Biased Assimilation, Homophily and the Dynamics of Polarization

Pranav Dandekar*   Ashish Goel†   David Lee‡

September 27, 2012

Abstract

Are we as a society getting more polarized, and if so, why? We try to answer this question through a model of opinion formation. Empirical studies have shown that homophily 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 either polarization, persistent disagreement or consensus depending on how biased individuals are. In other words, homophily alone, without biased assimilation, is not sufficient to polarize society. Quite interestingly, biased assimilation also provides insight into the following related question: do internet based recommender algorithms that show us personalized content contribute to polarization? We make a connection between biased assimilation and the polarizing effects of some random-walk based recommender algorithms that are similar in spirit to some commonly used recommender algorithms.

---

*Department of Management Science & Engineering, Stanford University, Stanford, CA. Email: ppd@stanford.edu.

†Departments of Management Science & Engineering and (by courtesy) Computer Science, Stanford University, Stanford, CA. Email: ashishg@stanford.edu.

‡Department of Electrical Engineering, Stanford University, Stanford, CA. Email: davidtlee@stanford.edu.

1 Introduction

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 [PR84, PR91] and in the extreme policies adopted by candidates for political office [Hil09]. McCarty *et al.* [MPR06] claim via rigorous analysis that America is polarized in terms of political attitudes and beliefs. Phenomena such as segregation in urban residential neighborhoods ([Sch71, BM06, BIKK12]), the rising popularity of overtly partisan television news networks [Bil09, Bil10], and the readership and linking patterns of blogs along partisan lines [AG05, HG07, GBK09, LSF10] 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 [Wol99, FAP05]. We adopt the view that polarization is not a property of a state of society; instead it is a property of the dynamics of interaction between individuals.

It has been argued that homophily, *i.e.*, greater interaction with like-minded individuals, results in polarization [BHK$^+$96, Sun02, GBK09]. Evidence in support of 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 which ultimately results in increased polarization.

A rich body of work attempts to explain polarization through variants of a well-known opinion formation model due to DeGroot [DeG74]. 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.*, [FJ90, Kra00, ACFO10, BKO11]) account for the empirical observation that in many cases there is persistent disagreement between individuals and consensus is never reached. However, we show that repeated averaging of opinions, which underlies these models, always results in opinions that are less divergent compared to 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.* [LRL79] showed through experiments 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.*, [MMBD93, MDL$^+$02, TL06]). 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.

1.1 Summary of Contributions

We propose a generalization of DeGroot's model that accounts for biased assimilation. Like De-Groot's model, our opinion formation process unfolds over an exogenously defined social network represented by a weighted undirected graph $G = (V, E)$. Each individual $i \in V$ has an opinion $x_i(t) \in [0, 1]$, which represents his degree of support at time step t for the position represented by 1. In order to weight confirming evidence more heavily relative to disconfirming evidence, opinions are updated as follows: individual i weights each neighbor j's opinion $x_j(t)$ by a factor $(x_i(t))^{b_i}$ and weights the opposing view $(1 - x_j(t))$ by a factor $(1 - x_i(t))^{b_i}$, where $b_i \geq 0$ is a *bias parameter*. Informally, b_i represents the bias with which i assimilates his neighbors opinions. When $b_i = 0$, our model reduces to DeGroot's, and corresponds to unbiased assimilation. Our biased opinion formation process mathematically reproduces the effect empirically observed by Lord *et al.* (Theorem 1).

We measure divergence of opinions in terms of the *network disagreement index* (NDI), which we define to be $\sum_{(i,j) \in E} w_{ij}(x_i(t) - x_j(t))^2$. It is similar to the notion of social cost used by Bindel *et al.* [BKO11]. We say that an opinion formation process is *polarizing* if the NDI at the end of the process is greater than that initially. We show that:

- (Theorem 2) DeGroot-like repeated averaging processes can never be polarizing, even if individuals are arbitrarily homophilous (*i.e.*, the underlying network is presented adversarially as opposed to based on a mathematical model).

- (Theorem 4) The biased opinion formation process over a simple model of networks with homophily results in polarization if individuals' bias parameter $b \geq 1$. If $b < 1$, the process results in either persistent disagreement or consensus depending on the degree of homophily.

In summary, we show that homophily alone, without biased assimilation, is not sufficient to polarize society. This conclusion disagrees with the literature (*e.g.*, [BHK+96, Sun02]) that proposes homophily as the predominant cause of polarization. As the reader might expect, there are many ways of mathematically measuring the divergence of opinions among individuals. Many of our results hold for more general measures of divergence, which we discuss in Section 6.

