ROCKWOOL Foundation Berlin
Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES 001/26

Affective Polarization, Media Outlets, and Opinion Dynamics

Sebastiano Della Lena, Luca Paolo Merlino, Yves Zenou

www.rfberlin.com JANUARY 2026

Affective Polarization, Media Outlets, and Opinion Dynamics

Authors

Sebastiano Della Lena, Luca Paolo Merlino, Yves Zenou

Reference

JEL Codes: C7, D7, D85

Keywords: Signed Networks, Opinion Dynamics, Affective Polarization, Group Antagonism, Information Campaigns, Targeting.

Recommended Citation: Sebastiano Della Lena, Luca Paolo Merlino, Yves Zenou (2026): Affective Polarization, Media Outlets, and Opinion Dynamics. RFBerlin Discussion Paper No. 001/26

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com

Affective Polarization, Media Outlets, and Opinion Dynamics*

Sebastiano Della Lena† Luca Paolo Merlino‡ Yves Zenou§

December 26, 2025

Abstract

We study opinion dynamics in a social network consisting of two groups. Agents update their opinions by conforming to members of their own group while rejecting the views of the opposing group (affective polarization), and by listening to a media outlet that may provide biased information. We characterize the long-run opinions and identify when affective polarization and media bias lead to ideological polarization, persistent disagreement, or failures of learning. We also derive when information interventions or censorship improve learning and reduce disagreement, and when they backfire: better information helps only under specific media bias configurations and when directed to the agents we identify as most effective at propagating it through the network.

JEL Classification Numbers: C7, D7, D85.
Keywords: Signed Networks, Opinion Dynamics, Affective Polarization, Group Antagonism, Information Campaigns, Targeting.

*We thank Bård Harstad, three anonymous referees, as well as the participants in AETW in Melbourne, Economic theory Festival in Brisbane, Network Science Conference 2023 in Virginia Tech, the Network Workshop in Marseilles in 2023, BiNoMa 2024, ASSET 2024, and in seminars at Universitat Autónoma de Barcelona, University of São Paulo, University of Siena, University Technology of Sydney (UTS), in particular Mikhail Anufriev, Francis Bloch, Juan Carlos Carbajal, Chen Cheng, Tony Cookson, Alberto Dalmazzo, Ben Golub, Matt O. Jackson, Weijia Li, Suraj Malladi, Fabrizio Panebianco, Lin Peng, Antonio Penta, Paolo Pin, Luis Pontes de Vasconcelos, Antonio Rosato, Evan Sadler, Benjamin Young, and Yiqing Xing, Ekaterina Zhuravskaya, for their very helpful comments. We gratefully acknowledge the financial support provided by grants 1258321N and G029621N from the Research Foundation Flanders (FWO), the Belgian National Bank, and grant DP240100158 from the Australian Research Council.

†Department of Economics, Monash University, Australia. E-mail: sebastiano.dellalena@monash.edu.
‡ECARES, Université libre de Bruxelles, Belgium. Email: luca.paolo.merlino@ulb.be
§Department of Economics, Monash University, Australia and the University of Southampton, UK. E-mail: yves.zenou@monash.edu.

1 Introduction

In 2013, NASA released data showing a clear decline in Arctic sea ice, underscoring the urgency of climate action. Guilbeault et al. (2018) conducted a laboratory experiment to study how liberals and conservatives interpreted this objective data. Liberals predicted a further decline in sea ice, while conservatives saw it as evidence of an increase due to a rebound in the most recent observations, illustrating polarization. The researchers then facilitated discussions between both groups in social networks resembling a social media platform and randomized participants in treatments with different information on the political affiliation of their neighbors. When displaying political logos (Republican or Democrat), participants remained entrenched in their original views and polarization persisted. However, when political logos were not shown, cross-party interactions eliminated belief polarization. By the end, both liberals and conservatives reached 90% accuracy in their forecasts, with political differences largely disappearing.[1] This shows how *affective polarization*, i.e., the emotional attachment to the own group and hostility toward the opposing one,[2] shapes opinions and biases views on critical issues such as climate change.

The aim of this paper is to deepen the understanding of these issues by developing an exogenous-network model of opinion dynamics with two distinct groups (e.g., Democrats and Republicans). The model examines how affective polarization—in which individuals internalize the opinions of their group positively and those of the other group negatively—and media outlets, represented as biased or unbiased sources of information, impact the long-run opinions.[3] We characterize the long-run opinions and establish conditions under which affective polarization and media influence contribute to ideological polarization or failure of learning, thereby guiding the formulation of effective policies.

Specifically, we investigate the opinion dynamics of individuals embedded in a signed network, where society is divided into two groups. We model affective polarization by assuming that individuals, at each time period, align their opinions with those of their own group members (i.e., they positively average the opinions of their in-group peers); conversely, they oppose the views of the other group (i.e., they negatively average the opinions of their out-group peers). Together, these dynamics constitute what we refer to

[1] Djourelova et al. (2024) also provide evidence that exposure to the same disaster event increases climate change and environmental concerns among liberals but decreases them among conservatives, widening the ideological gap by 11–17%.

[2] Affective polarization refers to the "us vs. them" mentality among people with different political beliefs, values, or attitudes, where opponents are not just seen as having different views but as morally wrong or even evil (Iyengar et al., 2012). See Boxell et al. (2024) for cross-country evidence.

[3] In our model, media outlets include both traditional mass media (e.g., CNN, Fox News) and social media platforms (e.g., Twitter, Facebook), which may provide biased or unbiased information. We assume affective polarization exists, the interaction network is fixed, and exposure to media is exogenous. The idea is that the opinions we focus on are not the reason the network exists so it can be taken as given.

as *in-group identity* and *out-group antagonism*. In addition to peer influences, individuals may also rely on a potentially biased source of information—such as partisan or neutral media outlets.

We examine three distinct scenarios in which individuals form opinions about an unknown state of the world. In the first scenario (case (i)), agents have *no* access to external information and base their opinions solely on the aggregation of the opinions of others. In the second scenario (case (ii)), individuals consider the opinions of other agents and have access to an *unbiased* source of information (e.g., an impartial media outlet). In the third scenario (case (iii)), individuals again consider the opinions of other agents but have access to a *group-specific biased* source of information (e.g., a partisan media outlet).

In case (i), when agents form an opinion without having access to any source of information, with affective polarization, we always observe *in-group consensus* and *out-group polarization*. The long-run opinion of each agent is determined solely by the weighted sum of the initial opinions of their direct and indirect connections. The weights, and thus the influences of each agent's initial opinion on others' long-run opinions, are given by the eigenvector centralities of the *identity-interaction network*, which is a signed, structurally balanced network.[4] Indeed, since each agent wants to be as close as possible to the opinions of agents of the same group but as far as possible from those of agents of the other group, the more central an agent j is in the network, the more agent i conforms to (deviate from) j's opinion in the long-run if she belongs to the same (other) group. Compared to standard models of opinion dynamics (e.g., DeGroot, 1974; Golub and Jackson, 2010), introducing affective polarization leads to polarization of opinion between the two groups, thereby preventing a consensus among all agents.

In case (ii), when agents have access to an unbiased source of information, long-run opinions are independent of initial opinions. Prior studies (Jadbabaie et al., 2012; Molavi et al., 2018) show that, in the absence of affective polarization, mild assumptions guarantee that agents reach a consensus and aggregate information effectively. By contrast, we show that affective polarization—whereby agents negatively incorporate out-group information—leads to a failure of learning the truth, and can also foster ideological polarization depending on the network structure. Long-run opinions converge to a vector proportional to weighted Katz–Bonacich centrality in the identity-interaction network, which contains both positive and negative links. Opinion leaders in this signed network—the individuals who better aggregate information—are generally not the same as in standard models:

[4]From the social-interaction network, we derive the identity-interaction matrix, which captures how each agent's opinion influences others depending on group identity. Its eigenvector centralities correspond closely to the status measure for signed networks of Bonacich and Lloyd (2004), and our framework provides a microfoundation for this measure in settings with negative relations. A network is structurally balanced when all in-group links are positive and all out-group links are negative (Harary, 1953), as observed, e.g., in online social networks (Guha et al., 2004).

agents with many inter-group interactions face greater exposure to negative sentiment and may therefore be the least accurate in aggregating information, despite being central in the social network. These predictions align with experimental evidence: users endorsing conspiracy theories tend to increase their engagement with conspiratorial content when exposed to debunking posts (Zollo et al., 2017), while antisemitism instigated anti-market culture among non-Jews in proximity to Jewish communities (Grosfeld et al., 2013).

In case (*iii*), when agents have a biased source of information, long-run opinions depend on the biases and how they propagate in the network. Thus, the long-run opinion of each agent is determined by a linear combination of the true state of the world and the biases. In standard models, biases of opposite sign tend to cancel each other out, while biases of the same sign tend to persist. By contrast, with affective polarization opposite-sign biases tend to exacerbate each other, amplifying polarization; instead, same-sign biases tend to cancel each other out, dampening the negative effect of affective polarization on opinions. Surprisingly, biases may even lead to long-run opinions closer to the truth than in case (*ii*), and reducing affective polarization does not necessarily improve information aggregation, consistent with experimental evidence (Levy, 2021).

Then, we study policies that aim to improve learning and reduce disagreement. Under affective polarization, information campaigns can backfire due to group antagonism. Improving exposure to unbiased information (case (*ii*)) for only one group moves it closer to the truth while pushing the other away, so this intervention may increase disagreement unless both sides are reached. If the media outlets are sufficiently biased (case (*iii*)), providing better information to the whole society is optimal only if the two groups are exposed to biases in opposite directions; if instead the biases are aligned, only the more biased one should be targeted. We also identify the agent to optimally target with more accurate information, i.e., the one whose beliefs propagate maximally through the network. Finally, we show that censorship can backfire as well: restricting extreme opinions improves learning only when groups are on opposite sides of the truth, but may worsen outcomes when both groups are on the same side. Overall, these results provide guidance for designing policies that mitigate the effects of affective polarization.

This paper contributes to the literature on non-Bayesian opinion dynamics (DeGroot, 1974; Golub and Jackson, 2010, 2012; Jadbabaie et al., 2012; Molavi et al., 2018) by introducing negative weights to model affective polarization. Relative to standard models, our framework explains how exposure to the opposing group or to partisan media can increase polarization (Bail et al., 2018; Zollo et al., 2017) and that under affective polarization information policies designed to curb misinformation may backfire.

Recent work studies negative relationships primarily as anti-conformism. For instance, Buechel et al. (2015) model agents who average neighbors' opinions but may misrepresent

their own via conformity or anti-conformity, while Zhang et al. (2018) and Grabisch et al. (2019) classify agents as conformist or anti-conformist. In contrast, our framework lets agents' desire to conform depend on both their own identity and that of the agent with whom they interact. Shi et al. (2019) also study convergence in signed networks. Our contribution is to introduce biased and unbiased media and study how they interact with antagonistic links. This allows us to identify when affective polarization causes learning failure and ideological polarization, and to derive novel policy implications. Finally, our paper also contributes to the literature on the spread of misinformation in networks (Bloch et al., 2018; Merlino et al., 2023; Della Lena, 2024) and polarization (Callander and Carbajal, 2022; Campbell et al., 2025) by explicitly modeling out-group antagonism.

We also contribute to the literature on identity (Akerlof and Kranton, 2000; Shayo, 2020) and affective polarization (Iyengar et al., 2019). While identity has been widely studied in economics, its role in opinion dynamics has received little attention. Work on affective polarization typically focuses on partisan identity, i.e., the tendency of partisans to view opponents negatively and co-partisans positively (Iyengar and Westwood, 2015), but lacks a formal model explaining how affective polarization and social media jointly shape ideological polarization. Understanding how affective polarization shapes opinions is key to interpreting our results. Our predictions align with recent empirical evidence (e.g., Guilbeault et al., 2018; Jenke, 2024; Lerman et al., 2024) and help reconcile seemingly contradictory findings on whether exposure to opinion leaders or partisan media mitigates or exacerbates polarization (Bail et al., 2018; Levy, 2021). Most importantly, the framework provides a basis for designing policies to counteract the negative effects of affective polarization and partisan media on ideological polarization.

The paper proceeds as follows. Section 2 presents the model. Section 3 states the main results. Section 4 explores policy implications. Section 5 concludes. Appendix A provides the microfoundations of our model. All proofs are in Appendix B.

2 Model

We consider a society where agents update their opinions about the true state of the world $\theta^* \in \Theta \subset \mathbb{R}$ by aggregating information from two sources: their social contacts (whose opinions they may wish to conform to or reject) and, when available, a (potentially biased) external source of information. Without loss of generality, we assume $\theta^* \neq 0$.

Agents Agents are divided into two groups, $\mathcal{C} := \{A, B\}$, with sizes n^A and n^B respectively, such that $n^A + n^B = n$. We order the agents such that $A = \{1, ..., n^A\}$ and $B = \{n^A + 1, ..., n\}$, and denote the entire set of agents as $N = \{1, 2, ..., n\}$.

Social Interactions Agents interact in a *network of social interactions*, represented by an $n \times n$ non-negative matrix $\boldsymbol{W}$. Each entry $w_{ij}^C \in [0,1]$ indicates the extent to which agent $i \in C$, with $C \in \{A, B\}$ pays attention to the opinion of agent j, for each $i, j \in N$. The *social-interaction matrix* $\boldsymbol{W}$ can be expressed as follows:

$$\boldsymbol{W} := \begin{bmatrix} \boldsymbol{W}^{AA} & \boldsymbol{W}^{AB} \\ \boldsymbol{W}^{BA} & \boldsymbol{W}^{BB} \end{bmatrix},$$

where $\boldsymbol{W}^{AB} := ((w_{ij}^A)_{i \in A})_{j \in B} \in [0,1]^{n^A \times n^B}$ represents the interactions between an agent i from group A and an agent j from group B. The other sub-matrices of $\boldsymbol{W}$ are interpreted similarly, capturing the interactions within and between groups. We assume that the network is strongly connected[5] and that $w_{ii}^C > 0$ for all $i \in C$, with $C \in \{A, B\}$, thereby ensuring that each agent pays attention to their own opinion to some degree.[6]

Media Outlets Each agent $i \in C$, with $C \in \{A, B\}$, may have access to a potentially biased source of information θ_i^C (e.g., a media outlet) that could reveal the true state of the world. We denote by w_i^C the weight that agent i assigns to θ_i^C—i.e, the degree to which i pays attention to or is influenced by the information source. Thus, $w_i^C > 0$ can also be interpreted as the exogenous probability that i believes their source is unbiased. If agents lack access to such a source (e.g., because the topic is not verifiable), then $w_i^C = 0$.

We focus on the group bias in agents' private signals, assuming that $\xi_i^A \equiv \xi^A$ for all $i \in A$ and $\xi_i^B \equiv \xi^B$ for all $i \in B$, i.e., all agents in a group have the same bias—for example, because they have access to the same source of information. Thus, for each agent $i \in C$, $\theta_i^C \equiv \theta^C = \theta^* + \xi^C$. Let $\boldsymbol{\xi} = \left(\boldsymbol{\xi}^A, \boldsymbol{\xi}^B\right)^\top$ denote the bias vector. Thus, the overall column vector of information sources can be defined as $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^A, \boldsymbol{\theta}^B\right)^\top = \theta^* \mathbf{1} + \boldsymbol{\xi}$, where $\mathbf{1}$ is the n-vector of ones. Similarly, all vectors in the model follow this structure, starting with the first n^A agents, followed by the $n^B = n - n^A$ agents. For example, $\mathbf{w} = \left(\mathbf{w}^A, \mathbf{w}^B\right)^\top$ is the vector representing exposure to possibly biased information.

Opinion Updating Let $\mu_{i,t}^C$ denote the opinion help by agent $i \in C$, with $C \in \{A, B\}$, at time t. The corresponding vector of all agents' opinions is given by $\boldsymbol{\mu}_t = \left(\boldsymbol{\mu}_t^A, \boldsymbol{\mu}_t^B\right)^\top$. Agents update their opinions taking a weighted combination of opinions from their own

[5] Strong connectivity requires that every node can reach every other node through a (possibly directed) sequence of links, regardless of whether interactions are positive or negative. While this assumption is not satisfied in platforms such as Twitter/X, large-scale directed networks typically contain a giant strongly connected component (Gabielkov et al., 2014). Our results apply to such components, which capture the part of the network where sustained opinion exchange is possible.

[6] We later use these properties to show the convergence of the opinion dynamics to a steady state. These are standard assumptions in the literature on opinion dynamics (see, e.g., Golub and Jackson, 2010). As we focus on the implications of affective polarization, we rely on these well-established results.

group, the other group, and the (possibly biased) external source of information.

Following Tajfel and Turner (1979) and Akerlof and Kranton (2000), we assume agents strongly identify with their own group, shaping whose opinions they seek to match: agents align with their in-group while tending to disagree with the out-group. The opinion of agent $i \in C$ with $C \in \{A, B\}$ then evolves according to the following equation:

$$\mu_{i,t}^{C} = \alpha_i^{C} \sum_{j \in C} w_{ij}^{C} \mu_{j,t-1}^{C} + \beta_i^{C} \sum_{z \in C^c} w_{iz}^{C} \mu_{z,t-1}^{C^c} + w_i^{C} \theta^{C}, \tag{1}$$

where $C^c := \mathcal{C} \setminus C$ represents the complement of group C (e.g., $A^c = B$), $\alpha_i^C \geq 0$ denotes the intensity of the agent's in-group identity, and $\beta_i^C \leq 0$ represents the intensity of out-group antagonism. We assume that $\alpha_i^C \sum_{j \in C} w_{ij}^C + |\beta_i^C| \sum_{z \in C^c} w_{iz}^C + w_i^C = 1$.[7]

Defining $\mathbf{\Lambda}^C := diag[(\alpha_i^C)_{i \in C}]$ and, $\mathbf{\Gamma}^C := diag[(\beta_i^C)_{i \in C}]$, for all $C = \{A, B\}$, the opinions of agents of groups A and B evolve according to the following equations:

$$\boldsymbol{\mu}_t^A = \mathbf{\Lambda}^A \boldsymbol{W}^{AA} \boldsymbol{\mu}_{t-1}^A + \mathbf{\Gamma}^A \boldsymbol{W}^{AB} \boldsymbol{\mu}_{t-1}^B + \mathbf{w}^A \odot \boldsymbol{\theta}^A, \tag{2}$$

$$\boldsymbol{\mu}_t^B = \mathbf{\Lambda}^B \boldsymbol{W}^{BB} \boldsymbol{\mu}_{t-1}^B + \mathbf{\Gamma}^B \boldsymbol{W}^{BA} \boldsymbol{\mu}_{t-1}^A + \mathbf{w}^B \odot \boldsymbol{\theta}^B, \tag{3}$$

where $\odot$ is the element-wise (Hadamard) product.

Identity-Interaction Matrix Given this process of opinion dynamics, from the *social-interaction matrix* $\boldsymbol{W}$ we derive the *identity-interaction matrix* $\tilde{\boldsymbol{W}}$, which captures how an agent's opinion influences another's based on their group identities, i.e.,

$$\tilde{\boldsymbol{W}} := \begin{bmatrix} \mathbf{\Lambda}^A \boldsymbol{W}^{AA} & \mathbf{\Gamma}^A \boldsymbol{W}^{AB} \\ \mathbf{\Gamma}^B \boldsymbol{W}^{BA} & \mathbf{\Lambda}^B \boldsymbol{W}^{BB} \end{bmatrix} \qquad \text{with} \qquad sign(\tilde{\boldsymbol{W}}) = \begin{bmatrix} + & - \\ - & + \end{bmatrix}.$$

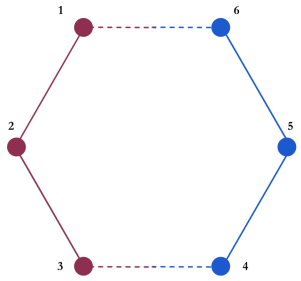
Each agent maintains positive (or zero) links with members of their own group and negative (or zero) links with members of the other group. By Theorem 3 in Harary (1953), the matrix $\tilde{\boldsymbol{W}}$ corresponds to a *structurally balanced* network—that is, a network in which the product of the signs of the edges of any possible cycle in $\tilde{\boldsymbol{W}}$ is positive.[8] While not needed for our main results, structural balance naturally emerges from negative intergroup sentiments implied by affective polarization.

