2002) *Republic.com* (Princeton Univ Press, Princeton, NJ).
18. DeGroot MH (1974) Reaching a consensus. *J Am Stat Assoc* 69(345):118–121.
19. Friedkin NE, Johnsen EC (1990) Social influence and opinions. *J Math Sociol* 15(3–4): 193–206.
20. Krause U (2000) A discrete nonlinear and non-autonomous model of consensus formation. *Communications in Difference Equations*, eds Elaydi S, Ladas G, Popenda J, Rakowski J (Gordon and Breach Science Publishers, Amsterdam), pp 227–236.
21. Acemoglu D, Como G, Fagnani F, Ozdaglar AE (2010) Opinion fluctuations and disagreement in social networks. arXiv:1009.2653.
22. Bindel D, Kleinberg JM, Oren S (2011) How bad is forming your own opinion? *Proc IEEE Symposium on Foundations of Computer Science*, ed Ostrovsky R (Institute of Electrical and Electronic Engineers Computer Society, Washington DC), pp 57–66.
23. Lord CG, Ross L, Lepper MR (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol* 37(11):2098–2109.
24. Miller AG, McHoskey JW, Bane CM, Dowd TG (1993) The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *J Pers Soc Psychol* 64(4):561–574.
25. Munro GD, et al. (2002) Biased assimilation of sociopolitical arguments: Evaluating the 1996 U.S. presidential debate. *Basic Appl Soc Psych* 24(1):15–26.
26. Taber CS, Lodge M (2006) Motivated skepticism in the evaluation of political beliefs. *Am J Pol Sci* 50(3):755–769.
27. Sunstein C (September 18, 2012) Breaking up the echo. *NY Times*, Op-Ed Section, p A25.
28. Centola D (2011) An experimental study of homophily in the adoption of health behavior. *Science* 334(6060):1269–1272.
29. Page L, Brin S, Motwani R, Winograd T (1999) *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report (Stanford InfoLab, Stanford, CA).
30. Heider F (1946) Attitudes and cognitive organization. *J Psychol* 21(1):107–112.
31. Festinger L (1957) *A Theory of Cognitive Dissonance* (Stanford Univ Press, Stanford, CA).
32. Baldassarri D, Gelman A (2008) Partisans without constraint: Political polarization and trends in American public opinion. *Am J Sociol* 114(2):408–446.
33. Golub B, Jackson MO (2011) How homophily affects the speed of learning and best response dynamics. *Q J Econ* 127(3):1287–1338.
34. Yildiz E, Acemoglu D, Ozdaglar AE, Saberi A, Scaglione A (2011) Discrete opinion dynamics with stubborn agents. *SSRN eLibrary*, in press.
35. Tsitsiklis JN (1984) Problems in decentralized decision making and computation. PhD thesis (Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA).
36. Lempel R, Moran S (2001) Salsa: The stochastic approach for link-structure analysis. *ACM Trans Inf Syst* 19(2):131–160.
37. Linden G, Smith B, York J (2003) Amazon.com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE* 7(1)76–80.

Supplemental Information:
Biased Assimilation, Homophily, and the Dynamics of Polarization

Pranav Dandekar* Ashish Goel† David T. Lee‡

*Department of Management Science & Engineering, Stanford University, Stanford, CA.

†Departments of Management Science & Engineering and (by courtesy) Computer Science, Stanford University, Stanford, CA.

‡Department of Electrical Engineering, Stanford University, Stanford, CA.

We provide supplemental information for the article "Biased Assimilation, Homophily, and the Dynamics of Polarization" submitted to Proceedings of the National Academy of Sciences. This document contains proofs of the theorems stated in the paper. Additionally, we state and prove a less restrictive version of Theorem 3 showing that in two-island networks with *non-homogeneous* opinions, if the initial opinions are sufficiently far apart and if $b \geq 1$, the biased opinion formation process produces polarization.

1 Proof of Theorem 1

Recall that

$$x(t+1) := \frac{wx(t) + (x(t))^b s}{w + (x(t))^b s + (1 - x(t))^b (1 - s)}$$

Equivalently,

$$\frac{x(t+1)}{1 - x(t+1)} = \frac{wx(t) + (x(t))^b s}{w(1 - x(t)) + (1 - x(t))^b (1 - s)} = \frac{w + (x(t))^{b-1} s}{w + (1 - x(t))^{b-1}(1 - s)} \frac{x(t)}{1 - x(t)} \tag{1.1}$$

First we will show that if $x(t) = \hat{x}$, then for all $t' > t, x(t') = \hat{x}$.

Lemma 1.1. *Assume $b \neq 1$. Fix $t \geq 0$. Let $x(t) = \hat{x}$. Then for all $t' > t, x(t') = \hat{x}$.*

Proof. To prove the lemma, it suffices to show that $x(t+1) = x(t) = \hat{x}$. Recall that

$$\hat{x} := \frac{s^{1/(1-b)}}{s^{1/(1-b)} + (1 - s)^{1/(1-b)}}$$

Or equivalently,

$$\left(\frac{\hat{x}}{1 - \hat{x}}\right)^{1-b} = \frac{s}{1 - s}$$

This implies that when $x(t) = \hat{x}$, $x(t)^{b-1} s = (1 - x(t))^{b-1}(1 - s)$. Substituting this in (1.1), we get that

$$\frac{x(t+1)}{1 - x(t+1)} = \frac{x(t)}{1 - x(t)}$$

Or equivalently, $x(t+1) = x(t)$. □

Next we will show that when $b > 1$, $\hat{x}$ is an unstable equilibrium.

Lemma 1.2. *Let $b > 1$. Fix $t \geq 0$.*

1. *If $x(t) > \hat{x}$, then $x(t+1) > x(t)$.*
2. *If $x(t) < \hat{x}$, then $x(t+1) < x(t)$.*

Proof. Again, recall that

$$\left(\frac{\hat{x}}{1 - \hat{x}}\right)^{1-b} = \frac{s}{1 - s}$$

Therefore, if $x(t) > \hat{x}$, it implies that

$$\frac{x(t)}{1 - x(t)} > \frac{\hat{x}}{1 - \hat{x}} \Rightarrow \left(\frac{x(t)}{1 - x(t)}\right)^{1-b} < \left(\frac{\hat{x}}{1 - \hat{x}}\right)^{1-b} = \frac{s}{1 - s} \text{ (since } b > 1)$$

Or equivalently, $(x(t))^{b-1}s > (1-x(t))^{b-1}(1-s)$. Substituting this in (1.1), we get that

$$\frac{x(t+1)}{1-x(t+1)} > \frac{x(t)}{1-x(t)}$$

Or equivalently, $x(t+1) > x(t)$.

By a similar argument, if $x(t) < \hat{x}$, then$(x(t))^{b-1}s < (1-x(t))^{b-1}(1-s)$. Again, substituting this in (1.1), we get that

$$\frac{x(t+1)}{1-x(t+1)} < \frac{x(t)}{1-x(t)}$$

Or equivalently, $x(t+1) < x(t)$. □

Next we will show that when $b > 1$, either $\lim_{t\to\infty} x(t) = 1$ or $\lim_{t\to\infty} x(t) = 0$.

Lemma 1.3. *Let $b > 1$. Fix $t \geq 0$.*

1. *If $x(t) > \hat{x}$, then $\lim_{t\to\infty} x(t) = 1$.*
2. *If $x(t) < \hat{x}$, then $\lim_{t\to\infty} x(t) = 0$.*

Proof. For the proof, we will assume that $x(t) > \hat{x}$ and show that $\lim_{t\to\infty} x(t) = 1$. The case when $x(t) < \hat{x}$ can be argued in an analogous way.

By definition, we know that for all $t \geq 0, x(t) \in [0,1]$. Further, from Lemma 1.2, we know that the sequence $\{x(t')_{t'\geq t}\}$ is strictly increasing. Since the sequence is strictly increasing and bounded, it must converge either to 1 or to some value in the interval $[x(t), 1)$. Consider the function $g : [0,1] \to \mathbb{R}$ defined as

$$g(y) := \frac{w + y^b s}{w + y^b s + (1-y)^b(1-s)} - y$$

Observe that for all $t \geq 0,\ x(t+1) - x(t) = g(x(t))$. Therefore,

(a) for all $y \in [x(t), 1),\ g(y) > 0$ (since, by Lemma 1.2, the sequence $\{x(t')_{t'\to t}\}$ is strictly increasing), and

(b) $g(1) = 0$.

