arXiv:2403.16940v1 [cs.SI] 25 Mar 2024

Dynamics of Affective Polarization: From Consensus to Partisan Divides

Buddhika Nettasinghe,[1*] Allon G. Percus,[2] Kristina Lerman[3]

[1]University of Iowa,
[2]Claremont Graduate University
[3]USC Information Sciences Institute

*To whom correspondence should be addressed; E-mail: buddhika-nettasinghe@uiowa.edu.

Politically divided societies are also often divided emotionally: people like and trust those with similar political views (in-group favoritism) while disliking and distrusting those with different views (out-group animosity). This phenomenon, called affective polarization, influences individual decisions, including seemingly apolitical choices such as whether to wear a mask or what car to buy. We present a dynamical model of decision-making in an affectively polarized society, identifying three potential global outcomes separated by a sharp boundary in the parameter space: consensus, partisan polarization, and non-partisan polarization. Analysis reveals that larger out-group animosity compared to in-group favoritism, i.e. *more hate than love*, is sufficient for polarization, while larger in-group favoritism compared to out-group animosity, i.e., *more love than hate*, is necessary for consensus. We also show that, counter-intuitively, increasing cross-party connections facilitates polarization, and that by emphasizing partisan differences, mass media creates self-

fulfilling prophecies that lead to polarization. Affective polarization also creates *tipping points* in the opinion landscape where one group suddenly reverses their trends. Our findings aid in understanding and addressing the cascading effects of affective polarization, offering insights for strategies to mitigate polarization.

1 Introduction

American society has grown more ideologically divided, with Democrats and Republicans not only disagreeing on policy issues but also making dramatically different choices about where to live and work, what products to buy, leisure activities to pursue (1) or sports teams to support (2). Surveys also reveal a growing emotional divide, with members of each party increasingly disliking and distrusting the opposing party (3, 4). This phenomenon, called affective polarization, is manifested in people expressing warm feelings, i.e., *in-group love*, towards their ideological allies but negative feelings and animosity, i.e., *out-group hate*, to members of the opposing party. Over the last decade, cross-party antipathy has grown and now exceeds in-group love (5, 6). The escalating partisan animosity poses a challenge to effective governing and the well-being of society. For example, during the COVID-19 pandemic individuals' trust and adherence to public health recommendations, like wearing a mask or getting vaccinated, were shaped by whether their own political party supported or opposed those recommendations (7), hindering an effective response to the pandemic.

Research has shown that demographics alone cannot account for the partisan divide in beliefs and behaviors (8, 9, 10). Instead, these phenomena arise from collective social dynamics. The tendency to associate with others who are similar, a process known as homophily, amplifies chance correlations between individual preferences and ideology, giving rise to a unified behavior within a group over time. This effect was used to explain the emergence of stereotypes like “latte-drinking liberals” and “bird-hunting conservatives” (*1*). The rise of online media has further amplified social cleavages by enabling people to align their information environments with their ideology. Similar to the mechanisms described above, these preferences tend to segregate people within ideologically-homogeneous communities, i.e., echo chambers (*11, 12*), which insulate them from opposing views and promote polarization. However, recent research has challenged this understanding (*13*), pointing to studies that show instead how increasing polarization can arise from exposure to opposing views.

This paper presents a model of information cascades in an affectively polarized social network composed of two groups (e.g., red and blue), where individuals within each group like and trust members of their own group (in-group love) and dislike and distrust members of the other group (out-group hate). When choosing between two possible choices (e.g., wear a mask or not, get vaccinated or not, which team to support in the Superbowl), individuals observe their social connections and attempt to *conform* to the choices of their in-group and *oppose* choices made by members of their out-group. Depending on the size of the minority and majority groups, homophily (preference of individuals to connect to others of the same group), and the levels of in-group conformity and out-group opposition, several different long-term outcomes can emerge, marked by a sharp boundary: global consensus (all individuals adopt the same choice), polarization (party-line division of choices) and non-partisan polarization in which each group’s choices are uniformly divided. We theoretically characterize the conditions under which such outcomes occur and provide numerical experiments that yield further insights.

Despite its simplicity, the model exhibits remarkably complex behaviors and reconciles seemingly contradictory findings from literature. The model explains how rapid collective transitions, or *tipping points* in the opinion landscape (*14*), can emerge in social systems. It shows that opposition to the choices by members of the other party, driven by out-group hate, is a potent driver of polarization. When out-group hate is stronger than in-group love, no consensus is feasible. This may explain why disagreement on issues between Democrats and Republicans accelerated since 2012, when out-group hate exceeded in-group love in the U.S. (*6*). The model also explains why conventional wisdom-based approaches aimed at reducing polarization, such as connecting people from opposite parties, often backfire (*15, 13*). Specifically, our results corroborate the findings in (*16*) showing that consensus between two antagonistic communities can be achieved only when they are loosely connected. Beyond this, our analysis provides a comprehensive explanation for role of out-group hate, in-group love, cross-party connections and the initial beliefs in shaping opinions. Our work suggests that emphasizing partisan differences, even when they are small or non-existent, can fuel polarization through a self-fulfilling prophecy. To counteract this, news media and social platforms could instead strive to diminish the perception of party-line differences to impede actual polarization. For example, fostering connections between similar individuals from opposing parties may be one of the few effective methods to facilitate consensus.

Our model is useful to understand the forms of divisions that emerge collectively from affective polarization, homophily and imbalanced party sizes and leads to new insights into polarization as well as methods to mitigate it. The theoretical tractability of the model, which yields closed-form expressions for its dynamics, reduces the need to rely on large scale simulations to obtain such insights and may lead to new solutions to control polarization.

2 A Model of Information Cascades with Affective Polarization

We present a dynamical model of how people make choices in a social network (e.g., to mask or support a sports team) by viewing the past choices of their in-group (e.g., members of their own party), which they approve of, as well as the choices of their out-group (e.g., cross-party members), which they oppose. The choice dynamics lead to an information cascade which reaches a steady state of partisan polarization or consensus depending on group sizes and the levels of in-group love and out-group hate.

Consider an undirected social network $G = (V, E)$ with $N = |V|$ individuals. Each individual (node) $v \in V$ has two binary attributes: a static binary attribute $R(v) \in \{0, 1\}$ and a dynamic binary attribute $H_k(v) \in \{0, 1\}$ where k denotes discrete-time. The static attribute represents the group (e.g., political) affiliation: v is red ($v \in \mathcal{R}$) if $R(v) = 1$; otherwise, v is blue ($v \in \mathcal{B}$). Let $N^{\mathcal{B}} = |\mathcal{B}|$ and $N^{\mathcal{R}} = |\mathcal{R}|$ denote the sizes of the two groups and $r = N^{\mathcal{R}}/N$ denote the fraction of red nodes. The dynamic attribute $H_k(v) \in \{0, 1\}$ represents v's choice at time k (e.g., wearing a mask vs not wearing a mask).

At each time k (where $k = 0, 1, 2, \dots$), a node $X_k \in V$ chosen uniformly at random updates its choice by observing the choices of its neighbors.Let

$$\begin{aligned}
d_k^{in,0}(X_k) &= \sum_{(X_k,u)\in E} \mathbb{1}(R(u) = R(X_k) \wedge H_k(u) = 0)/d(X_k) \\
d_k^{in,1}(X_k) &= \sum_{(X_k,u)\in E} \mathbb{1}(R(u) = R(X_k) \wedge H_k(u) = 1)/d(X_k) \\
d_k^{out,0}(X_k) &= \sum_{(X_k,u)\in E} \mathbb{1}(R(u) \neq R(X_k) \wedge H_k(u) = 0)/d(X_k) \\
d_k^{out,1}(X_k) &= \sum_{(X_k,u)\in E} \mathbb{1}(R(u) \neq R(X_k) \wedge H_k(u) = 1)/d(X_k)
\end{aligned} \tag{1}$$

denote the number of in-group and out-group neighbors with choice-0 and choice-1 at time k normalized by the total number of neighbors $d(X_k)$. Node X_k updates its choice at $k+1$ according to:

$$H_{k+1}(X_k) = \begin{cases} 0 & \text{if } \alpha\left(d_k^{in,1}(X_k) - d_k^{in,0}(X_k)\right) - \beta\left(d_k^{out,1}(X_k) - d_k^{out,0}(X_k)\right) < -\delta \\ 1 & \text{if } \alpha\left(d_k^{in,1}(X_k) - d_k^{in,0}(X_k)\right) - \beta\left(d_k^{out,1}(X_k) - d_k^{out,0}(X_k)\right) > \delta \\ H_k(X_k) & \text{otherwise,} \end{cases} \tag{2}$$

where $\alpha, \beta, \delta \in [0, 1]$ are constant model parameters. Choices of all other nodes except $X_k \in V$ remain unchanged: for all $u \neq X_k$, $H_{k+1}(u) = H_k(u)$.

The above stylized model aims to capture the dynamics of choices in an affectively polarized society. Consider a red node v deciding whether to wear a mask during the pandemic. The red neighbors (in-group) that wear masks push v towards masking, whereas the red neighbors who do not wear masks push v towards not-masking. The out-group (blue) neighbors have the opposite effect: blue masking neighbors push node v towards not-masking, whereas blue non-masking neighbors push the node towards masking. The relative strengths of these effects, *in-group love* and *out-group hate*, are quantified by α and β, respectively. If the combined effect of out-group hate and in-group love exceeds δ in favor of a certain choice (1 or 0), then v adopts it. If not, it keeps it current choice. Thus, δ quantifies the level of *inertia* of a person, or the degree of social proof, including from the out-group, required to change the choice. Also note from Eq. 2 that, among the neighbors of v belonging to each group, only the difference between how many chose choice-0 and choice-1 matters and not the ratio. Even with the normalization in Eq. 1, 50 out of a total of 100 masking blue neighbors will create a greater out-group effect for a red node than when one out of two blue neighbors masks.

To analyze the dynamics, we examine the fraction of nodes in each group that have adopted choice-1 at time k. Formally, we define the state of the system at time k as the column vector $\theta_k = [\theta_k^{\mathcal{B}}, \theta_k^{\mathcal{R}}]'$ where,

$$\theta_k^{\mathcal{B}} = \frac{\sum_{v \in V} \mathbb{1}(R(v) = 0 \wedge H_k(v) = 1)}{\sum_{v \in V} \mathbb{1}(R(v) = 0)}, \quad \theta_k^{\mathcal{R}} = \frac{\sum_{v \in V} \mathbb{1}(R(v) = 1 \wedge H_k(v) = 1)}{\sum_{v \in V} \mathbb{1}(R(v) = 1)}. \tag{3}$$

Since the node X_k is chosen randomly at time k to update its choice, the trajectory of the system $\theta_k = [\theta_k^{\mathcal{B}}, \theta_k^{\mathcal{R}}]', k = 0, 1, 2, \dots$ is also a random process. We show that the discrete-time stochastic trajectory $\theta_k, k = 0, 1, 2, \dots$ can be approximated using the continuous-time deterministic trajectory of a differential equation under a few assumptions. This differential equation representation of the stochastic model, called the *limit mean differential equation* can thus be used to analyze the emergence of various patterns in the social network over sufficiently large time horizons. We will focus on two cases of practical interest: a fully connected network and a stochastic block model.