As a running example, Figure 1 depicts a ring network of six agents, $A = \{1, 2, 3\}$

[7] This normalization naturally results if agents share a given amount of time/attention to the different sources of information, but they incorporate negatively the opinions of members from the opposing group. Negative links result, e.g., if agents using the Bayes rule to update their beliefs. See Appendix A.

[8] There is a *path* among distinct nodes in the ordered sequence $S := \{1, 2, \ldots, K-1, K\}$ in $\tilde{\boldsymbol{W}}$ if $\tilde{w}_{k,k+1}^{C_k} \neq 0$ for each $k \in \{1, \ldots, K-1\}$, where C_k is the group of agent k. The *sign of the path* is the sign of $\prod_{k=1}^{K-1} \tilde{w}_{k,k+1}^{C_k}$. A *cycle* is a path that begins and ends at the same node, and its *sign* is defined as the sign of the associated path.

and $B = \{4, 5, 6\}$, with the corresponding social- and identity-interaction matrices; the identity-interaction matrix is structurally balanced (see Appendix C for more details).



$$\boldsymbol{W} = \frac{1-w}{3}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \tilde{\boldsymbol{W}} = \frac{1-w}{3}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Figure 1: Ring network of six agents with groups $A = \{1, 2, 3\}$ (red) and $B = \{4, 5, 6\}$ (blue) with $\alpha_i^C = -\beta_i^C = 1$ and $w_i^C = w$ for all $i \in C$ with $C \in \{A, B\}$. Dotted edges indicate negative links. The social-interaction matrix $\boldsymbol{W}$ and the identity-interaction matrix $\tilde{\boldsymbol{W}}$ are shown alongside the network.

We can now express equations (2) and (3) more compactly as follows:

$$\boldsymbol{\mu}_t = \tilde{\boldsymbol{W}}\boldsymbol{\mu}_{t-1} + \mathbf{w} \odot \boldsymbol{\theta}. \tag{4}$$

The first term on the right-hand side of (4) captures the influence of interpersonal opinion exchanges in shaping beliefs, while the second term reflects the impact of the (un)biased source of information. Hence, for any agent $i \in C$ with $C \in \{A, B\}$, equation (1) becomes

$$\mu_{i,t}^C = \sum_{j \in C} \tilde{w}_{ij}^C \mu_{j,t-1}^C + \sum_{z \in C^c} \tilde{w}_{iz}^C \mu_{z,t-1}^{C^c} + w_i^C(\theta^* + \xi^C). \tag{5}$$

Long run opinions We are interested in the long-run opinions of agents belonging to the two groups. We denote the steady state value of the variables by suppressing the index t. Thus, $\mu_i^C := \lim_{t\to\infty} \mu_{i,t}^C$ is agent's i opinion in the long run and $\boldsymbol{\mu} := \left(\boldsymbol{\mu}^C\right)_{C \in \mathcal{C}}$ represents the vector of long-run opinions in society. Furthermore, let $\bar{\mu}^C := \frac{\sum_{i \in C} \mu_i^C}{n^C}$ represent the average long-run opinion within each group $C \in \{A, B\}$.

Centralities Let $\tilde{\boldsymbol{M}} := \left(\boldsymbol{I} - \tilde{\boldsymbol{W}}\right)^{-1} = \sum_{k=0}^{+\infty} \tilde{\boldsymbol{W}}^k$ denote the Leontief inverse of the identity-interaction matrix, which, under our assumptions, is well defined when agents have access to media, as we show below. The generic element $\tilde{m}_{ij}^C$ measures the cumulative influence of agent j on agent $i \in C$, with $C \in \{A, B\}$, through all walks in the network, accounting for both positive in-group links and negative out-group links. We define the vector of *weighted (signed) Katz–Bonacich centralities* in the identity-interaction network as $\tilde{\boldsymbol{b}} := \boldsymbol{b}(\tilde{\boldsymbol{W}}) = \tilde{\boldsymbol{M}}\mathbf{w}$, with generic element for agent $i \in C$ given by:

$$\tilde{b}_i^C = \sum_{j \in A} \tilde{m}_{ij}^C w_j^A + \sum_{k \in B} \tilde{m}_{ik}^C w_k^B. \tag{6}$$

To emphasize the role of group interactions in bias propagation, we also define the contributions from each group separately:

$$\tilde{b}_i^{CA} = \sum_{j \in A} \tilde{m}_{ij}^C w_j^A, \quad \text{and} \quad \tilde{b}_i^{CB} = \sum_{k \in B} \tilde{m}_{ik}^C w_k^B, \tag{7}$$

so that $\tilde{b}_i^C = \tilde{b}_i^{CA} + \tilde{b}_i^{CB}$ and the total centrality vector can be compactly expressed as

$$\tilde{\boldsymbol{b}} = \begin{bmatrix} \tilde{\boldsymbol{b}}^A \\ \tilde{\boldsymbol{b}}^B \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{b}}^{AA} + \tilde{\boldsymbol{b}}^{AB} \\ \tilde{\boldsymbol{b}}^{BA} + \tilde{\boldsymbol{b}}^{BB} \end{bmatrix}. \tag{8}$$

As we show below, if instead no agent has media access, i.e., $w_i^C = 0$ for all $i \in N$, the assumption $\alpha_i^C \sum_{j \in C} w_{ij}^C + |\beta_i^C| \sum_{z \in C^c} w_{iz}^C = 1$, together with structurally balance, implies that $\tilde{\boldsymbol{W}}$ has its largest eigenvalue equal to 1. Thus, we can define the left eigenvector corresponding to the eigenvalue 1 of matrix $\tilde{\boldsymbol{W}}$ as $\tilde{\boldsymbol{\pi}} := \boldsymbol{\pi}(\tilde{\boldsymbol{W}}) = (\tilde{\pi}_i^C)_{i \in N}$. Intuitively, $\tilde{\pi}_i^C$ captures the influence of agent i in the signed network. Its signs identify the two groups: positive entries correspond to the own group and negative entries to the other.

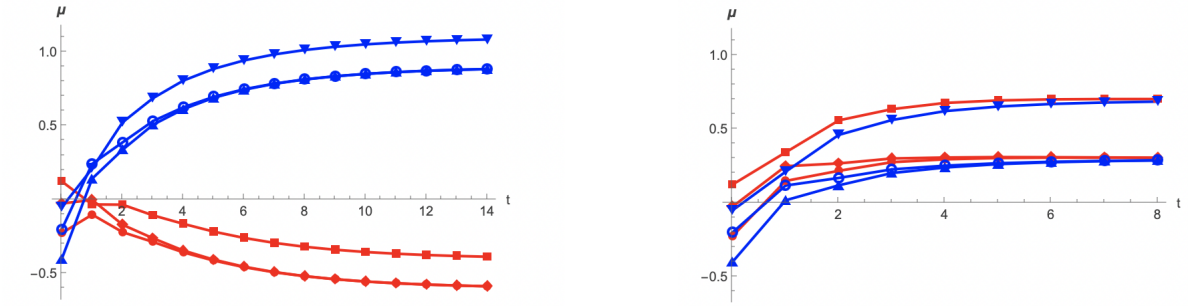
Two points are worth noting. First, eigenvector and Katz-Bonacich centralities capture different notions of network influence: eigenvector centrality emphasizes recursive importance through connections to other central nodes, while the weighted Katz-Bonacich incorporates the exposure to exogenous sources of information (the media outlets) and discounts more heavily influence that travels through longer walks. Second, while the measures on unsigned matrices are always positive (Jackson, 2008), those based on the *identity-interaction* matrix $\tilde{\boldsymbol{W}}$ can be positive or negative due to positive in-group and negative out-group links; thus, a node influential in $\boldsymbol{W}$ need not be influential in $\tilde{\boldsymbol{W}}$. Table C-1 in Appendix C illustrates this for the ring network.

Ideological polarization and disagreement We say that society exhibits *ideological polarization* when, in the long run, the opinions of the two groups systematically differ. Recall that μ_i^C denotes the long-run opinion of agent $i \in C$, with $C \in \{A, B\}$. Then:

Definition 1 *A society exhibits* ***ideological polarization*** *if there exists a value y such that $\text{sign}(\mu_i^A - y) \neq \text{sign}(\mu_j^B - y)$ for all $i \in A$ and $j \in B$. We measure the* ***degree*** *of ideological polarization by the distance between the two groups' average opinions, $|\bar{\mu}^A - \bar{\mu}^B|$.*

We say that a group or society exhibits disagreement when its members hold differing opinions. If all members share the same opinion, the variance equals zero, indicating *consensus*. Conversely, when opinions diverge, the variance becomes positive, signaling disagreement. Formally:

Definition 2 *A society (group) exhibits* ***disagreement*** *if long-run opinions are not identical in it, i.e., if* $Var[\boldsymbol{\mu}] > 0$ *(*$Var[\boldsymbol{\mu}^C] > 0$ *for group* $C \in \{A, B\}$*). We measure the* ***degree*** *of disagreement by the variance of long-run opinions, i.e.,* $Var[\boldsymbol{\mu}]$ *(*$Var[\boldsymbol{\mu}^C]$*).*



(a) Ideological polarization and disagreement ($\xi^A = -\xi^B$)

(b) Disagreement without ideological polarization ($\xi^A = \xi^B$)

Figure 2: Ideological polarization vs disagreement: Opinion dynamics for the blue and red groups in the ring network of Figure 1 with the same initial opinions but different media biases.

To illustrate these concepts, Figure 2 shows two examples of opinion dynamics in the ring network of Figure 1. In both cases, long-run disagreement arises within and across groups, but ideological polarization is present only in panel (a).

Microfoundations In Appendix A, we provide several microfoundations for the updating rule (5). In the first, we interpret $\boldsymbol{\mu}_t$ as *opinions*. Bayesian agents exhibit two cognitive biases: *persuasion bias*, whereby they fail to properly account for repetition in the information they receive (DeMarzo et al., 2003); and *zero-sum thinking*, the belief that gains for one group necessarily come at the expense of the other (e.g., Chinoy et al., 2025). Under zero-sum thinking, when an agent observes an out-group member expressing a positive view about a policy, they interpret it as evidence that the policy benefits the out-group at their own group's expense and adjust their opinion in the opposite direction. A positive view expressed by an in-group member is taken as evidence that the policy benefits their own group, raising their opinion. Under these biases, after receiving a private noisy signal about the policy's effect, each agent updates their opinion via peer interactions and public information; then (5) follows.

If we interpret $\boldsymbol{\mu}_t$ as *behaviors*, *actions*, *voiced opinions*, or *attitudes*, (5) corresponds to myopic best-response dynamics in network games with both positive and negative spillovers, arising from in-group conformity and out-group differentiation (e.g., political voiced opinions) or peer pressure (e.g., recycling).

3 Main results

In this section, we compare opinion dynamics with and without antagonism in three distinct informational environments:

(*i*) **No Information** ($w_i^C = 0$ for each $i \in C$, $C \in \{A, B\}$) Agents have no access to external information. This corresponds to debates on issues without an objective truth (e.g., ethical dilemmas such as abortion) or where data are unavailable (e.g., a vaccine before large-scale rollout).

(*ii*) **Unbiased Source** ($w_i^C > 0$ for $i \in C$, $C \in \{A, B\}$, $\xi^A = \xi^B = 0$) Agents observe a neutral, identity-independent source, such as factual or scientific evidence. This represents settings where all individuals receive the same objective information—for example, scientific data on vaccine efficacy in reducing COVID-19 mortality.

(*iii*) **Biased Sources** ($w_i^C > 0$ for $i \in C$, $C \in \{A, B\}$, $\xi^A \neq 0$, $\xi^B \neq 0$) Agents consume partisan information aligned with their group identity, reflecting selective media exposure—for instance, Republicans following Fox News and Democrats or Liberals preferring MSNBC or CNN.

3.1 Opinion dynamics with affective polarization

Our main result is summarized in the following theorem.

Theorem 1 *The opinions dynamics defined in equation (4) always converges to a unique long-run opinion vector* $\boldsymbol{\mu} := \lim_{t\to\infty} \boldsymbol{\mu}_t$. *Moreover,*

*(i) (**No information**) If there is no source of information, for all* $i \in A$ *and* $j \in B$, *long-run opinions are equal to*

$$\begin{cases} \mu_i^A \equiv \mu^A = \tilde{\boldsymbol{\pi}}^\top \boldsymbol{\mu}_0, \\ \mu_j^B \equiv \mu^B = -\mu^A. \end{cases} \tag{9}$$

*(ii) (**Unbiased information**) If agents have access to an unbiased source of information, for each* $i \in C$, *with* $C \in \{A, B\}$, *long-run opinions are given by*

$$\mu_i^C = \tilde{b}_i^C \theta^*. \tag{10}$$

*(iii) (**Biased information**) If agents have access to a biased source of information, for each* $i \in C$, *with* $C \in \{A, B\}$, *long-run opinions are equal to*

$$\mu_i^C \;=\; \tilde{b}_i^C\, \theta^* + \;\tilde{b}_i^{CA}\, \xi^A + \tilde{b}_i^{CB}\, \xi^B. \tag{11}$$

Opinions always converge to a unique value. This follows from the strong connectivity of the network, the normalization of individuals' weights over information sources, and the fact that agents place a positive weight on their own past opinion and/or the media outlet. Under these assumptions, the spectral properties of the identity-interaction matrix ensure that both the eigenvector and Katz-Bonacich centralities are well-defined.[9]

Consider case (i). When there is no available source of information (or no true state of the world), affective polarization generates a sharp divergence in opinions between agents belonging to different groups. Regardless of the network structure or the initial distribution of opinions, *in-group consensus* and *out-group polarization* emerge. Hence, while an agent's network position (captured by their eigenvector centrality) determines how their initial opinion contributes to the group's long-run opinion, all agents within the same group ultimately share identical views, with no within-group disagreement.[10]

This pattern is consistent with highly polarized moral debates, where no objective or testable truth exists and opinions reflect values, emotions, group identity, and affective polarization. Evidence from social media confirms this for debates about abortion: following the June 2022 overturning of *Roe vs. Wade*, Lerman et al. (2024) show that liberals and conservatives formed two distinct clusters on Twitter, with supportive interactions within groups and negative sentiment across them.

Next, consider case (ii), where agents value information about the state of the world and are exposed to an *unbiased* source, θ^*. Despite receiving accurate information, affective polarization causes agents to reject the opinions of the opposing group. As a result, even identical views across groups are weighted negatively, distorting perceptions of the true state of the world. In this setting, each agent's long-run opinion equals θ^* times her Katz–Bonacich centrality in $\tilde{\mathbf{W}}$, which is strictly below one. This centrality therefore measures how efficiently the agent aggregates information. The closer, in this sense, she is to members of the other group—measured by the cumulative strength of direct and indirect paths connecting them—the lower her centrality, as negative links attenuate the contributions along those paths. Intuitively, an agent with more connections to the out-group relies more heavily on negatively-weighted information, which pulls her opinion away from the truth. The next proposition formalizes these facts.

Proposition 1 *When agents are exposed to an unbiased source of information (case (ii)), affective polarization leads to learning failure. Additionally, agents with more (direct and indirect) exposure to the other group are farther away from the truth.*

[9] In our model, $\tilde{\boldsymbol{W}}$ has spectral radius equal to one in case (i), guaranteeing that the eigenvector centralities are well-defined, and lower than one in cases (ii) and (iii), guaranteeing that the Neumann series, $\sum_{k=0}^{+\infty} \tilde{\boldsymbol{W}}^k$, leading to Katz-Bonacich centralities converges.

[10] A similar result has been obtained by Shi et al. (2019).

Case (ii) captures contexts where individuals, regardless of identity, have access to credible information (e.g., scientific evidence or neutral media). Here, affective polarization does not necessarily induce ideological polarization, as in case (i), but it hampers collective learning and information aggregation.

Political and scientific debates often illustrate the pattern of case (ii): partisan hostility or group rivalry prevents agents from fully processing valid information, leading to persistent disagreement despite unbiased evidence. For example, the retweet network of COVID-19 discussions during the first months of the pandemic contained many small clusters, and these divisions did not coalesce into the kind of entrenched, binary ideological camps seen in case (i) (Lerman et al., 2024). This empirical pattern is consistent with the heterogeneous opinion structure $\boldsymbol{\mu} = \tilde{\boldsymbol{b}} \, . \, \theta^*$, where opinion variation is driven by agents' differing network centralities rather than a simple group-wide split.

Lastly, consider case (iii). In this scenario, long-run opinions also depend on the biases, ξ^C, and how they propagate in the network, as governed by the matrix $\tilde{\boldsymbol{M}}$. Specifically, the long-run opinion of agent i is a linear combination of the true state of the world, θ^*, and the biases, as expressed by equation (11). These biases are internalized by agents either positively or negatively, depending on whether they originate from their own or the opposing group. Consequently, from a social learning perspective, it is more desirable for the two groups to be biased in the same direction (e.g., $\xi^A = \xi^B > 0$) than in opposite directions (e.g., $\xi^A = -\xi^B > 0$). In other words, when media outlets share *biases of the same sign* but differ in narrative,[11] they can mitigate the adverse effects of out-group antagonism, reducing ideological polarization and disagreement and improving learning than if the biases were of opposite sign. Interestingly, this may lead to long-run opinions being closer to the truth than in case (ii), where no bias exists.

Proposition 2 *When agents are exposed to biased information (case (iii)), affective polarization reinforces biases of opposite sign and mitigates those of the same sign. Moreover, agents' opinions may be closer to the truth than under unbiased information.*

Studies on COVID-19 show that exposure to biased information from partisan media increased political polarization (Jungkunz, 2021), as individuals disproportionately relied on news aligned with their ideological identity, reinforcing disagreement across political groups (Strydhorst et al., 2023). Research documenting these patterns highlights the role of partisan information environments in amplifying political divides during the pandemic

[11] A real-world example is trade policy: Republican-leaning media (e.g., Fox News) highlight harm to U.S. workers, while Democratic-leaning outlets (e.g., MSNBC) emphasize social and environmental dumping. A second example comes from antitrust debates: left-leaning media stress inequality and worker exploitation, whereas right-leaning outlets focus on market fairness. In both cases, the two media outlets disseminate information through different narratives, but biased in the same directions.

(Motta et al., 2020). Additionally, affective polarization drives selective belief formation. Combining observational data and a survey experiment, Jenke (2024) shows that individuals are inclined to accept in-party misinformation and to reject out-party misinformation. Thus, antagonism can act as a filter against out-party misinformation, which should be taken into account when designing policies, as we show in Section 4.