For the purpose of contradiction, assume that $\lim_{t\to\infty} x(t) = a$, where $x(t) \leq a < 1$. This implies, for every $\epsilon > 0$, there exists a $t(\epsilon)$ such that for all $t' \geq t(\epsilon),\ x(t'+1) - x(t') < \epsilon$, or equivalently, that for all $t' \geq t(\epsilon),\ g(x(t')) < \epsilon$.

Let $\min_{y\in[x(t),a]} g(y) = c$. It implies for all $y \in [x(t), a],\ g(y) \geq c$. From (a), it follows that $c > 0$. Setting $\epsilon = c$, our analysis implies the following two properties of g: (1) for all $t \geq 0, g(x(t)) \geq c$, and (2) for all $t' \geq t(\epsilon), g(x(t')) < c$, which contradict each other. This completes the proof by contradiction. □

Using a similar argument we can show that when $b < 1$, $\hat{x}$ is a stable equilibrium.

Lemma 1.4. *Let $b < 1$. Fix $t \geq 0$.*

1. *If $x(t) > \hat{x}$, then $x(t+1) < x(t)$.*
2. *If $x(t) < \hat{x}$, then $x(t+1) > x(t)$.*

Lemma 1.5. *Let $b < 1$. Then, $\lim_{t\to\infty} x(t) = \hat{x}$.*

2 Proof of Theorem 2

Recall that since $b_i = 0$, the opinion of node i at time $t+1$ is given by

$$x_i(t+1) = \frac{w_{ii}x_i(t) + \sum_{j \in N(i)} w_{ij}x_j(t)}{w_{ii} + d_i} \tag{2.1}$$

where recall that $d_i := \sum_{j \in N(i)} w_{ij}$ is the weighted degree of node i. Let L_G be the weighted laplacian matrix of G. Recall that L_G is given by

$$(L_G)_{ij} = \begin{cases} d_i, & \text{if } i = j \\ -w_{ij}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Now consider the vector $L_G\mathbf{x}(t)$. The ith entry of the vector is given by

$$\begin{aligned}(L_G\mathbf{x}(t))_i = d_i x_i(t) - \sum_{j \in N(i)} w_{ij}x_j(t) &= d_i x_i(t) + w_{ii}x_i(t) - \left(w_{ii}x_i(t) + \sum_{j \in N(i)} w_{ij}x_j(t) \right) \\ &= (d_i + w_{ii})(x_i(t) - x_i(t+1)) \text{ (from (2.1))}\end{aligned}$$

Equivalently, in matrix notation,

$$\mathbf{x}(t+1) = (I - DL_G)\mathbf{x}(t) \tag{2.2}$$

where, D is a diagonal matrix such that $D_{ii} = 1/(d_i + w_{ii})$. Note that since G is connected, $d_i > 0$, and therefore D_{ii} is finite. Consider the difference $\eta(G, \mathbf{x}(t+1)) - \eta(G, \mathbf{x}(t))$. Observe that for a vector $\mathbf{y} \in [0,1]^n$, $\eta(G, \mathbf{y}) = \mathbf{y}^\top L_G \mathbf{y}$. Therefore, we have that

$$\begin{aligned}\eta(G, \mathbf{x}(t+1)) - \eta(G, \mathbf{x}(t)) &= (\mathbf{x}(t+1))^\top L_G(\mathbf{x}(t+1)) - (\mathbf{x}(t))^\top L_G \mathbf{x}(t) \\ &= (\mathbf{x}(t))^\top (I - DL_G)^\top L_G (I - DL_G)\mathbf{x}(t) - (\mathbf{x}(t))^\top L_G \mathbf{x}(t) \text{ (from (2.2))} \\ &= (\mathbf{x}(t))^\top \left((L_G - L_G D L_G)(I - DL_G) - L_G \right) \mathbf{x}(t) \text{ (since } L_G \text{ is symmetric)} \\ &= (\mathbf{x}(t))^\top \left(L_G - L_G D L_G - L_G D L_G - L_G D L_G D L_G - L_G \right) \mathbf{x}(t) \\ &= (\mathbf{x}(t))^\top \left(L_G D L_G D L_G - 2 L_G D L_G \right) \mathbf{x}(t) \\ &= (\mathbf{x}(t))^\top L_G^\top D^{1/2} ((D^{1/2} L_G D^{1/2} - 2I)) D^{1/2} L_G \mathbf{x}(t) \text{ (since } L_G \text{ is symmetric)} \\ &= \mathbf{y}^\top (D^{1/2} L_G D^{1/2} - 2I)\mathbf{y} \text{ (where } \mathbf{y} := D^{1/2} L_G \mathbf{x}(t))\end{aligned}$$

Thus, in order to show that $\eta(G, \mathbf{x}(t+1)) - \eta(G, \mathbf{x}(t)) \leq 0$, it suffices to show that for all vectors $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{y}^\top D^{1/2} L_G D^{1/2} \mathbf{y} \leq 2||\mathbf{y}||_2^2$. We prove this as Lemma 2.1.

Lemma 2.1. *Consider an arbitrary weighted undirected graph $G = (V, E, w)$ over n nodes. Let L_G be the weighted laplacian matrix of G. Let D be an $n \times n$ diagonal matrix such that for $i = 1, \ldots, n$, $D_{ii} = 1/(d_i + w_{ii})$, where $d_i = \sum_{j \in N(i)} w_{ij}$ is the weighted degree of i in G. Let $\mathbf{y} \in \mathbb{R}^n$ be an arbitrary vector. Then, $\mathbf{y}^\top D^{1/2} L_G D^{1/2} \mathbf{y} \leq 2||\mathbf{y}||_2^2$.*

Proof. For $i = 1, \ldots, n$, let $r_i := d_i + w_{ii}$. Let $P := D^{1/2} L_G D^{1/2}$. Then,

$$P_{ij} = \begin{cases} \frac{d_i}{r_i}, & i = j \\ \frac{-w_{ij}}{\sqrt{r_i r_j}}, & (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Then, we have that

$$\begin{aligned}
\mathbf{y}^\top P\mathbf{y} = \sum_{i,j} P_{ij}y_iy_j &= \sum_{i=1}^{n} P_{ii}y_i^2 + 2\sum_{(i,j)\in E} P_{ij}y_iy_j = \sum_i \frac{d_i}{r_i}y_i^2 - 2\sum_{(i,j)\in E}\frac{w_{ij}}{\sqrt{r_ir_j}}y_iy_j \\
&= \sum_i \left(\frac{1}{r_i}y_i^2 \sum_{j\in N(i)} w_{ij}\right) - 2\sum_{(i,j)\in E}\frac{w_{ij}}{\sqrt{r_ir_j}}y_iy_j \\
&= \sum_{(i,j)\in E} w_{ij}\left(\frac{y_i^2}{r_i} + \frac{y_j^2}{r_j}\right) - 2\sum_{(i,j)\in E}\frac{w_{ij}}{\sqrt{r_ir_j}}y_iy_j \\
&= \sum_{(i,j)\in E} w_{ij}\left(\frac{y_i}{\sqrt{r_i}} - \frac{y_j}{\sqrt{r_j}}\right)^2 \\
&= -\sum_{(i,j)\in E} w_{ij}\left(\frac{y_i}{\sqrt{r_i}} + \frac{y_j}{\sqrt{r_j}}\right)^2 + 2\sum_i \frac{d_i}{r_i}y_i^2 \\
&\le -\sum_{(i,j)\in E} w_{ij}\left(\frac{y_i}{\sqrt{r_i}} + \frac{y_j}{\sqrt{r_j}}\right)^2 + 2\sum_i y_i^2 \text{ (since } d_i \le r_i) \\
&\le 2||\mathbf{y}||_2^2
\end{aligned}$$

□

3 Proof of Theorem 3

To prove the theorem, we begin by making three simple observations that hold for all $b \ge 0$. The first observation follows directly from the symmetry of nodes in each set V_1 and V_2.