2.1 Dynamics of the Model in a Fully Connected Network

We first consider a fully connected social network $G = (V, E)$, where each node $v \in V$ can observe the state of the system $\theta_k = [\theta_k^{\mathcal{B}}, \theta_k^{\mathcal{R}}]'$ at any time k. This occurs, for example, when people are informed about the prevalence of masking within each political party via daily news broadcasts and make their decisions to mask accordingly.

In such a graph, the piece-wise interpolation[1] of the discrete-time trajectory $\theta_k = [\theta_k^{\mathcal{B}}, \theta_k^{\mathcal{R}}]', k = 0, 1, 2, \dots$ can be approximated using the continuous-time trajectory $\theta(t) = [\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]', t \geq 0$ of the following differential equation as the number of nodes in the graph N is large:

$$\begin{bmatrix} \dot{\theta}^{\mathcal{B}} \\ \dot{\theta}^{\mathcal{R}} \end{bmatrix} = \begin{bmatrix} \left(1 - \theta^{\mathcal{B}}\right) p_\theta^{\mathcal{B}}(0 \to 1) - \theta^{\mathcal{B}} p_\theta^{\mathcal{B}}(1 \to 0) \\ \left(1 - \theta^{\mathcal{R}}\right) p_\theta^{\mathcal{R}}(0 \to 1) - \theta^{\mathcal{R}} p_\theta^{\mathcal{R}}(1 \to 0) \end{bmatrix}, \tag{4}$$

where,

$$\begin{aligned}
p_\theta^{\mathcal{B}}(0 \to 1) &= \mathbb{1}\left(\alpha(1-r)\left(2\theta^{\mathcal{B}} - 1\right) - \beta r\left(2\theta^{\mathcal{R}} - 1\right) > \delta\right), \\
p_\theta^{\mathcal{B}}(1 \to 0) &= \mathbb{1}\left(\alpha(1-r)\left(2\theta^{\mathcal{B}} - 1\right) - \beta r\left(2\theta^{\mathcal{R}} - 1\right) < -\delta\right), \\
p_\theta^{\mathcal{R}}(0 \to 1) &= \mathbb{1}\left(\alpha r\left(2\theta^{\mathcal{R}} - 1\right) - \beta(1-r)\left(2\theta^{\mathcal{B}} - 1\right) > \delta\right), \\
p_\theta^{\mathcal{R}}(1 \to 0) &= \mathbb{1}\left(\alpha r\left(2\theta^{\mathcal{R}} - 1\right) - \beta(1-r)\left(2\theta^{\mathcal{B}} - 1\right) < -\delta\right).
\end{aligned}$$

The intuition behind the differential equation in Eq. 4 is as follows. In a fully connected network, each node is a neighbor of all other nodes. Thus, the node-level statistics in Eq. 1 can be written using the population statistics in Eq. 3. For a blue node X_k, we can write $d_k^{in,1}(X_k) = \theta_k^{\mathcal{B}}, d_k^{in,0}(X_k) = 1 - \theta_k^{\mathcal{B}}, d_k^{out,1}(X_k) = \theta_k^{\mathcal{R}}, d_k^{out,0}(X_k) = 1 - \theta_k^{\mathcal{R}}$. According to Eq. 2, a blue node X_k picks choice-1 when $\alpha(1-r)\left(2\theta_k^{\mathcal{B}} - 1\right) - \beta r\left(2\theta_k^{\mathcal{R}} - 1\right) > \delta$, i.e., positive

[1] The piece-wise interpolation of $\theta_k, k = 0, 1, 2, \dots$ refers to the continuous time trajectory $\theta^{\frac{1}{N}}(t) = \theta_k$ for $t \in \left[\frac{k}{N}, \frac{k+1}{N}\right)$ for discrete time $k = 0, 1, 2, \dots$

influence from the presence of choice-1 among in-group neighbors is larger than the negative influence from the presence of choice-1 among out-group neighbors by a margin of at least δ. Similarly, a blue node picks choice-0 when $\alpha(1-r)\left(2\theta_k^{\mathcal{B}}-1\right)-\beta r\left(2\theta_k^{\mathcal{R}}-1\right)<-\delta$. Since a fraction $1-\theta_k^{\mathcal{B}}$ of blue nodes have choice-0 and a fraction $\theta_k^{\mathcal{B}}$ of blue nodes have choice-1, the expected rate of change of blue nodes with choice-1 $\theta_k^{\mathcal{B}}$ can thus be written as $\dot{\theta}^{\mathcal{B}}$ in Eq. 4, and similarly for $\dot{\theta}^{\mathcal{R}}$. When the network is large, the stochastic dynamics converge to the deterministic differential equation in Eq. 4 according to stochastic averaging theory. The formal proof of convergence is given in Supplementary Information (SI) A. Thus, for any initial state $\theta(0)=[\theta^{\mathcal{B}}(0),\theta^{\mathcal{R}}(0)]'$, the continuous-time trajectory $\theta(t)=\theta(0)+\int_0^t\dot{\theta}(s)ds, t\geq 0$ obtained using Eq. 4 approximates the stochastic model dynamics $\theta_k=[\theta_k^{\mathcal{B}},\theta_k^{\mathcal{R}}]', k=0,1,2,\ldots$.

In the remainder of the paper, we rely on the differential equation in Eq. 4 and its generalizations to explore how polarized information cascades emerge in affectively polarized populations.

2.2 Dynamics of the Model on a Social Network with Communities

Next, we consider the case where the network $G=(V,E)$ is sampled from a stochastic block model with two communities. Specifically, each node is connected to a node in the same party with probability ρ and a node in the other party with probability $1-\rho$, where $\rho\in(0,1)$ is a constant model parameter. Thus, ρ quantifies the level of *homophily* (*17*) of the individuals in the population: $\rho>0.5$ implies that individuals are more likely to connect with others of the same party (homophily), whereas $\rho<0.5$ implies that individuals tend to mostly connect with members of the other party (heterophily). When $\rho=0.5$, the graph can be viewed as an Erdős-Rényi random graph with each edge being formed with a probability of 0.5.

Alternatively, ρ can be interpreted in the following way: each individual looks at a fraction ρ of their in-group members and a fraction $1-\rho$ of their out-group members and makes a decision based on their choices. Thus, ρ might also be used to represent the balance of information an individual receives from the news media in terms of how well they represent the two parties: $\rho > 0.5$ means the news consumed by an individual over-represents views of the in-group (relative to its size), while $\rho < 0.5$ means that the news over-represents the views of the out-group (relative to its size). When $\rho = 0.5$, each group is represented in the news proportionate to its group size.

The dynamics of the system $\theta_k = [\theta_k^{\mathcal{B}}, \theta_k^{\mathcal{R}}]', k = 0, 1, 2, \dots$ in a stochastic block model network can be approximated using the continuous-time trajectory of Eq. 4 with α replaced by $\alpha\rho$ and β replaced by $\beta(1-\rho)$. In other words, the homophily ρ amplifies the effects of in-group love while reducing the effects of out-group hate. The exact differential equation for the stochastic block model is stated in SI B.

3 Results

We analyze dynamics of the model and obtain insights about information cascades in an affectively polarized society. We first focus on a fully connected population with no inertia (i.e., $\delta = 0$) that starts from an initial state with no party-dependency ($\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$). The case $\delta = 0$ describes a highly reactive population where individuals choices are driven by the direction of the net effect of in-group neighbors and out-group animosity and not the amount. Then, we extend the results to more general settings with homophily, and party-dependent initial states ($\theta^{\mathcal{B}}(0) \neq \theta^{\mathcal{R}}(0)$).

3.1 Emergence of Polarization in a Fully Connected Network

Consider the case where choice-1 is initially equally popular in both groups ($\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$). This describes the early COVID-19 pandemic, when Democrats and Republicans were equally cautious about the disease and chose to mask. Remarkably, the long-term outcomes that emerge from a symmetric initial state can be characterized by just two quantities: the ratio of in-group love to out-group hate α/β and the ratio of group sizes $r/(1-r)$.

Theorem 1 (*Information Cascades in a Fully Connected Network with Affective Polarization*)**.** *Consider Eq. 4 which represents the dynamics of the state of the population $\theta(t) = [\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]'$ under the proposed model in a fully connected graph. Let $\delta = 0$ (i.e., no inertia) and $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$ (i.e., initial state is party independent). Then, the following statements characterize the asymptotic state of the system for various different values of α (level of in-group love), β (level of out-group hate) and r (fraction of red nodes in the network):*

- *Case 1: Let $\frac{\beta}{\alpha} < \frac{r}{1-r} < \frac{\alpha}{\beta}$. If $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) > 0.5$, then $\lim_{t\longrightarrow\infty} \theta(t) = [\theta^{\mathcal{B}}_*, \theta^{\mathcal{R}}_*]' = [1, 1]'$. If $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) < 0.5$, then $\lim_{t\longrightarrow\infty} \theta(t) = [\theta^{\mathcal{B}}_*, \theta^{\mathcal{R}}_*]' = [0, 0]'$ i.e., there is no polarization and both groups fully adopt the choice that was initially more popular.*

- *Case 2: Let $\frac{r}{1-r} > \frac{\alpha}{\beta}$ and $\frac{r}{1-r} > \frac{\beta}{\alpha}$. If $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) > 0.5$, then $\lim_{t\longrightarrow\infty} \theta^{\mathcal{R}}(t) = [\theta^{\mathcal{B}}_*, \theta^{\mathcal{R}}_*]' = [1, 0]'$. If $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) < 0.5$, then $\lim_{t\longrightarrow\infty} \theta(t) = [\theta^{\mathcal{B}}_*, \theta^{\mathcal{R}}_*]' = [0, 1]'$ i.e., there is party-line polarization and the red-group (which is the majority) fully adopt the choice that was initially popular while the blue-group fully adopt the other choice.*

- *Case 3: Let $\frac{r}{1-r} < \frac{\alpha}{\beta}$ and $\frac{r}{1-r} < \frac{\beta}{\alpha}$. If $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) > 0.5$, then $\lim_{t\longrightarrow\infty} \theta^{\mathcal{R}}(t) = [\theta^{\mathcal{B}}_*, \theta^{\mathcal{R}}_*]' = [0, 1]'$. If $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) < 0.5$, then $\lim_{t\longrightarrow\infty} \theta(t) = [\theta^{\mathcal{B}}_*, \theta^{\mathcal{R}}_*]' = [1, 0]'$ i.e., there is party-line polarization and the blue-group (which is the majority) fully adopt the choice that was initially popular while the red-group fully adopt the other choice.*

- *Case 4: Let $\frac{\beta}{\alpha} > \frac{r}{1-r} > \frac{\alpha}{\beta}$. If $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) > 0.5$, then $\lim_{t\longrightarrow\infty} \theta(t) = [\theta^{\mathcal{B}}_*, \theta^{\mathcal{R}}_*]' = [0.5, 0.5]'$. i.e., there is non-partisan polarization with half of each group adopting choice-1 and the remaining half adopting choice-0.*