3.2 Opinion dynamics without affective polarization

To highlight our findings, we compare Theorem 1 with a benchmark model without out-group antagonism, in which $\beta_i^C \geq 0$ for all $i \in C$, with $C \in \{A, B\}$. For this benchmark, let $\tilde{\boldsymbol{W}}^+ := (|\tilde{w}_{ij}^C|)_{i,j\in N}$, and define $\tilde{\boldsymbol{\pi}}^+ := \boldsymbol{\pi}(\tilde{\boldsymbol{W}}^+) = (\tilde{\pi}_i^{C+})_{i\in N}$ the associated eigenvector centrality and $\tilde{\boldsymbol{b}}^+ := \boldsymbol{b}(\tilde{\boldsymbol{W}}^+)$ the associated Katz–Bonacich centrality vector, where the elements are defined as in equations (6), (7), and (8). Since the functional forms of these centralities are unchanged, the superscript "+" simply identifies the corresponding objects in the model without out-group antagonism.[12]

Theorem 2 *The opinion dynamics defined in equation (4) with $\alpha_i^C > 0, \beta_i^C \geq 0$, for all $i \in C$, with $C \in \{A, B\}$, always converges to a unique $\boldsymbol{\mu}$. Furthermore:*

*(i) (**No information**) (Golub and Jackson, 2010) If there is no media outlet, then $\mu_i^C \equiv \mu$ for all $i \in C$, with $C \in \{A, B\}$, where*

$$\mu = \sum_{j\in N} \tilde{\pi}_j^{C+} \mu_{j,0}^C. \tag{12}$$

*(ii) (**Unbiased information**) (Jadbabaie et al., 2012) If agents have access to an unbiased source of information, then $\mu_i^C \equiv \mu$ for all $i \in C$, with $C \in \{A, B\}$, where*

$$\mu = \theta^*. \tag{13}$$

*(iii) (**Biased information**) (Friedkin and Johnsen, 1990) If agents have access to a biased source of information, then, for each $i \in C$, with $C \in \{A, B\}$, long-run opinions are equal to*

$$\mu_i^C \;=\; \tilde{b}_i^{C+}\,\theta^* + \;\tilde{b}_i^{CA+}\,\xi^A + \tilde{b}_i^{CB+}\,\xi^B. \tag{14}$$

Theorem 2(i) corresponds to the standard DeGroot model of opinion updating based on neighbors' average opinions (DeGroot, 1974; Golub and Jackson, 2010). Without out-group antagonism, agents' opinions always converge to a single long-run value, resulting

[12] See the proof of Theorem 2 for the formal definition of these centralities.

in societal *consensus*. As shown by Golub and Jackson (2010), long-run opinions are a weighted average of the initial opinions, where each agent's weight corresponds to their eigenvector centrality (see equation (12)). Furthermore, if initial opinions are independently distributed with mean θ^*, and the network is balanced and satisfies the minimal out-dispersion property, long-run opinions converge to the true state of the world.[13] This highlights the critical role of out-group antagonism in the emergence of ideological polarization in the society, especially when agents have no access to external information.

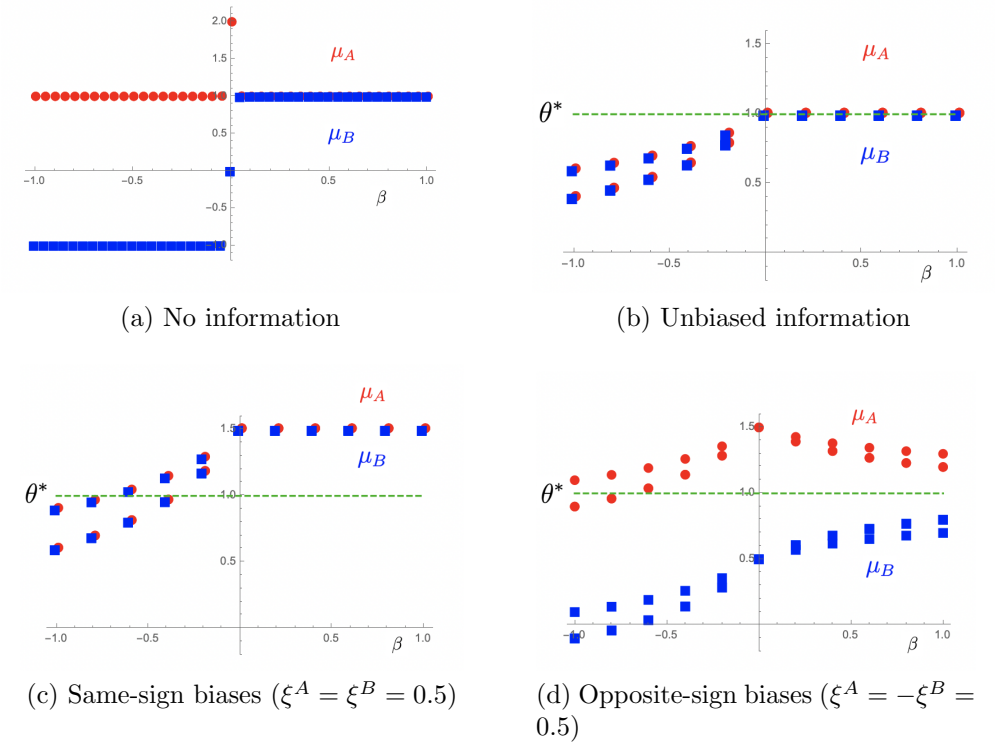
Theorem 2(*ii*) parallels the findings of Jadbabaie et al. (2012). They show that when agents have access to an unbiased source of information, consensus on the truth is always reached, regardless of initial opinions or network structure (provided that the network is strongly connected). Comparing the long-run opinions in (13) with those in (10) (Theorem 1) reveals that group antagonism introduces distortions that hinder both learning of the true state of the world and the achievement of consensus. Once out-group antagonism is introduced, long-run opinions deviate from the truth θ^*, and this divergence increases with the strength of antagonism.

Theorem 2(*iii*) corresponds to a setting with a *biased* information source, consistent with the findings of Friedkin and Johnsen (1990). The functional form of long-run opinions in (14) remains as in Theorem 1, but the Katz–Bonacich centralities are now computed over a non-signed network, implying that agents internalize the biases of both groups with positive weights. Consequently, in the absence of out-group antagonism, negatively correlated biases enable agents to internalize opposing perspectives through social interactions, partially offsetting their own bias and reducing its overall effect. When such opposing biases balance each other, the average opinion in society moves closer to the truth. Conversely, when biases are aligned in the same direction, they reinforce one another and persist. Hence, without out-group antagonism, negatively correlated group biases tend to yield more accurate societal beliefs. By contrast, while out-group antagonism always distorts learning with unbiased media (case (*ii*)), when biases are present (case (*iii*)), the impact of antagonism depends on the correlation of group biases: negatively correlated biases imply more polarization than positively correlated biases.

Figure 3 compares long-run opinions from Theorem 1 (with group antagonism) and Theorem 2 (without antagonism) for different values of β_i^C and when $\theta^* = 1$. Each panel, based on the ring network in Figure 1, shows how opinions vary with the information structure and the strength of inter-group interactions: $\beta = 0$ represents no interaction

[13] A network is balanced if no *family* can receive infinitely more weight from the remaining agents than it gives. It satisfies the minimal out-dispersion property if any sufficiently large finite family allocates at least some minimal weight to nearly all of society. In Golub and Jackson (2010), a family is defined as a collection of agents that may be changing and growing as the society expands. For formal definitions and a detailed discussion, see Golub and Jackson (2010).

between groups; $\beta > 0$ corresponds to the benchmark without antagonism (Theorem 2), while $\beta < 0$ captures out-group antagonism (Theorem 1), with more negative values indicating stronger antagonism.



(a) No information

(b) Unbiased information

(c) Same-sign biases ($\xi^A = \xi^B = 0.5$)

(d) Opposite-sign biases ($\xi^A = -\xi^B = 0.5$)

Figure 3: Long-run opinions in the ring network with and without out-group antagonism. The green dotted line (if present) represents the true state of the world, $\theta^* = 1$.

The general pattern displayed in Figure 3 mirrors the experimental evidence in Guilbeault et al. (2018), where displaying political logos prevents information aggregation and consensus, while removing the logos—and thus eliminating antagonistic reactions—allows subjects to better aggregate information and their opinions tend to converge. Our model displays a similar qualitative contrast. With group antagonism, disagreement persists and long-run opinions may move further from the truth and display polarization; without antagonism, social interactions favor opinion aggregation and consensus even under different informational environments. In this sense, our model shows how antagonism interacts with the structure of information to determine whether social interactions promote consensus or disagreement.

In particular, Panel (a) of Figure 3 (no information) shows that out-group antagonism is crucial for the emergence of ideological polarization when agents lack external information. In panel (b) (unbiased information), once antagonism appears ($\beta < 0$), long-run opinions begin to deviate from the truth θ^*, with stronger antagonism leading to greater divergence. Panels (c) and (d) of Figure 3 illustrate the case of biased information. In

panel (c), the biases have the same sign ($\xi^A = \xi^B = 0.5$). When affective polarization is present ($\beta < 0$), out-group antagonism partially counteracts positively correlated misinformation, bringing long-run opinions closer to the truth θ^*; in contrast, when $\beta > 0$, the two biases reinforce each other and generate a consensus far away from θ^*. In panel (d), the groups receive biases of opposite sign ($\xi^A = 0.5$ and $\xi^B = -0.5$). Under conformity ($\beta > 0$), these opposing distortions cancel out, and the average opinion aligns closely with the truth. When affective polarization emerges ($\beta < 0$), however, antagonism prevents this cancellation, disagreement rises, and the average opinion moves further from θ^* as β becomes more negative.

Table 1 summarizes these results, highlighting the impact of the information structure and the presence or absence of out-group antagonism.

	Out-Group Antagonism	**No Out-Group Antagonism**
(*i*) No information	*In-group* consensus *Out-group* polarization	Consensus
(*ii*) Unbiased source of information	Disagreement and failure of learning	Consensus and learning
(*iii*.1) Same-sign biased sources of information	Reduced distance from the truth	No reduced distance from the truth
(*iii*.2) Opposite-sign biased sources of information	Increased Polarization	Reduced distance from the truth

Table 1: Opinion dynamics with and without out-group antagonism

4 Policy implications

In this section, we examine how policies that enhance the accuracy of information available to one or both groups influence outcomes. We also consider interventions targeting specific individuals and censorship. Our analysis focuses on two key outcomes: the distance from the truth and the degree of disagreement between groups.[14]

4.1 Information campaigns

In many countries, information campaigns or media coverage on issues such as vaccination, climate change, or immigration often lead individuals to interpret messages differently and respond in opposing ways (Djourelova et al., 2024; Egorov et al., 2025; Grossman et al., 2020; Schneider-Strawczynski and Valette, 2025). Our model captures these heterogeneous

[14]Designing policies to maximize welfare requires committing to a specific microfoundation, as each assumes a different utility function, leading to distinct welfare definitions and policy implications.

reactions. In particular, it allows us to derive the conditions under which providing better information improves learning.

Consider first the case in which information campaigns are unbiased (Theorem 1(*ii*)), such as government advertising on the societal benefits of vaccination or recycling. It is natural to assume that only one group pays attention to the campaign, namely the group politically aligned with the government. Formally, the campaign increases the attention that all agents in group $C \in \{A, B\}$ pay to the unbiased information source—while all the other weights are proportionally reduced—such that for all $i \in C$, $w_i^{C\prime} > w_i^C$, where $w_i^{C\prime}$ denotes the post-campaign weight assigned to the unbiased source. For all agents $j \in C^c$, the weights remain unchanged, i.e., $w_j^{C^c\prime} = w_j^{C^c}$. The next proposition shows that, due to affective polarization, the campaign has opposite effects on the two groups.

Proposition 3 *If agents are exposed to an unbiased source of information (case (ii)), increasing this exposure for one group moves its opinions closer to the truth but pushes those of the other group further away.*

This proposition implies that focusing on individuals already inclined toward the desired behavior—for example, motivating environmentally conscious individuals to recycle more (as in Ellen et al., 1991)—may backfire by eliciting opposite reactions among others. Thus, the effect of such a policy on average learning in society is ambiguous. Hence, information campaigns should aim to persuade both groups by offering differentiated messages tailored to distinct political orientations, as shown by (Kidwell et al., 2013) in the context of recycling.

We now turn to information campaigns in the presence of biased media outlets (Theorem 1(*iii*)). In this setting, we examine the effects of a policy aimed at reducing media bias. The next proposition characterizes the conditions under which providing marginally more accurate information to one of the two groups—i.e., reducing $|\xi^A|$ or $|\xi^B|$—enhances learning and lowers disagreement in society as a whole.

Proposition 4 *If agents are exposed to biased sources of information (case (iii)), for each $C \in \{A, B\}$, there exist two weakly increasing piecewise-linear functions, $f^C(\cdot)$ (convex) and $g^C(\cdot)$ (concave), such that:*

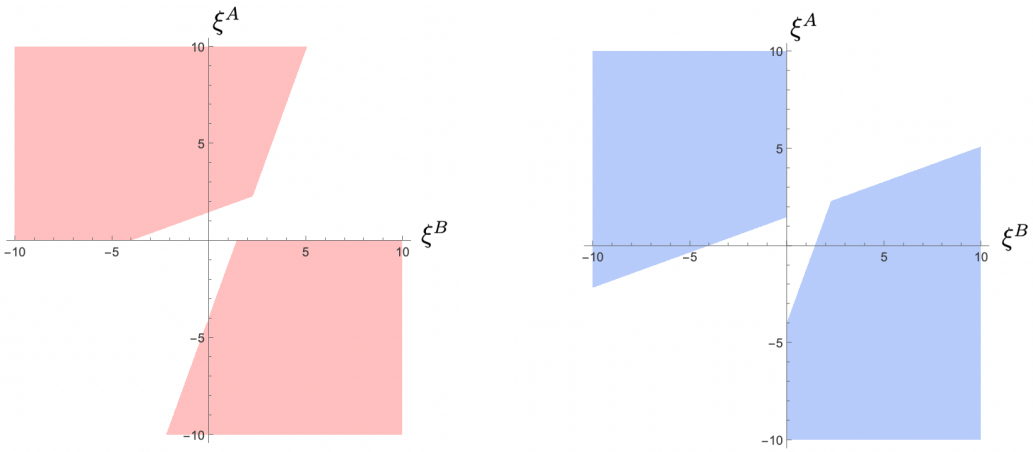
(*a*) *If $\xi^C > 0$, providing marginally more accurate information to group C improves learning for both groups and decreases disagreement in society if and only if $\xi^C > f^C(\xi^{C^c})$.*

(*b*) *If $\xi^C < 0$, providing marginally more accurate information to group C improves learning for both groups and decreases disagreement in society if and only if $\xi^C < g^C(\xi^{C^c})$.*

Proposition 4 characterizes when providing more accurate information to a group improves learning for both groups and reduces societal disagreement. The key insight is that interventions are most effective when the targeted group's bias is sufficiently extreme compared to the other group. For a group C with positive bias ($\xi^C > 0$, part (a)), improving information accuracy benefits society when $\xi^C > f^C(\xi^{C^c})$. For a group with negative bias ($\xi^C < 0$, part (b)), the intervention is beneficial when $\xi^C < g^C(\xi^{C^c})$.

The functions $f^C(\cdot)$ and $g^C(\cdot)$ define the critical thresholds for successful intervention. These functions delimit the parameter regions where information provision has a positive effect for learning in group A, learning in group B, and overall disagreement. For each outcome, the corresponding condition defines to a line in the (ξ^A, ξ^B) plane, since opinions are linear in biases. The upper and lower envelopes of these lines define, respectively, the convex function $f^C(\cdot)$ and the concave function $g^C(\cdot)$, separating regions where greater information accuracy improves outcomes from those where it may be counterproductive.

These regions are illustrated by the colored areas in Figure 4. In panel (a), the red region identifies the parameter values for which improving information for group A enhances average learning in both groups and reduces disagreement. The blue region in panel (b) identifies the analogous parameter values for interventions targeting group B. The white regions indicate cases in which the intervention backfires along at least one dimension.



(a) Red: $\xi^A > f^A(\xi^B)$, $\xi^A < g^A(\xi^B)$ (b) Blue: $\xi^A > (f^B)^{-1}(\xi^B)$, $\xi^A < (g^B)^{-1}(\xi^B)$

Figure 4: Regions where providing more accurate information to group A (panel (a)) or group B (panel (b)) improves average learning in both groups and reduces disagreement.

When biases are positively correlated (ξ^A and ξ^B have the same sign), the conditions of Proposition 4 conflict. For instance, if both are positive, we need $\xi^A > f^A(\xi^B)$ if targeting group A, a constraint that holds only if biases differ sufficiently (the red region in Figure 5). Analogously, we need $\xi^B > f^B(\xi^A)$ if targeting group B (or, $\xi^A < (f^B)^{-1}(\xi^B)$, the

blue region of Figure 5). Hence, improving both groups' information would generate counteracting forces. Intuitively, correcting one group's bias makes it more moderate, but the other group—observing this through group antagonism—reacts by becoming more extreme, potentially worsening overall learning.

When biases are negatively correlated (ξ^A and ξ^B have opposite signs), the conditions in Proposition 4 align. For instance, with $\xi^A > 0$, if both $\xi^A > f^A(\xi^B)$ and $\xi^B < g^B(\xi^A)$ (i.e., $\xi^A > (g^B)^{-1}(\xi^B)$) hold, then improving information for either or both groups enhances learning and reduces disagreement (the purple regions in Figure 5). This occurs because correcting one group's upward bias complements correcting the other group's downward bias—both corrections push the average opinion toward the truth, making interventions beneficial when biases are sufficiently large.

Corollary 1 formalizes the key insight on simultaneous interventions.

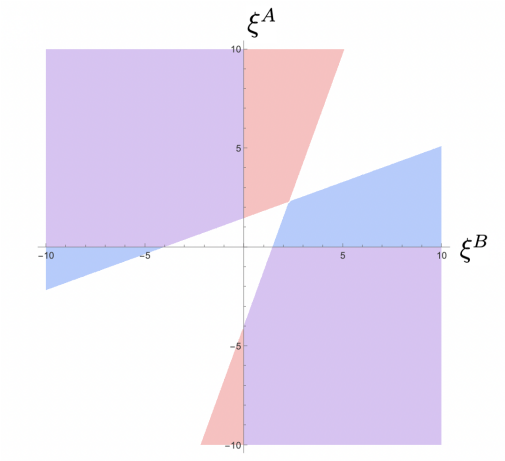


Figure 5: Optimal information-provision policy in the (ξ^B, ξ^A) space. Regions show when providing more accurate information improves learning and reduce disagreement. Purple areas: improving information for either or both groups A and B is beneficial. Red areas: improving only A's information is effective without backfire. Blue areas: improving only B's information is effective without backfire. White areas: improving information for either group backfires on learning or disagreement.

Corollary 1 *If agents are exposed to biased sources of information (case (iii)), providing better information to both groups improves learning and disagreement for the whole society if and only if* $sign(\xi^B) \neq sign(\xi^A)$ *and the magnitude of biases is large enough—i.e.,* $\xi^A > \max\{f^A(\xi^B), (g^B)^{-1}(\xi^B)\}$ *when* $\xi^A > 0$.

Overall, Proposition 4 and Corollary 1 show that when group biases have opposite signs, improving information for one or both groups improves learning and reduces disagreement, provided that the biases are sufficiently large. When biases share the same sign, such interventions are more likely to backfire: better information benefits society only when the gap between biases is wide enough and the more biased group is targeted.

These results shed light on how malevolent actors seek to harm learning and social cohesion. Misinformation campaigns circulate false or misleading content, distort facts, or highlight divisive narratives, as seen in foreign influence operations, online propaganda, and anti-vaccination misinformation (e.g., Broniatowski et al., 2018; Zhuravskaya et al., 2020; Simchon et al., 2022). In our model, such interventions correspond to increasing the bias received by one or both groups—i.e., increasing $|\xi^A|$ and/or $|\xi^B|$. Whenever reducing these biases would improve learning and decrease disagreement, pushing them in the opposite direction produces the opposite outcomes. Therefore, whenever biases have opposite sign and are strong enough, malevolent agents would spread even more misinformation.

Understanding these dynamics informs the design of effective anti-misinformation policies. Identifying the most vulnerable groups allows policymakers to limit attacks that seek to increase polarization.

4.2 Targeting

We analyze a policy that targets a single agent rather than an entire group. In practice, broad information campaigns may be infeasible or too costly, as they require reaching a large and heterogeneous audience. In such cases, influencing a specific individual can be a more efficient way to shape beliefs and foster information diffusion. For instance, Alatas et al. (2020) show that engaging influential figures on social media significantly increased the reach and impact of pro-vaccination messages in Indonesia.

However, identifying which individual should be targeted is not straightforward. Following Ballester et al. (2006), we define the *key player* as the agent for whom improving information more effectively brings the average opinion in society closer to the truth.