Lemma 3.1. *Consider nodes $i, j \in V$ such that either both $i, j \in V_1$ or both $i, j \in V_2$. Then for all $t \ge 0$, $x_i(t) = x_j(t)$.*

The next observation allows us to focus on only analyzing the equilibrium opinion of nodes in V_1.

Lemma 3.2. *Consider a node $i \in V_1$ and a node $j \in V_2$. Then, for all $t \ge 0$, $x_i(t) = 1 - x_j(t)$.*

Proof of Lemma 3.2. By induction.

Induction hypothesis: Assume that the statement holds for some $t \ge 0$.

Base case: The statement holds for $t = 0$ by assumption in the theorem statement.

We will now show that the statement holds for $t + 1$.

$$\frac{x_i(t+1)}{1 - x_i(t+1)} = \frac{(x_i(t))^b}{(1 - x_i(t))^b}\frac{s_i(t)}{d_i - s_i(t)} \tag{3.1}$$

where $d_i = n(p_s + p_d)$ and, by Lemma 3.1, $s_i(t) = n(p_s x_i(t) + p_d x_j(t))$. On the other hand,

$$\frac{x_j(t+1)}{1 - x_j(t+1)} = \frac{(x_j(t))^b}{(1 - x_j(t))^b}\frac{s_j(t)}{d_j - s_j(t)} \tag{3.2}$$

where $s_j(t) = n(p_s x_j(t) + p_d x_i(t))$, and $d_j = n(p_s + p_d) = d_i$. By the induction hypothesis, we know that $x_i(t) = 1 - x_j(t)$. It follows that $S_i(t) = d_i - s_j(t)$. Substituting this into (3.1), we get

$$\frac{x_i(t+1)}{1-x_i(t+1)} = \frac{(x_i(t))^b}{(1-x_i(t))^b}\frac{s_i(t)}{d_i - s_i(t)} = \frac{(1-x_j(t))^b}{(x_j(t))^b}\frac{d_j - s_j(t)}{s_j(t)} = \frac{1-x_j(t+1)}{x_j(t+1)}$$

where the last equality follows from (3.2). It follows that $x_i(t+1) = 1 - x_j(t+1)$.

This completes the inductive proof. □

Lemma 3.2 implies that if we prove the theorem statement for nodes in V_1, we get the proof for nodes in V_2 for free. So, in the rest of the proof, we only make statements about nodes in V_1. The third observation lower bounds the opinions of nodes in V_1.

Lemma 3.3. *Consider a node $i \in V_1$. For all $t \geq 0$, $x_i(t) \in [\frac{1}{2}, 1]$.*

Proof of Lemma 3.3. It is easy to see that for all $t \geq 0$, $x_i(t) \leq 1$. We will prove that $x_i(t) \geq \frac{1}{2}$ by induction over t.

Base case: The statement holds for $t = 0$ by assumption in the theorem statement.

Induction hypothesis: Assume that the lemma statement holds for some $t \geq 0$, *i.e.*, assume that $x_i(t) \geq \frac{1}{2}$ for some $t \geq 0$.

We will show that the lemma statement holds for $t+1$.

$$\begin{aligned}
\frac{x_i(t+1)}{1-x_i(t+1)} &= \frac{(x_i(t))^b}{(1-x_i(t))^b}\frac{S_i(t)}{d_i - s_i(t)} \\
&\geq \frac{(x_i(t))^b}{(1-x_i(t))^b} \text{ (since } s_i(t) > d_i - s_i(t)) \\
&\geq 1 \text{ (since } x_i(t) \geq \frac{1}{2} \text{ by the induction hypothesis, and } b \geq 0)
\end{aligned}$$

This implies $x_i(t+1) \geq \frac{1}{2}$, completing the inductive proof. □

Recall that i's opinion at time $t+1$ is given by

$$x_i(t+1) = \frac{(x_i(t))^b s_i(t)}{(x_i(t))^b s_i(t) + (1-x_i(t))^b (d_i - s_i(t))}$$

where $s_i(t) = n(p_s x_i(t) + p_d(1 - x_i(t)))$, and $d_i = n(p_s + p_d)$. Now consider the equation

$$x_i(t+1) = x_i(t) \tag{3.3}$$

We will show that if $b \geq 1$ or $b < \frac{2}{h_G+1}$, (3.3) has no solution in $(\frac{1}{2}, 1)$, whereas if $1 > b \geq \frac{2}{h_G+1}$, there exists a unique solution to (3.3) in $(\frac{1}{2}, 1)$.

Lemma 3.4. *Consider a node $i \in V_1$. Fix $t \geq 0$.*

(a) If $b \geq 1$, for every $x_i(t) \in (\frac{1}{2}, 1)$, $x_i(t+1) > x_i(t)$.

(b) If $1 > b \geq \frac{2}{h_G+1}$, there exists a unique solution, say $\hat{x}$, to Eq.(3.3) in $(\frac{1}{2}, 1)$.

(c) If $b < \frac{2}{h_G+1}$, for every $x_i(t) \in (\frac{1}{2}, 1)$, $x_i(t+1) < x_i(t)$.

Proof of Lemma 3.4. Consider the function $f : [0,1] \to \mathbb{R}$ defined as

$$f(y;b) := \begin{cases} 1, & y \in [0,1], b = 1 \\ 0, & y \in [0,1], b = 2 \\ \frac{2}{b} - 1, & y = \frac{1}{2}, b > 0 \\ \frac{(y)^{2-b} - (1-y)^{2-b}}{y(1-y)^{1-b} - y^{1-b}(1-y)}, & \text{otherwise} \end{cases} \tag{3.4}$$

We will first prove a few properties of f and then use those properties to prove Lemma 3.4.

Proposition 3.1. *1. For all $b > 0$, f is continuous over $[0,1]$.*

2. If $0 < b < 1$, f is strictly increasing over $[\frac{1}{2}, 1]$.

3. If $b \geq 1$, for all $y \in [0,1)$, $f(y;b) \leq 1$.

Proof. 1. Observe that f is continuous when $b = 1$ or $b = 2$. So, we only need to show that f is continuous at $y = \frac{1}{2}$ when $b \neq 1$ and $b \neq 2$. Let $p(y;b) := (y)^{2-b} - (1-y)^{2-b}$ and $q(y;b) := y(1-y)^{1-b} - y^{1-b}(1-y)$. Observe that when $b \neq 1$ and $b \neq 2$, both p and q are differentiable on $[0,1]$. For $y \in [0,1]$,

$$p'(y;b) = (2-b)(y^{1-b} + (1-y)^{1-b}); q'(y;b) = (1-y)^{1-b} - (1-b)y(1-y)^{-b} - (1-b)y^{-b}(1-y) + y^{1-b}$$

Therefore,

$$\lim_{y \to 1/2} \frac{p'(y;b)}{q'(y;b)} = \lim_{y \to 1/2} \frac{(2-b)(y^{1-b} + (1-y)^{1-b})}{(1-y)^{1-b} - (1-b)y(1-y)^{-b} - (1-b)y^{-b}(1-y) + y^{1-b}} = \frac{2}{b} - 1 \tag{3.5}$$

So, we have that

$$\lim_{y \to 1/2} f(y;b) = \lim_{y \to 1/2} \frac{p(y;b)}{q(y;b)} = \lim_{y \to 1/2} \frac{p'(y)}{q'(y)} \text{ (using L'Hôpital's rule)} = \frac{2}{b} - 1 \text{ (from (3.5))} = f(\frac{1}{2}; b)$$

Therefore, when $b \neq 1$ and $b \neq 2$, f is continuous at $\frac{1}{2}$.