The notion of biased assimilation also provides insight into the following related question: do internet based recommender algorithms that show us personalized content contribute to polarization? We analyze the polarizing effects of three recommender algorithms—SimpleSALSA, SimplePPR, and SimpleICF—that are similar in spirit to three well-known algorithms from the literature: SALSA [LM01], Personalized PageRank [PBMW99], and Item-based Collaborative Filtering [LSY03]. 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 score, is always polarizing (Theorem 5). On the other hand, Simple-SALSA and SimpleICF, 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 (Theorem 6). Designing algorithms and online social systems that reduce polarization, for example, by counteracting biased assimilation is a promising research direction.

2 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 will 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 two 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 simply 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.

2.1 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 2.1 (Network Disagreement Index (NDI)). *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*

$$\eta(G, \mathbf{x}) := \sum_{(i,j) \in E} w_{ij}(x_i - x_j)^2 \tag{2.1}$$

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 NDI is similar to the notion of social cost used by Bindel *et al.* [BKO11]. 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 [Hei46, Fes57]. The NDI captures the phenomenon of *issue radicalization, i.e.*, pre-existing groups of individuals becoming progressively more extreme. Admittedly, it does not entirely capture an aspect of polarization called *issue alignment* [BG08] whereby individuals with diverse opinions organize into ideologically coherent, but opposing factions. However, there is significant empirical evidence [MPR06, BG08, Cas12] that issue radicalization is more prevalent compared to 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 Section 6.

2.2 DeGroot's Repeated Averaging Model

In his seminal work on opinion formation, DeGroot [DeG74] proposed a model 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 2.2 (DeGroot's Repeated Averaging Process). *The opinion of individual i at time $t+1$, $x_i(t+1)$, is given by*

$$x_i(t + 1) = \frac{w_{ii}x_i(t) + s_i(t)}{w_{ii} + d_i} \tag{2.2}$$

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.

2.3 Biased Opinion Formation Model

We generalize DeGroot's model to account for *biased assimilation*. Biased assimilation is a well-known phenomenon in social psychology described by Lord *et al.* [LRL79] 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.* [LRL79] showed through experiments that biased assimilation of mixed or inconclusive evidence does indeed result in more extreme opinions.

In order 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

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

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

Definition 2.3 (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*

$$x_i(t+1) = \frac{w_{ii} x_i(t) + (x_i(t))^{b_i} s_i(t)}{w_{ii} + (x_i(t))^{b_i} s_i(t) + (1 - x_i(t))^{b_i}(d_i - s_i(t))} \tag{2.3}$$

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 $b_i = 0$, (2.3) is identical to (2.2), *i.e.*, DeGroot's averaging 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 $\beta_i : [0,1] \to [0,1]$ is non-decreasing.

Connection with Urn Models. Urn models are an elegant abstraction that have been used to analyze the properties of a wide variety of probabilistic processes. DeGroot's model of weighted averaging has the following analogous urn dynamic: $x_i(t)$ denotes the fraction of `RED` balls in

individual i's urn at time t, and $1 - x_i(t)$ denotes the corresponding fraction of BLUE balls. At each time step, i chooses a neighbor j with probability proportional to w_{ij} and chooses a ball uniformly at random from j's urn. Individual i adds that ball to his urn and discards a ball chosen uniformly at random from his urn. When the bias parameter $b_i = 1$, the biased opinion formation process can be interpreted as the following variant of the above urn dynamic: as before, i chooses a neighbor j with probability proportional to w_{ij} and chooses a ball uniformly at random from j's urn. In addition, i also chooses a ball uniformly at random from his own urn. If the colors of the two balls match, i puts them both into his urn and discards a ball chosen uniformly at random from his urn. If the colors do not match, the two balls are returned to their respective urns.

2.4 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.* [LRL79]. 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 will 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, $w_{ii} = w$. Let $s \in (0,1)$ denote the (time-invariant) weighted average of the opinions of all sources in the individual's environment. Then, from (2.3), the individual's opinion at time $t + 1$ is given by

$$x(t+1) = \frac{wx(t) + (x(t))^b s}{w + (x(t))^b s + (1 - x(t))^b (1 - s)} \tag{2.4}$$

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

$$\hat{x}(s,b) := \frac{s^{1/(1-b)}}{s^{1/(1-b)} + (1-s)^{1/(1-b)}} \tag{2.5}$$

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 theorem is proved in Appendix A. 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 toward one of the extremes, and the individual 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 towards 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 proven in the theorem; as $b \to 1$, $\hat{x} \to s$. When the individual is connected to other individuals in a social network, we will show that the biased opinion formation process produces polarization even when $b = 1$.