Formally, let $\boldsymbol{\xi}'(i)$ denote a vector of biases such that agent $i \in C$ receives marginally more accurate information—that is, ξ_i^C decreases in magnitude while maintaining its sign—whereas all other agents in group C retain the same bias, $\xi_j^C = \xi^C$ for all $j \neq i$ and $C \in \{A, B\}$. Let $\bar{\mu}\big(\boldsymbol{\xi}'(i)\big)$ denote the long-run average opinion in society after agent i receives improved information.

The *key player* i^* is defined as the agent who minimizes the distance between the average opinion and the truth:

$$\min_{\substack{i \in C \\ C \in \mathcal{C}}} \left\{ \left| \bar{\mu}\big(\boldsymbol{\xi}'(i)\big) - \theta^* \right| \right\}.$$

Finally, define the *out-degree* (weighted) Bonacich centrality as the column sum of the Leontief inverse—which accounts for all walks of all lengths originating from agent

i—weighted by the attention given to the media outlet, w_i^C. Formally:

$$\tilde{b}_i^{[out]C} := \left(\sum_{j \in A} \tilde{m}_{ji}^A + \sum_{k \in B} \tilde{m}_{ki}^B \right) w_i^C.$$

Proposition 5 *Suppose agents are exposed to biased sources of information (case (iii)).*

(*a*) *The key player i^* is the agent with the largest absolute (out-degree) Bonacich centrality $|\tilde{b}_{i^*}^{[out]C}|$, chosen from group $C \in \{A, B\}$, whose sign matches ξ^C when $\bar{\mu} > \theta^*$ and opposes it when $\bar{\mu} < \theta^*$. If no such agent exists, improving information for any agent worsens the average learning in society.*

(*b*) *There exists $\bar{\xi} > 0$ such that the key-player policy reduces disagreement in the whole society if and only if $|\xi_{i^*}^C| > \bar{\xi}$.*

The key-player policy identifies the agent whose improved information most effectively brings society's average opinion closer to the truth. The sign of an agent's out-degree Bonacich centrality $\tilde{b}_i^{[out]C}$ indicates whether their influence propagates primarily through in-group connections (positive sign) or out-group connections (negative sign). When influence flows primarily through in-group paths, other agents assign positive weight to this person's opinion on average; when it flows primarily through out-group paths, they assign negative weight on average, due to the dominance of negative inter-group links.

The optimal intervention depends on the aggregate bias and the position of the key player in the network. If society overestimates the truth ($\bar{\mu} > \theta^*$), we need to target an agent whose corrected opinion will pull the average down. This requires, e.g., someone with positive out-degree centrality (influence through in-group paths) and positive bias: reducing their upward bias directly reduces the average opinion. Conversely, if society underestimates ($\bar{\mu} < \theta^*$), we need to pull the average up. This requires, e.g., someone with negative out-degree centrality (influence through out-group paths) and positive bias: reducing their upward bias increases the average opinion.

Hence, the magnitude of the Bonacich centrality captures how strongly an agent's information affects others, while its sign indicates the direction of this influence. The policy therefore targets the agent with the largest absolute Bonacich centrality, among those whose influence moves the others' opinion towards the truth.

By construction, targeting the key player improves the average learning across society. However, its effect on each group's learning is ambiguous: one group may move further from the truth even as the societal average improves. As we discuss in the proof of Proposition 5, because group averages depend linearly on individual biases, adjusting the key player's bias has the same qualitative effect on group-level learning as shifting the bias of all agents in her group in the same direction, mirroring the logic of Proposition 4.

Part (b) extends this parallel to disagreement. However, disagreement depends non-linearly on individual biases, so the resulting condition in part (b) is novel, although the requirement that the key player's bias be sufficiently large in magnitude remains in the same spirit as Proposition 4.

Finally, the key-player framework also provides insight into how a malevolent actor could worsen learning. By the logic of Proposition 5, such an agent would target the same key player, providing them with highly biased or misleading information. Identifying the key player thus reveals which individuals are most critical to protect from misinformation, as influence over them propagates strongly through the network and can significantly increase polarization or move the average opinion away from the truth.

4.3 Censorship

Another form of intervention is censorship, that is, limiting the diffusion of information or opinions that deviate excessively from the truth. Such policies are often motivated by the goal of curbing misinformation or protecting public discourse from extreme or false claims. In what follows, we analyze the effects of censorship and show that, even when applied "fairly"—that is, by removing extreme opinions independently of their political alignment with those in power—it may backfire and worsen aggregate learning.

The results in the previous sections have direct implications for censorship when applied to media outlets. Conceptually, a benevolent censor would reduce media bias—an instance of the information policy in Proposition 4. The effects then depend on the joint configuration of group biases. When biases are negatively correlated, limiting the bias of one group—that is, censoring its most distorted media outlet—improves learning for both groups and reduces disagreement, as the direct and indirect effects of the policy reinforce each other. When biases are positively correlated, however, such interventions can backfire: reducing the bias of one group may trigger opposite reactions in the other, worsening overall disagreement, particularly when the two groups are exposed to similarly distorted media outlets. Thus, even well-intentioned censorship of extreme information can inadvertently increase polarization or hinder learning, depending on the underlying alignment of biases in society.

Another form of censorship targets individuals directly by constraining their publicly stated opinions, e.g., on social media. Such interventions are relatively easy to implement online and have indeed been used in practice through account moderation or content removal. We now analyze this case, where agents are prevented from expressing views that deviate too far from the truth.

To study this kind of censorship, we focus on the case in which the policy maker censors all opinions above a certain threshold $\bar{\bar{\mu}} > \theta^*$. Denote the long-run opinion when

censorship is in place by $\hat{\boldsymbol{\mu}}$ and define $[x]^+ := \max\{x, 0\}$; then:

Proposition 6 *If the policymaker imposes an upper censorship level $\bar{\bar{\mu}} > \theta^*$, the long-run opinions with censorship $\hat{\boldsymbol{\mu}}$ satisfy*

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} - \tilde{\boldsymbol{M}}\tilde{\boldsymbol{W}}[\hat{\boldsymbol{\mu}} - \bar{\bar{\mu}}\mathbf{1}]^+. \tag{15}$$

Moreover, there exists a threshold k such that if $\max_{j\in B} \mu_j^B < k < \bar{\bar{\mu}} < \max_{i\in A} \mu_i^A$, then censorship applies only to agents in group A. In this case, equation (15) implies that:

- *Censorship moves the average opinions of group A closer to the truth if and only if $\bar{\mu}^A > \theta^*$.*
- *Censorship moves the average opinions of group B closer to the truth if and only if $\bar{\mu}^B < \theta^*$.*

The effect of censorship depends on the groups' initial positions relative to the truth. Censoring a group shifts its opinion downward, and, because of out-group antagonism, induces an upward reaction in the other group. If the two groups are sufficiently polarized, censoring members of the group that overstates the truth makes both adjustments point toward θ^*, improving learning for everyone. Conversely, when the average opinions of both groups lie above the truth, censoring either one pulls that group closer while pushing the other further away; the opposite holds when both are below the truth.

In short, in the presence of group antagonism, censorship is beneficial for the whole society only if the two groups are sufficiently polarized; otherwise, it can backfire. This is consistent with the broader logic of the model and the previous policy sections.

5 Conclusion

Affective polarization (a deep distrust and hostility toward the opposing political group) has become a defining feature of contemporary public debate. For example, former President Obama referred to the rise of "negative partisanship" in the 2024 U.S. elections, noting that citizens are motivated less by support for specific policies than by opposition to the other side (The Economist, 2024). In such environments, arguments are filtered through group identity, and discussions are driven as much by loyalty and hostility as by facts. These patterns motivate a formal framework for understanding identity-driven opinion dynamics.

We study how affective polarization and media exposure influence ideological polarization and disagreement in a society of two groups exchanging opinions on a network.

Without media exposure, affective polarization always generates long-run ideological polarization. With exposure to unbiased media, affective polarization prevents consensus and induces persistent disagreement, and the distortion in learning is larger for agents with more inter-group contacts. With biased media, affective polarization amplifies biases of opposite sign, deepening polarization, but attenuates biases of the same sign, which can bring opinions closer to the truth.

Our model shows that the impact of information interventions depends on the alignment and magnitude of group biases and the network of interactions. Providing more accurate information to all individuals improves learning and reduces disagreement only when biases point in opposite directions; when biases align, interventions are more likely to backfire unless they target the most extreme group. Targeting influential agents allows efficient correction of aggregate beliefs, while censorship or limiting extreme opinions is effective only when groups start on opposite sides of the truth; otherwise, it may increase polarization. These insights underscore that policies must account for social influence and preexisting biases: ignoring affective polarization can lead to misleading predictions, as illustrated during the COVID-19 pandemic, where partisan animosity was tightly linked to attitudes toward health measures and vaccination (Druckman et al., 2021). Addressing these challenges requires more than reducing misinformation; it calls for carefully designed interventions that account for group antagonism and bias alignment.

Motivated by the literature on affective polarization, this paper takes it as given and focuses on how its effects on opinion dynamics. Additionally, we assumed that the network was exogenous, positioning it as a foundational step toward understanding the impact of out-group antagonism on long-run opinions. Introducing network endogeneity would represent a significant advancement, as it would allow for the study of endogenous out-group antagonism and the co-evolution of opinions and network formation. We leave these extensions as a direction for future research.

Appendix

A Microfoundation/interpretation

We provide here various microfoundations for the updating rule described in (5). While we interpret $\boldsymbol{\mu}_t$ as individuals' *opinions* or *beliefs*, it can also represent *behavior* or *attitude*, as emphasized by Golub and Jackson (2010, 2012).

A.1 Opinions with persuasion bias and zero-sum thinking

When interpreting $\boldsymbol{\mu}_t$ as *opinions*, we build on DeMarzo et al. (2003). At time $t = 0$, agents receive private noisy signals, observe the opinions of their neighbors, and update their own opinions by assigning weights to these observed opinions. These weights are determined at $t = 1$ and remain constant thereafter, capturing *persuasion bias*—the tendency of agents to neglect the correlation induced by repeated information over time.

We depart from DeMarzo et al. (2003) in two important respects. First, we allow for group-specific public information sources. Second, we assume that agents possess incomplete information not only about the precision of others' signals but also about how these signals relate to their own expected utility. This perspective aligns with evidence that individuals often interpret policy debates through a zero-sum lens. For example, in trade policy, while economists generally emphasize that free trade raises overall welfare, public opposition frequently reflects the belief that its benefits accrue mainly to economic elites while ordinary workers bear the losses (e.g., Ali et al., 2025).]

In particular, we assume that agents form opinions about the average effect of a policy ϕ. Let $u_i^C(\phi)$ denote the utility of individual $i \in C$ under policy p. The objective expected value of policy p is group-independent and equals $U^C(\phi) := \int_{i \in C} u_i^C(\phi) di = \theta^*$ for all $C \in \{A, B\}$. That is, although individual utilities $u_i^C(\phi)$ may vary across agents, the average impact of the policy is identical across groups.

Agents, however, lack full information about how the policy affects the two groups and hold a misspecified utility function that reflects *zero-sum thinking*—the belief that gains for one individual or group necessarily come at the expense of others.[15] Formally, the expected utility of an individual $i \in C$ is given by

$$\mathbb{E}_i^C\big[U^C(\phi)\big] = -\,\mathbb{E}_i^C\big[U^{C^c}(\phi)\big],$$

[15] Zero-sum thinking is treated here as a psychological trait, as in Bergeron et al. (2023), Chinoy et al. (2025), and Gavrilets and Seabright (2025). Ali et al. (2025) shows instead how zero-sum thinking can manifest even with completely rational voters.

so that any policy believed to benefit the out-group is perceived as harmful to the in-group.

At time $t = 0$, each agent $i \in C$ with $C \in \{A, B\}$ receives a private noisy signal s_i^C about the effect of the policy on their own utility. This signal is given by $s_i^C = u_i^C(\phi) + \epsilon_i^C$, where ϵ_i^C is normally distributed with mean zero. Since the group-level expected utility satisfies $U^C(\phi) = \theta^*$, all private signals are unbiased. That is, for any $i, j \in C$, $\mathbb{E}[s_i^C] = \mathbb{E}[u_i^C(\phi)] = \mathbb{E}[u_j^C(\phi)] = U^C(\phi) = \theta^*$. Given uninformative priors, each agent sets their initial opinion to $\mu_{i,0}^C = s_i^C$. These private signals thus set the initial conditions for the opinion dynamics; for all subsequent periods ($t \geq 1$), updates are driven by peer interactions and the recurring public information source as in DeMarzo et al. (2003).

In particular, at time $t = 1$, each agent $i \in C$ observes the initial opinions of their neighbors and interprets them as noisy signals about the expected utility of the policy. Specifically, for each in-group neighbor $j \in C$, agent i treats $\mu_{j,0}^C$ as a signal about the group-level expected utility $U^C(\phi) = \theta^*$ and assigns it a subjective precision $\tau_{ij}^{CC} = \frac{1}{\mathrm{Var}_i^C[\epsilon_j^C]}$. For each out-group neighbor $k \in C^c$, agent i interprets $\mu_{k,0}^{C^c}$ as a signal about the out-group's expected utility $U^{C^c}(\phi)$, which, under zero-sum thinking, they believe equals $-U^C(\phi) = -\theta^*$. Accordingly, agent i treats $\mu_{k,0}^{C^c}$ as a signal about $-U^C(\phi)$ and assigns it a subjective precision $\tau_{ik}^{CC^c} = \frac{1}{\mathrm{Var}_i^C[\epsilon_k^{C^c}]}$.[16]

Agents of each group C also observe signal θ^C about their group's expected utility from the policy, which they perceive as unbiased with subjective precision $\tau_{i\theta}^C$. Therefore, for $t \geq 1$, Bayesian agent $i \in C$ updates opinions according to:

$$\mu_{i,t}^C = \sum_{j \in C} \tilde{w}_{ij}^C \mu_{j,t-1}^C + \sum_{z \in C^c} \tilde{w}_{iz}^C \mu_{z,t-1}^{C^c} + w_i^C \theta^C,$$

where, for all $k \in N \equiv A \cup B$,

$$\tilde{w}_{ik}^C = \begin{cases} \dfrac{\tau_{ik}^{CC}}{\sum_{j \in A} \tau_{ij}^{CA} + \sum_{z \in B} \tau_{iz}^{CB} + \tau_{i\theta}^C} & \text{if } k \in C, \\ \dfrac{\tau_{ik}^{CC^c}}{\sum_{j \in A} \tau_{ij}^{CA} + \sum_{z \in B} \tau_{iz}^{CB} + \tau_{i\theta}^C} \cdot (-1) & \text{if } k \in C^c, \end{cases}$$

[16] Negative weights on the opinions of the out-group (i.e., $\beta_i^C < 0$) may also arise if agents systematically pay attention to different experts or information sources (Sethi and Yildiz, 2016), or rely on different models to interpret the same experiences (Haghtalab et al., 2021). In this view, belief distortion reflects not hostility but inference under limited or group-biased exposure: agents filter out (what they perceive as) the bias in their neighbors' opinions. Alternatively, negative weights can be justified as a rational response when agents believe that the signals of out-group members are less precise, for example when trying to learn a drifting state of the world (Dasaratha et al., 2023).

and[17]

$$w_i^C = \frac{\tau_{i\theta}^C}{\sum_{j\in A}\tau_{ij}^{CA} + \sum_{z\in B}\tau_{iz}^{CB} + \tau_{i\theta}^C}.$$

In sum, persuasion bias and zero-sum thinking jointly yield the linear updating rule (5), embedding both positive and negative influence in a microfounded way. This provides the bridge between psychologically grounded belief formation and the long-run opinion dynamics we analyze.[18]

A.2 Behaviors/actions/attitudes and network games

When interpreting $\boldsymbol{\mu}_t$ as *behaviors*, *actions*, or *attitudes*, the updating rule described in (5) corresponds to the myopic best-response dynamics of network games with linear best replies with both positive and negative spillovers. In doing so, our model not only incorporates but also extends several key contributions from the existing network-game literature (Ballester et al., 2006; Bramoullé et al., 2014; Jackson and Zenou, 2015).[19]

Coordination/anti-coordination game In many settings, individuals face social pressure to express opinions that signal loyalty to their own group and distance from rivals. For instance, in debates on trade, climate change, or COVID vaccination, publicly endorsing a position associated with the opposite political camp may carry identity costs, even when privately agreed with. This motivates a framework in which voiced opinions or attitudes balance in-group conformity, out-group differentiation, and closeness to one's own beliefs about the state of the world. We formalize this as follows. For each $i \in C$, consider the following utility function:

[17] This reduced-form approach captures the idea that the media outlet provides in each period a signal centered at θ^C with precision $\tau_{i\theta}^C$, as in Jadbabaie et al. (2012) or Della Lena (2024). We abstract from noise to simplify the exposition, as it does not affect the steady-state characterization (see Theorem 2).

[18] An alternative formulation has each agent i form subjective expectations about how agent j's signal relates to the truth θ^*. Agent i believes $\mathbb{E}_i[s_j] = f_{ij}(\theta^*)$, where $f_{ij} : \mathbb{R} \to \mathbb{R}$ is a (group-specific) distortion function capturing perceived bias. Denoting the subjective precision as $\tau_{ij} = Var_i[\epsilon_j]^{-1}$, the updating rule becomes:

$$\mu_{i,t} = \sum_{k\in N} \frac{\tau_{ik}}{\sum_{j\in N}\tau_{ij} + \tau_{i\theta}} f_{ik}^{-1}(\mu_{k,t-1}) + \frac{\tau_{i\theta}}{\sum_{j\in N}\tau_{ij} + \tau_{i\theta}} f_{i\theta}^{-1}(\theta).$$

This formulation nests the zero-sum case when $f_{ij}(\theta^*) = \theta^*$ for in-group agents and $f_{ij}(\theta^*) = -\theta^*$ for other agents. Agents perceive their own group receiving unbiased signals centered at θ^* as *well-informed*, while perceiving the other group receiving biased signals centered at $-\theta^*$ as *ill-informed*.

[19] Given our assumptions about the network, in cases with external information the condition $\alpha_i^C \sum_{j\in C} w_{ij}^C + |\beta_i^C| \sum_{z\in C^c} w_{iz}^C + w_i^C = 1$ ensures the existence and uniqueness of an equilibrium. When there is no external information, there may be multiple Nash equilibria; however, fixing initial beliefs selects a unique trajectory under best-reply dynamics and thus a unique limiting opinion profile.

$$u_i^C\left(\boldsymbol{\mu};\theta^*\right) = -\hat{\alpha}_i^C \cdot \underbrace{\sum_{j\in C} w_{ij}^C\left(\mu_i^C-\mu_j^C\right)^2}_{\text{in-group identity}} - \hat{\beta}_i^C \cdot \underbrace{\sum_{z\in C^c} w_{iz}^C\left(\mu_i^C-\left(-\mu_z^{C^c}\right)\right)^2}_{\text{out-group antagonism}} - \hat{w}_i^C \cdot \underbrace{\mathbb{E}_i\left[\left(\mu_i^C-\theta^*\right)^2\right]}_{\text{distance from the truth}}, \tag{A-1}$$

The first term, weighted by $\hat{\alpha}_i^C > 0$, captures the preference to align with the voiced opinions/attitudes of the agents in their own group. The second term, weighted by $\hat{\beta}_i^C > 0$, captures out-group antagonism and the psychological loss of not expressing an oppositional attitude toward the out-group. This loss is low when the agent's voiced opinion counteracts the out-group's stance and increases whenever the agent's opinion aligns with, or supports, the out-group's stance. The last term, weighted by $\hat{w}_i^C \geq 0$, reflects the cost of voicing an opinion or holding an attitude different from their true one.

The interaction between identity and cognitive dissonance (e.g., Festinger, 1957) explains these dynamics: individuals reduce the discomfort of holding conflicting beliefs or attitudes by conforming to the opinions of their own group and distancing themselves from the ideas embraced by a group they disdain.