2. Assume $0 < b < 1$. Fix $y_1, y_2 \in [\frac{1}{2}, 1]$ such that $y_1 > y_2$. We will show that $f(y_1;b) > f(y_2;b)$. For conciseness of expression, define $\bar{y_1} := 1 - y_1$ and $\bar{y_2} := 1 - y_2$. Then

$$y_1 y_2 - y_1 \bar{y_2} > (y_1 y_2)^{1-b} - (y_1 \bar{y_2})^{1-b} \tag{3.6}$$

Similarly,

$$\bar{y_1} y_2 - \bar{y_1} \bar{y_2} > (\bar{y_1} y_2)^{1-b} - (\bar{y_1} \bar{y_2})^{1-b} \tag{3.7}$$

Adding (3.6) and (3.7), we get

$$y_1 y_2 - y_1 \bar{y_2} + \bar{y_1} y_2 - \bar{y_1} \bar{y_2} > (y_1 y_2)^{1-b} - (y_1 \bar{y_2})^{1-b} + (\bar{y_1} y_2)^{1-b} - (\bar{y_1} \bar{y_2})^{1-b}$$

Or equivalently,

$$(y_1 y_2 - \bar{y_1} \bar{y_2}) - \left((y_1 y_2)^{1-b} - (\bar{y_1} \bar{y_2})^{1-b}\right) > (y_1 \bar{y_2} - \bar{y_1} y_2) - \left((y_1 \bar{y_2})^{1-b} - (\bar{y_1} y_2)^{1-b}\right) \tag{3.8}$$

Moreover, since $y_1, y_2 \in [\frac{1}{2}, 1]$ and $y_1 > y_2$,

$$y_1 y_2 - \bar{y_1} \bar{y_2} > 0; (y_1 y_2)^{1-b} - (\bar{y_1} \bar{y_2})^{1-b} > 0; y_1 \bar{y_2} - \bar{y_1} y_2 > 0; (y_1 \bar{y_2})^{1-b} - (\bar{y_1} y_2)^{1-b} > 0 \tag{3.9}$$

(3.8) and (3.9) imply that

$$\frac{y_1y_2 - \bar{y_1}\bar{y_2}}{y_1\bar{y_2} - \bar{y_1}y_2} > \frac{(y_1y_2)^{1-b} - (\bar{y_1}\bar{y_2})^{1-b}}{(y_1\bar{y_2})^{1-b} - (\bar{y_1}y_2)^{1-b}}$$

Rearranging, we get

$$\frac{(y_1)^{2-b} - \bar{y_1}^{2-b}}{y_1\bar{y_1}^{1-b} - y_1^{1-b}\bar{y_1}} = f(y_1; b) > \frac{(y_2)^{2-b} - \bar{y_2}^{2-b}}{y_2\bar{y_2}^{1-b} - y_2^{1-b}\bar{y_2}} = f(y_2; b)$$

3. Since f is symmetric about $y = \frac{1}{2}$, we will prove the theorem for $y \in [\frac{1}{2}, 1)$. Fix $y \in [\frac{1}{2}, 1)$. Observe that when $b \geq 1$, $(1-y)^{1-b} \geq y^{1-b}$ (since $y \geq 1-y$). Equivalently

$$y(1-y)^{1-b} \geq y^{2-b} \tag{3.10}$$

For the same reason,

$$y^{1-b}(1-y) \leq (1-y)^{2-b} \tag{3.11}$$

From (3.10) and (3.11), it follows that

$$y(1-y)^{1-b} - y^{1-b}(1-y) \geq (y)^{2-b} - (1-y)^{2-b}$$

or equivalently, $f(y; b) \leq 1$.

□

Using these properties of f we will prove Lemma 3.4.

1. If $b \geq 1$, then for all $y \in [0, 1)$, $f(y; b) \leq 1$ (by Proposition 3.1) $< h_G$. Therefore, for $y \in [\frac{1}{2}, 1)$,

$$\begin{aligned}
\frac{(y)^{2-b} - (1-y)^{2-b}}{y(1-y)^{1-b} - y^{1-b}(1-y)} &< h_G \\
\Leftrightarrow y^{2-b} - (1-y)^{2-b} &< h_G(y(1-y)^{1-b} - y^{1-b}(1-y)) \\
\Leftrightarrow y^{2-b} + h_G y^{1-b}(1-y) &< (1-y)^{2-b} + h_G y(1-y)^{1-b} \\
\Leftrightarrow y^{1-b}(y + (1-y)h_G) &< (1-y)^{1-b}((1-y) + h_G y) \\
\Leftrightarrow \frac{y}{1-y} &< \left(\frac{y}{1-y}\right)^b \cdot \frac{(1-y) + h_G y}{y + (1-y)h_G}
\end{aligned}$$

For $y = x_i(t)$, the right hand side of the last inequality above is equal to $x_i(t+1)/(1-x_i(t+1))$, implying that $x_i(t+1) > x_i(t)$.

2. If $1 > b \geq \frac{2}{h_G+1}$, then observe that $f(\frac{1}{2}; b) = \frac{2}{b} - 1 \leq h_G < f(1; b) = \infty$. Since f is a continuous function (by Proposition 3.1), therefore, by the intermediate value theorem, there must exist a $\hat{y} \in [\frac{1}{2}, 1)$ such that $f(\hat{y}; b) = h_G$. Equivalently,

$$\frac{(\hat{y})^{2-b} - (1-\hat{y})^{2-b}}{\hat{y}(1-\hat{y})^{1-b} - \hat{y}^{1-b}(1-\hat{y})} = h_G$$

Rearranging the above expression, we get

$$\frac{\hat{y}}{1-\hat{y}} = \left(\frac{\hat{y}}{1-\hat{y}}\right)^b \cdot \frac{(1-\hat{y}) + h_G\hat{y}}{\hat{y} + (1-\hat{y})h_G}$$

Again, for $\hat{y} = x_i(t)$, we have that $x_i(t+1) = x_i(t)$. The uniqueness of $\hat{x}$ follows from the fact that, by Proposition 3.1, f is strictly increasing over $(\frac{1}{2}, 1]$.

3. If $b < \frac{2}{h_G+1}$, then for all $y \in [\frac{1}{2}, 1]$, $f(y;b) \geq f(\frac{1}{2};b)$ (by Proposition 3.1) $= \frac{2}{b} - 1 > h_G$. In other words,

$$\frac{(y)^{2-b} - (1-y)^{2-b}}{y(1-y)^{1-b} - y^{1-b}(1-y)} > h_G$$

Again, rearranging the above expression, we get

$$\frac{y}{1-y} > \left(\frac{y}{1-y}\right)^b \cdot \frac{(1-y) + h_G y}{y + (1-y)h_G}$$

Again, for $y = x_i(t)$, the right hand side of the last inequality above is equal to $x_i(t+1)$, implying that $x_i(t+1) < x_i(t)$.

This concludes the proof of Lemma 3.4. □

Next we will prove Theorem 3 for the case of persistent disagreement, the cases of polarization and consensus are limiting cases of that case as $b \to 1$ and $b \to 2/(h_G+1)$ respectively. We will show that when $1 > b \geq \frac{2}{h_G+1}$, the value $\hat{x}$ defined in Lemma 3.4(b) is a stable equilibrium. The other two cases can be formally proven using an argument similar to the one below. Next we will show that when $1 > b \geq \frac{2}{h_G+1}$, the sequence $\{x_i(t)\}$ is bounded.

Lemma 3.5. *Consider a node $i \in V_1$. Let $1 > b \geq \frac{2}{h_G+1}$. Let $\hat{x} \in (\frac{1}{2}, 1)$ be the solution to (3.3).*

1. *If $x_0 < \hat{x}$, then for all $t > 0$, $x_i(t) < \hat{x}$.*
2. *If $x_0 > \hat{x}$, then for all $t > 0$, $x_i(t) > \hat{x}$.*

Proof of Lemma 3.5. We will prove statement (1). Statement (2) can be proven using a similar argument.

Proof by induction.

Induction hypothesis: Assume that the lemma statement holds for some $t \geq 0$, *i.e.*, assume that $x_i(t) < \hat{x}$ for some $t \geq 0$.

Base case: The statement holds for $t = 0$ by assumption.