3   DeGroot's Repeated Averaging Process is not Polarizing

It is easy to see that if DeGroot's process was asynchronous, *i.e.*, individuals update their opinion one at a time, each opinion update can only lower the network disagreement index (NDI). However, here we will 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. 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.

Theorem 2. *Consider an arbitrary weighted undirected graph $G = (V, E, w)$. Assume that G is connected. Let $\mathbf{x}(t) \in [0, 1]^n$ 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.*

The theorem is proved in Appendix B. Observe that in the limit as $w_{ii} \to \infty$, individual i can be viewed as being a *zealot* [YAO+11],*i.e.*, an individual with an unchanging opinion. So our result also holds for repeated averaging in the presence of zealots.

A possible criticism of this result is that it holds for this particular definition of the NDI which may not always capture the intuitive notion of polarization. For example, consider a network partitioned into two densely connected opposing factions with sparse cross linkages. One might consider such a network to be polarized, even though the network disagreement index for it is small. An alternate measure that does capture the divergence of opinions in the above example is the *global disagreement index* (GDI) defined below.

Definition 3.1 (Global Disagreement Index (GDI)). *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*

$$\gamma(\mathbf{x}) := \sum_{i<j} (x_i - x_j)^2 \tag{3.1}$$

Observe that it is possible to assign edge weights w_{ij} such that DeGroot's repeated averaging process increases the GDI since the latter is independent of the weights. However, we show that a variant of repeated averaging, based on the well-known flocking model for decentralized consensus [Tsi84], can only decrease the GDI. We consider a repeated averaging process where 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 3.2 (Flocking Process). *Let $\epsilon \in [0,1]$. For $t \geq 0$, let $S(t) \subseteq V$ be an arbitrary set of individuals. 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*

$$x_i(t+1) = \begin{cases} (1-\epsilon)x_i(t) + \epsilon s(t), & \text{if } i \in S(t) \\ x_i(t), & \text{otherwise} \end{cases} \tag{3.2}$$

Next we show that each opinion update in the flocking process can only lower the GDI.

Theorem 3. *Let $\mathbf{x}(t) \in [0,1]^n$ 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.*

The theorem is proved in Appendix B.

4   Polarization due to Biased Assimilation

In this section we state and prove our main result: in a simple model of networks with homophily, the biased opinion formation process may result in either polarization, persistent disagreement, or consensus depending on how biased the individuals are. We model homophilous networks using a deterministic variant of *multi-type random networks* [GJ11]. Multi-type 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 multi-type random networks where $P_{ii} > P_{ij}$ for all i, j, is called is the islands model, and is used to model homophily (since an individual is more likely to be connected with individuals of the same type). We will analyze the biased opinion formation process over a deterministic variant of the islands model, which we call a *two-island network*.

Definition 4.1. *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[1].*

- *$p_s > p_d$.*

For a two-island network, we define the *degree of homophily* as follows.

Definition 4.2. *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.*

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 4. *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]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.

Algorithm 1 SimpleSALSA

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

---

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 $\hat{x} \in (\frac{1}{2}, 1)$ 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 $i \in V$, $\lim_{t \to \infty} x_i(t) = \frac{1}{2}$.*

The theorem is proved in Appendix C. Let us analyze the implications of this theorem. Let $\eta(G, \mathbf{x}(t)) \to \eta_\infty$ as $t \to \infty$, *i.e.*, let η_∞ be the NDI at equilibrium. 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 $x_0 < \hat{x}$; so the opinion formation process may not be polarizing, but it doesn't produce consensus either. Finally, when individuals have low bias (*i.e.*, $b < 2/(h_G + 1)$, $\eta_\infty = 0 < \eta(G, \mathbf{x}(0))$, *i.e.*, the opinion formation process is depolarizing, since the network reaches consensus in equilibrium.

This illustrates the importance of the bias parameter in causing polarization. Also, observe that $b = 1$ corresponds to the urn dynamic described in Section 2.3, and hence the above result shows that that urn dynamic causes polarization for arbitrarily small degree of homophily.

5 Recommender Systems and Polarization

Recommender systems are widely used on the Internet to present personalized information (*e.g.*, search results, new articles, products) to individuals. This personalization is typically done by algorithms that use an individual's the past behavior (*e.g.*, history of browsing and purchases) and 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 [Sun02] that this personalization of information has an echo-chamber effect where individuals are only exposed to information they agree with, and this ultimately leads to increased polarization. In this section we investigate this question: do recommender systems have a polarizing effect? We analyze three simple random-walk based recommender algorithms— SimpleSALSA (Algorithm 1), SimplePPR (Algorithm 2) and SimpleICF(Algorithm 3)— that are similar in spirit to three well-known recommender algorithms from the literature: SALSA [LM01], Personalized PageRank [PBMW99], and item-based collaborative filtering [LSY03], respectively.