Let $\boldsymbol{\mu}_{-i}$ denote the action of all agents other than i and define $\kappa := \hat{\alpha}_i^C \sum_{j\in C} w_{ij}^C + \hat{\beta}_i^C \sum_{z\in C^c} w_{iz}^C + \hat{w}_i^C$. Each agent $i \in C$, with $C \in \{A, B\}$ chooses μ_i^C to maximize equation (A-1), yielding the best-reply function:[20]

$$\mu_i^C(\boldsymbol{\mu}_{-i}) = \frac{1}{\kappa}\left[\hat{\alpha}_i^C \sum_{j\in C} w_{ij}^C \mu_j^C - \hat{\beta}_i^C \sum_{z\in C^c} w_{iz}^C \mu_z^{C^c} + \hat{w}_i^C \theta^C\right].$$

If $\alpha_i^C = \hat{\alpha}_i^C/\kappa$, $\beta_i^C = -\hat{\beta}_i^C/\kappa$, and $w_i^C = \hat{w}_i^C/\kappa$, we obtain

$$\begin{aligned}\mu_i^C(\boldsymbol{\mu}_{-i}) =& \alpha_i^C \sum_{j\in C} w_{ij}^C \mu_j^C + \beta_i^C \sum_{z\in C^c} w_{iz}^C \mu_z^{C^c} + w_i^C \theta^C \\ =& \sum_{j\in C} \tilde{w}_{ij}^C \mu_j^C + \sum_{z\in C^c} \tilde{w}_{iz}^C \mu_z^{C^c} + w_i^C(\theta^* + \xi^C).\end{aligned} \tag{A-2}$$

When each agent myopically responds to their peers, we can aggregate these best replies for all agents in groups A and B, yielding the opinion dynamics described in (5).

[20] To derive the best response below, we use the assumption that the signal is precise, although not necessarily correct. This implies that $\mathbb{E}_i\left[\left(\mu_i^C-\theta^*\right)^2\right] = \left(\mu_i^C-\theta^C\right)^2$. The F.O.C. are $-2\hat{\alpha}_i^C \sum_{j\in C} w_{ij}^C\left(\mu_i^C - \mu_j^C\right) - 2\hat{\beta}_i^C \sum_{z\in C^c} w_{iz}^C\left(\mu_i^C + \mu_z^{C^c}\right) - 2\hat{w}_i^C\left(\mu_i^C - \theta^C\right) = 0$. The S.O.C., $-2\hat{\alpha}_i^C \sum_{j\in C} w_{ij}^C - 2\hat{\beta}_i^C \sum_{z\in C^c} w_{iz}^C - 2\hat{w}_i^C < 0$, are always satisfied since $\hat{\alpha}_i^C > 0$, $\hat{\beta}_i^C > 0$, and $\hat{w}_i^C \geq 0$.

Network games with strategic complements and substitutes Following Ballester et al. (2006) and Bramoullé et al. (2014), let the utility function for each agent $i \in C$ be:

$$u_i^C(\boldsymbol{\mu}; \theta^C) = w_i^C \theta^C \mu_i^C - \frac{1}{2}(\mu_i^C)^2 + \mu_i^C \left(\alpha_i^C \sum_{j \in C} w_{ij}^C \mu_j^C + \beta_i^C \sum_{z \in C^c} w_{iz}^C \mu_z^{C^c} \right), \tag{A-3}$$

where the first two terms determine i's optimal action in isolation; $\alpha_i^C > 0$ reflects the intensity of positive spillovers from other agents within the same group, representing *strategic complementarities* in actions among agents of the same group; in contrast, $\beta_i^C < 0$ indicates the intensity of the negative spillovers from agents in the other group, capturing *strategic substitutes* in actions between agents from different groups. Contrary to Ballester et al. (2006) and Bramoullé et al. (2014), we allow for heterogeneous spillovers and negative values of the action x_i. In practice, we do not impose any restrictions on the relative magnitude of these spillovers, permitting individuals to experience both positive and negative spillovers in a heterogeneous manner. The only restriction is the normalization $\alpha_i^C \sum_{j \in C} w_{ij}^C + |\beta_i^C| \sum_{z \in C^c} w_{iz}^C + w_i^C = 1$ for each $i \in N$.

The utility function (A-3) can be exemplified by the case of recycling: while evidence suggests positive peer effects, where individuals are more likely to recycle if their friends do (Johansson, 2016), recycling also functions as a local public good (Kinateder and Merlino, 2017, 2022), which can lead to negative spillovers. In our model, agents experience peer pressure from their friends, motivating them to recycle, but are more individualistic when considering the recycling efforts of others, which tends to reduce their own effort via the free-riding effect.

Let $\boldsymbol{\mu}_{-i}$ denote the action of all agents other than i. Each agent $i \in C$, with $C \in \{A, B\}$ chooses μ_i^C to maximizes equation (A-3), yielding the best-reply function:[21]

$$\begin{aligned} \mu_i^C(\boldsymbol{\mu}_{-i}) =& \alpha_i^C \sum_{j \in C} w_{ij}^C \mu_j^C + \beta_i^C \sum_{z \in C^c} w_{iz}^C \mu_z^{C^c} + w_i^C \theta^C \\ =& \sum_{j \in C} \tilde{w}_{ij}^C \mu_j^C + \sum_{z \in C^c} \tilde{w}_{iz}^C \mu_z^{C^c} + w_i^C(\theta^* + \xi^C). \end{aligned} \tag{A-4}$$

When each agent myopically responds to their peers, we can aggregate these best replies for all agents in groups A and B, yielding the opinion dynamics described in (5).

Finally, note that equations (A-2) and (A-4) also result from the best replies of any combination of the two network games in equation (A-1) and equation (A-3). As in Boucher et al. (2024) and Ushchev and Zenou (2020), we can envision scenarios where peer pressure, conformism, and anti-conformism coexist.

[21] The first order condition is $w_i^C \theta^C - \mu_i^C + \alpha_i^C \sum_{j \in C} w_{ij}^C \mu_j^C + \beta_i^C \sum_{z \in C^c} w_{iz}^C \mu_z^{C^c} = 0$; the second-order condition is always satisfied.

B Proofs

Proof of Theorem 1

Note that equation (4) can be written as:

$$\begin{bmatrix} \boldsymbol{\mu}_t^A \\ \boldsymbol{\mu}_t^B \end{bmatrix} = \tilde{\boldsymbol{W}} \begin{bmatrix} \boldsymbol{\mu}_{t-1}^A \\ \boldsymbol{\mu}_{t-1}^B \end{bmatrix} + \begin{bmatrix} \mathbf{w}^A(\theta^* + \xi^A) \\ \mathbf{w}^B(\theta^* + \xi^B) \end{bmatrix}. \tag{B-1}$$

Since the matrix $\tilde{\boldsymbol{W}}$ has conformable partitions (i.e., each partition has the same number of rows and columns), its powers have positive terms on the diagonal blocks and negative terms on the off-diagonal blocks. For example,

$$\tilde{\boldsymbol{W}}^2 = \begin{bmatrix} (\tilde{\boldsymbol{W}}^{AA})^2 + \tilde{\boldsymbol{W}}^{AB}\tilde{\boldsymbol{W}}^{BA} & \tilde{\boldsymbol{W}}^{AA}\tilde{\boldsymbol{W}}^{AB} + \tilde{\boldsymbol{W}}^{AB}\tilde{\boldsymbol{W}}^{BB} \\ \tilde{\boldsymbol{W}}^{BA}\tilde{\boldsymbol{W}}^{AA} + \tilde{\boldsymbol{W}}^{BB}\tilde{\boldsymbol{W}}^{BA} & \tilde{\boldsymbol{W}}^{BA}\tilde{\boldsymbol{W}}^{AB} + (\tilde{\boldsymbol{W}}^{BB})^2 \end{bmatrix}.$$

Hence, $sign(\tilde{\boldsymbol{W}}^2) = sign(\tilde{\boldsymbol{W}}) = \begin{bmatrix} + & - \\ - & + \end{bmatrix}$. Iterating, this property holds for all t, so that

$$sign(\tilde{\boldsymbol{W}}^t) = \begin{bmatrix} + & - \\ - & + \end{bmatrix} \quad \text{for all } t \in \mathbb{N}. \tag{B-2}$$

Case (i) First, consider the case in which $w_i^C = 0$ for all $i \in N$. Equation (B-1) becomes $\begin{bmatrix} \boldsymbol{\mu}_t^A \\ \boldsymbol{\mu}_t^B \end{bmatrix} = \tilde{\boldsymbol{W}} \begin{bmatrix} \boldsymbol{\mu}_{t-1}^A \\ \boldsymbol{\mu}_{t-1}^B \end{bmatrix}$. In more compact form, $\boldsymbol{\mu}_t = \tilde{\boldsymbol{W}}\boldsymbol{\mu}_{t-1}$. By iterating the process, we obtain $\boldsymbol{\mu}_t = \tilde{\boldsymbol{W}}^t \boldsymbol{\mu}_0$, where $\tilde{\boldsymbol{W}}^t$ is the power t of the matrix $\tilde{\boldsymbol{W}}$. Assuming that the network is strongly connected with positive self-loops ensures that the matrix $\tilde{\boldsymbol{W}}$ is irreducible. Moreover, since $\lambda_1 = 1$ is the dominant eigenvalue of $\tilde{\boldsymbol{W}}$, then $\tilde{\boldsymbol{W}}$ is power convergent and we can define $\tilde{\boldsymbol{W}}^\infty := \lim_{t\to\infty} \tilde{\boldsymbol{W}}^t$. We have that

$$\boldsymbol{\mu} := \lim_{t\to\infty} \boldsymbol{\mu}_t = \tilde{\boldsymbol{W}}^\infty \boldsymbol{\mu}_0. \tag{B-3}$$

Since $\tilde{\boldsymbol{W}}$ is structurally balanced, there exists a diagonal matrix $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_n)$ with $d_i \in \{+1, -1\}$ such that $\boldsymbol{D}\,\tilde{\boldsymbol{W}}\,\boldsymbol{D} = \tilde{\boldsymbol{W}}^+$. Hence, $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{W}}^+$ are similar and share the same eigenvalues. Given our assumptions on $\boldsymbol{W}$, $\tilde{\boldsymbol{W}}^+$ is a nonnegative, irreducible, aperiodic, and row-stochastic (since in case (i) $w_i^C = 0$ for all $i \in N$) matrix.
Thus, by the Perron–Frobenius theorem, the spectral radius is $\rho(\tilde{\boldsymbol{W}}^+) = 1$ with the dominant eigenvalue $\lambda_1 = 1$, while all other eigenvalues satisfy $|\lambda_i| < 1$. Since similarity preserves spectral properties, the same holds for $\tilde{\boldsymbol{W}}$. Moreover, there exist strictly positive

right and left eigenvectors satisfying $\tilde{\boldsymbol{W}}^{+}\mathbf{1} = \mathbf{1}$ and $\boldsymbol{\pi}^{\top}\tilde{\boldsymbol{W}}^{+} = \boldsymbol{\pi}^{\top}$, where $\boldsymbol{\pi}^{\top}$ can be normalized to 1. Perron–Frobenius also implies that $(\tilde{\boldsymbol{W}}^{+})^{\infty} := \lim_{t\to\infty}(\tilde{\boldsymbol{W}}^{+})^{t} = \mathbf{1}\,\boldsymbol{\pi}^{\top}$. Because $\tilde{\boldsymbol{W}}^{t} = \boldsymbol{D}(\tilde{\boldsymbol{W}}^{+})^{t}\boldsymbol{D}$ for all t, taking limits we obtain $\tilde{\boldsymbol{W}}^{\infty} = \boldsymbol{D}\,(\mathbf{1}\boldsymbol{\pi}^{\top})\,\boldsymbol{D}$.
Thus, defining $\tilde{\mathbf{1}} := \boldsymbol{D}\mathbf{1}$ and $\tilde{\boldsymbol{\pi}}^{\top} := \boldsymbol{\pi}^{\top}\boldsymbol{D}$, substituting into (B-3), the long-run opinion is given by $\boldsymbol{\mu} := \left(\tilde{\mathbf{1}}\ \tilde{\boldsymbol{\pi}}^{\top}\right)\boldsymbol{\mu}_0$, so that $\mu_i^A \equiv \mu^A$, $\mu_j^B \equiv \mu^B$ for each $i \in A$ and $j \in B$, with

$$\begin{aligned}\mu^A =& 1 \cdot \left(\sum_{j\in A}\tilde{\pi}_j^A\mu_{j,0}^A + \sum_{k\in B}\tilde{\pi}_k^B\mu_{k,0}^B\right)\\ =& \sum_{j\in A}\tilde{\pi}_j^A\mu_{j,0}^A + \sum_{k\in B}\tilde{\pi}_k^B\mu_{k,0}^B\\ =& (\tilde{\boldsymbol{\pi}}^A)^{\top}\boldsymbol{\mu}_0^A + (\tilde{\boldsymbol{\pi}}^B)^{\top}\boldsymbol{\mu}_0^B = -\mu^B.\end{aligned}$$

Note that $\tilde{\mathbf{1}}$ and $\tilde{\boldsymbol{\pi}}^{\top}$ are the right and left eigenvectors of $\tilde{\boldsymbol{W}}$ associated with the leading eigenvalue $\lambda_1 = 1$. Let $\mathbf{u}^{\top}$ and $\mathbf{v}$ denote the left and right eigenvectors of $\tilde{\boldsymbol{W}}$ for $\lambda_1 = 1$, so that $\mathbf{u}^{\top}\tilde{\boldsymbol{W}} = \mathbf{u}^{\top}$ and $\tilde{\boldsymbol{W}}\mathbf{v} = \mathbf{v}$. Substituting $\tilde{\boldsymbol{W}} = \boldsymbol{D}\tilde{\boldsymbol{W}}^{+}\boldsymbol{D}$ gives $\mathbf{u}^{\top}(\boldsymbol{D}\tilde{\boldsymbol{W}}^{+}\boldsymbol{D}) = \mathbf{u}^{\top}$ and $(\boldsymbol{D}\tilde{\boldsymbol{W}}^{+}\boldsymbol{D})\mathbf{v} = \mathbf{v}$. Multiplying the first equation on the right by $\boldsymbol{D}$ and the second on the left by $\boldsymbol{D}$, and using $\boldsymbol{D}^2 = \boldsymbol{I}$, yields $(\mathbf{u}^{\top}\boldsymbol{D})\tilde{\boldsymbol{W}}^{+} = \mathbf{u}^{\top}\boldsymbol{D}$ and $\tilde{\boldsymbol{W}}^{+}(\boldsymbol{D}\mathbf{v}) = \boldsymbol{D}\mathbf{v}$. Since the left and right Perron eigenvectors of $\tilde{\boldsymbol{W}}^{+}$ are $\boldsymbol{\pi}^{\top}$ and $\mathbf{1}$, it follows that $\mathbf{u}^{\top}\boldsymbol{D} = \boldsymbol{\pi}^{\top}$ and $\boldsymbol{D}\mathbf{v} = \mathbf{1}$, hence $\mathbf{u}^{\top} = \boldsymbol{\pi}^{\top}\boldsymbol{D} = \tilde{\boldsymbol{\pi}}^{\top}$ and $\mathbf{v} = \boldsymbol{D}\mathbf{1} = \tilde{\mathbf{1}}$.

Cases *(ii)*-*(iii)* Consider the cases when $w_i > 0$ for at least some $i \in N$. Iterating (B-1), we obtain:

$$\begin{aligned}\boldsymbol{\mu}_{t+1} =& \tilde{\boldsymbol{W}}\boldsymbol{\mu}_t + \mathbf{w}\odot\boldsymbol{\theta},\\ \boldsymbol{\mu}_{t+2} =& \tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{W}}\boldsymbol{\mu}_t + \tilde{\mathbf{w}}\odot\boldsymbol{\theta}\right) + \mathbf{w}\odot\boldsymbol{\theta},\\ &\ldots\\ \boldsymbol{\mu}_{t+T} =& \tilde{\boldsymbol{W}}^{T}\boldsymbol{\mu}_t + \sum_{\tau=0}^{T-1}\tilde{\boldsymbol{W}}^{\tau}\mathbf{w}\odot\boldsymbol{\theta},\end{aligned}$$

so that

$$\lim_{T\to\infty}(\boldsymbol{\mu}_{t+T}) = \lim_{T\to\infty}\left(\tilde{\boldsymbol{W}}^{T}\boldsymbol{\mu}_t + \sum_{\tau=0}^{T-1}\tilde{\boldsymbol{W}}^{\tau}\mathbf{w}\odot\boldsymbol{\theta}\right) = \left(\boldsymbol{I} - \tilde{\boldsymbol{W}}\right)^{-1}\mathbf{w}\odot\boldsymbol{\theta}. \tag{B-4}$$

For cases (ii) and (iii), we assume at least one $w_i^C > 0$. The model's normalization, $\alpha_i^C\sum_{j\in C}w_{ij}^C + |\beta_i^C|\sum_{z\in C^c}w_{iz}^C + w_i^C = 1$, implies that the row sums of the matrix $\tilde{\boldsymbol{W}}^{+}$—which contains the absolute values of matrix $\tilde{\boldsymbol{W}}$—are $\sum_j|\tilde{w}_{ij}| = 1 - w_i^C \le 1$. Thus, $\tilde{\boldsymbol{W}}^{+}$ is a substochastic matrix. Since the network $\boldsymbol{W}$ is assumed to be strongly connected, the

non-negative matrix $\tilde{\boldsymbol{W}}^{+}$ is irreducible. As $\tilde{\boldsymbol{W}}^{+}$ is an irreducible and substochastic matrix with at least one row sum strictly less than 1 (since at least one $w_i^C > 0$), its spectral radius $\rho(\tilde{\boldsymbol{W}}^{+})$ is strictly less than 1. Because $\rho(\tilde{\boldsymbol{W}}) \leq \rho(\tilde{\boldsymbol{W}}^{+})$, we have $\rho(\tilde{\boldsymbol{W}}) < 1$. This ensures that $lim_{T\to\infty}\tilde{\boldsymbol{W}}^{T} = \mathbf{0}$ and the Neumann series $\sum_{\tau=0}^{\infty}\tilde{\boldsymbol{W}}^{\tau}$ converges to $(\boldsymbol{I}-\tilde{\boldsymbol{W}})^{-1}$. Given this, noting that $\left(\boldsymbol{I}-\tilde{\boldsymbol{W}}\right)$ is always invertible establishes the result.
Thus, if the source of information is unbiased, i.e., $\theta_i = \theta^*$ for all $i \in N$, (B-4) becomes $\boldsymbol{\mu} = \left(\boldsymbol{I}-\tilde{\boldsymbol{W}}\right)^{-1}\mathbf{w}\theta^* = \tilde{\boldsymbol{b}}\theta^*$. Therefore the long-run opinion of agent i is

$$\mu_i^C = \left(\sum_{j\in A}\tilde{m}_{ij}^C w_j^A + \sum_{k\in B}\tilde{m}_{ik}^C w_k^B\right)\theta^* = \tilde{b}_i^C\theta^*.$$

If instead the source of information is biased, observing that $\tilde{\boldsymbol{M}} := \left(\boldsymbol{I}-\tilde{\boldsymbol{W}}\right)^{-1}$, we can write equation (B-4) as

$$\boldsymbol{\mu} = \tilde{\boldsymbol{M}}\begin{bmatrix}\mathbf{w}^A(\theta^*+\xi^A)\\ \mathbf{w}^B(\theta^*+\xi^B)\end{bmatrix} = \tilde{\boldsymbol{M}}\mathbf{w}\theta^* + \tilde{\boldsymbol{M}}\left(\mathbf{w}\odot\boldsymbol{\xi}\right) = \tilde{\boldsymbol{b}}\theta^* + \tilde{\boldsymbol{M}}\left(\mathbf{w}\odot\boldsymbol{\xi}\right).$$