The limiting states in Cases 1-3 (consensus and polarization along party lines) are locally asymptotically stable stationary states of the system in Eq. (4) *whereas the limiting state in*

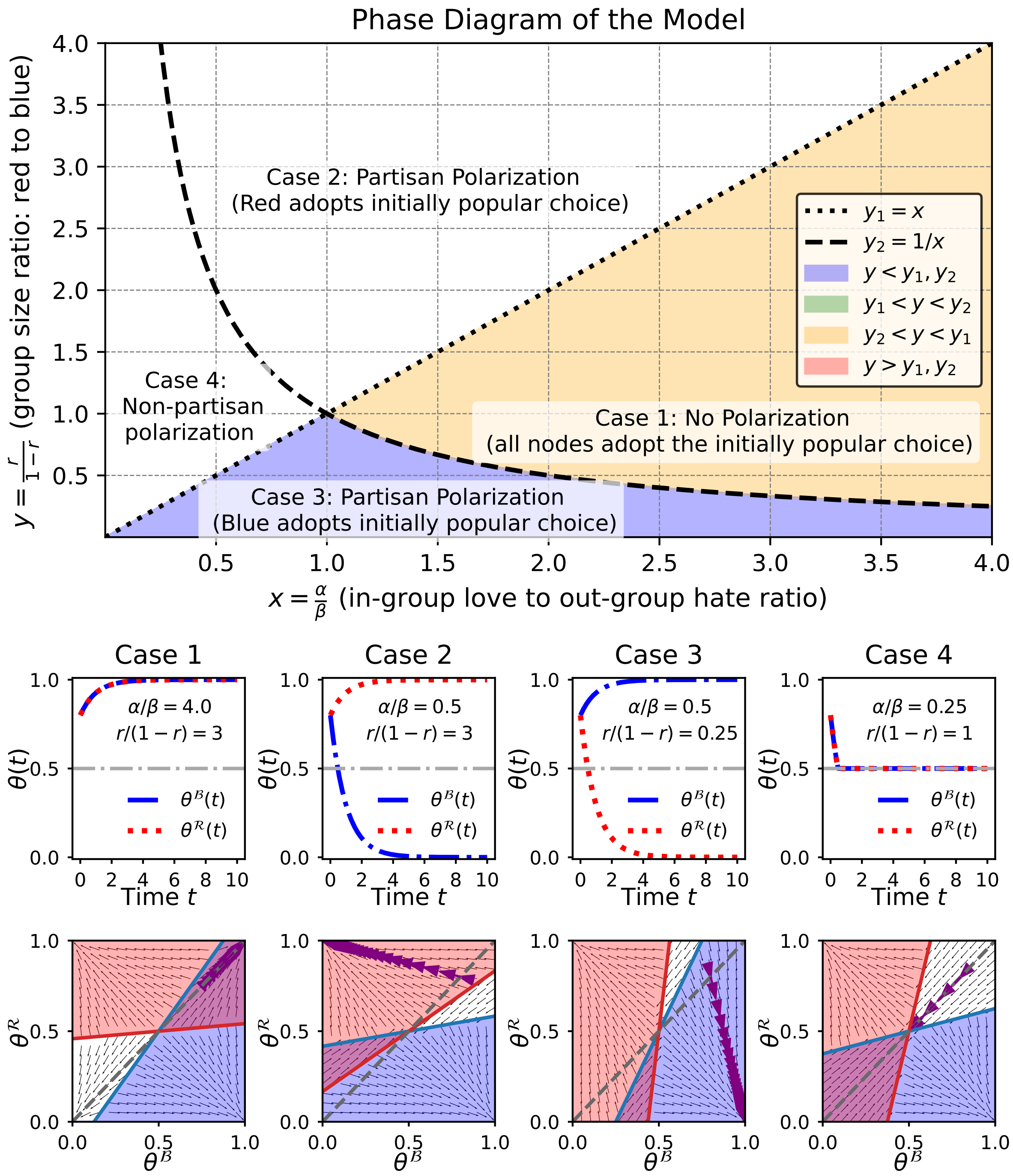


Figure 1: Phase diagram of the model (top) and four example trajectories. The four different regions of the phase diagram (defined by the ratio of in-group love to out-group hate and the ratio of group sizes) lead to different long-term outcomes in a fully connected network when both groups start from the same initial state ($\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$). The long-term outcomes are: **(Case 1, yellow)** No Polarization, **(Case 2, red / Case 3, blue)** Partisan Polarization, **(Case 4, green)** Non-Partisan Polarization. Example trajectories in both time-domain and state space are shown below the phase diagram. The blue and red color areas in state space indicate regions where $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$ increase (i.e., regions where $p^{\mathcal{B}}_{\theta}(0 \to 1) = 1$ and $p^{\mathcal{R}}_{\theta}(0 \to 1) = 1$ according to Eq. 4). The black arrows in state space plots indicate the path of the differential equation Eq. 4. The purple arrows map the time domain trajectory to the state space.

Case 4 is an unstable stationary state of Eq. (4).

3.1.1 Insights from Theorem 1

The four cases in Theorem 1 shed light on the forms of polarization that can emerge in an emotionally divided population starting from a state with no group-level differences: (case 1) global consensus, where all nodes ultimately adopt the same choice, (case 2 and 3) party-line polarization, where the choices are split along party lines, and (case 4) non-partisan polarization, where each group is split evenly between the two choices. Below we consider additional insights from Theorem 1.

Out-group hate is necessary for polarization: Note from Fig. 1, that if β is approximately zero, then the network will always be in Case 1 which achieves consensus from any party-independent initial state $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) \neq 0.5$.

Larger out-group hate relative to in-group love is sufficient for polarization: When individual choices are driven more by a desire to oppose the out-group than a desire to conform to the in-group, some form of polarization is unavoidable regardless of group sizes. As a result, in the region to the left of the vertical line at $\alpha/\beta = 1$ in Fig. 1, consensus is not possible. If out-group hate is very high compared to in-group love ($\alpha/\beta \approx 0$ corresponding to case 4), then each group will be evenly split between the two choices. When the disparity between α and β is not too large compared to group size disparity (i.e., $\beta/\alpha < r/(1-r)$ or $\alpha/\beta > r/(1-r)$), polarization will emerge with the majority adopting the initially more popular choice and the minority adopting the other choice (Case 2 and Case 3 in Theorem 1). Further, party-line polarization is stable: a small deviation will push the system back to the polarized state as indicated by the arrows pointing to the polarized state in the state space plots of Fig. 1. Additional examples trajectories in the cases where polarization emerge are given in SI Fig. S2.

Larger in-group love relative to out-group hate leads to consensus as long as the group imbalance is not too large: When the two groups have the same size (i.e., $r = 0.5$), Case 1 of Theorem 1 shows that even a slightly larger in-group love compared to the out-group hate (i.e., $\alpha > \beta$) is sufficient for the network to adopt the initially popular choice, leading to consensus (see row i of SI Fig. S3 for an example). Even with unequal group sizes, consensus can be achieved with larger in-group love as long as the group imbalance is not large enough to push the system into Case 2 or Case 3. In other words, when α is sufficiently large compared to β , consensus can be achieved even when group sizes are not highly unequal (see row ii of SI Fig. S3 for an example). Further, note that when β is negligible compared to α, consensus is always achieved when both groups start from the same initial state (grey diagonal line in state space plots). This highlights our claim that out-group hate is crucial for any form of polarization to occur from a party independent initial state $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$. However, even with high in-group love $\alpha > \beta$, a large enough group imbalance ($r/(1-r) > \alpha/\beta$ or $r/(1-r) < \beta/\alpha$) can lead to polarization (as shown in row iii of SI Fig. S3). This observation emphasizes that *more love than hate is necessary but not sufficient for consensus*.

Majority cannot fully adopt the initially unpopular choice: When $r > 0.5$ (region above $y = 1$ line in Fig. 1) and $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) > 0.5$ (i.e., choice-1 is initially more popular), there cannot be a case where all of the red-group adopts choice-0. In general, starting from a state $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$ in a fully connected network, the majority cannot adopt the initially less popular choice.

Small perturbations from non-partisan polarization (case 4) can lead to party-line polarization but not to consensus: Consider Case 4 in Theorem 1 where the population is evenly split between the two choices, regardless of group membership. This stationary state $\theta^{\mathcal{B}}(t) = \theta^{\mathcal{R}}(t) = 0.5$ is unstable, and a small change in $\theta^{\mathcal{B}}(t)$ or $\theta^{\mathcal{R}}(t)$ can lead the population to polarize along party lines. This can be seen from state space plot corresponding to Case 4 in Fig. 1: a small deviation from $\theta^{\mathcal{B}}(t) = \theta^{\mathcal{R}}(t) = 0.5$ caused by a change of either $\theta^{\mathcal{B}}(t)$ or $\theta^{\mathcal{R}}(t)$ will lead to party-line polarization. For example, if just a few red nodes switch to choice-1 from choice-0, $\theta^{\mathcal{B}}(t)$ will converge to 1 and $\theta^{\mathcal{R}}(t)$ to 0.

Thus, even on a fully mixed population containing a majority and a minority that are not initially polarized, out-group hate and in-group love alone can lead to the emergence of a wide array of cascading choices.

3.2 Implications for Networks with Echo Chambers

Next, we consider the case where the network $G = (V, E)$ is sampled from a stochastic block model with two communities, where each node is connected to ρ fraction of their in-group members and $1 - \rho$ fraction of their out-group members, and ρ gives the homophily of the network. Recall from Sec. 2.2 that the dynamics of the model with homophily can be obtained by replacing α and β in Eq. (4) with $\alpha\rho$ and $\beta(1 - \rho)$, respectively. Consequently, replacing α and β in Theorem 1 and Fig. 1 with $\alpha\rho$ and $\beta(1 - \rho)$ leads to a characterization of the forms of polarization that can emerge in the presence of in-group love, out-group hate, homophily as well as a minority/majority division of the population. This is illustrated in SI Fig. S1. We now discuss some insights on how these factors can collectively affect the emergence of polarization.

Neutral homophily is indistinguishable from the fully-connected graph: When people are neither homophilic nor heterophilic ($\rho = 0.5$), the continuous-time trajectory in a stochastic block model is the same as the continuous-time trajectory in a fully connected graph given in Eq. (4) (since both sides of the inequalities inside indicator functions in Eq. (4) would be multiplied by 0.5). Thus, Theorem 1 as well as insights discussed in Sec. 3.1 are applicable not only to fully connected graphs but also to Erdős-Rényi random graphs where edges are formed in an independent and identically distributed manner.