We will show that the lemma statement holds for $t+1$.

$$\frac{x_i(t+1)}{1 - x_i(t+1)} = \frac{(x_i(t))^b}{(1-x_i(t))^b} \frac{s_i(t)}{d_i - s_i(t)} < \frac{(\hat{x})^b}{(1-\hat{x})^b} \frac{s_i(t)}{d_i - s_i(t)} \text{ (since } \frac{1}{2} < x_i(t) < \hat{x}, \text{ and } b > 0)$$

Observe that since $x_i(t) < \hat{x}$ and $p_s > p_d$, $s_i(t) = n(p_s x_i(t) + p_d(1 - x_i(t))) < n(p_s\hat{x} + p_d(1-\hat{x}))$. Therefore,

$$\frac{s_i(t)}{d_i - s_i(t)} < \frac{p_s\hat{x} + p_d(1-\hat{x})}{p_s(1-\hat{x}) + p_d\hat{x}}$$

As a result,

$$\frac{x_i(t+1)}{1 - x_i(t+1)} < \frac{(\hat{x})^b}{(1-\hat{x})^b} \frac{p_s\hat{x} + p_d(1-\hat{x})}{p_s(1-\hat{x}) + p_d\hat{x}} = \frac{\hat{x}}{1-\hat{x}} \text{ (by definition of } \hat{x})$$

This implies $x_i(t+1) < \hat{x}$. This completes the inductive proof. □

Next we will prove that when $1 > b \geq \frac{2}{h_G+1}$, the sequence $\{x_i(t)\}$ is monotone.

Lemma 3.6. *Consider a node $i \in V_1$. Let $1 > b \geq \frac{2}{h_G+1}$. Let $\hat{x} \in (\frac{1}{2}, 1)$ be the solution to (3.3).*

1. *If $x_0 < \hat{x}$, the sequence $\{x_i(t)\}$ is strictly increasing.*

2. *If $x_0 > \hat{x}$, the sequence $\{x_i(t)\}$ is strictly decreasing.*

Proof of Lemma 3.6. We will prove statement (1); statement (2) can be proven using a similar argument.

Assume $x_0 < \hat{x}$. Then, from Lemma 3.5, we know that for all $t \geq 0, x_i(t) < \hat{x}$. Fix $t \geq 0$. Let $x_i(t) = y < \hat{x}$. Recall that by definition of $\hat{x}$, if $x_i(t) = \hat{x}$, $x_i(t+1) = x_i(t)$. Equivalently, $f(\hat{x}; b) = h_G$, where f is defined by (3.4). From Proposition 3.1, we know that f is strictly increasing over the interval $(\frac{1}{2}, \hat{x})$. Therefore, $f(y; b) < f(\hat{x}; b) = h_G$. Equivalently,

$$\frac{(y)^{2-b} - (1-y)^{2-b}}{y(1-y)^{1-b} - y^{1-b}(1-y)} < h_G$$

Rearranging, we get

$$\frac{y}{1-y} < \left(\frac{y}{1-y}\right)^b \cdot \frac{(1-y) + h_G y}{y + (1-y)h_G} = \frac{x_i(t+1)}{1 - x_i(t+1)}$$

Equivalently, $x_i(t+1) > x_i(t)$. □

Using the fact that the sequence $\{x_i(t)\}$ is monotone and bounded, next we will prove that it converges to $\hat{x}$.

Lemma 3.7. *Consider a node $i \in V_1$. Let $1 > b \geq \frac{2}{h_G+1}$. Let $\hat{x} \in (\frac{1}{2}, 1)$ be the solution to (3.3). Then, $\lim_{t\to\infty} x_i(t) = \hat{x}$.*

Proof. For the proof, we will assume that the initial opinion $x_i(0) = x_0 \leq \hat{x}$. The case when $x_0 > \hat{x}$ can be argued in an analogous way.

Observe that if $x_0 = \hat{x}$, then by Lemma 3.4, it follows that for all $t \geq 0$, $x_i(t+1) = \hat{x}$, and we are done. So let us assume that $\frac{1}{2} < x_0 < \hat{x}$. From Lemma 3.5 and Lemma 3.6, we know that the sequence $\{x_i(t)\}$ is strictly increasing and bounded. This implies that the sequence must converge either to $\hat{x}$ or to some value in the interval $[x_0, \hat{x})$. Consider the function $g : [0, 1] \to \mathbb{R}$ defined as

$$g(y) := \frac{y^b(h_G y + (1-y))}{y^b(h_G y + (1-y) + (1-y)^b(h_G(1-y) + y)} - y$$

Observe that for all $t \geq 0$, $x_i(t+1) - x_i(t) = g(x_i(t))$. Therefore,

(a) for all $y \in (\frac{1}{2}, \hat{x})$, $g(y) > 0$ (since, by Lemma 3.6, the sequence $\{x_i(t)\}$ is strictly increasing), and

(b) $g(\hat{x}) = 0$ (by definition of $\hat{x}$).

For the purpose of contradiction, assume that $\lim_{t\to\infty} x_i(t) = a$, where $x_0 \leq a < \hat{x}$. This implies, for every $\epsilon > 0$, there exists a $t(\epsilon)$ such that for all $t \geq t(\epsilon)$, $x_i(t+1) - x_i(t) < \epsilon$, or equivalently, that for all $t \geq t(\epsilon)$, $g(x_i(t)) < \epsilon$.

Let $\min_{y\in[x_0,a]} g(y) = c$. It implies for all $y \in [x_0, a]$, $g(y) \geq c$. From (a), it follows that $c > 0$. Setting $\epsilon = c$, our analysis implies the following two properties of g: (1) for all $t \geq 0, g(x_i(t)) \geq c$, and (2) for all $t \geq t(\epsilon), g(x_i(t)) < c$, which contradict each other. This completes the proof by contradiction. □

This completes the proof of Theorem 3.

4 Two-island Networks with Non-homogeneous Opinions

In this section, we prove a less restrictive version of the polarization result in Theorem 3, which does not require that the initial opinions in each island be homogeneous. We show that in a two-island network, if the bias parameter $b \geq 1$ and the initial opinions of the two islands are sufficiently far apart relative to the homophily index h_G, then the biased opinion formation process results in polarization.

Theorem 4.1. *Let $G = (V_1, V_2, E, w)$ be a (n, n, p_s, p_d)-two island network. For all $(i,j) \in E$, let $w_{ij} = 1$. Fix $\epsilon \in (0, \frac{1}{2}]$. Assume for all $i \in V_1$, $x_i(0) \geq \frac{1}{2} + \epsilon$ and for all $i \in V_2$, $x_i(0) \leq \frac{1}{2} - \epsilon$. Assume for all $i \in V$, the bias parameter $b_i = b \geq 1$. Then, if $\epsilon > \frac{1}{2h_G}$, for all $i \in V_1$, $\lim_{t\to\infty} x_i(t) = 1$, and for all $i \in V_2$, $\lim_{t\to\infty} x_i(t) = 0$.*

Proof. We will show that the opinions of individuals in V_1 are strictly increasing whereas that of individuals in V_2 are strictly decreasing.

Lemma 4.1. *Fix $t \geq 0$. Then,*

1. *For all $i \in V_1$, if $x_i(t) \in [\frac{1}{2} + \epsilon, 1)$, then $x_i(t+1) > x_i(t)$.*
2. *For all $i \in V_2$, if $x_i(t) \in (0, \frac{1}{2} - \epsilon]$, then $x_i(t+1) < x_i(t)$.*

Proof. We will prove Statement 1 of the lemma. Statement 2 can be proven using an analogous argument. Fix an individual $i \in V_1$.

$$\begin{aligned}
\frac{x_i(t+1)}{1 - x_i(t+1)} &= \frac{w_{ii}x_i(t) + x_i(t)^b s_i(t)}{w_{ii}(1 - x_i(t)) + (1 - x_i(t))^b (d_i - s_i(t))} \\
&= \frac{w_{ii}x_i(t) + x_i(t)^b \left(\sum_{j \in N(i) \cap V_1} x_j(t) + \sum_{j \in N(i) \cap V_2} x_j(t)\right)}{w_{ii}(1 - x_i(t)) + (1 - x_i(t))^b \left(\sum_{j \in N(i) \cap V_1} (1 - x_j(t)) + \sum_{j \in N(i) \cap V_2} (1 - x_j(t))\right)}
\end{aligned}$$