Therefore the long-run opinion of each agent $i \in N$ is given by:

$$\begin{aligned}\mu_i^C &= \left(\sum_{j\in A}\tilde{m}_{ij}^C w_j^A + \sum_{k\in B}\tilde{m}_{ik}^C w_k^B\right)\theta^* + \sum_{j\in A}\tilde{m}_{ij}^C w_j^A\xi^A + \sum_{k\in B}\tilde{m}_{ik}^C w_k^B\xi^B\\ &= \tilde{b}_i^C\theta^* + \tilde{b}_i^{CA}\xi^A + \tilde{b}_i^{CB}\xi^B.\end{aligned}$$

By equation (B-2), it is trivial to see that $\tilde{b}_i^{CC} > 0$ and $\tilde{b}_i^{CC^c} < 0$ always. ■

Proof of Proposition 1

By Theorem 1, long–run opinions under an unbiased source satisfy

$$\boldsymbol{\mu} = (\boldsymbol{I}-\tilde{\boldsymbol{W}})^{-1}\mathbf{w}\,\theta^* = \tilde{\boldsymbol{b}}\,\theta^*.$$

As we have already shown, under structural balance, the signed matrix admits the decomposition $\tilde{\boldsymbol{W}} = \boldsymbol{D}\,\tilde{\boldsymbol{W}}^{+}\,\boldsymbol{D}$, where $\boldsymbol{D}$ is diagonal with entries ± 1 and $\tilde{\boldsymbol{W}}^{+} \geq 0$ has the same absolute values as $\tilde{\boldsymbol{W}}$. Since $(\boldsymbol{I}-\tilde{\boldsymbol{W}}^{+})\mathbf{1} = \mathbf{w}$, in the corresponding unsigned network,

$$\boldsymbol{\mu} = (\boldsymbol{I}-\tilde{\boldsymbol{W}}^{+})^{-1}\mathbf{w}\,\theta^* = \tilde{\boldsymbol{b}}^{+}\,\theta^* = \mathbf{1}\,\theta^*.$$

Note that $\boldsymbol{D} = \boldsymbol{D}^{-1} = \boldsymbol{D}^{\top}$ and $\boldsymbol{D}^k = \boldsymbol{D}$ for k odd while $\boldsymbol{D}^k = \boldsymbol{I}$ for k even. Then, since $\boldsymbol{D}^2 = \boldsymbol{I}$ it follows that $(\boldsymbol{D}\tilde{\boldsymbol{W}}^{+}\boldsymbol{D})^t = \boldsymbol{D}(\tilde{\boldsymbol{W}}^{+})^t\boldsymbol{D}$. Thus:

$$(\boldsymbol{I}-\tilde{\boldsymbol{W}})^{-1}=\sum_{t=0}^{\infty}(\tilde{\boldsymbol{W}})^{t}=\sum_{t=0}^{\infty}(\boldsymbol{D}\tilde{\boldsymbol{W}}^{+}\boldsymbol{D})^{t}=\sum_{t=0}^{\infty}\boldsymbol{D}(\tilde{\boldsymbol{W}}^{+})^{t}\boldsymbol{D}=\boldsymbol{D}(\boldsymbol{I}-\tilde{\boldsymbol{W}}^{+})^{-1}\boldsymbol{D}.$$

As $\tilde{\boldsymbol{M}}^{+}=(\boldsymbol{I}-\tilde{\boldsymbol{W}}^{+})^{-1}$ has strictly positive entries, $\tilde{\boldsymbol{M}}=(\boldsymbol{I}-\tilde{\boldsymbol{W}})^{-1}$ has the same absolute values but matches the sign pattern induced by $\boldsymbol{D}$. Thus, since $\mathbf{w}$ has all positive entries, each element of $(\boldsymbol{I}-\tilde{\boldsymbol{W}})^{-1}\mathbf{w}$ is always less or equal than the corresponding element of $(\boldsymbol{I}-\tilde{\boldsymbol{W}}^{+})^{-1}\mathbf{w}$. Hence, every Katz-Bonacich coefficient satisfies $\tilde{b}_i^C\leq 1$, so that $\tilde{b}_i^C|\theta^*|\leq|\theta^*|$, for all $i\in C$, with $C\in\{A,B\}$ and the learning gap for agent $i\in C$ is $1-\tilde{b}_i^C$.

Moreover, for agent $i\in C$, the exposure (direct and indirect) to the other group is captured by $\sum_{j\in C^c}|\tilde{\boldsymbol{M}}_{ij}|$. Thus, given that $\tilde{\boldsymbol{M}}=\boldsymbol{D}\tilde{\boldsymbol{M}}^{+}\boldsymbol{D}$ the more direct and indirect paths each agent has with agents belonging to the other group the lower her centrality and thus the farther away from the truth. ■

Proof of Proposition 2

Let us write the long-run opinion of a generic agent $i\in C$ when ξ^A and ξ^B are different from zero and affective polarization is present:

$$\mu_i^C=\underbrace{\tilde{b}_i^C}_{<1}\theta^*+\underbrace{\tilde{b}_i^{CC}}_{+}\xi^C+\underbrace{\tilde{b}_i^{CC^c}}_{-}\xi^{C^c}.$$

From this equation, we see that:

- If $\text{sign}(\xi^A)=\text{sign}(\xi^B)$ (positively correlated biases), the negative coefficient $\tilde{b}_i^{CC^c}\in(-1,0)$ partially cancels the contribution of ξ^{C^c}, leading to bias mitigation.
- If $\text{sign}(\xi^A)\neq\text{sign}(\xi^B)$ (negatively correlated biases), the negative sign of $\tilde{b}_i^{CC^c}\in(-1,0)$ reverses the contribution of ξ^{C^c}, aligning both bias components in the same direction and exacerbating the total bias.

Since $|\tilde{b}_i^C\theta^*|<|\theta^*|$ always holds, long-run opinions can be closer to the truth θ^* whenever the bias terms $\tilde{b}_i^{CC}\xi^C+\tilde{b}_i^{CC^c}\xi^{C^c}$ reduce the gap between $\tilde{b}_i^C\theta^*$ and θ^*. ■

Proof of Theorem 2

First, define the identity-interaction matrix without group antagonism, $\tilde{\boldsymbol{W}}^{+}$. Let $\boldsymbol{\Gamma}^{C+} := diag[(|\beta_i^C|)_{i\in C}]$, for all $C \in \{A, B\}$. Then, all entries of $\boldsymbol{\Gamma}^{C+}$ are nonnegative. We have

$$\tilde{\boldsymbol{W}}^{+} := \begin{bmatrix} \boldsymbol{\Lambda}^{A}\boldsymbol{W}^{AA} & \boldsymbol{\Gamma}^{A+}\boldsymbol{W}^{AB} \\ \boldsymbol{\Gamma}^{B+}\boldsymbol{W}^{BA} & \boldsymbol{\Lambda}^{B}\boldsymbol{W}^{BB} \end{bmatrix} \qquad \text{with} \qquad sign(\tilde{\boldsymbol{W}}^{+}) = \begin{bmatrix} + & + \\ + & + \end{bmatrix}.$$

Then, $\tilde{\boldsymbol{M}}^{+} := \left(\boldsymbol{I} - \tilde{\boldsymbol{W}}^{+}\right)^{-1} = \sum_{k=0}^{+\infty}(\tilde{\boldsymbol{W}}^{+})^{k}$ with elements $\tilde{m}_{ij}^{+}$. Thus, $\tilde{\boldsymbol{b}}^{+} := \boldsymbol{b}(\tilde{\boldsymbol{W}}^{+}) = \tilde{\boldsymbol{M}}^{+}\mathbf{w}$ is the vector of *weighted Katz-Bonacich centralities* (Ballester et al., 2006) in the identity-interaction network without group antagonism. The corresponding individual weighted Katz-Bonacich centralities is given by

$$\tilde{b}_i^{C+} := \sum_{j\in A} \tilde{m}_{ij}^{C+} w_j^{A} + \sum_{k\in B} \tilde{m}_{ik}^{C+} w_k^{B}.$$

The contributions from each group is thus defined as:

$$\tilde{b}_i^{CA+} := \sum_{j\in A} \tilde{m}_{ij}^{C+} w_j^{A} \quad \text{and} \quad \tilde{b}_i^{CB+} := \sum_{k\in B} \tilde{m}_{ik}^{C+} w_k^{B},$$

so that $\tilde{b}_i^{C+} = \tilde{b}_i^{CA+} + \tilde{b}_i^{CB+}$ and the total centrality vector can be expressed compactly as

$$\tilde{\boldsymbol{b}}^{+} = \begin{bmatrix} \tilde{\boldsymbol{b}}^{A+} \\ \tilde{\boldsymbol{b}}^{B+} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{b}}^{AA+} + \tilde{\boldsymbol{b}}^{AB+} \\ \tilde{\boldsymbol{b}}^{BA+} + \tilde{\boldsymbol{b}}^{BB+} \end{bmatrix}.$$

The rest of the proof can be derived by adapting that of Theorem 1. ■

Proof of Proposition 3

Consider case (ii), with unbiased information ($\xi^A = \xi^B = 0$). By Theorem 1(ii), $|\mu_i^C| < |\theta^*|$ for all $i \in C$ and $C \in \{A, B\}$. Moreover, the stead-state opinion vector satisfies:

$$\left(\boldsymbol{I} - \tilde{\boldsymbol{W}}\right)\boldsymbol{\mu} = \mathbf{w} \odot \boldsymbol{\theta}^{*}. \tag{B-5}$$

Let us denote with $\mathbf{e}_i^C$ the standard basis vector for agent i in group C, that is, the vector with a 1 in position i and zeros elsewhere. Then by differencing equation (B-5) with respect to w_i^C we get

$$\left(\boldsymbol{I}-\tilde{\boldsymbol{W}}\right)\frac{\partial\boldsymbol{\mu}}{\partial w_i^C}-\frac{\partial\tilde{\boldsymbol{W}}}{\partial w_i^C}\boldsymbol{\mu}=\theta^*\mathbf{e}_i^C$$

$$\Rightarrow\ \frac{\partial\boldsymbol{\mu}}{\partial w_i^C}=\boldsymbol{M}\left(\theta^*\mathbf{e}_i^C+\frac{\partial\tilde{\boldsymbol{W}}}{\partial w_i^C}\boldsymbol{\mu}\right)$$

$$\frac{\partial\boldsymbol{\mu}}{\partial w_i^C}=\boldsymbol{M}\left(\theta^*\mathbf{e}_i^C-\frac{1}{1-w_i^C}\left(\sum_{j\in A}\tilde{w}_{ij}^C\mu_j^A+\sum_{k\in B}\tilde{w}_{ik}^C\mu_k^B\right)\mathbf{e}_i^C\right)$$

The last equality follows from the fact that, for each i, the sum of the absolute values of the entries in row i of $\tilde{\boldsymbol{W}}$ satisfies $\sum_{j\in A}|\tilde{w}_{ij}^C|+\sum_{k\in B}|\tilde{w}_{ik}^C|=1-w_i^C$. Therefore, a marginal increase in w_i^C affect only the i-th row by rescaling all weights proportionally. As a result, for each $j\in N$, $\frac{\partial\tilde{w}_{ij}^C}{\partial w_i^C}=-\frac{\tilde{w}_{ij}^C}{1-w_i^C}$.

Using the steady state condition $\mu_i^C=\sum_{j\in A}\tilde{w}_{ij}^C\mu_j^A+\sum_{k\in B}\tilde{w}_{ik}^C\mu_k^B+w_i^C\theta^*=\tilde{b}_i^C\theta^*$ we get

$$\frac{\partial\boldsymbol{\mu}}{\partial w_i^C}=\boldsymbol{M}\left(\theta^*\mathbf{e}_i^C-\frac{\mu_i^C-w_i^C\theta^*}{1-w_i^C}\mathbf{e}_i^C\right)=\boldsymbol{M}\left(\frac{\theta^*-\mu_i^C}{1-w_i^C}\,\mathbf{e}_i^C\right)=\left(\frac{1-\tilde{b}_i^C}{1-w_i^C}\theta^*\right)\boldsymbol{M}\mathbf{e}_i^C$$

Since $\tilde{b}_i^C<1$ and $w_i^C<1$, the scalar $\frac{1-\tilde{b}_i^C}{1-w_i^C}\theta^*$ has the same sign as θ^*. Hence, a marginal increase in w_i^C shifts long-run opinions in the direction of θ^* or away from it according to the sign pattern of $\tilde{\boldsymbol{M}}$, which has the same block-sign structure as $\tilde{\boldsymbol{W}}$ (see proof of Proposition 1). Thus, agents in the same group as i move toward θ^*, while agents in the other group move away from it.

Finally, since the result holds for each $i\in C$ individually, a simultaneous increase in w_i^C for all agents in group C yields an aggregate effect that preserves the same sign pattern for all agents' long-run opinions, which proves the proposition. ■

Proof of Proposition 4

By Theorem 1(iii), we have that $(\mu_i^C-\theta^*)=(\tilde{b}_i^C-1)\theta^*+\tilde{b}_i^{CA}\xi^A+\tilde{b}_i^{CB}\xi^B$, so the average distance from the truth is

$$\bar{\mu}^C-\theta^*=\frac{1}{n^C}\sum_{i\in C}(\mu_i^C-\theta^*)=\frac{1}{n^C}\sum_{i\in C}\left[(\tilde{b}_i^C-1)\theta^*+\tilde{b}_i^{CA}\xi^A+\tilde{b}_i^{CB}\xi^B\right],$$

where, for each $i\in C$, $\tilde{b}_i^{CA}\geq 0$ and $\tilde{b}_i^{CB}\leq 0$ if $C=A$; the inequalities are reversed if $C=B$.

Suppose now that $\xi^A>0$. Then, giving group A more accurate information is equivalent

to reducing ξ^A. Hence, a lower ξ^A reduces the average distance from the truth in group A if and only if $\bar{\mu}^A > \theta^*$, which implies $\frac{1}{n^A}\sum_{i\in A}\left[(\tilde{b}_i^A - 1)\theta^* + \tilde{b}_i^{AA}\xi^A + \tilde{b}_i^{AB}\xi^B\right] > 0$. Rearranging, the condition implies that $\xi^A > l^A(\xi^B)$, where

$$l^A(\xi^B) := \underbrace{\frac{\sum_{i\in A}(1-\tilde{b}_i^A)}{\sum_{i\in A}\tilde{b}_i^{AA}}}_{>0}\theta^* + \underbrace{\frac{-\sum_{i\in A}\tilde{b}_i^{AB}}{\sum_{i\in A}\tilde{b}_i^{AA}}}_{>0}\xi^B. \tag{B-6}$$

Similarly, reducing ξ^A reduces the distance from the truth of an agent $i \in B$ if and only if $\bar{\mu}^B < \theta^*$, which implies $\xi^A > l^B(\xi^B)$, where

$$l^B(\xi^B) := \underbrace{\frac{\sum_{i\in B}(1-\tilde{b}_i^B)}{\sum_{i\in B}\tilde{b}_i^{BA}}}_{<0}\theta^* + \underbrace{\frac{\sum_{i\in B}\tilde{b}_i^{BB}}{-\sum_{i\in B}\tilde{b}_i^{BA}}}_{>0}\xi^B. \tag{B-7}$$

If $\xi^A < 0$, the inequalities are reversed, yielding $\xi^A < l^A(\xi^B)$ and $\xi^A < l^B(\xi^B)$.
Figure B-1 shows the effect of marginally reducing $|\xi^A|$ on average learning across groups in the (ξ^B, ξ^A) space.

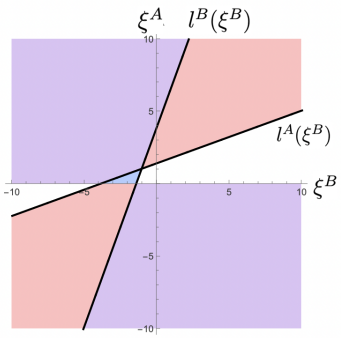


Figure B-1: Effect on group learning of providing more accurate information to group A. Violet: both groups move closer to truth. Red: only group A moves closer. Blue: only group B moves closer. White: neither group moves closer. Solid lines represent (B-6) and (B-7).

Suppose that $\xi^B > 0$. Then, giving group B more accurate information is equivalent to reducing ξ^B. Hence, a lower ξ^B reduces the average distance from the truth in group B if and only if $\bar{\mu}^B > \theta^*$, which leads to $\xi^A < l^B(\xi^B)$, which we can also write as $\xi^B > (l^B)^{-1}(\xi^A)$ as l^B is a linear function with positive slope.
A smaller ξ^B reduces the distance from the truth of an agent $i \in A$ if and only if $\bar{\mu}^A < \theta^*$, which leads to $\xi^A < l^A(\xi^B)$, which we can also write as $\xi^B > (l^A)^{-1}(\xi^A)$ as l^A is a linear function with positive slope.
If $\xi^B < 0$, the inequalities are reversed, yielding $\xi^A > l^A(\xi^B)$ and $\xi^A > l^B(\xi^B)$ (or $\xi^B < (l^A)^{-1}(\xi^A)$ and $\xi^B < (l^B)^{-1}(\xi^A)$).
Figure B-2 shows the effect of marginally reducing $|\xi^B|$ on average learning across groups

in the (ξ^B, ξ^A) space.

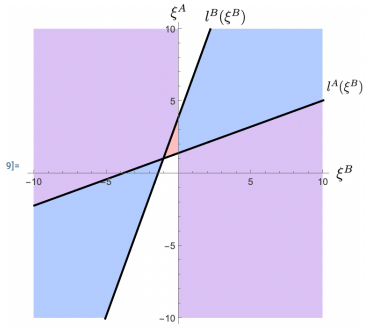


Figure B-2: Effect on group learning of providing more accurate information to group B. Violet: both groups move closer to truth. Red: only group A moves closer. Blue: only group B moves closer. White: neither group moves closer. Solid lines represent (B-6) and (B-7).

Concerning disagreement, by Theorem 1(iii), we can write the long run opinions as

$$\boldsymbol{\mu} = \tilde{\boldsymbol{b}}\theta^* + (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top \xi^A + (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top \xi^B.$$

Hence,

$$\begin{aligned} Var\left[\boldsymbol{\mu}\right] =& (\theta^*)^2 Var[\tilde{\boldsymbol{b}}] + (\xi^A)^2 Var[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top] + (\xi^B)^2 Var[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top] + \\ &+ 2\xi^A\theta^* Cov\left[\tilde{\boldsymbol{b}}, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top\right] + 2\xi^B\theta^* Cov\left[\tilde{\boldsymbol{b}}, (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top\right] + \\ &+ 2\xi^A\xi^B Cov\left[\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top)\right]. \end{aligned}$$

Noting that $\tilde{\boldsymbol{b}} = (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top + (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top$, after algebraic manipulations, we get

$$\begin{aligned} Var\left[\boldsymbol{\mu}\right] =& (\theta^*)^2 Var\left[\tilde{\boldsymbol{b}}\right] + \xi^A(\xi^A + 2\theta^*) Var\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top\right] + \xi^B(\xi^B + 2\theta^*) Var\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top\right] + \\ &+ 2\left(\theta^*(\xi^A + \xi^B) + \xi^A\xi^B\right) Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top)\right]. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{\partial Var[\boldsymbol{\mu}]}{\partial \xi^A} &= 2(\xi^A + \theta^*) Var\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top\right] + 2\left(\theta^* + \xi^B\right) Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top)\right], \\ \frac{\partial Var[\boldsymbol{\mu}]}{\partial \xi^B} &= 2(\xi^B + \theta^*) Var\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top\right] + 2\left(\theta^* + \xi^A\right) Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top)\right]. \end{aligned}$$

If ξ^A is positive, giving more accurate information to group A means a reduction in ξ^A. Hence, disagreement in society decreases if and only if $\frac{\partial Var[\boldsymbol{\mu}]}{\partial \xi^A} > 0$, and vice versa if ξ^A is negative. Hence, the following holds.