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 will 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 5.1. *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 assume that i can only buy

---

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

---

a book 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.

Since, we are interested in analyzing the polarizing effects of recommender systems, we will 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 5.2. *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 in Section 2.3 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.

Definition 5.3. *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.*

In order to analyze the recommender algorithms, we assume a generative model for G, which we describe next.

5.1   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 each individual $i \in V_1$, we draw x_i independently from a distribution over $[0, 1]$ with a probability density function (pdf) $g(\cdot)$.

Algorithm 3 SimpleICF

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

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.*, $\text{Var}(x_i) > 0$. 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.

5.2 Analysis

Next we prove our results about the polarizing effects of each of the three algorithms. Our results hold with probability 1 in the limit as $n \to \infty$. First we invoke 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 5.1. *In the limit as $n \to \infty$, with probability 1,*

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

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

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

(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. $\qquad \square$

We use Lemma 5.1 to prove our results. First we show that SimplePPR (Algorithm 2) is polarizing with respect to i even if i is unbiased.

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

Next we show that SimpleSALSA and SimpleICF are polarizing only if i is biased.

Theorem 6. *In the limit as $n \to \infty$,*

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 Theorem 5 and Theorem 6 are proved in Appendix D.

6   Discussion of Various Measures of Opinion Divergence

Recall that we define an opinion formation process to be polarizing if it results in an increased divergence of opinions. Here we describe a number of alternate measures of divergence, and discuss how many of our results hold for these measures. A generalization of the global disagreement index (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 3, 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 $\mathbf{x}$.

A stronger definition of divergence is one based on second order stochastic dominance, which is defined over distributions, but can be easily modified to work with vectors. Informally, a distribution F is second order stochastically dominated by a distribution G if F is a mean-preserving spread of G. Let us say an opinion formation process is polarizing if the final opinion vector is dominated (second order stochastically) by the initial opinion vector, and is depolarizing if the final vector dominates the initial vector. According to this definition, a single opinion update in the DeGroot and flocking processes is in general neither polarizing nor depolarizing. However, both these processes have been shown to converge to consensus under fairly general conditions ([DeG74, Tsi84]). Thus, under those conditions, both these processes are depolarizing in equilibrium. Moreover, our results on the three recommender algorithms (Theorem 5 and Theorem 6) also hold under this definition of divergence.

Consider the following even stronger definition of polarization: a process is polarizing if at each time step, it pushes the opinions of individuals away from the average and is depolarizing if it brings their opinions closer to the average. Under this definition too, the DeGroot and flocking processes are neither polarizing nor depolarizing. However, under all three definitions, the biased opinion formation process is polarizing on a two-island network when $b \geq 1$.

7   Conclusion

In this paper we attempted to explain polarization in society through a model of opinion formation. We generalized DeGroot's repeated averaging model to account for biased assimilation. We showed that DeGroot-like repeated averaging processes can never be polarizing, even if individuals are arbitrarily homophilous. We also showed that in a two-island network, our biased opinion formation process may result in either polarization (if $b \geq 1$), persistent disagreement (if $1 > b \geq 2/(h+1)$), or consensus (if $b < 2/(h+1)$). In other words, homophily alone, without biased assimilation, is not sufficient to polarize society. We used biased assimilation to provide insight into the polarizing effects of three recommender algorithms: SimpleSALSA, SimplePPR and SimpleICF. We showed that for a simple, natural model of the underlying user-item graph, SimpleSALSA and SimpleICF are polarizing only if individuals are biased whereas SimplePPR is polarizing even if individuals are unbiased.

One direction for further investigation is to study through human subject experiments how the degree of homophily and the strength of biased assimilation affect whether individuals interacting over a network polarize or arrive at a consensus? Our analysis of recommender algorithms is a first step toward designing algorithms and online social systems that counteract polarization and facilitate greater consensus between individuals over complex and vexing social, economic and political issues. We view this as a promising and important direction for further research.

References

[ACFO10] Daron Acemoglu, Giacomo Como, Fabio Fagnani, and Asuman E. Ozdaglar. Opinion fluctuations and disagreement in social networks. *CoRR*, abs/1009.2653, 2010.

[AG05] Lada Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *In LinkKDD 05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[BG08] D. Baldassarri and A. Gelman. Partisans without Constraint: Political Polarization and Trends in American Public Opinion. *American Journal of Sociology*, 114(2):408–446, 2008.