Highlighting the choices of the out-group in social networks may lead to polarization: A typical approach to reducing partisan divisions calls for increasing the number of cross-party links. For example, consider the case where the two parties are approximately equal in size ($r \approx 0.5$) and $\alpha > \beta$, which corresponds to Case 1 of Fig. 1 where $\frac{\beta}{\alpha} < \frac{r}{1-r} < \frac{\alpha}{\beta}$. Thus, when an individual looks at the entire population (i.e., a fully connected graph) or an unbiased sample of the population (i.e., an Erdős-Rényi random graph), universal consensus is achieved. Then, consider the case where the individual observes others in a biased manner, where each in-group member is observed with probability ρ and each out-group member with probability $1 - \rho$. If $\rho < 0.5$, the out-group will be over-represented compared to its size, amplifying the effect of out-group hate while reducing the effect of in-group love. Thus, the population could move to the red (Case 2) or blue regions (Case 3) of Fig. 1 where $\frac{\alpha\rho}{\beta(1-\rho)}, \frac{\beta(1-\rho)}{\alpha\rho} > \frac{r}{1-r}$ or $\frac{\alpha\rho}{\beta(1-\rho)}, \frac{\beta(1-\rho)}{\alpha\rho} < \frac{r}{1-r}$ i.e., partisan polarization can emerge starting from a uniform initial state where the choice is equally popular in both groups. Even a small increase in the number of cross-party links is likely to give rise to polarization (Case 2 or Case 3) from a non-polarized state (Case 1) when $\frac{\alpha\rho}{\beta(1-\rho)} \approx \frac{r}{1-r}$ or $\frac{\beta(1-\rho)}{\alpha\rho} \approx \frac{r}{1-r}$ (i.e., near the boundaries of Case 1 in the phase diagram of Fig. 1 with x-axis re-scaled as $\frac{\alpha\rho}{\beta(1-\rho)}$). Thus, *merely increasing the number of cross-party connections among the two groups may in fact facilitate polarization instead of consensus by amplifying the effect of out-group hate*. Figure 2 shows two different trajectories of $\theta(t)$ where the two groups start from the same initial state. Consensus is achieved for a homophilic network ($\rho = 0.7$), where individuals get more information about the in-group, while polarization emerges in an unbiased network ($\rho = 0.5$). This is because decreasing ρ from 0.7 to 0.5, pushes the network to Case 2 in Fig. 1 (with x-axis re-scaled as $\frac{\alpha\rho}{\beta(1-\rho)}$).

In fact, increased exposure to the out-group (i.e., decreasing ρ) can bring divisions to a

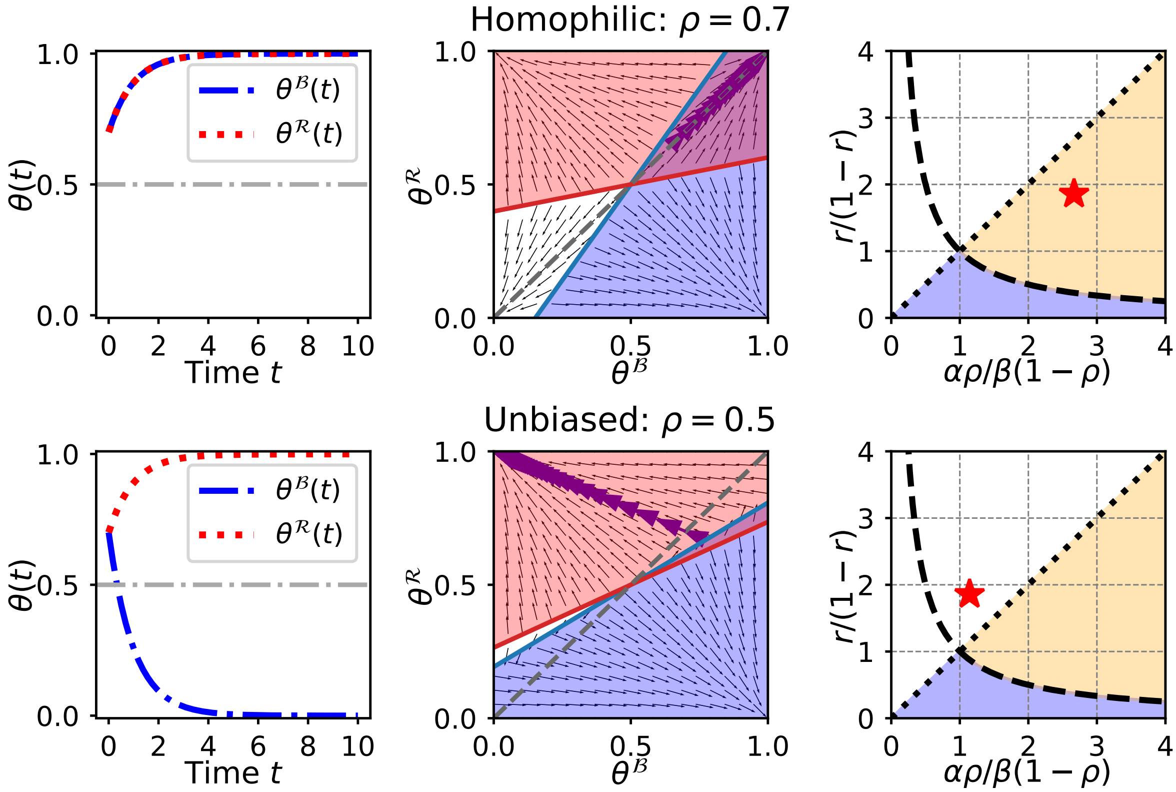


Figure 2: An illustration of how decreasing homophily can cause a party-line polarization. Both figures correspond to $\alpha = 0.8, \beta = 0.7$ (larger in-group favoritism compared to out-group animosity) and $r = 0.65$ (a majority red group). First row corresponds to a homophilic network (inter-group links are more likely to form than intra-group links) with $\rho = 0.7$ whereas second row corresponds to an unbiased network (all links are equally likely to form). Note that decreasing ρ from 0.7 (homophily) to 0.5 (unbiased) increases the effect of out-group hate and decreases the effect of in-group love on the choices, and pushes the social network from Case 1 (consensus) to Case 3 (party-line polarization) in Fig. 1 (with x-axis re-scaled as $\frac{\alpha\rho}{\beta(1-\rho)}$).

society already at global consensus. See SI Fig. S4 for an example. Note that global consensus remains at higher homophily (Case 1 in Fig. 1), and decreasing ρ to 0.5 makes the network unbiased but amplifies out-group hate, pushing it to Case 3, where the majority stays in the initial state but the minority adopts the choice that no one had chosen at the beginning. Further decreasing homophily makes the network highly heterophilic, where both groups focus largely on the out-group, pushing it to Case 4. As this state is unstable, a small deviation causes polarization with one group adopting choice-1 and the other adopting choice-0. Thus, in a society with multiple ideologies, choices being driven by what the *"opposition does"* more than what *"our own group does"* can lead to divisive (Case 2 and Case 3 in Fig. 1) and even unpredictable (Case 4 in Fig. 1) division of choices for a society that was initially united. In practice, such situations occur when partisan information sources (e.g., news organizations) emphasize the choices, decisions and actions of the out-group more than those of the in-group.

Relatedly, recall from Eq. 4 that when the two groups are approximately equal in size (i.e., $r \approx 0.5$) and $\rho = 0.5$ (unbiased network), people's choices are driven by $\theta(t) = [\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]'$ i.e., the prevalences of choice-1 in the in-group and out-group. If the popularity of choices is misrepresented in the information they receive at some time instant, that itself could lead to polarization. For example, consider *latte drinking* as the choice and assume that it is equally prevalent among liberals and conservatives. However, if conservatives are selectively exposed to latte-drinking liberals, giving the perception that latte drinking is highly prevalent among them, that may cause them to give up lattes due to the out-group hate effect, and that in turn would lead liberals to further embrace it. Once this divergence takes off, it will be further amplified by the in-group love, leading to the eventual polarization of a seemingly non-partisan choice (*1*). Thus, even if a choice is not initially polarized, making it appear to be so in the news or on social media by selectively emphasizing the out-group, can eventually lead to polarization in the form of a self-fulfilling prophecy. This serves as one possible explanation of why even traits that are historically non-partisan, such as the preferred choice of beverage, leisure activity, vocabulary, etc., can start to diverge along party lines when the prevalence of that trait in the opposite party is emphasized in the digital news (*13*).

3.3 Group-dependent Initial States

When choices are not initially identically distributed in the two groups, several interesting phenomena can emerge. The differential equation in Eq. 4 (and its generalization to stochastic block models) can be used to study such phenomena as well. We begin by stating a result which characterizes conditions that lead to consensus from a party-dependent initial state.

Theorem 2 (Consensus from Party-Dependent Initial States)**.** *Consider dynamics of the model on a fully connected graph given in Eq. 4 with $\delta = 0$ (i.e., no inertia). Consensus emerges from a group-dependent initial state $\theta^{\mathcal{B}}(0) \neq \theta^{\mathcal{R}}(0)$ if and only if,*

1. $\frac{\beta}{\alpha} < \frac{r}{1-r} < \frac{\alpha}{\beta}$*, and,*

2. *the initial state satisfies* $\frac{\beta r}{\alpha(1-r)} < \frac{2\theta^{\mathcal{B}}(0)-1}{2\theta^{\mathcal{R}}(0)-1} < \frac{\alpha r}{\beta(1-r)}$.

The first condition of Theorem 2 states that the system has to be in Case 1 of Fig. 1, which ensures that consensus is a stable steady state of the system. The second condition of Theorem 2 states that initial distribution of the choices within the groups cannot be too different from each other. The two conditions collectively ensure that consensus is reachable from the initial state. Any parameter configuration (α, β, r) or an initial state that does not satisfy the two conditions will give rise to polarization. The result further highlights the difficulties that lie in the path towards consensus in an affectively polarized society: even with high in-group love and balanced group sizes, the initial differences between the two parties can lead to polarized choices. In order to avoid this, social and news media through which people estimate the choice distributions must avoid emphasizing the differences between groups of different political ideologies.