Suppose that $\xi^A > 0$. Then, giving group A more accurate information reduces the disagreement in society if and only if $\xi^A > d^A(\xi^B)$, where

$$d^A(\xi^B) := -\theta^* \left(1 + \frac{Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top)\right]}{Var\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top\right]} \right) + \underbrace{\frac{-Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top)\right]}{Var\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top\right]}}_{>0} \xi^B. \tag{B-8}$$

If $\xi^A < 0$, the condition is reversed. Figure B-3 shows the effect of marginally reducing $|\xi^A|$ on overall disagreement in society within the (ξ^B, ξ^A) space.

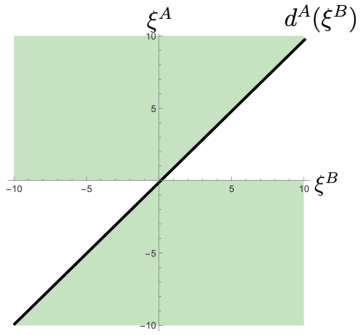


Figure B-3: Effect on disagreement of providing more accurate information to group A. Green: disagreement decreases. White: disagreement increases. Solid line represents (B-8).

If ξ^B is positive, giving more accurate information to group B means a reduction in ξ^B. Hence, disagreement in society decreases if and only if $\frac{\partial Var[\boldsymbol{\mu}]}{\partial \xi^B} > 0$, and vice versa if ξ^B is negative. Hence, the following hold.
Suppose that $\xi^B > 0$. Then, giving group B more accurate information, i.e., decreasing ξ^B, reduces the disagreement in society if and only if $\xi^A < d^B(\xi^B)$, where

$$d^B(\xi^B) := -\theta^* \left(1 + \frac{Var\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top\right]}{Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top)\right]} \right) + \underbrace{\frac{Var\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top\right]}{-Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top)\right]}}_{>0} \xi^B, \tag{B-9}$$

which we can also write as $\xi^B > (d^B)^{-1}(\xi^A)$ as d^B is a linear function with positive slope. The condition is reversed if $\xi^B < 0$. Figure B-4 shows the effect of marginally reducing $|\xi^B|$ on overall disagreement in society within the (ξ^B, ξ^A) space.

We can summarize the conditions under which providing more accurate information to group A improves learning for group A and group B and reduce disagreement as:

$$\xi^A > f^A(\xi^B) := \max\left\{l^A(\xi^B), l^B(\xi^B), d^A(\xi^B)\right\} \qquad \text{if} \quad \xi^A > 0, \tag{B-10}$$

$$\xi^A < g^A(\xi^B) := \min\left\{l^A(\xi^B), l^B(\xi^B), d^A(\xi^B)\right\} \qquad \text{if} \quad \xi^A < 0. \tag{B-11}$$

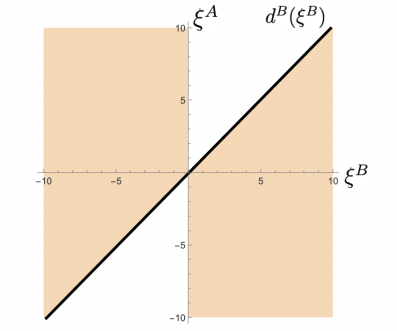


Figure B-4: Effect on disagreement of providing more accurate information to group B. Orange: disagreement decreases. White: disagreement increases. Solid line represents (B-9).

Similarly, we summarize the conditions under which providing more accurate information to group B improves learning for group A and group B and reduce disagreement as:

$$\xi^B > f^B(\xi^A) := \max\left\{(l^A)^{-1}(\xi^A), (l^B)^{-1}(\xi^A), (d^B)^{-1}(\xi^A)\right\} \quad \text{if} \quad \xi^B > 0, \tag{B-12}$$

$$\xi^B < g^B(\xi^A) := \min\left\{(l^A)^{-1}(\xi^A), (l^B)^{-1}(\xi^A), (d^B)^{-1}(\xi^A)\right\} \quad \text{if} \quad \xi^B < 0. \tag{B-13}$$

Proposition 4 follows directly from conditions (B-10)–(B-13). Since the functions l^A, l^B, d^A, and d^B are linear with positive slope, it follows that, for each $C \in A, B$, the functions f^C and g^C are increasing, with f^C convex and g^C concave. ■

Proof of Corollary 1

The corollary follows from two observations.
(i) When ξ^A and ξ^B have opposite signs, the conditions (B-10)–(B-13) under which providing better information to group A or to group B improves learning and decreases disagreement in each group are compatible; when they have the same sign, these conditions are incompatible.
(ii) Likewise, the conditions (B-10)–(B-13) under which providing better information to group A or to group B reduces disagreement move in the same direction when ξ^A and ξ^B have opposite signs, and in opposite directions otherwise.
Hence, when ξ^A and ξ^B have opposite signs, the relevant constraint is the more stringent of $\xi^A > f^A(\xi^B)$ and $\xi^B < g^B(\xi^A)$, or equivalently $\xi^A > (g^B)^{-1}(\xi^B)$.
This establishes Corollary 1. ■

Proof of Proposition 5

To prove part *(a)*, remember that the average opinion in society is

$$\bar{\mu}\left(\boldsymbol{\xi}'(i)\right) := \frac{1}{n}\mathbf{1}^{\top}\boldsymbol{\mu}\left(\boldsymbol{\xi}'(i)\right) = \frac{1}{n}\left(\sum_{C\in\{A,B\}}\sum_{z\in C}\left(\sum_{j\in A}\tilde{m}_{zj}^{C}w_{j}^{A}(\theta^{*}+\xi_{j}^{A})+\sum_{k\in B}\tilde{m}_{zk}^{C}w_{k}^{B}(\theta^{*}+\xi_{j}^{B})\right)\right).$$

It follows that

$$\frac{\partial\bar{\mu}\left(\boldsymbol{\xi}'(i)\right)}{\partial\xi_{i}^{C}} = \frac{1}{n}\left(\sum_{j\in A}\tilde{m}_{ji}^{A}w_{i}^{C}+\sum_{k\in B}\tilde{m}_{ki}^{B}w_{i}^{C}\right) = \frac{1}{n}\tilde{b}_{i}^{C[out]}. \tag{B-14}$$

Then, we have the following:

1. Consider $\bar{\mu} > \theta^{*}$, which is equivalent to $\frac{1}{n}\mathbf{1}^{\top}\left(\tilde{\boldsymbol{b}}\theta^{*}+\tilde{\boldsymbol{M}}\left(\mathbf{w}\odot\boldsymbol{\xi}\right)\right) > \theta^{*}$. Then:
   (a) If $\xi^{C} > 0$, to move the average opinion closer to the truth, provide more accurate information (i.e., decrease ξ^{C}) to an agent with positive $\tilde{b}_{i}^{C[out]}$; then, pick the agent i with the largest $\tilde{b}_{i}^{C[out]}$ (if more than one, pick one of them at random).
   (b) If $\xi^{C} < 0$, to move the average opinion closer to the truth, provide more accurate information (i.e., increase ξ^{C}) to an agent with negative $\tilde{b}_{i}^{C[out]}$; then, pick the agent i with the smallest $\tilde{b}_{i}^{C[out]}$ (if more than one, pick one of them at random).

   This is equivalent to giving more accurate information to an agent in S, where

$$S = \arg\max_{\substack{i\in C\\ C\in\{A,B\}}}\left|\tilde{b}_{i}^{C[out]}\right| \quad \text{s.t. } sign\left(\tilde{b}_{i}^{C[out]}\right) = sign(\xi^{C}).$$

   If such an agent exists, denote them by i^{*}. If S is empty, providing more accurate information to any agent moves the average opinion further from the truth.
2. Consider $\bar{\mu} < \theta^{*}$, which is equivalent to $\frac{1}{n}\mathbf{1}^{\top}\left(\tilde{\boldsymbol{b}}\theta^{*}+\tilde{\boldsymbol{M}}\left(\mathbf{w}\odot\boldsymbol{\xi}\right)\right) < \theta^{*}$. Then:
   (a) If $\xi^{C} > 0$, to move the average opinion closer to the truth, provide more accurate information (i.e., decrease ξ^{C}) to an agent with negative $\tilde{b}_{i}^{C[out]}$; then, pick the agent i with the smallest $\tilde{b}_{i}^{C[out]}$ (if more than one, pick one of them at random).
   (b) If $\xi^{C} < 0$, to move the average opinion closer to the truth, provide more accurate information (i.e., increase ξ^{C}) to an agent with positive $\tilde{b}_{i}^{C[out]}$; then, pick the agent i with the largest $\tilde{b}_{i}^{C[out]}$ (if more than one, pick one of them at random).

   This is equivalent to giving more accurate information to an agent in S', where

$$S' = \arg\max_{\substack{i\in C\\ C\in\{A,B\}}}\left|\tilde{b}_{i}^{C[out]}\right| \quad \text{s.t. } sign\left(\tilde{b}_{i}^{C[out]}\right) \neq sign(\xi^{C}).$$

If such an agent exists, denote them by i^*. If S is empty, providing more accurate information to any agent moves the average opinion further from the truth.

Consider the effect of providing more accurate information to i^* on the average opinion of group A and B. By (B-14), it follows that

$$\frac{\partial \bar{\mu}^A\left(\boldsymbol{\xi}'(i^*)\right)}{\partial \xi^C_{i^*}} = \frac{1}{n^A}\sum_{j\in A} \tilde{m}^A_{ji^*} w^C_{i^*} = \frac{1}{n^A}\tilde{b}^{CA[out]}_{i^*},$$
$$\frac{\partial \bar{\mu}^B\left(\boldsymbol{\xi}'(i^*)\right)}{\partial \xi^C_{i^*}} = \frac{1}{n^B}\sum_{k\in B} \tilde{m}^B_{ki^*} w^C_{i^*} = \frac{1}{n^B}\tilde{b}^{CB[out]}_{i^*}.$$

Suppose $S \cup S'$ is not empty, $i^* \in A$, and $\xi^A > 0$. Then, giving better information to the key player i^* is equivalent to reducing ξ^A_i. Clearly, this reduces μ^A_i. As this reduces (increases) μ^A_j (μ^B_j) for all $j \in A$ ($j \in B$) connected to i^*, $\bar{\mu}^A$ decrease while $\bar{\mu}^B$ increases. This reduces the distance from the truth of the average opinion in group A if and only if $\bar{\mu}^A > \theta^*$, which is equivalent to (B-6), and of the average opinion in group B if and only if $\bar{\mu}^B < \theta^*$, which is equivalent to (B-7). Analogous arguments for the other cases show that the conditions under which providing better information to the agent i^* reduces the distance between the group's average opinion and the truth are identical to those characterizing the effect of improving information for all agents in group A, as derived in Proposition 4.

As for part *(b)*, without loss of generality, suppose that $i^* \in A$. Then,

$$\boldsymbol{\mu} = \tilde{\boldsymbol{b}}\theta^* + (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top \xi^A + \begin{pmatrix} (\tilde{m}^A_{ki^*} w^A_{i^*})_{k\in A} \\ (\tilde{m}^B_{li^*} w^A_{i^*})_{l\in B} \end{pmatrix} \Delta\xi^A_{i^*} + (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top \xi^B.$$

We can then write disagreement as:

$$\begin{aligned} Var\left[\boldsymbol{\mu}(\Delta\xi^A_{i^*})\right] =& (\theta^*)^2 Var[\tilde{\boldsymbol{b}}] + (\xi^A)^2 Var[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top] + (\Delta\xi^A_{i^*})^2 Var\left[\begin{pmatrix} (\tilde{m}^A_{ki^*} w^A_{i^*})_{k\in A} \\ (\tilde{m}^B_{li^*} w^A_{i^*})_{l\in B} \end{pmatrix}\right] + \\ &+ (\xi^B)^2 Var[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top] + 2\xi^A\xi^B Cov\left[\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top)\right] + \\ &+ 2\xi^A\theta^* Cov\left[\tilde{\boldsymbol{b}}, (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top\right] + 2\xi^B\theta^* Cov\left[\tilde{\boldsymbol{b}}, (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top\right] + \\ &+ 2\theta^*\Delta\xi^A_{i^*} Cov\left[\tilde{\boldsymbol{b}}, \begin{pmatrix} (\tilde{m}^A_{ki^*} w^A_{i^*})_{k\in A} \\ (\tilde{m}^B_{li^*} w^A_{i^*})_{l\in B} \end{pmatrix}\right] + \\ &+ 2\xi^A\Delta\xi^A_{i^*} Cov\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, \begin{pmatrix} (\tilde{m}^A_{ki^*} w^A_{i^*})_{k\in A} \\ (\tilde{m}^B_{li^*} w^A_{i^*})_{l\in B} \end{pmatrix}\right] + \\ &+ 2\xi^B\Delta\xi^A_{i^*} Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, \begin{pmatrix} (\tilde{m}^A_{ki^*} w^A_{i^*})_{k\in A} \\ (\tilde{m}^B_{li^*} w^A_{i^*})_{l\in B} \end{pmatrix}\right]. \end{aligned}$$

Therefore,

$$\frac{\partial Var[\boldsymbol{\mu}]}{\partial \Delta\xi^A_{i^*}} = 2\Delta\xi^A_{i^*} \cdot Var\left[\begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right] + 2\theta^* \cdot Cov\left[\tilde{\boldsymbol{b}}, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right] +$$

$$+ 2\xi^A \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right] + 2\xi^B \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right].$$

As we consider marginal changes in $\xi^A_{i^*}$, $\Delta\xi^A_{i^*}$ is infinitesimal, so that $\frac{\partial Var[\boldsymbol{\mu}]}{\partial \Delta\xi^A_{i^*}}$ reduces to

$$\frac{\partial Var[\boldsymbol{\mu}]}{\partial \Delta\xi^A_{i^*}} = 2\theta^* \cdot Cov\left[\tilde{\boldsymbol{b}}, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right] + 2\xi^A \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right] +$$

$$+ 2\xi^B \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right].$$

Noting that $\tilde{\boldsymbol{b}} = (\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top + (\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top$, we can write

$$\frac{\partial Var[\boldsymbol{\mu}]}{\partial \Delta\xi^A_{i^*}} = 2(\theta^* + \xi^A) \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right] +$$

$$+ 2(\theta^* + \xi^B) \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right].$$

Hence, giving more accurate information to player i^* is equivalent to reducing $\xi^A_{i^*}$ if and only if $\xi^A > 0$. In that case, disagreement in society decreases if $\frac{\partial Var[\boldsymbol{\mu}]}{\partial \Delta\xi^A_{i^*}} > 0$, which means

$$(\theta^* + \xi^A) \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right] > -(\theta^* + \xi^B) \cdot Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right],$$

or, as the covariance $Cov\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right]$ is positive,

$$\xi^A > -\theta^* - (\theta^* + \xi^B) \cdot \frac{Cov\left[(\tilde{\boldsymbol{b}}^{AB}, \tilde{\boldsymbol{b}}^{BB})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right]}{Cov\left[(\tilde{\boldsymbol{b}}^{AA}, \tilde{\boldsymbol{b}}^{BA})^\top, \begin{pmatrix}(\tilde{m}^A_{ki^*}w^A_{i^*})_{k\in A}\\(\tilde{m}^B_{li^*}w^A_{i^*})_{l\in B}\end{pmatrix}\right]} := \bar{\xi}.$$

A similar argument holds if $i^* \in B$. To conclude, if S is not empty, there exists a threshold $\bar{\xi}$ such that targeting the agent $i \in C$ with more accurate information decreases

disagreement if and only if $\xi_{i^*}^C > \bar{\xi}$. This concludes the proof of Proposition 5. ■

Proof of Proposition 6

Consider the updating rule with an upper–censorship threshold $\bar{\bar{\mu}} > \theta^*$. Censorship removes the portion of any opinion exceeding $\bar{\bar{\mu}}$, so the stated opinion used by others is $\hat{\mu}_{i,t} - [\hat{\mu}_{i,t} - \bar{\bar{\mu}}]^+$. In vector form, the censored dynamics are

$$\hat{\boldsymbol{\mu}}_{t+1} = \tilde{\boldsymbol{W}}\hat{\boldsymbol{\mu}}_t + \mathbf{w} \odot \boldsymbol{\theta} - \tilde{\boldsymbol{W}}\left[\hat{\boldsymbol{\mu}}_t - \bar{\bar{\mu}}\mathbf{1}\right]^+. \tag{B-15}$$

Since $\rho(\tilde{\boldsymbol{W}}) < 1$, the uncensored updating rule converges to a unique steady state. Censorship keeps opinions within a fixed range at each time t, limiting extremes without creating new ones. Thus, the censored dynamics also converge to a unique steady state. Let $\hat{\boldsymbol{\mu}} := \lim_{t\to\infty} \hat{\boldsymbol{\mu}}_t$. Taking limits in (B-15) yields $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{W}}\hat{\boldsymbol{\mu}} + \mathbf{w} \odot \boldsymbol{\theta} - \tilde{\boldsymbol{W}}\left[\hat{\boldsymbol{\mu}} - \bar{\bar{\mu}}\mathbf{1}\right]^+$, from which we obtain the implicit fixed-point representation

$$\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{M}}(\mathbf{w} \odot \boldsymbol{\theta}) - \tilde{\boldsymbol{M}}\tilde{\boldsymbol{W}}\left[\hat{\boldsymbol{\mu}} - \bar{\bar{\mu}}\mathbf{1}\right]^+. \tag{B-16}$$

Since the long–run opinion without censorship is $\boldsymbol{\mu} = \tilde{\boldsymbol{M}}(\mathbf{w} \odot \boldsymbol{\theta})$, substituting into (B-16) yields equation (15).