Observe that $\sum_{j \in N(i) \cap V_1} x_j(t) \geq np_s\left(\frac{1}{2} + \epsilon\right)$ and $\sum_{j \in N(i) \cap V_2} (1 - x_j(t)) \leq np_d$. Therefore,

$$\begin{aligned}
\frac{x_i(t+1)}{1 - x_i(t+1)} &\geq \frac{w_{ii}x_i(t) + x_i(t)^b \left(np_s\left(\frac{1}{2} + \epsilon\right) + 0\right)}{w_{ii}(1 - x_i(t)) + (1 - x_i(t))^b \left(np_s\left(\frac{1}{2} - \epsilon\right) + np_d\right)} \\
&= \frac{w_{ii}x_i(t) + x_i(t)^b \left(\frac{1}{2} + \epsilon\right)}{w_{ii}(1 - x_i(t)) + (1 - x_i(t))^b \left(\frac{1}{2} - \epsilon + \frac{1}{h_G}\right)} \\
&> \frac{w_{ii}x_i(t) + x_i(t)^b}{w_{ii}(1 - x_i(t)) + (1 - x_i(t))^b} \text{ (since } \epsilon > \frac{1}{2h_G}) \\
&> \frac{x_i(t)}{1 - x_i(t)} \text{ (since } x_i(t) > \frac{1}{2} \text{ and } b \geq 1)
\end{aligned}$$

Or equivalently, $x_i(t+1) > x_i(t)$. □

Next we will show that for an individual $i \in V_1$, $x_i(t) \in [\frac{1}{2} + \epsilon, 1]$ for all $t \geq 0$, and for an individual $i \in V_2$, $x_i(t) \in [0, \frac{1}{2} - \epsilon]$ for all $t \geq 0$.

Lemma 4.2. 1. *Fix individual $i \in V_1$. For all $t \geq 0$, $x_i(t) \in [\frac{1}{2} + \epsilon, 1]$.*

2. *Fix individual $i \in V_2$. For all $t \geq 0$, $x_i(t) \in [0, \frac{1}{2} - \epsilon]$.*

Proof. We will prove Statement 1 of the lemma. Statement 2 can be proven using an analogous argument. Proof by induction on t.

Base case: The statement holds for $t = 0$ by assumption.

Induction hypothesis: Assume that the statement holds for some $t \geq 0$.

We will show that the statement holds for $t + 1$. If $x_i(t) = 1$, then $x_i(t') = 1$ for all $t' \geq t$, and we are done. So let us assume $x_i(t) < 1$. Then, by Lemma 4.1, $x_i(t + 1) > x_i(t)$. Therefore, $x_i(t + 1) \in [\frac{1}{2} + \epsilon, 1]$. Therefore, the statement holds for $t + 1$. This concludes the proof by induction. □

Next we will show that for an individual $i \in V_1$, $\lim_{t\to\infty} x_i(t) = 1$. The corresponding statement for individuals in V_2 can be proven using an analogous argument.

Lemma 4.3. *Fix an individual $i \in V_1$. Then, $\lim_{t\to\infty} x_i(t) = 1$.*

Proof. The proof is along the same lines as that for Lemma 3.7. Again, observe that if $x_i(t) = 1$, then for all $t' \geq t, x_i(t') = 1$, and we are done. Define a function $g : [\frac{1}{2} + \epsilon, 1] \to \mathbb{R}$, as follows:

$$g(y) := \frac{w_{ii}y + y^b\left(\frac{1}{2} + \epsilon\right)}{w_{ii} + y^b\left(\frac{1}{2} + \epsilon\right) + (1 - y)^b\left(\frac{1}{2} - \epsilon + \frac{1}{h_G}\right)} - y$$

Observe that for all $t \geq 0$, for all $x_i(t) \in [\frac{1}{2} + \epsilon, 1)$, $x_i(t + 1) - x_i(t) \geq g(x_i(t)) > 0$. Moreover, $g(1) = 0$. For the purpose of contradiction, assume that $\lim_{t\to\infty} x_i(t) = a$, where $\frac{1}{2} + \epsilon \leq a < 1$. This implies, for every $\delta > 0$, there exists a $t(\delta)$ such that for all $t \geq t(\delta)$, $x_i(t + 1) - x_i(t) < \delta$, which implies that for all $t \geq t(\delta)$, $g(x_i(t)) < \delta$.

Let $\min_{y\in[\frac{1}{2}+\epsilon,a]} g(y) = c$. It implies for all $y \in [\frac{1}{2} + \epsilon, a]$, $g(y) \geq c$. Since $g(y) > 0$ for $y \in [\frac{1}{2} + \epsilon, 1)$, it follows that $c > 0$. Setting $\delta = c$, our analysis implies the following two properties of g: (1) for all $t \geq 0, g(x_i(t)) \geq c$, and (2) for all $t \geq t(\delta), g(x_i(t)) < c$, which contradict each other. This completes the proof by contradiction. □

□

5 Proof of Theorem 4

Let $|S(t)| = k$. Then, the opinion update under the flocking process can be written in matrix form as

$$\mathbf{x}(t + 1) = (1 - \epsilon)\mathbf{x}(t) + \epsilon A(t)\mathbf{x}(t)$$

where $A(t)$ is a $n \times n$ matrix given by

$$A_{ij}(t) = \begin{cases} \frac{1}{k}, & \text{if } i \in S(t), j \in S(t) \\ 1, & \text{if } i = j \text{ and } i \notin S(t) \\ 0, & \text{otherwise} \end{cases}$$

Observe that $A(t)$ is doubly-stochastic. Then

$$\begin{aligned} \gamma(\mathbf{x}(t + 1)) &= \gamma((1 - \epsilon)\mathbf{x}(t) + \epsilon A(t)\mathbf{x}(t)) \text{ (by definition of } \mathbf{x}(t + 1)) \\ &\leq (1 - \epsilon)\gamma(\mathbf{x}(t)) + \epsilon\gamma(A(t)\mathbf{x}(t)) \text{ (since } \gamma \text{ is convex in } \mathbf{x}) \\ &\leq (1 - \epsilon)\gamma(\mathbf{x}(t)) + \epsilon\gamma(\mathbf{x}(t)) \text{ (by Proposition 5.1)} \\ &= \gamma(\mathbf{x}(t)) \end{aligned}$$

Proposition 5.1. $\gamma(A(t)\mathbf{x}(t)) \leq \gamma(\mathbf{x}(t))$.

Proof. Let $\mathbf{y} := A(t)\mathbf{x}(t)$. Since $A(t)$ is doubly stochastic, it follows by a famous theorem by Hardy, Littlewood and Polya, that $\mathbf{x}(t)$ majorizes $\mathbf{y}$. Moreover, $\gamma(\mathbf{x})$ is a convex symmetric function. Therefore, it is a Schur-convex function. By definition, a function $f : \mathbb{R}^n \to \mathbb{R}$ is Schur-convex if $f(\mathbf{x}_1) \geq f(\mathbf{x}_2)$ whenever $\mathbf{x}_1$ majorizes $\mathbf{x}_2$. Therefore, $\gamma(\mathbf{y}) \leq \gamma(\mathbf{x}(t))$. □

6 Proofs of Theorems on Recommender Systems and Polarization

In this section we prove Theorem 5 and Theorem 6 from the main paper. Both theorems rely on the following technical lemma that invokes the Strong Law of Large Numbers to show that the random quantities we care about all take their expected values with probability 1 as $n \to \infty$.

Lemma 6.1. *In the limit as $n \to \infty$, with probability 1,*

(a) for all $i \in V_1$, $|N(i)| \to k$,

(b) for all $i \in V_1$, $\sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} Z_{ij_1} \to x_i k$,

(c) for all $i \in V_1$, $\sum_{\substack{j_1 \in V_2 \\ j_2 \text{ is BLUE}}} Z_{ij_2} \to (1 - x_i)k$,

(d) for all $j \in V_2$, $|N(j)| \to \frac{mk}{2n}$,

(e) for every pair of RED books $j, j' \in V_2, M_{jj'} = \sum_{i \in V_1} Z_{ij} Z_{ij'} \to \frac{mk^2(\frac{1}{4} + Var(x_1))}{n^2}$,

(f) for every pair of BLUE books $j, j' \in V_2, M_{jj'} = \sum_{i \in V_1} Z_{ij} Z_{ij'} \to \frac{mk^2(\frac{1}{4} + Var(x_1))}{n^2}$, and

(g) for every RED book j and every BLUE book j', $M_{jj'} = \sum_{i \in V_1} Z_{ij} Z_{ij'} \to \frac{mk^2(\frac{1}{4} - Var(x_1))}{n^2}$.