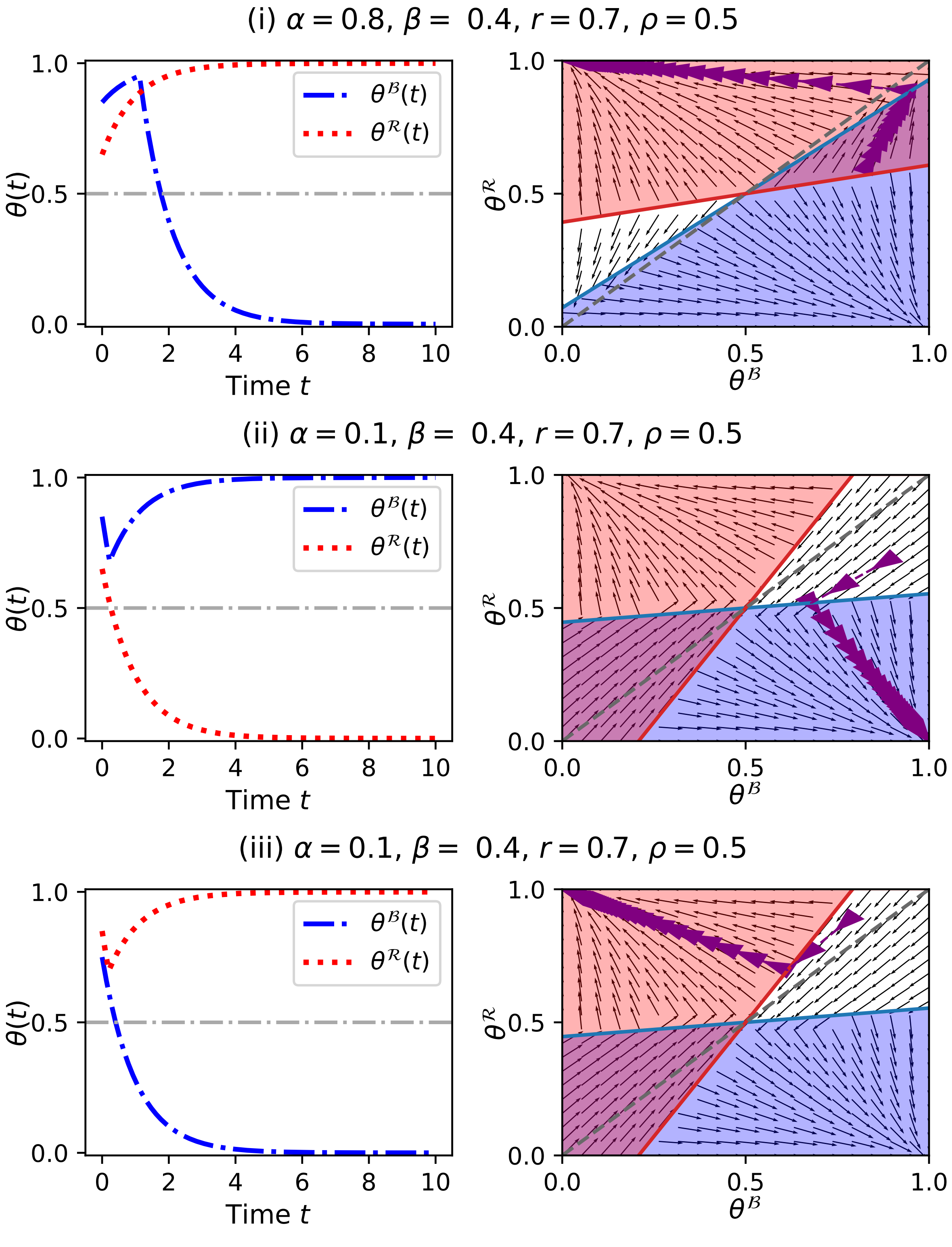


Figure 3: An illustration of three cases where the two groups start at different initial states i.e., $\theta^{\mathcal{B}}(0) \neq \theta^{\mathcal{R}}(0)$, and one group reverses its direction. In cases i and ii, the minority blue group reverses its direction. In case iii, the majority red group reverses its direction. The blue and red lines in state space indicate the *tipping points* in opinion landscape where the respective group reverses its trend when the trajectory reaches it. The proposed model can demonstrate a variety of such phenomena when the initial states are different for the two groups.

A group can flip: When the groups start from different initial states, their trajectories can change direction. For example, consider the three cases in Fig. 3. In case i of Fig. 3, in-group love is higher than out-group hate (i.e., $\alpha > \beta$) and choice-1 is initially more prevalent within each group but to a different degree. Due to higher in-group love, each group initially begins to embrace the choice-1 that is more popular within it. However, as this choice becomes more popular in the majority red group, the opposition intensifies in the minority blue group, which starts to adopt choice-0, leading to the eventual polarization. Interestingly, the *flip* occurs when the population is very closer to consensus. This represents how political negotiations in an affectively polarized society can very unexpectedly break down even when they are on the verge of reaching bi-partisan agreements: the high presence of the same choice in both groups amplifies the effect of out-group hate. More precisely, in-group love is high enough to get closer to consensus (due to the satisfied second condition of Theorem 2), but it is not high enough to make consensus a stable stationary state (due to violated first condition). More in-group love would drive both groups to consensus by focusing on unity within their own party rather than on hate towards the other party. Case ii and case iii of Fig. 3 show scenarios with higher out-group hate where both conditions of Theorem 2 are violated. In case ii, choice-1 is initially more prevalent in both groups but they both initially start adopting choice-0 due to higher out-group hate. However, as choice-0 becomes the more prevalent among the majority, the minority blue group starts adopting choice-1. Eventually, the trajectories converge in the opposite direction. Case iii of Fig. 3 shows a similar scenario where the majority red group reverses the trend. The theoretical tractability of the model Eq. 4 helps identify the exact trajectories for any initial state as seen from Fig. 3.

The majority can eventually fully adopt the initially less popular choice: Unlike the setting where both groups start in the same initial state, the majority can fully adopt the initially less popular choice when the two groups start in different initial states. For example, SI Fig. S5 shows an example of a case where choice-1 is initially more popular among both groups: $\theta^{\mathcal{B}}(0) = 0.9$ and $\theta^{\mathcal{R}}(0) = 0.6$. Also, 60% of the nodes in the network are red, making it the majority. However, the red group eventually abandons choice-1 due to the out-group hate effect resulting from the high popularity of choice-1 among the blue group (despite a smaller β). In other words, due to high initial unity of the minority blue group, the majority red group is driven more by a desire to oppose the blue party than to unite within their party. The minority blue group fully adopts choice-1 due to the higher in-group love effect created collectively by larger α and the high initial popularity of choice-1 within their group.

4 Conclusion

This paper introduced a dynamical model of decision making in a society where people trust the choices of those with same political views while distrusting the choices of those with opposing political views. The model is theoretically tractable and reveals the conditions for the emergence of consensus and partisan divisions from the initial state where there are no divisions. Our analysis highlights the importance of inter-group animosity in driving partisan division. Not only does out-group hate enable party-line polarization, but when it is larger than in-group love, consensus is no longer achievable. In particular, *more hate than love is sufficient for partisan divisions while more love than hate is necessary for consensus*. When partisan mass media emphasize the choices of the out-group more than in-group (i.e., focusing on the other more than own group), it amplifies the effects of out-group hate and facilitates the emergence of polarization. This may create self-fulfilling prophesies where the perceptions of polarization actually give rise to polarization and explains why, counter to our intuition, cross-party exposure facilitates polarization rather than deterring it. High out-group hate can shatter consensus even when both parties are on the brink of agreement, a trend that is becoming increasingly common within emotionally polarized societies.

The model and its theoretical tractability will also be useful to computational social scientists and network scientists to model the implications of affective polarization in future research and to gain insights on how to avoid its adverse implications on society.

References

1. D. DellaPosta, Y. Shi, M. Macy, Why do liberals drink lattes? *American Journal of Sociology* **120**, 1473–1511 (2015).

2. J. Wick, Taylor swift has driven some far-right pundits to do the unthinkable: Cheer for san francisco. *Los Angeles Times* (2024).

3. S. Iyengar, G. Sood, Y. Lelkes, Affect, not ideology: A social identity perspective on polarization. *Public opinion quarterly* **76**, 405–431 (2012).

4. S. Iyengar, S. J. Westwood, Fear and loathing across party lines: New evidence on group polarization. *American journal of political science* **59**, 690–707 (2015).

5. J. N. Druckman, J. Levy, 18. affective polarization in the american public. *Handbook on politics and public opinion* p. 257 (2022).

6. E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, *et al.*, Political sectarianism in america. *Science* **370**, 533–536 (2020).

7. G. Grossman, S. Kim, J. M. Rexer, H. Thirumurthy, Political partisanship influences behavioral responses to governors' recommendations for covid-19 prevention in the united states. *Proceedings of the National Academy of Sciences* **117**, 24144–24153 (2020).

8. S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, S. J. Westwood, The origins and consequences of affective polarization in the united states. *Annual review of political science* **22**, 129–146 (2019).

9. S. W. Webster, A. I. Abramowitz, The ideological foundations of affective polarization in the us electorate. *American Politics Research* **45**, 621–647 (2017).

10. S. Whitt, A. B. Yanus, B. McDonald, J. Graeber, M. Setzler, G. Ballingrud, M. Kifer, Tribalism in america: behavioral experiments on affective polarization in the trump era. *Journal of Experimental Political Science* **8**, 247–259 (2021).

11. D. Nikolov, D. F. Oliveira, A. Flammini, F. Menczer, Measuring online social bubbles. *PeerJ computer science* **1**, e38 (2015).

12. W. Chen, D. Pacheco, K.-C. Yang, F. Menczer, Neutral bots probe political bias on social media. *Nature communications* **12**, 5580 (2021).

13. P. Törnberg, How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences* **119**, e2207159119 (2022).

14. S. Thurner, New forms of collaboration between the social and natural sciences could become necessary for understanding rapid collective transitions in social systems. *Perspectives on Psychological Science* p. 17456916231201135 (2023).

15. C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, A. Volfovsky, Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115**, 9216–9221 (2018).

16. P. Siedlecki, J. Szwabiński, T. Weron, The interplay between conformity and anticonformity and its polarizing effect on society. *arXiv preprint arXiv:1603.07556* (2016).

17. M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks. *Annual review of sociology* **27**, 415–444 (2001).

18. J. Cortes, Discontinuous dynamical systems. *IEEE Control systems magazine* **28**, 36–73 (2008).

19. V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, Texts and Readings in Mathematics (Springer, 2023), second edn.

Supplementary Information

A Proof of Convergence and Uniqueness

A.1 Outline of the Proof and Preliminaries

High-level idea of the proof: The proof relies on the fact that the dynamics of $\theta_k, k = 0, 1, \dots$ are Markovian and the expected value of the next state given the previous state $\mathbb{E}_k\left\{\theta_{k+1}\right\} = \mathbb{E}\{\theta_{k+1}|\theta_k\}, k = 0, 1, \dots$ can be written as,

$$\begin{bmatrix}\mathbb{E}_k\left\{\theta^{\mathcal{B}}_{k+1}\right\}\\ \mathbb{E}_k\left\{\theta^{\mathcal{R}}_{k+1}\right\}\end{bmatrix} = \begin{bmatrix}\theta^{\mathcal{B}}_{k}\\ \theta^{\mathcal{R}}_{k}\end{bmatrix} + \frac{1}{N} \times \begin{bmatrix}\left(1-\theta^{\mathcal{B}}_{k}\right)p^{\mathcal{B}}_{\theta}(0\to 1) - \theta^{\mathcal{B}}_{k}p^{\mathcal{B}}_{\theta}(1\to 0)\\ \left(1-\theta^{\mathcal{R}}_{k}\right)p^{\mathcal{R}}_{\theta}(0\to 1) - \theta^{\mathcal{R}}_{k}p^{\mathcal{R}}_{\theta}(1\to 0)\end{bmatrix}, \tag{5}$$

where $p^{\mathcal{B}}_{\theta}(0 \to 1), p^{\mathcal{B}}_{\theta}(1 \to 0), p^{\mathcal{R}}_{\theta}(0 \to 1), p^{\mathcal{R}}_{\theta}(1 \to 0)$ were defined in Eq. 4. Therefore, the Markovian dynamics of the proposed model can be expressed as,