To ensure that censorship applies only to agents in group A let explicitly write the censored long-run opinions for the two groups

$$\begin{bmatrix} \hat{\boldsymbol{\mu}}^A \\ \hat{\boldsymbol{\mu}}^B \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^A \\ \boldsymbol{\mu}^B \end{bmatrix} - \begin{bmatrix} \tilde{\boldsymbol{M}}^{AA} & \tilde{\boldsymbol{M}}^{AB} \\ \tilde{\boldsymbol{M}}^{BA} & \tilde{\boldsymbol{M}}^{BB} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{W}}^{AA} & \tilde{\boldsymbol{W}}^{AB} \\ \tilde{\boldsymbol{W}}^{BA} & \tilde{\boldsymbol{W}}^{BB} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}}^A - \bar{\bar{\mu}}\mathbf{1}_{n^A} \\ \hat{\boldsymbol{\mu}}^B - \bar{\bar{\mu}}\mathbf{1}_{n^B} \end{bmatrix}^+.$$

Therefore the long-run opinions of agents in group B satisfies

$$\hat{\boldsymbol{\mu}}^B = \boldsymbol{\mu}^B - \left(\tilde{\boldsymbol{M}}^{BA}\tilde{\boldsymbol{W}}^{AA} + \tilde{\boldsymbol{M}}^{BB}\tilde{\boldsymbol{W}}^{BA}\right)\left[\hat{\boldsymbol{\mu}}^A - \bar{\bar{\mu}}\mathbf{1}_{n^A}\right]^+ - \left(\tilde{\boldsymbol{M}}^{BA}\tilde{\boldsymbol{W}}^{AB} + \tilde{\boldsymbol{M}}^{BB}\tilde{\boldsymbol{W}}^{BB}\right)\left[\hat{\boldsymbol{\mu}}^B - \bar{\bar{\mu}}\mathbf{1}_{n^B}\right]^+.$$

Thus, the censorship does not bite on agents of group B if

$$\max_{j\in B}\left\{\boldsymbol{\mu}^B - \left(\tilde{\boldsymbol{M}}^{BA}\tilde{\boldsymbol{W}}^{AA} + \tilde{\boldsymbol{M}}^{BB}\tilde{\boldsymbol{W}}^{BA}\right)\left[\hat{\boldsymbol{\mu}}^A - \bar{\bar{\mu}}\mathbf{1}_{n^A}\right]^+\right\} < \bar{\bar{\mu}}.$$

The matrix $\boldsymbol{K} := -\left(\tilde{\boldsymbol{M}}^{BA}\tilde{\boldsymbol{W}}^{AA} + \tilde{\boldsymbol{M}}^{BB}\tilde{\boldsymbol{W}}^{BA}\right)$ has all positive entries, so that there exists $\bar{k} > 0$ such that the sum each row's entries are less than $\bar{k}$. Hence, the vector $\boldsymbol{K}[\hat{\boldsymbol{\mu}}^A - \bar{\bar{\mu}}\mathbf{1}_{n^A}]^+$ is bounded by $\bar{k}\left[\max_{i\in A}\mu_i^A - \bar{\bar{\mu}}\right]^+$. Therefore, a sufficient condition to

ensure that the censorship does not affect agents in group B is

$$\max_{j\in B}\mu_j^B + \bar{k}\left[\max_{i\in A}\mu_i^A - \bar{\bar{\mu}}\right]^+ < \bar{\bar{\mu}}.$$

The censorship applies to agents in group A when $\max_{i\in A}\mu_i^A > \bar{\bar{\mu}}$. Thus the condition for which the censorship affect only agents in A (and no agent in B) is

$$\max_{j\in B}\mu_j^B + \bar{k}\max_{i\in A}\mu_i^A - \bar{k}\bar{\bar{\mu}} < \bar{\bar{\mu}} \quad \text{and} \quad \bar{\bar{\mu}} < \max_{i\in A}\mu_i^A$$
$$\Rightarrow \quad k := \frac{\max_{j\in B}\mu_j^B + \bar{k}\max_{i\in A}\mu_i^A}{1+\bar{k}} < \bar{\bar{\mu}} < \max_{i\in A}\mu_i^A.$$

Thus, when the condition is satisfied $[\hat{\boldsymbol{\mu}} - \bar{\bar{\mu}}\mathbf{1}]^+$ has positive entries only for agents in A. Recall that $\tilde{\boldsymbol{M}} = (\boldsymbol{I}-\tilde{\boldsymbol{W}})^{-1} = \boldsymbol{D}(\boldsymbol{I}-\tilde{\boldsymbol{W}}^+)^{-1}\boldsymbol{D}$, therefore $\tilde{\boldsymbol{M}}\tilde{\boldsymbol{W}} = \boldsymbol{D}(\boldsymbol{I}-\tilde{\boldsymbol{W}}^+)^{-1}\tilde{\boldsymbol{W}}^+\boldsymbol{D}$, which has the same block-sign structure as $\tilde{\boldsymbol{W}}$.
Therefore, censorship *reduces* long-run opinions in group A and *raises* long-run opinions in group B:

$$\hat{\mu}_i^A \le \mu_i^A \quad \text{for all } i\in A, \text{ and } \hat{\mu}_j^B \ge \mu_j^B \quad \text{for all } j\in B,$$

with strict inequalities whenever some agent in A is actually censored. Thus:

- Reducing $\bar{\mu}^A$ moves it closer to θ^* if $\bar{\mu}^A > \theta^*$, and farther away if $\bar{\mu}^A < \theta^*$. Thus, censorship brings the average opinion of group A closer to the truth if and only if $\bar{\mu}^A > \theta^*$.
- Increasing $\bar{\mu}^B$ moves it closer to θ^* if $\bar{\mu}^B < \theta^*$, and farther away if $\bar{\mu}^B > \theta^*$. Thus, censorship brings the average opinion of group B closer to the truth if and only if $\bar{\mu}^B < \theta^*$.

This concludes the proof of Proposition 6. ■

C Ring Network Example

Network Structure and Weights Consider the six-agent ring in Figure 1, with groups $A = \{1,2,3\}$ and $B = \{4,5,6\}$, and parameters $\alpha_i^C = -\beta_i^C = 1$ and $w_i^C = w$ for all $i \in C$, $C \in A, B$. Agents' total attention is normalized to one. In these examples, we assume that agents allocate attention equally across their own past opinion, the opinions of their two neighbors, and—when present—the external signal. Each agent has two neighbors and a self-loop. When no external source is present ($w_i^C = 0$), attention is divided among these three channels, so each receives a weight of $1/3$. When an external source is introduced ($w_i^C > 0$), there is an additional information channel, and attention is divided equally

across four channels. Thus every interpersonal link, the self loop, and the external source of information all receive a weight of 1/4.

In the parameterization used in the main text, where the external weight is denoted w, this corresponds to assigning $(1-w)/3$ to each interpersonal link and w to the external source, as shown in Figure 1.

Matrix Representations and Leontief Inverses The social-interaction and identity-interaction matrices can be explicitly written as follows. When no external source is present ($w = 0$):

$$\mathbf{W} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \tilde{\mathbf{W}} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The resulting matrix $\tilde{\mathbf{W}}$ is a signed matrix, with negative entries reflecting antagonistic interactions across groups, and it is structurally balanced in the sense that each agent's positive and negative ties are arranged in accordance with with group membership.

When an external source is present ($w = \frac{1}{4}$):

$$\mathbf{W} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \tilde{\mathbf{W}} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The corresponding Leontief inverses, $\mathbf{M} = (\mathbf{I} - \mathbf{W})^{-1}$ and $\tilde{\mathbf{M}} = (\mathbf{I} - \tilde{\mathbf{W}})^{-1}$, are:

$$\mathbf{M} = \begin{bmatrix} 1.8 & 0.7 & 0.3 & 0.2 & 0.3 & 0.7 \\ 0.7 & 1.8 & 0.7 & 0.3 & 0.2 & 0.3 \\ 0.3 & 0.7 & 1.8 & 0.7 & 0.3 & 0.2 \\ 0.2 & 0.3 & 0.7 & 1.8 & 0.7 & 0.3 \\ 0.3 & 0.2 & 0.3 & 0.7 & 1.8 & 0.7 \\ 0.7 & 0.3 & 0.2 & 0.3 & 0.7 & 1.8 \end{bmatrix}, \quad \tilde{\mathbf{M}} = \begin{bmatrix} 1.8 & 0.7 & 0.3 & -0.2 & -0.3 & -0.7 \\ 0.7 & 1.8 & 0.7 & -0.3 & -0.2 & -0.3 \\ 0.3 & 0.7 & 1.8 & -0.7 & -0.3 & -0.2 \\ -0.2 & -0.3 & -0.7 & 1.8 & 0.7 & 0.3 \\ -0.3 & -0.2 & -0.3 & 0.7 & 1.8 & 0.7 \\ -0.7 & -0.3 & -0.2 & 0.3 & 0.7 & 1.8 \end{bmatrix}.$$

Multiplying these matrices by the vector $(1/4) \cdot \mathbf{1}$ gives the Katz-Bonacich centralities $\boldsymbol{b}$ and $\tilde{\boldsymbol{b}}$.

Social and Identity Centralities Table C-1 reports Katz-Bonacich and eigenvector centralities for the ring network. Under the social-interaction matrix $\mathbf{W}$, all agents have

identical values because the network is perfectly symmetric: each agent interacts with two neighbors and themselves. This uniformity disappears when considering the identity-interaction matrix $\tilde{\mathbf{W}}$, which introduces negative links to capture out-group antagonism.

These negative links reduce the effective influence of agents exposed to the opposing group, creating variation in $\tilde{\boldsymbol{b}}$. Agents 2 and 5, connected only to in-group neighbors, have higher centralities, while agents 1, 3, 4, and 6—linked to out-group members—have lower values.

Eigenvector centralities $\tilde{\boldsymbol{\pi}}$ also reflect alignment with group identity, highlighting how identity-based interactions reshape both the distribution and interpretation of influence compared to standard social networks.

Nodes	$\boldsymbol{b}(\boldsymbol{W})$	$\boldsymbol{b}(\tilde{\boldsymbol{W}})$	$\boldsymbol{\pi}(\boldsymbol{W})$	$\boldsymbol{\pi}(\tilde{\boldsymbol{W}})$
1	1	0.4	0.16	-0.16
2	1	0.6	0.16	-0.16
3	1	0.4	0.16	-0.16
4	1	0.4	0.16	0.16
5	1	0.6	0.16	0.16
6	1	0.4	0.16	0.16

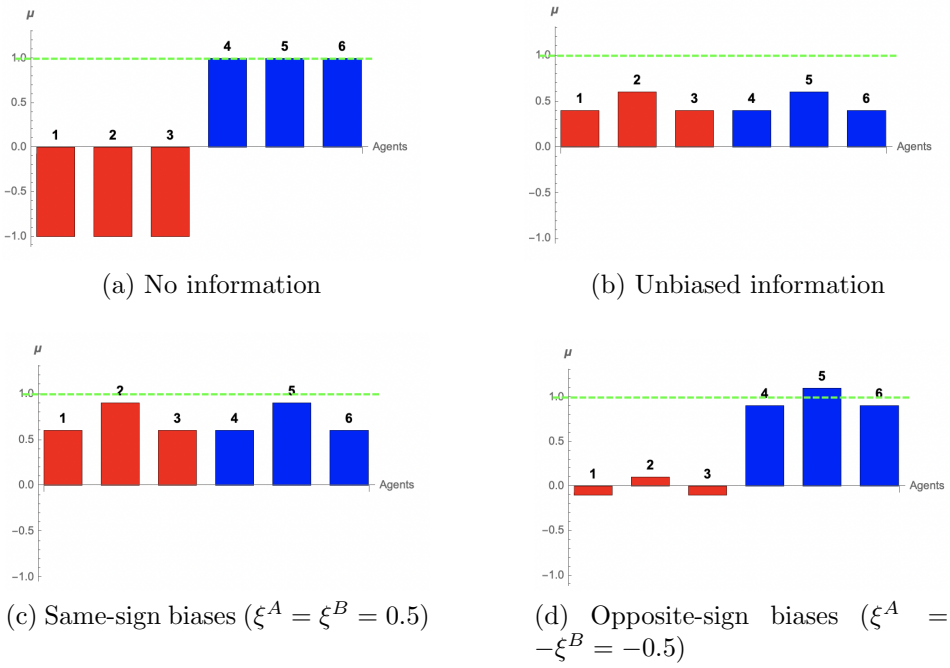
Table C-1: Katz-Bonacich and eigenvector centralities for both the social interaction matrix $\mathbf{W}$ and the (signed) identity-interaction matrix $\tilde{\mathbf{W}}$ of ring network of six agents of Figure C-1.

Long-Run Opinions The structure of the identity-interaction matrix $\tilde{\mathbf{W}}$, which incorporates antagonistic out-group links, generates heterogeneity in long-run opinions even in a symmetric social network.

In case (i), with no external information, each group converges to an internal consensus. The disagreement between groups reflects the combination of initial opinions and the structure of identity-interaction matrix, as encoded in $\tilde{\boldsymbol{\pi}}$.

In case (ii), with an unbiased external source, we have disagreement but no ideological polarization across groups. Agents less exposed to out-group interactions (2, 5) are closer to the truth, creating within-group disagreement. Because the social network is symmetric across groups, both groups display the same degree of internal disagreement and the same average opinion.

Case (iii) considers biased external sources. When biases align across groups ($\xi^A = \xi^B = 0.5$), long-run opinions converge closer to the truth than in (ii). When biases are opposed ($\xi^A = -\xi^B = -0.5$), out-group antagonism amplifies divergence and the degree of ideological polarization and disagreement is high.



(a) No information

(b) Unbiased information

(c) Same-sign biases ($\xi^A = \xi^B = 0.5$)

(d) Opposite-sign biases ($\xi^A = -\xi^B = -0.5$)

Figure C-1: Long-run opinions for the ring network of Figure 1, beginning from $(\mu_{i,0})_{i\in A} = 0$ and $(\mu_{i,0})_{i\in B} = 2$. Group A is depicted in Red and Group B is depicted in Blue. The green dotted line represents the true state of the world, $\theta^* = 1$.

References

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753.

Alatas, V., Chandrasekhar, A. G., Mobius, M., Olken, B. A., and Paladines, C. (2020). Designing effective celebrity public health messaging: Results from a nationwide Twitter experiment in Indonesia. Mimeo.

Ali, S. N., Mihm, M., and Siga, L. (2025). The political economy of zero-sum thinking. *Econometrica*, 93(1):41–70.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., and Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the USA*, 115(37):9216–9221.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. Wanted: The key player. *Econometrica*, 74(5):1403–1417.

Bergeron, A., Carvalho, J.-P., Henrich, J., Nunn, N., and Weigel, J. (2023). Zero-sum thinking, the evolution of effort-suppressing beliefs, and economic development. Mimeo.

Bloch, F., Demange, G., and Kranton, R. (2018). Rumors and social networks. *International Economic Review*, 59(2):421–448.

Bonacich, P. and Lloyd, P. (2004). Calculating status with negative relations. *Social Networks*, 26(4):331–338.

Boucher, V., Rendall, M., Ushchev, P., and Zenou, Y. (2024). Toward a general theory of peer effects. *Econometrica*, 92(2):543–565.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2024). Cross-country trends in affective polarization. *Review of Economics and Statistics*, 106(2):557–565.

Bramoullé, Y., Kranton, R., and D'amours, M. (2014). Strategic interaction and networks. *American Economic Review*, 104(3):898–930.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., and Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10):1378–1384.

Buechel, B., Hellmann, T., and Klößner, S. (2015). Opinion dynamics and wisdom under conformity. *Journal of Economics and Dynamics Control*, 52:240–257.

Callander, S. and Carbajal, J. C. (2022). Cause and effect in political polarization: A dynamic analysis. *Journal of Political Economy*, 130(4):825–880.

Campbell, A., Leister, M., Ushchev, P., and Zenou, Y. (2025). The polarization paradox: Why more connections can divide us. CEPR Discussion Paper No. 20729 .

Chinoy, S., Nunn, N., Sequeira, S., and Stantcheva, S. (2025). Zero-sum thinking and the roots of US political differences. *American Economic Review,* forthcoming.

Dasaratha, K., Golub, B., and Hak, N. (2023). Learning from neighbors about a changing state. *Review of Economic Studies*, 90(5):2326–2369.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Della Lena, S. (2024). The spread of misinformation in networks with individual and social learning. *European Economic Review*, 168:104804.

DeMarzo, P. M., Vayanos, D., and Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *Quarterly Journal of Economics*, 118(3):909–968.

Djourelova, M., Durante, R., Motte, E., and Patacchini, E. (2024). Media slant and public policy views. *American Economic Association Papers and Proceedings*, 114:684–689.

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., and Ryan, J. B. (2021). Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour*, 5:28–38.

Egorov, G., Guriev, S., Mironov, M., and Zhuravskaya, E. (2025). Political information and network effects. *Journal of the European Economic Association*, page jvaf052.

Ellen, P. S., Wiener, J. L., and Cobb-Walgren, C. (1991). The role of perceived consumer effectiveness in motivating environmentally conscious behaviors. *Journal of Public Policy and Marketing*, 10(2):102–117.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Friedkin, N. E. and Johnsen, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206.

Gabielkov, M., Rao, A., and Legout, A. (2014). Studying social networks at scale: Macroscopic anatomy of the Twitter social graph. In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, pages 277–288.

Gavrilets, S. and Seabright, P. (2025). The evolution of zero-sum and positive-sum worldviews. *Proceedings of the National Academy of Sciences of the USA*, 122(32):e2504339122.

Golub, B. and Jackson, M. O. (2010). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49.

Golub, B. and Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *Quarterly Journal of Economics*, 127(3):1287–1338.

Grabisch, M., Poindron, A., and Rusinowska, A. (2019). A model of anonymous influence with anti-conformist agents. *Journal of Economics and Dynamics Control*, 109:103773.

Grosfeld, I., Rodnyansky, A., and Zhuravskaya, E. (2013). Persistent antimarket culture: A legacy of the pale of settlement after the Holocaust. *American Economic Journal: Economic Policy*, 5(3):189–226.

Grossman, G., Kim, S., Rexer, J., and Thirumurthy, H. (2020). Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proceedings of the National Academy of Sciences of the USA*, 117:24144–24153.

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. (2004). Propagation of trust and distrust. *Proceedings of the 13th International Conference on World Wide Web*, May:403–412.

Guilbeault, D., Becker, J., and Centola, D. (2018). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences of the USA*, 115(39):9714–9719.

Haghtalab, N., Jackson, M. O., and Procaccia, A. D. (2021). Belief polarization in a complex world: A learning theory perspective. *Proceedings of the National Academy of Sciences of the USA*, 118(19):e2010144118.

Harary, F. (1953). On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143–146.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22:129–146.

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

Iyengar, S. and Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707.

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton University Press, Princeton.

Jackson, M. O. and Zenou, Y. (2015). Games on networks. In: P. Young and S. Zamir (Eds.), *Handbook of Game Theory Vol. 4*, Amsterdam: Elsevier Publisher, pp. 91–157.

Jadbabaie, A., Molavi, P., Sandroni, A., and Tahbaz-Salehi, A. (2012). Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210–225.

Jenke, L. (2024). Affective polarization and misinformation belief. *Political Behavior*, 46:825–884.

Johansson, K. (2016). Understanding recycling behavior: A study of motivational factors behind waste recycling. *WIT Transactions on Ecology & the Environment*, 202:401–414.

Jungkunz, S. (2021). Political polarization during the COVID-19 pandemic. *Frontiers in Political Science*, 3:622512.

Kidwell, B., Farmer, A., and Hardesty, D. M. (2013). Getting liberals and conservatives to go green: Political ideology and congruent appeals. *Journal of Consumer Research*, 40(2):350–367.

Kinateder, M. and Merlino, L. P. (2017). Public goods in endogenous networks. *American Economic Journal: Microeconomics*, 9(3):187–212.

Kinateder, M. and Merlino, L. P. (2022). Local public goods with weighted link formation. *Games and Economic Behavior*, 132:316–327.

Lerman, K., Feldman, D., He, Z., and Rao, A. (2024). Affective polarization and dynamics of information spread in online networks. *Nature NPJ Complexity*, 1(1):8.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–870.

Merlino, L. P., Pin, P., and Tabasso, N. (2023). Debunking rumors in networks. *American Economic Journal: Microeconomics*, 15(1):467–496.

Molavi, P., Tahbaz-Salehi, A., and Jadbabaie, A. (2018). A theory of non-Bayesian social learning. *Econometrica*, 86(2):445–490.

Motta, M., Stecula, D., and Farhart, C. (2020). How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US. *Canadian Journal of Political Science*, 53(2):335–342.

Schneider-Strawczynski, S. and Valette, J. (2025). Media coverage of immigration and the polarization of attitudes. *American Economic Journal: Applied Economics*, 17(1):337–368.

Sethi, R. and Yildiz, M. (2016). Communication with unknown perspectives. *Econometrica*, 84(6):2029–2069.

Shayo, M. (2020). Social identity and economic policy. *Annual Review of Economics*, 12:355–389.

Shi, G., Altafini, C., and Baras, J. S. (2019). Dynamics over signed networks. *SIAM Review: Society for Industrial and Applied Mathematics*, 61(2):229–257.

Simchon, A., Brady, W. J., and Van Bavel, J. J. (2022). Troll and divide: The language of online polarization. *PNAS Nexus*, 1(1):pgac019.

Strydhorst, N., Morales-Riech, J., and Landrum, A. R. (2023). Exploring partisans' biased and unreliable media consumption and their misinformed health-related beliefs. *Harvard Kennedy School Misinformation Review*.

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. In Austin, W. G. and Worchel, S., editors, *The Social Psychology of Intergroup Relations*, pages 33–37. Brooks/Cole, Monterey, CA.

The Economist (2024). When politics is about hating the other side, democracy suffers.

Ushchev, P. and Zenou, Y. (2020). Social norms in networks. *Journal of Economic Theory*, 185:104969.

Zhang, B., Cao, Z., Qin, C.-Z., and Yang, X. (2018). Fashion and homophily. *Operations Research*, 66(6):1486–1497.

Zhuravskaya, E., Petrova, M., and Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, 12(1):415–438.

Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., and Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLOS One*, 12(7):e0181821.