[BHK$^+$96] R. S. Baron, S. I. Hoppe, C. F. Kao, B. Brunsman, B. Linneweh, and D. Rogers. Social corroboration and opinion extremity. *Journal of Experimental Social Psychology*, 32:537–560, 1996.

[BIKK12] Christina Brandt, Nicole Immorlica, Gautam Kamath, and Robert Kleinberg. An analysis of one-dimensional schelling segregation. In *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, pages 789–804, New York, NY, USA, 2012. ACM. Available from: http://doi.acm.org/10.1145/2213977.2214048, doi:10.1145/2213977.2214048.

[Bil09] Bill Carter. With Rivals Ahead, Doubts for CNN's Middle Road. New York Times, April 2009. Available from: http://www.nytimes.com/2009/04/27/business/media/27cnn.html.

[Bil10] Bill Carter. CNN Fails to Stop Fall in Ratings. New York Times, March 2010. Available from: http://www.nytimes.com/2010/03/30/business/media/30cnn.html.

[BKO11] David Bindel, Jon M. Kleinberg, and Sigal Oren. How Bad is Forming Your Own Opinion? In Rafail Ostrovsky, editor, *FOCS*, pages 57–66. IEEE, 2011.

[BM06] Elizabeth Eve Bruch and Robert D. Mare. Neighborhood choice and neighborhood change. *American Journal of Sociology*, 112(3):667–709, 2006.

[Cas12] Cass Sunstein. Breaking Up The Echo. New York Times, September 2012. Available from: http://www.nytimes.com/2012/09/18/opinion/balanced-news-reports-may-only-inflame.html.

[DeG74] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):pp. 118–121, 1974. Available from: http://www.jstor.org/stable/2285509.

[FAP05]   Morris P. Fiorina, Samuel J. Abrams, and Jeremy C. Pope. *Culture War? The Myth of a Polarized America.* Pearson Education Inc., New York, 2005.

[Fes57]   Leon Festinger. *A Theory of Cognitive Dissonance.* Stanford University Press, June 1957.

[FJ90]   N. E. Friedkin and E. C. Johnsen. Social Influence and Opinions. *Journal of Mathematical Sociology*, 15(3-4), 1990.

[GBK09]   Eric Gilbert, Tony Bergstrom, and Karrie Karahalios. Blogs are echo chambers: Blogs are echo chambers. In *HICSS*, pages 1–10, 2009.

[GJ11]   Benjamin Golub and Matthew O. Jackson. How homophily affects the speed of learning and best response dynamics. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1443787, 2011.

[Hei46]   Fritz Heider. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112, 1946. doi:10.1080/00223980.1946.9917275.

[HG07]   Eszter Hargittai and Jason Gallo. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice*, 134:67–86, 2007.

[Hil09]   Steven Hill. Divided We Stand: The Polarization of American Politics. *National Civic Review*, 94:3–14, 2009.

[Kra00]   Ulrich Krause. A discrete nonlinear and non-autonomous model of consensus formation. In *Communications in Difference Equations*. Gordon and Breach, 2000.

[LM01]   R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, April 2001. Available from: http://doi.acm.org/10.1145/382979.383041, doi:10.1145/382979.383041.

[LRL79]   Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979. Available from: http://www.psych.umn.edu/courses/spring06/borgidae/psy5202/readings/lord,%20ross%20&%20lepper%20(1979).pdf.

[LSF10]   Eric Lawrence, John Sides, and Henry Farrell. Self-segregation or deliberation? blog readership, participation, and polarization in american politics. *Perspectives on Politics*, 8(1):141–157, 2010.

[LSY03]   G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76 – 80, jan/feb 2003. doi:10.1109/MIC.2003.1167344.

[MDL+02]   Geoffrey D. Munro, Peter H. Ditto, Lisa K. Lockhart, Angela Fagerlin, Mitchell Gready, and Elizabeth Peterson. Biased assimilation of sociopolitical arguments: Evaluating the 1996 u.s. presidential debate. *Basic and Applied Social Psychology*, 24(1):15–26, 2002. Available from: http://www.tandfonline.com/doi/abs/10.1207/S15324834BASP2401_2, arXiv:http://www.tandfonline.com/doi/pdf/10.1207/S15324834BASP2401_2, doi:10.1207/S15324834BASP2401_2.

[MMBD93] Arthur G. Miller, John W. McHoskey, Cynthia M. Bane, and Timothy G. Dowd. The attitude polarization phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change. *Journal of Personality and Social Psychology*, 64(4):561–574, 1993.

[MPR06] Nolan McCarty, Keith T. Poole, and Howard Rosenthal. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press, Cambridge, Massachusetts, 2006.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999. Available from: http://ilpubs.stanford.edu:8090/422/.