$$\theta_{k+1} = \theta_k + \frac{1}{N}(g(\theta_k) + M_{k+1}) \tag{6}$$

where

$$g(\theta_k) = \begin{bmatrix}g^{\mathcal{B}}(\theta_k)\\ g^{\mathcal{R}}(\theta_k)\end{bmatrix} = \begin{bmatrix}\left(1-\theta^{\mathcal{B}}_{k}\right)p^{\mathcal{B}}_{\theta}(0\to 1) - \theta^{\mathcal{B}}_{k}p^{\mathcal{B}}_{\theta}(1\to 0)\\ \left(1-\theta^{\mathcal{R}}_{k}\right)p^{\mathcal{R}}_{\theta}(0\to 1) - \theta^{\mathcal{R}}_{k}p^{\mathcal{R}}_{\theta}(1\to 0)\end{bmatrix},$$

and M_k is a martingale difference noise sequence. Eq. 6 can be viewed as a stochastic approximation with constant step size $1/N$. Thus, for large N, the discrete time trajectory $\theta_k, k = 0, 1, 2, \ldots$ evolves without jumps and it converges to the trajectory of the limit mean differential in Eq. 4. For such constant step-size stochastic approximation algorithms, typical proof approach is to invoke a form of law of large-numbers and establish that the interpolated trajectory of Eq. 6 converges weakly (in distribution) to a differential equation of the form $\dot{\theta}(t) = g(\theta(t))$ as the step size $1/N$ tends to 0. However, since $g(\cdot)$ is a discontinuous function, this typical approach that establishes (weak) convergence to an ordinary differential equation does not work. We establish the weak convergence of the interpolated stochastic trajectory of the model to a (deterministic) differential inclusion of the form $\dot{x}(t) \in h(x(t))$ where $h(\cdot)$ is a set-valued map constructed using the discontinuous $g(\cdot)$. Any trajectory of the form $x(t) = x(0) + \int_0^t y(s)ds$ satisfying $y(t) \in h(x(t))$ for all t is called a solution to the differential inclusion $\dot{x}(t) \in h(x(t))$. Such solutions are called Filippov solutions to the discontinuous dynamical system $\dot{\theta}(t) = g(\theta(t))$ (or Caratheodory solution of the differential inclusion $\dot{x}(t) \in h(x(t))$) [2] We then show that due to the piece-wise continuous form of $g(\theta_k)$, the solution is unique in all cases Except Case 4 of Theorem 1.

Required results from literature: The proof relies on two results from literature related to discontinuous dynamical systems that we state below. Let the distance between a continuous trajectory $z(\cdot)$ and the solution set $\mathcal{S}_T$ of a differential inclusion $\dot{x}(t) \in h(x(t))$ be defined as,

$$l(z(\cdot), \mathcal{S}_T) \overset{\text{def}}{=} \inf_{y(\cdot) \in \mathcal{S}_T} \sup_{t \in [0,T]} ||z(t) - y(t)||. \tag{7}$$

The following result from (*19*) is used to establish the weak-convergence of the sample paths.

[2] See (*18*) for a detailed introduction to discontinuous dynamical systems and their solution concepts.

Lemma 3 ((*19*)[Adapted from Theorem 9.4). *] Consider the stochastic approximation,*

$$x_{k+1} = x_k + a\left(g(x_k) + M_{k+1}\right), k \geq 0 \tag{8}$$

where $g(\cdot)$ is measurable and satisfies $||g(x)|| \leq C(1 + ||x||)$ for some $C > 0$. Let

$$h(x) = \bigcap_{\epsilon>0} \bar{co}\left(g(y) : ||y - x|| < \epsilon\right). \tag{9}$$

where $\bar{co}$ denotes the convex closure. Then,

$$l(x^a(\cdot)|_{[t',t'+T]}, \mathcal{S}_T) \xrightarrow{a\downarrow 0} 0 \tag{10}$$

uniformly in t' where $x^a(t)$ is the interpolated trajectory of the stochastic approximation algorithm and $\mathcal{S}_T$ is the solution set of the differential inclusion

$$\dot{x}(t) \in h(x(t)). \tag{11}$$

We will also use (*18*)[Proposition 5] to establish the uniqueness of the solutions to the differential inclusion. At a high-level, (*18*)[Proposition 5] states that the Filippov solution of a piece-wise continuous differential equation with a discontinuous right-hand side (i.e., the solutions to the differential inclusion constructed using that differential equation as in 9) is unique if the trajectories that approach the boundary of a continuous region either slides along the boundary or cross into the next region.

A.2 Proof of Convergence

Consider the model proposed in Sec. 2. Note that the value of $\theta^{\mathcal{B}}_{k+1} - \theta^{\mathcal{B}}_{k}$ can take three different values under three events:

Event 1: $\theta^{\mathcal{B}}_{k+1} - \theta^{\mathcal{B}}_{k} = \frac{1}{N^{\mathcal{B}}}$ in the event that X_{k+1} is a blue node that takes action-0 at time k and switches to action-1 at time $k+1$

Event 2: $\theta^{\mathcal{B}}_{k+1} - \theta^{\mathcal{B}}_{k} = -\frac{1}{N^{\mathcal{B}}}$ in the event that X_{k+1} is a blue node that takes action-1 at time k

and switches to action-0 at time $k+1$

Event 3: $\theta^{\mathcal{B}}_{k+1} - \theta^{\mathcal{B}}_{k} = 0$ in any event other than Event 1 and Event 3.

Let $\mathbb{P}_k\{\cdot\}, \mathbb{E}_k\{\cdot\}$ denote the probability measure and expected value conditional on all events that have occurred till time k. Consider the Event 1 first. Note that the probability that X_k is a blue node with choice-0 at time k is $\mathbb{P}_k\left\{R(X_{k+1}) = 0 \wedge H_k(X_{k+1}) = 0\right\} = \frac{N^{\mathcal{B}}(1-\theta^{\mathcal{B}}_k)}{N}$. For a fully connected graph, note that the probability that a random blue node with choice-0 at time k switches to the action choice-1 at time $k+1$ can be written as:

$$\mathbb{P}_k\left\{H_{k+1}(X_{k+1}) = 1 | R(X_{k+1}) = 0 \wedge H_k(X_{k+1}) = 0\right\} \tag{12}$$

$$= \mathbb{P}_k\left\{\alpha\left(d^{in,0}_k(X_k) - d^{in,1}_k(X_k)\right) - \beta\left(d^{out,0}_k(X_k) - d^{out,1}_k(X_k)\right) > 0 | R(X_k) = 0 \wedge H_k(X_k) = 0\right\} \tag{13}$$

$$= \mathbb{1}\left(\alpha(1-r)\left(2\theta^{\mathcal{B}}_k - 1\right) - \beta r\left(2\theta^{\mathcal{R}}_k - 1\right) > 0\right) \tag{14}$$

$$= p^{\mathcal{B}}_{\theta}(0 \to 1) \tag{15}$$

Similarly, we also obtain,

$$p^{\mathcal{B}}_{\theta}(1 \to 0) = \mathbb{1}\left(\alpha(1-r)\left(2\theta^{\mathcal{B}}_k - 1\right) - \beta r\left(2\theta^{\mathcal{R}}_k - 1\right) < 0\right). \tag{16}$$

Therefore, conditional on all events that have occurred till time k, the expected value of $\theta^{\mathcal{B}}_{k+1}$ can be written as:

$$\mathbb{E}_k\left\{\theta^{\mathcal{B}}_{k+1}\right\} = \theta^{\mathcal{B}}_k + \frac{1}{N^{\mathcal{B}}} \times \frac{N^{\mathcal{B}}\left(1-\theta^{\mathcal{B}}_k\right)}{N} \times p^{\mathcal{B}}_{\theta}(0 \to 1) - \frac{1}{N^{\mathcal{B}}} \times \frac{N^{\mathcal{B}}\theta^{\mathcal{B}}_k}{N} \times p^{\mathcal{B}}_{\theta}(1 \to 0) \tag{17}$$

Following similar arguments for the red-group yields similar expressions for $\mathbb{E}_k\left\{\theta^{\mathcal{R}}_{k+1}\right\}$, which yields

$$\begin{bmatrix} \mathbb{E}_k\left\{\theta^{\mathcal{B}}_{k+1}\right\} \\ \mathbb{E}_k\left\{\theta^{\mathcal{R}}_{k+1}\right\} \end{bmatrix} = \begin{bmatrix} \theta^{\mathcal{B}}_k \\ \theta^{\mathcal{R}}_k \end{bmatrix} + \frac{1}{N} \times \begin{bmatrix} \left(1-\theta^{\mathcal{B}}_k\right) p^{\mathcal{B}}_{\theta}(0 \to 1) - \theta^{\mathcal{B}}_k p^{\mathcal{B}}_{\theta}(1 \to 0) \\ \left(1-\theta^{\mathcal{R}}_k\right) p^{\mathcal{R}}_{\theta}(0 \to 1) - \theta^{\mathcal{R}}_k p^{\mathcal{R}}_{\theta}(1 \to 0) \end{bmatrix}, \tag{18}$$

where,

$$p^{\mathcal{B}}_{\theta}(0 \to 1) = \mathbb{1}\left(\alpha(1-r)\left(2\theta^{\mathcal{B}} - 1\right) - \beta r\left(2\theta^{\mathcal{R}} - 1\right) > 0\right), \quad p^{\mathcal{B}}_{\theta}(1 \to 0) = 1 - p^{\mathcal{B}}_{\theta}(0 \to 1)$$

$$p^{\mathcal{R}}_{\theta}(0 \to 1) = \mathbb{1}\left(\alpha r\left(2\theta^{\mathcal{R}} - 1\right) - \beta(1-r)\left(2\theta^{\mathcal{B}} - 1\right) > 0\right), \quad p^{\mathcal{R}}_{\theta}(1 \to 0) = 1 - p^{\mathcal{R}}_{\theta}(0 \to 1).$$

Thus, the evolution of the state can be expressed as Eq. (6). Note $g(\cdot)$ in Eq. (6) satisfies the linear growth condition since it is a piece-wise linear function taking values in $[-1, 1]^2$. The set of discontinuities are defined by the states $\theta(t) = [\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]'$ that satisfy

$$\alpha(1-r)\left(2\theta^{\mathcal{B}} - 1\right) - \beta r\left(2\theta^{\mathcal{R}} - 1\right) = 0 \tag{19}$$

or

$$\alpha r\left(2\theta^{\mathcal{R}} - 1\right) - \beta(1-r)\left(2\theta^{\mathcal{B}} - 1\right) = 0. \tag{20}$$

Thus, Lemma 3 implies that the stochastic trajectory of the proposed model converges to the solution set of the differential inclusion

$$\dot{\theta}(t) \in h(\theta(t)) = [h^{\mathcal{B}}(\theta(t)), h^{\mathcal{R}}(\theta(t))]' \tag{21}$$

where

$$h^{\mathcal{B}}(\theta(t)) = \begin{cases} [-\theta^{\mathcal{B}}(t), 1-\theta^{\mathcal{B}}(t)] & \text{if } \alpha(1-r)\left(2\theta^{\mathcal{B}} - 1\right) - \beta r\left(2\theta^{\mathcal{R}} - 1\right) = 0 \\ \{g^{\mathcal{B}}(\theta)\} & \text{otherwise,} \end{cases} \tag{22}$$

$$h^{\mathcal{R}}(\theta(t)) = \begin{cases} [-\theta^{\mathcal{R}}(t), 1-\theta^{\mathcal{R}}(t)] & \text{if } \alpha r\left(2\theta^{\mathcal{R}} - 1\right) - \beta(1-r)\left(2\theta^{\mathcal{B}} - 1\right) = 0 \\ \{g^{\mathcal{R}}(\theta)\} & \text{otherwise,} \end{cases} \tag{23}$$

which is the Filippov solution set of the discontinuous differential equation Eq. 4.