Proof. Recall that as $n \to \infty$, $m = f(n) \to \infty$. So statements (a) through (g) follow from the Strong Law of Large Numbers. □

Using Lemma 6.1, we will first prove Theorem 6.

6.1 Proof of Theorem 6

Lemma 6.2. *In the limit as $n \to \infty$, SimpleSALSA is polarizing with respect to i if and only if i is biased.*

Proof. Assume without loss of generality that $x_i > \frac{1}{2}$.

Let p_r be the probability that SimpleSALSA recommends a RED book. The proof consists of two steps: first we show that $p_r > \frac{1}{2}$ and $p_r \leq x_i$, and then we show that if $p_r > \frac{1}{2}$ and $p_r \leq x_i$,

SimpleSALSA is polarizing with respect to i if and only if i is biased.

$$
\begin{aligned}
p_r &= \sum_{j\in V_2: j_2 \text{ is } \texttt{RED}} \mathbb{P}[i \xrightarrow{3} j] \\
&= \sum_{\substack{j_1\in N(i)\\ j_1 \text{ is } \texttt{RED}}} \mathbb{P}[i \xrightarrow{1} j_1] \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2\in N(i)\\ j_2 \text{ is } \texttt{BLUE}}} \mathbb{P}[i \xrightarrow{1} j_2] \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \mathbb{P}[j_2 \xrightarrow{2} j] \\
&= \sum_{\substack{j_1\in N(i)\\ j_1 \text{ is } \texttt{RED}}} \frac{1}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2\in N(i)\\ j_2 \text{ is } \texttt{BLUE}}} \frac{1}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \mathbb{P}[j_2 \xrightarrow{2} j] \\
&= \sum_{\substack{j_1\in V_2\\ j_1 \text{ is } \texttt{RED}}} \frac{Z_{ij_1}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2\in V_2\\ j_2 \text{ is } \texttt{BLUE}}} \frac{Z_{ij_2}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \mathbb{P}[j_2 \xrightarrow{2} j] \\
&= \sum_{\substack{j_1\in V_2\\ j_1 \text{ is } \texttt{RED}}} \frac{Z_{ij_1}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \sum_{i'\in N(j_1)\cap N(j)} \frac{1}{|N(j_1)|}\frac{1}{|N(i')|} + \sum_{\substack{j_2\in V_2\\ j_2 \text{ is } \texttt{BLUE}}} \frac{Z_{ij_2}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \sum_{i'\in N(j_2)\cap N(j)} \frac{1}{|N(j_2)|}\frac{1}{|N(i')|} \\
&= \sum_{\substack{j_1\in V_2\\ j_1 \text{ is } \texttt{RED}}} \frac{Z_{ij_1}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \sum_{i'\in V_1} \frac{Z_{i'j_1}Z_{i'j}}{|N(j_1)||N(i')|} + \sum_{\substack{j_2\in V_2\\ j_2 \text{ is } \texttt{BLUE}}} \frac{Z_{ij_2}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \sum_{i'\in V_1} \frac{Z_{i'j_2}Z_{i'j}}{|N(j_2)||N(i')|}
\end{aligned}
$$

By Lemma 6.1, in the limit as $n \to \infty$, with probability 1,

$$\sum_{\substack{j_1\in V_2\\ j_1 \text{ is } \texttt{RED}}} \frac{Z_{ij_1}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \sum_{i'\in V_1} \frac{Z_{i'j_1}Z_{i'j}}{|N(j_1)||N(i')|} \to x_i \frac{1}{k\cdot mk/2n} n \frac{mk^2(\frac{1}{4}+\mathrm{Var}(x_1))}{n^2} = x_i\left(\frac{1}{2}+2\mathrm{Var}(x_1)\right)$$

and

$$\sum_{\substack{j_2\in V_2\\ j_2 \text{ is } \texttt{BLUE}}} \frac{Z_{ij_1}}{|N(i)|} \sum_{\substack{j\in V_2\\ j \text{ is } \texttt{RED}}} \sum_{i'\in V_1} \frac{Z_{i'j_2}Z_{i'j}}{|N(j_2)||N(i')|} \to (1-x_i) \frac{1}{k\cdot mk/2n} n \frac{mk^2(\frac{1}{4}-\mathrm{Var}(x_1))}{n^2} = (1-x_i)\left(\frac{1}{2}-2\mathrm{Var}(x_1)\right)$$

Therefore, in the limit as $n \to \infty$, with probability 1,

$$p_r \to x_i\left(\frac{1}{2}+2\mathrm{Var}(x_1)\right) + (1-x_i)\left(\frac{1}{2}-2\mathrm{Var}(x_1)\right)$$

Since $x_i > \frac{1}{2}$ (by assumption), and $\mathrm{Var}(x_1) > 0$ (by assumption), we have that

$$p_r > \frac{1}{2} \text{ and } p_r \le x_i \tag{6.1}$$

First, assume that i is unbiased. Let p be the probability that i accepts the recommendation. Therefore, the probability that the recommended book was `RED` given that i accepted the recommendation is given by

$$\frac{p_r p}{p_r p + (1-p_r)p} = p_r \le x_i$$

Therefore, SimpleSALSA is not polarizing.

Now, assume that i is biased. This implies i accepts the recommendation of a `RED` book with probability x_i and that of a `BLUE` book with probability $1-x_i$. Therefore, the probability that the recommended book was `RED` given that i accepted the recommendation is given by

$$\frac{p_r x_i}{p_r x_i + (1-x_i)(1-p_r)} > \frac{p_r x_i}{p_r x_i + p_r(1-x_i)} \text{ (since } p_r > \frac{1}{2}\text{, from (6.1))} = x_i$$

Therefore, by definition, SimpleSALSA is polarizing. Recall that our definition of a biased individual in this section corresponds to $b = 1$. Consider the general case, where i accepts the recommendation of a RED book with probability x_i^b and accepts that of a BLUE book with probability $(1 - x_i)^b$, where $b \geq 0$. Then, the probability that the recommended book was RED given that i accepted the recommendation is given by

$$\frac{p_r x_i^b}{p_r x_i^b + (1 - x_i)^b (1 - p_r)}$$

If $b \geq 1$, then

$$\begin{aligned}
\frac{p_r x_i^b}{p_r x_i^b + (1 - x_i)^b (1 - p_r)} &> \frac{p_r x_i}{p_r x_i + (1 - x_i)(1 - p_r)} \text{ (since } x_i > \frac{1}{2} \text{ and } b \geq 1) \\
&> \frac{p_r x_i}{p_r x_i + p_r (1 - x_i)} \text{ (since } p_r > \frac{1}{2}, \text{ from (6.1))} \\
&= x_i
\end{aligned}$$

This shows that SimpleSALSA is polarizing for any $b \geq 1$.

□

Lemma 6.3. *In the limit as $n \to \infty$ and as $T \to \infty$, SimpleICF is polarizing with respect to i if and only if i is biased.*

Proof. Assume without loss of generality that $x_i > \frac{1}{2}$.