[PR84] Keith T. Poole and Howard Rosenthal. The polarization of american politics. *The Journal of Politics*, 46(4):1061–1079, 1984.

[PR91] Keith T. Poole and Howard Rosenthal. Patterns of congressional voting. *American Journal of Political Science*, 35(1):228–278, 1991.

[Sch71] Thomas C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1:143–186, 1971.

[Sun02] Cass R. Sunstein. *Republic.com*. Princeton University Press, Princeton, New Jersey, 2002.

[TL06] Charles S. Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769, 2006. Available from: http://dx.doi.org/10.1111/j.1540-5907.2006.00214.x, doi:10.1111/j.1540-5907.2006.00214.x.

[Tsi84] John N. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, Department of EECS, MIT, Cambridge, MA, November 1984.

[Wol99] Alan Wolfe. *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, 1999.

[YAO+11] Ercan Yildiz, Daron Acemoglu, Asuman E. Ozdaglar, Amin Saberi, and Anna Scaglione. Discrete Opinion Dynamics with Stubborn Agents. *SSRN eLibrary*, 2011. doi:10.2139/ssrn.1744113.

A 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,

$$\frac{x(t+1)}{1 - x(t+1)} = \frac{wx(t) + (x(t))^b s}{w(1 - x(t)) + (1 - x(t))^b (1 - s)} = \frac{w + (x(t))^{b-1} s}{w + (1 - x(t))^{b-1} (1 - s)} \frac{x(t)}{1 - x(t)} \qquad \text{(A.1)}$$

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

15

Lemma A.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 (A.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 A.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

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

Or equivalently, $(x(t))^{b-1}s > (1-x(t))^{b-1}(1-s)$. Substituting this in (A.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 (A.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 A.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 $\lim_{t\to\infty} x(t) = 0$.*

16

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 A.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 A.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. $\square$

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

Lemma A.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 A.5. *Let $b < 1$. Then, $\lim_{t\to\infty} x(t) = \hat{x}$.*

B   Proofs of Section 3

Proof of Theorem 2. Recall that since $b_i = 0$, the opinion of node i at time $t + 1$ is given by

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

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

$$(L_G\mathbf{x}(t))_i = d_i x_i(t) - \sum_{j \in N(i)} w_{ij}x_j(t) = d_i x_i(t) + w_{ii}x_i(t) - \left(w_{ii}x_i(t) + \sum_{j \in N(i)} w_{ij}x_j(t) \right)$$

$$= (d_i + w_{ii})(x_i(t) - x_i(t+1)) \text{ (from (B.1))}$$

17

Equivalently, in matrix notation,

$$\mathbf{x}(t+1) = (I - DL_G)\mathbf{x}(t) \tag{B.2}$$

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

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

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 B.1. $\qquad\square$

Lemma B.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

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

18

$\square$

Proof of Theorem 3. Let $|S(t)| = k$. Then, the opinion update (3.2) 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

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

Proposition B.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, $\gamma(\mathbf{y}) \leq \gamma(\mathbf{x}(t))$. $\square$

$\square$

C Proof of Theorem 4

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

Lemma C.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 $t \geq 0$, $x_i(t) = x_j(t)$.*

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

Lemma C.2. *Consider a node $i \in V_1$ and a node $j \in V_2$. Then, for all $t \geq 0$, $x_i(t) = 1 - x_j(t)$.*

Proof of Lemma C.2. By induction.

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

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

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

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

where $d_i = n(p_s + p_d)$ and, by Lemma C.1, $s_i(t) = n(p_s x_i(t) + p_d x_j(t))$. On the other hand,

$$\frac{x_j(t+1)}{1 - x_j(t+1)} = \frac{(x_j(t))^b}{(1 - x_j(t))^b} \frac{s_j(t)}{d_j - s_j(t)} \tag{C.2}$$

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 (C.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 (C.2). It follows that $x_i(t+1) = 1 - x_j(t+1)$.

This completes the inductive proof. $\qquad\square$

Lemma C.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 C.3. *Consider a node $i \in V_1$. For all $t \geq 0$, $x_i(t) \in [\frac{1}{2}, 1]$.*

Proof of Lemma C.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. $\qquad\square$

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

$$x_i(t+1) = \frac{(x_i(t))^b s_i(t)}{(x_i(t))^b s_i(t) + (1 - x_i(t))^b (d_i - s_i(t))} \text{ (by (2.3))}$$

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

$$x_i(t+1) = x_i(t) \tag{C.3}$$

We will show that if $b \geq 1$ or $b < \frac{2}{h_G + 1}$, (C.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 (C.3) in $(\frac{1}{2}, 1)$.