A.3 Proof of Uniqueness

To establish the uniqueness of the Filippov solution, we note that any solution to the Eq. 21 which approaches a point of discontinuity except $(0.5, 0.5)$ crosses the boundary and move to the next region. The only setting in which a trajectory approaches $(0.5, 0.5)$ is the Case 4 of Theorem 1 with $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$. Thus, according to (*18*)[Proposition 5], all trajectories except Case 4 with $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$ are unique.

Uniqueness of trajectories can also be seen from state space plots for the four cases in Fig. 1 as well. Note that the only form of trajectory that approaches the boundary but does not cross in to the other region is the trajectory starting with $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$ in Case 4. All other initial states therefore have unique trajectories.

B Dynamics of the Model on a Network with Communities

When the graph $G = (V, E)$ is a stochastic block model with in-group link probability ρ and out-group link probability of $1 - \rho$, the piece-wise interpolation of the discrete-time trajectory $\theta_k = [\theta_k^{\mathcal{B}}, \theta_k^{\mathcal{R}}]', k = 0, 1, 2, \dots$ can be approximated using the continuous-time trajectory $\theta(t) = [\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]', t \geq 0$ of the following differential equation as the number of nodes in the graph N is large:

$$\begin{bmatrix} \dot{\theta}^{\mathcal{B}} \\ \dot{\theta}^{\mathcal{R}} \end{bmatrix} = \begin{bmatrix} \left(1 - \theta^{\mathcal{B}}\right) p_\theta^{\mathcal{B}}(0 \to 1) - \theta^{\mathcal{B}} p_\theta^{\mathcal{B}}(1 \to 0) \\ \left(1 - \theta^{\mathcal{R}}\right) p_\theta^{\mathcal{R}}(0 \to 1) - \theta^{\mathcal{R}} p_\theta^{\mathcal{R}}(1 \to 0) \end{bmatrix}, \tag{24}$$

where,

$$\begin{aligned} p_\theta^{\mathcal{B}}(0 \to 1) &= \mathbb{1}\left(\alpha\rho(1-r)\left(2\theta^{\mathcal{B}} - 1\right) - \beta(1-\rho)r\left(2\theta^{\mathcal{R}} - 1\right) > \delta\right) \\ p_\theta^{\mathcal{B}}(1 \to 0) &= \mathbb{1}\left(\alpha\rho(1-r)\left(2\theta^{\mathcal{B}} - 1\right) - \beta(1-\rho)r\left(2\theta^{\mathcal{R}} - 1\right) < -\delta\right) \\ p_\theta^{\mathcal{R}}(0 \to 1) &= \mathbb{1}\left(\alpha\rho r\left(2\theta^{\mathcal{R}} - 1\right) - \beta(1-\rho)(1-r)\left(2\theta^{\mathcal{B}} - 1\right) > \delta\right), \\ p_\theta^{\mathcal{R}}(1 \to 0) &= \mathbb{1}\left(\alpha\rho r\left(2\theta^{\mathcal{R}} - 1\right) - \beta(1-\rho)(1-r)\left(2\theta^{\mathcal{B}} - 1\right) < -\delta\right) \end{aligned}$$

Consequently, the analogous version of the Fig. 1 for stochastic block models is shown in Fig. S1.

C Additional Results

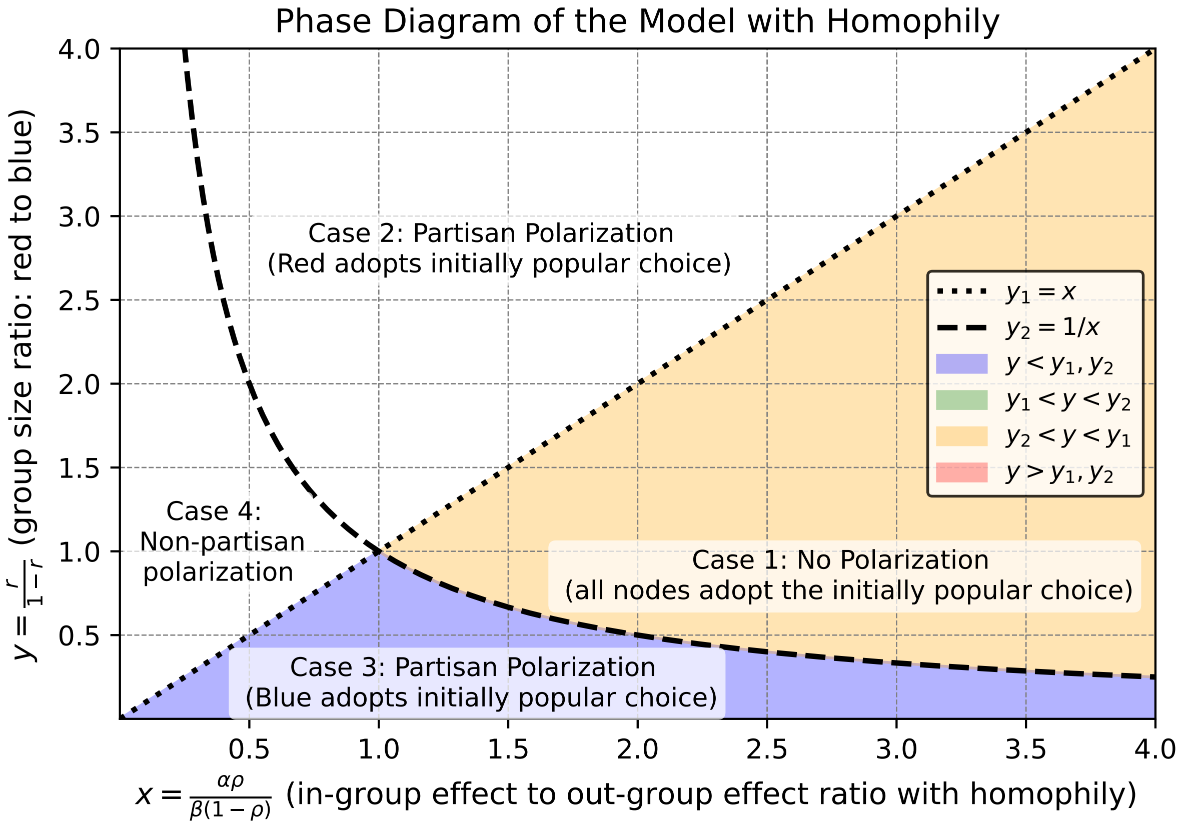
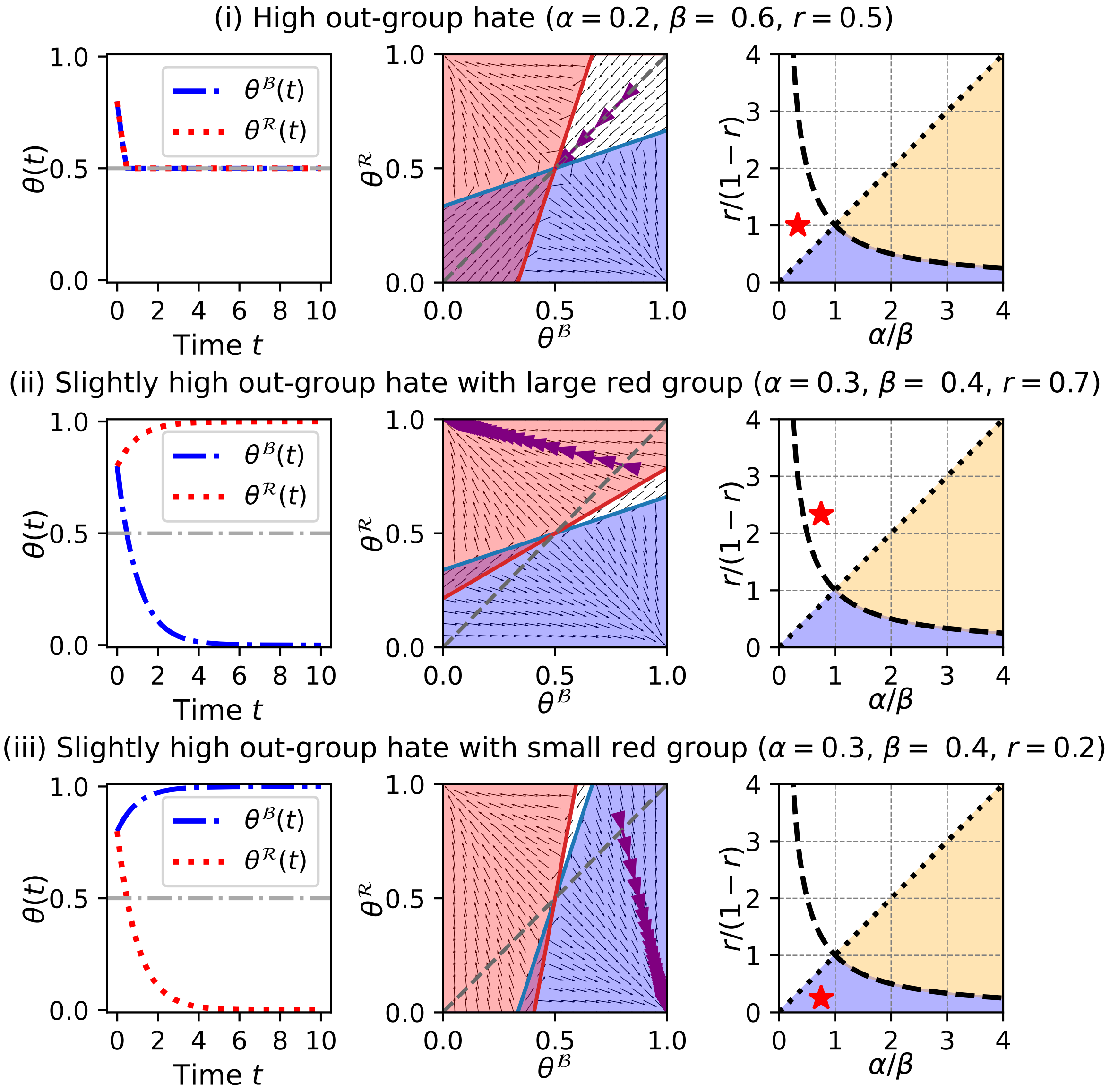


Figure S1: The four different regions of the model parameters (in-group conformity α, out-group dissent β, homophily ρ and fraction of red-nodes r) that lead to different asymptotic behaviors in a stochastic block model type graph starting from an initial state where the distribution of choices is the same for both parties i.e., i.e., $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$). This figure is similar to the analogous figure for the fully connected graph (Fig. 1) except that the in-group effect is amplified by ρ (probability observing each in-group member) and the out-group effect is amplified by $1 - \rho$ (probability of observing each out-group member).