Let p_r be the probability that SimpleICF recommends a RED book. For a node $j \in N(i)$, let $q_{j\texttt{RED}}$ be the probability that after T two-step random walks starting at j, the node with the largest value of `count(j)`, *i.e.*, j^*, is RED, and $q_{j\texttt{BLUE}}$ be the corresponding probability that j^* is BLUE. Then,

$$\begin{aligned}
p_r &= \sum_{\substack{j_1 \in N(i) \\ j_1 \text{ is RED}}} \mathbb{P}[i \xrightarrow{1} j_1] q_{j_1\texttt{RED}} + \sum_{\substack{j_2 \in N(i) \\ j_2 \text{ is BLUE}}} \mathbb{P}[i \xrightarrow{1} j_2] q_{j_2\texttt{RED}} \\
&= \sum_{\substack{j_1 \in N(i) \\ j_1 \text{ is RED}}} \frac{1}{|N(i)|} q_{j_1\texttt{RED}} + \sum_{\substack{j_2 \in N(i) \\ j_2 \text{ is BLUE}}} \frac{1}{|N(i)|} q_{j_2\texttt{RED}} \\
&= \sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} \frac{Z_{ij_1}}{|N(i)|} q_{j_1\texttt{RED}} + \sum_{\substack{j_2 \in V_2 \\ j_2 \text{ is BLUE}}} \frac{Z_{ij_1}}{|N(i)|} q_{j_2\texttt{RED}}
\end{aligned}$$

Consider T two-step random walks starting at a node $j_1 \in N(i)$. Observe that $q_{j_1\texttt{RED}}$ is exactly the probability that after these T random walks, there exists a RED node, say j, such that `count(j)` > `count(j')` for all BLUE nodes j'. However, as $T \to \infty$,

$$\mathbb{P}[\text{for all BLUE books } j' \in V_2,\ \texttt{count(j)} > \texttt{count(j')}] = \mathbb{P}[\text{for all BLUE books } j' \in V_2,\ \mathbb{P}[j_1 \xrightarrow{2} j] > \mathbb{P}[j_1 \xrightarrow{2} j']]$$

since as $T \to \infty$, `count(j)` $\to T \cdot \mathbb{P}[j_1 \xrightarrow{2} j]$ (by the Strong Law of Large Numbers). Therefore,

$$q_{j_1\texttt{RED}} = \mathbb{P}[\text{for all BLUE books } j' \in V_2,\ \mathbb{P}[j_1 \xrightarrow{2} j] > \mathbb{P}[j_1 \xrightarrow{2} j']]$$

Observe that for two RED books j_1 and j,

$$\mathbb{P}[j_1 \xrightarrow{2} j] = \sum_{i' \in N(j_1) \cap N(j)} \frac{1}{|N(j_1)|} \frac{1}{|N(i')|} = \sum_{i' \in V_1} \frac{Z_{i'j_1} Z_{i'j}}{|N(j_1)||N(i')|}$$

By Lemma 6.1, in the limit as $n \to \infty$, with probability 1,

$$\mathbb{P}[j_1 \xrightarrow{2} j] \to \frac{1}{k}\frac{1}{mk/2n}\frac{mk^2(\frac{1}{4}+\text{Var}(x_1))}{n^2} = \frac{1}{n}\left(\frac{1}{2}+2\text{Var}(x_1)\right)$$

Similarly, for a BLUE book j', in the limit as $n \to \infty$, with probability 1,

$$\mathbb{P}[j_1 \xrightarrow{2} j'] \to \frac{1}{k}\frac{1}{mk/2n}\frac{mk^2(\frac{1}{4}-\text{Var}(x_1))}{n^2} = \frac{1}{n}\left(\frac{1}{2}-2\text{Var}(x_1)\right)$$

Since $\text{Var}(x_1) > 0$, in the limit as $n \to \infty$, $\mathbb{P}[j_1 \xrightarrow{2} j] > \mathbb{P}[j_1 \xrightarrow{2} j']$ with probability 1. Therefore, $q_{j_1\texttt{RED}} = 1$. By symmetry $q_{j_2\texttt{RED}} = 1 - q_{j2\texttt{BLUE}} = 0$. Moreover, by Lemma 1, in the limit as $n \to \infty$, $\sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} \frac{Z_{ij_1}}{|N(i)|} = x_i$, with probability 1. Therefore, as $n \to \infty$,

$$p_r = x_i \tag{6.2}$$

The rest of the analysis is identical to Lemma 6.2. □

This completes the proof of Theorem 6.

6.2 Proof of Theorem 5

Assume, without loss of generality, that $x_i > \frac{1}{2}$.

Let p_r be the probability that SimplePPR recommends a RED book to i. This probability is exactly equal to the probability that after T three-step random walks starting at i there exists a RED node, say j, such that such that count(j) > count(j') for all BLUE nodes j'. However, as $T \to \infty$,

$$\mathbb{P}[\text{for all BLUE books } j' \in V_2,\ \texttt{count(j)} > \texttt{count(j')}] = \mathbb{P}[\text{for all BLUE books } j' \in V_2,\ \mathbb{P}[i \xrightarrow{3} j] > \mathbb{P}[i \xrightarrow{3} j']]$$

since as $T \to \infty$, $\texttt{count(j)} \to T \cdot \mathbb{P}[i \xrightarrow{3} j]$ with probability 1 (by the Strong Law of Large Numbers). Therefore,

$$p_r = \mathbb{P}[\text{for all BLUE books } j' \in V_2,\ \mathbb{P}[i \xrightarrow{3} j] > \mathbb{P}[i \xrightarrow{3} j']]$$

For a RED book $j \in V_2$,

$$\mathbb{P}[i \xrightarrow{3} j] = \sum_{\substack{j_1 \in N(i) \\ j_1 \text{ is RED}}} \mathbb{P}[i \xrightarrow{1} j_1]\mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2 \in N(i) \\ j_2 \text{ is BLUE}}} \mathbb{P}[i \xrightarrow{1} j_2]\mathbb{P}[j_2 \xrightarrow{2} j]$$

$$\mathbb{P}[i \xrightarrow{3} j] = \sum_{\substack{j_1 \in N(i) \\ j_1 \text{ is RED}}} \frac{1}{|N(i)|}\mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2 \in N(i) \\ j_2 \text{ is BLUE}}} \frac{1}{|N(i)|}\mathbb{P}[j_2 \xrightarrow{2} j]$$

$$\mathbb{P}[i \xrightarrow{3} j] = \sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} \frac{Z_{ij_1}}{|N(i)|}\mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2 \in V_2 \\ j_2 \text{ is BLUE}}} \frac{Z_{ij_2}}{|N(i)|}\mathbb{P}[j_2 \xrightarrow{2} j]$$

As we showed in the proof of Lemma 6.3, in the limit as $n \to \infty$,

$$\mathbb{P}[j_1 \xrightarrow{2} j] \to \frac{1}{n}\left(\frac{1}{2}+2\text{Var}(x_1)\right) \text{ and (by symmetry) } \mathbb{P}[j_2 \xrightarrow{2} j] \to \frac{1}{n}\left(\frac{1}{2}-2\text{Var}(x_1)\right)$$

with probability 1. Moreover, by Lemma 1, in the limit as $n \to \infty$, $\sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is } \texttt{RED}}} \frac{Z_{ij_1}}{|N(i)|} \to x_i$, with probability 1. Therefore, with probability 1,

$$\mathbb{P}[i \xrightarrow{3} j] \to \frac{x_i}{n}\left(\frac{1}{2} + 2\text{Var}(x_1)\right) + \frac{1-x_i}{n}\left(\frac{1}{2} - 2\text{Var}(x_1)\right)$$

Similarly, for a BLUE book $j' \in V_2$, in the limit as $n \to \infty$, with probability 1,

$$\mathbb{P}[i \xrightarrow{3} j'] \to \frac{x_i}{n}\left(\frac{1}{2} - 2\text{Var}(x_1)\right) + \frac{1-x_i}{n}\left(\frac{1}{2} + 2\text{Var}(x_1)\right)$$

Since $x_i > \frac{1}{2}$ and $\text{Var}(x_1) > 0$,

$$\mathbb{P}[i \xrightarrow{3} j] > \mathbb{P}[i \xrightarrow{3} j']$$

with probability 1. In other words, $p_r = 1$. Consider the general definition of a biased individual, where individual i accepts the recommendation of a RED book with probability x_i^b and accepts that of a BLUE book with probability $(1-x_i)^b$, where $b \geq 0$. Then, the probability that the recommended book was RED given that i accepted the recommendation is given by

$$\frac{p_r x_i^b}{p_r x_i^b + (1-x_i)^b(1-p_r)}$$

Since $p_r = 1$, the probability that a book recommended by SimplePPR was RED given that it was accepted is exactly p_r for all $b \geq 0$. Therefore, SimplePPR is polarizing for all $b \geq 0$.