Lemma C.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.(C.3) in $(\frac{1}{2}, 1)$.

20

(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 C.4. Consider the function $f : [0,1] \to \mathbb{R}$ defined as

$$f(y;b) := \begin{cases} 1, & y \in [0,1], b = 1 \\ 0, & y \in [0,1], b = 2 \\ \frac{2}{b} - 1, & y = \frac{1}{2}, b > 0 \\ \frac{(y)^{2-b}-(1-y)^{2-b}}{y(1-y)^{1-b}-y^{1-b}(1-y)}, & \text{otherwise} \end{cases} \tag{C.4}$$

We will first prove a few properties of f and then use those properties to prove Lemma C.4.

Proposition C.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,

$$\lim_{y \to 1/2} \frac{p'(y;b)}{q'(y;b)} = \lim_{y \to 1/2} \frac{(2-b)(y^{1-b} + (1-y)^{1-b})}{(1-y)^{1-b} - (1-b)y(1-y)^{-b} - (1-b)y^{-b}(1-y) + y^{1-b}} = \frac{2}{b} - 1 \tag{C.5}$$

So, we have that

$$\lim_{y \to 1/2} f(y;b) = \lim_{y \to 1/2} \frac{p(y;b)}{q(y;b)} = \lim_{y \to 1/2} \frac{p'(y)}{q'(y)} \text{ (using L'Hôpital's rule)} = \frac{2}{b} - 1 \text{ (from (C.5))} = f(\frac{1}{2}; b)$$

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

$$y_1 y_2 - y_1 \bar{y}_2 > (y_1 y_2)^{1-b} - (y_1 \bar{y}_2)^{1-b} \tag{C.6}$$

Similarly,

$$\bar{y}_1 y_2 - \bar{y}_1 \bar{y}_2 > (\bar{y}_1 y_2)^{1-b} - (\bar{y}_1 \bar{y}_2)^{1-b} \tag{C.7}$$

Adding (C.6) and (C.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,

$$(y_1 y_2 - \bar{y}_1 \bar{y}_2) - \left((y_1 y_2)^{1-b} - (\bar{y}_1 \bar{y}_2)^{1-b}\right) > (y_1 \bar{y}_2 - \bar{y}_1 y_2) - \left((y_1 \bar{y}_2)^{1-b} - (\bar{y}_1 y_2)^{1-b}\right) \tag{C.8}$$

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

$$y_1 y_2 - \bar{y}_1 \bar{y}_2 > 0; (y_1 y_2)^{1-b} - (\bar{y}_1 \bar{y}_2)^{1-b} > 0; y_1 \bar{y}_2 - \bar{y}_1 y_2 > 0; (y_1 \bar{y}_2)^{1-b} - (\bar{y}_1 y_2)^{1-b} > 0 \tag{C.9}$$

21

(C.8) and (C.9) imply that

$$\frac{y_1 y_2 - \bar{y}_1 \bar{y}_2}{y_1 \bar{y}_2 - \bar{y}_1 y_2} > \frac{(y_1 y_2)^{1-b} - (\bar{y}_1 \bar{y}_2)^{1-b}}{(y_1 \bar{y}_2)^{1-b} - (\bar{y}_1 y_2)^{1-b}}$$

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

$$y(1-y)^{1-b} \geq y^{2-b} \tag{C.10}$$

For the same reason,

$$y^{1-b}(1-y) \leq (1-y)^{2-b} \tag{C.11}$$

From (C.10) and (C.11), it follows that

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

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

$\square$

Using these properties of f we will prove Lemma C.4.

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

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

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

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

$$\Leftrightarrow y^{1-b}(y + (1-y)h_G) < (1-y)^{1-b}((1-y) + h_G y)$$

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

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 C.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 C.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 C.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 C.4. $\qquad\square$

Next we will prove Theorem 4 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 C.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 C.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 (C.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 C.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 $t + 1$.

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

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,

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

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

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

Lemma C.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 (C.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 C.6. We will prove statement (1); statement (2) can be proven using a similar argument.

Assume $x_0 < \hat{x}$. Then, from Lemma C.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 (C.4). From Proposition C.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, $x_i(t+1) > x_i(t)$. $\qquad\square$

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

Lemma C.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 (C.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 C.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 C.5 and Lemma C.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 C.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. $\qquad\square$

This completes the proof of Theorem 4.

24

D Proofs of Section 5

Proof of Theorem 6.