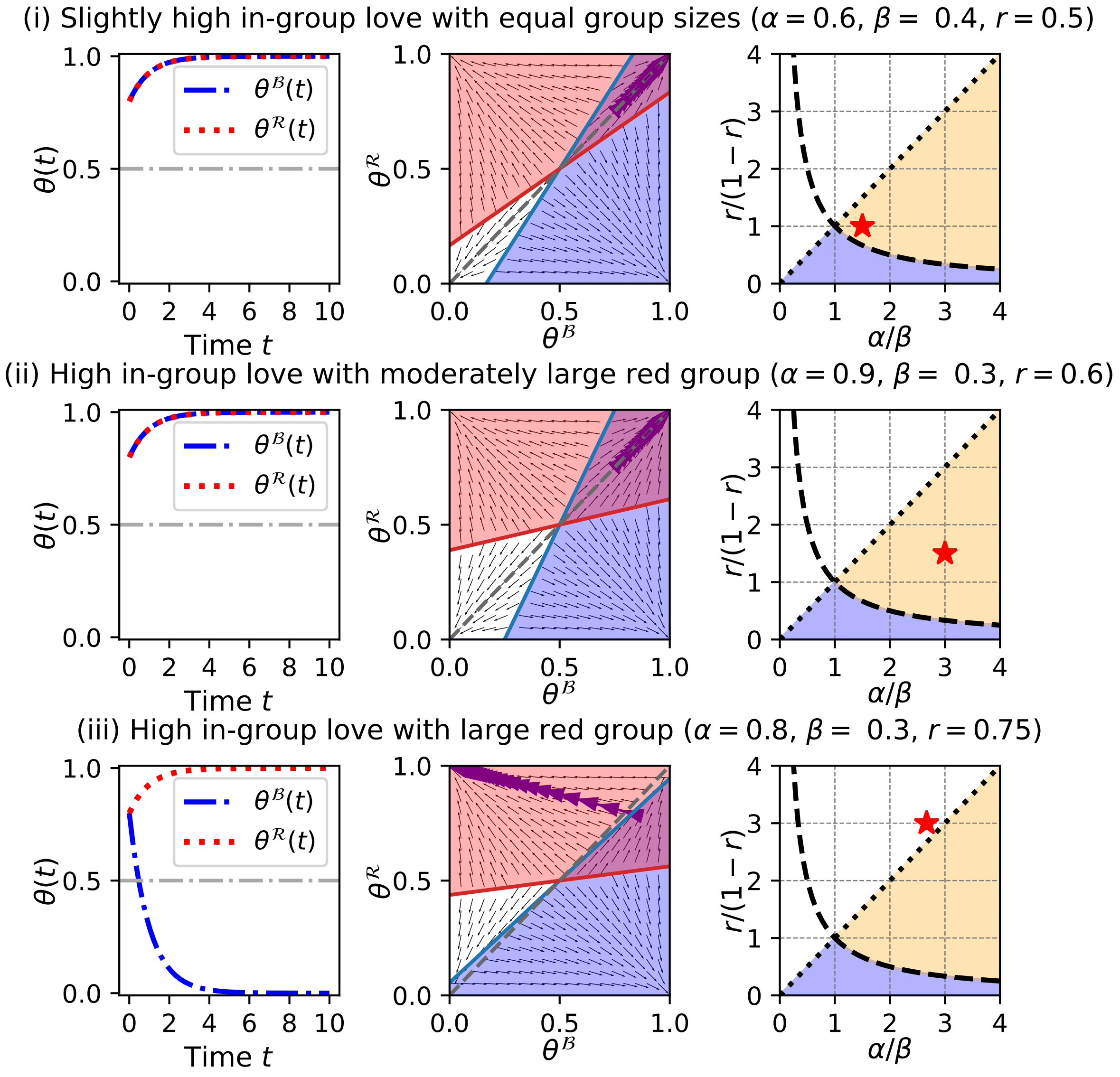


(i) High out-group hate ($\alpha = 0.2$, $\beta =$ 0.6, $r = 0.5$)

(ii) Slightly high out-group hate with large red group ($\alpha = 0.3$, $\beta =$ 0.4, $r = 0.7$)

(iii) Slightly high out-group hate with small red group ($\alpha = 0.3$, $\beta =$ 0.4, $r = 0.2$)

Figure S2: Example trajectories of the state when the out-group hate β is larger than in-group love α. The trajectories $[\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]$ over time (left column) and in the state space (middle column) show the evolution of $\theta(t) = [\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]$. The blue and red colors in middle column indicate regions where $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$ increase (i.e., regions where $p_{\theta}^{\mathcal{B}}(0 \to 1) = 1$ and $p_{\theta}^{\mathcal{R}}(0 \to 1) = 1$ according to Eq. 4). The black arrows in state space plots (middle column) indicate the path of the differential equation Eq. 4. The yellow arrows corresponds to the time domain trajectory (in left column). The figure shows how either uniform (row i) or party-line polarization (row ii and row-iii) can emerge when people are driven largely by their opposition to the out-group than their adherence to the in-group. Further, uniform polarization that emerges in the presence of very high out-group hate is unstable since some black arrows point away from $[0.5, 0.5]$ as seen from the state space plots (middle column) of row-i. In this case, small deviations from non-partisan polarization can lead to partisan polarization .



(i) Slightly high in-group love with equal group sizes ($\alpha = 0.6$, $\beta =$ 0.4, $r = 0.5$)

(ii) High in-group love with moderately large red group ($\alpha = 0.9$, $\beta =$ 0.3, $r = 0.6$)

(iii) High in-group love with large red group ($\alpha = 0.8$, $\beta =$ 0.3, $r = 0.75$)

Figure S3: Example trajectories of the state when out-group hate β is less than in-group love α. The trajectories over time (left column) and in the state space (middle column) show the evolution of $\theta(t) = [\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)]$. The blue and red regions in middle column indicate regions where $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$ increase (i.e., areas where $p_{\theta}^{\mathcal{B}}(0 \rightarrow 1) = 1$ and $p_{\theta}^{\mathcal{R}}(0 \rightarrow 1) = 1$ according to Eq. 4). The black arrows in the middle column indicate the path of the differential equation Eq. 4. The yellow arrows correspond to the time domain trajectory (left column). The figure shows how larger in-group love is necessary but not sufficient for the emergence of consensus. In particular, when the disparity between the sizes of the two groups is not too large compared to the disparity between α and β, consensus emerges.

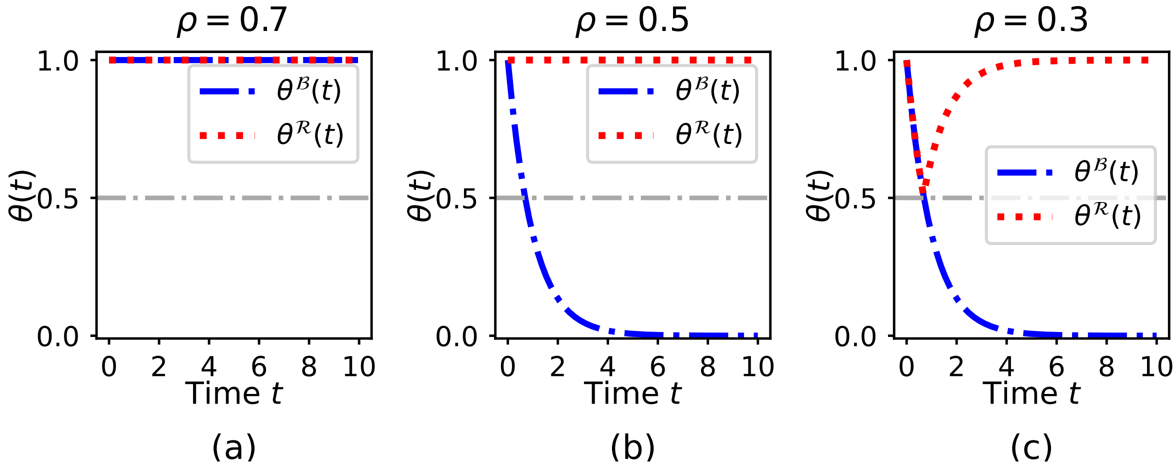


Figure S4: An illustration of how decreasing homophily can cause a party-line polarization from an initial state of global consensus. Figures correspond to $\alpha = 0.8, \beta = 0.7$ (larger in-group favoritism compared to out-group animosity) and $r = 0.65$ (a majority red group). Decreasing ρ from 0.7 (homophily) to 0.5 pushes the social network from Case-1 (consensus) to Case-3 (party-line polarization) in Fig. S1. Further decreasing ρ to 0.3 pushes the network to Case-4 which corresponds to an unstable state, where a small deviation leads to party-line polarization.

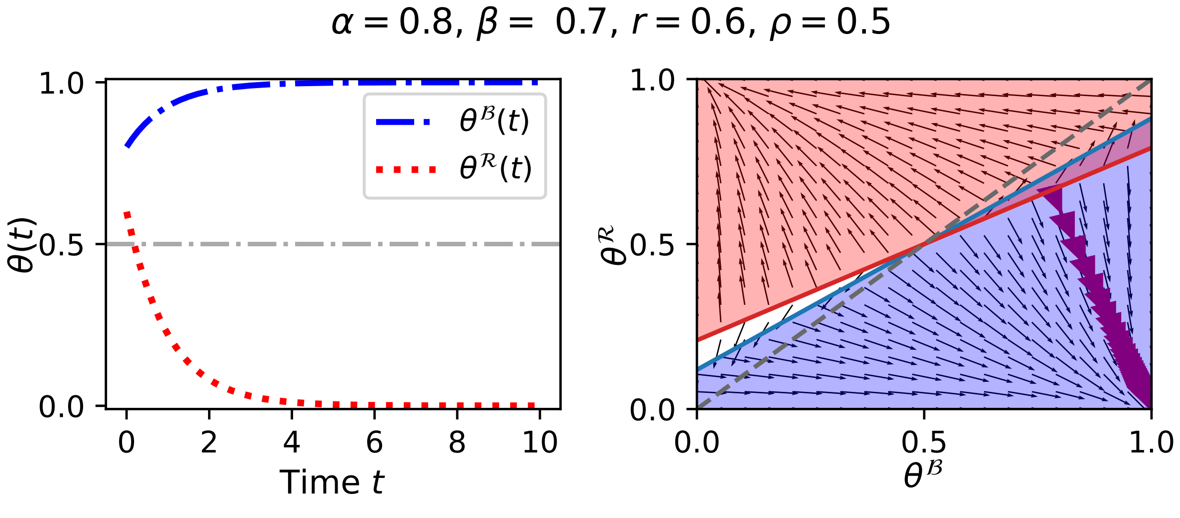


Figure S5: An illustration of a case where the two groups start with different popularity levels of the choices within them i.e., $\theta^{\mathcal{B}}(0) \neq \theta^{\mathcal{R}}(0)$, and the majority group adopts the choice that was initially less popular within it. The choice-1 is initially more popular within both groups with $\theta^{\mathcal{B}}(0) = 0.8, \theta^{\mathcal{R}}(0) = 0.6$. However, the majority red-group eventually adopts the choice that was initially less popular (i.e., choice-0 which had a 40% popularity) within it.