Lemma D.1. *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 $p_r \le x_i$, and then we show that if $p_r > \frac{1}{2}$ and $p_r \le x_i$, SimpleSALSA is polarizing with respect to i if and only if i is biased.

$$p_r = \sum_{\substack{j \in V_2 : j_2 \text{ is RED}}} \mathbb{P}[i \xrightarrow{3} j]$$

$$= \sum_{\substack{j_1 \in N(i) \\ j_1 \text{ is RED}}} \mathbb{P}[i \xrightarrow{1} j_1] \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2 \in N(i) \\ j_2 \text{ is BLUE}}} \mathbb{P}[i \xrightarrow{1} j_2] \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \mathbb{P}[j_2 \xrightarrow{2} j]$$

$$= \sum_{\substack{j_1 \in N(i) \\ j_1 \text{ is RED}}} \frac{1}{|N(i)|} \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \mathbb{P}[j_1 \xrightarrow{2} j] + \sum_{\substack{j_2 \in N(i) \\ j_2 \text{ is BLUE}}} \frac{1}{|N(i)|} \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \mathbb{P}[j_2 \xrightarrow{2} j]$$

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

$$= \sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} \frac{Z_{ij_1}}{|N(i)|} \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \sum_{i' \in N(j_1) \cap N(j)} \frac{1}{|N(j_1)|} \frac{1}{|N(i')|} + \sum_{\substack{j_2 \in V_2 \\ j_2 \text{ is BLUE}}} \frac{Z_{ij_2}}{|N(i)|} \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \sum_{i' \in N(j_2) \cap N(j)} \frac{1}{|N(j_2)|} \frac{1}{|N(i')|}$$

$$= \sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} \frac{Z_{ij_1}}{|N(i)|} \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \sum_{i' \in V_1} \frac{Z_{i'j_1} Z_{i'j}}{|N(j_1)||N(i')|} + \sum_{\substack{j_2 \in V_2 \\ j_2 \text{ is BLUE}}} \frac{Z_{ij_2}}{|N(i)|} \sum_{\substack{j \in V_2 \\ j \text{ is RED}}} \sum_{i' \in V_1} \frac{Z_{i'j_2} Z_{i'j}}{|N(j_2)||N(i')|}$$

By Lemma 5.1, in the limit as $n \to \infty$, with probability 1,

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

and

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

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

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

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

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

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

$$\frac{p_r x_i}{p_r x_i + (1 - x_i)(1 - p_r)} > \frac{p_r x_i}{p_r x_i + p_r(1 - x_i)} \quad \left(\text{since } p_r > \frac{1}{2}, \text{ from (D.1)}\right) = x_i$$

Therefore, by definition, SimpleSALSA is polarizing. $\qquad\square$

Lemma D.2. *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 $q_{j\text{RED}}$ 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 $q_{j\text{BLUE}}$ be the corresponding probability that j^* is BLUE. Then,

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

$$= \sum_{\substack{j_1 \in N(i) \\ j_1 \text{ is RED}}} \frac{1}{|N(i)|} q_{j_1 \text{RED}} + \sum_{\substack{j_2 \in N(i) \\ j_2 \text{ is BLUE}}} \frac{1}{|N(i)|} q_{j_2 \text{RED}}$$

$$= \sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} \frac{Z_{ij_1}}{|N(i)|} q_{j_1 \text{RED}} + \sum_{\substack{j_2 \in V_2 \\ j_2 \text{ is BLUE}}} \frac{Z_{ij_1}}{|N(i)|} q_{j_2 \text{RED}}$$

Consider T two-step random walks starting at a node $j_1 \in N(i)$. Observe that $q_{j_1 \text{RED}}$ 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 $T \to \infty$,

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

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,

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

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

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

26

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, $q_{j_1\text{RED}} = 1$. By symmetry $q_{j_2\text{RED}} = 1 - q_{j_2\text{BLUE}} = 0$. Moreover, by Lemma 5.1, 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 $n \to \infty$,

$$p_r = x_i \tag{D.2}$$

The rest of the analysis is identical to Lemma D.1. □

This completes the proof of Theorem 6. □

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 $T \to \infty$,

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

since as $T \to \infty$, 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 D.2, 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 5.1, in the limit as $n \to \infty$, $\sum_{\substack{j_1 \in V_2 \\ j_1 \text{ is RED}}} \frac{Z_{ij_1}}{|N(i)|} \to x_i$, with probability 1. Therefore, with probability 1,

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

27

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

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

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

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

with probability 1. In other words, $p_r = 1$. So, the probability that a book recommended by SimplePPR was RED given that it was accepted is exactly p_r regardless of whether i is biased or unbiased. Therefore, SimplePPR is polarizing.

$\square$