Department of Economics- FEA/USP

Confirmation Bias in Social Networks

MARCOS ROSS FERNANDES

WORKING PAPER SERIES Nº 2023-02

DEPARTMENT OF ECONOMICS, FEA-USP
WORKING PAPER Nº 2023-02

Confirmation Bias in Social Networks

Marcos Ross Fernandes (marcosross@usp.br)

Abstract:

In this study, I propose a theoretical social learning model to investigate how confirmation bias affects opinions when agents exchange information over a social network. Hence, besides exchanging opinions with friends, agents observe a public sequence of potentially ambiguous signals and interpret it according to a rule that includes confirmation bias. First, this study shows that regardless of level of ambiguity both for people or networked society, only two types of opinions can be formed, and both are biased. However, one opinion type is less biased than the other depending on the state of the world. The size of both biases depends on the ambiguity level and relative magnitude of the state and confirmation biases. Hence, long-run learning is not attained even when people impartially interpret ambiguity. Finally, analytically confirming the probability of emergence of the less-biased consensus when people are connected and have different priors is difficult. Hence, I used simulations to analyze its determinants and found three main results: i) some network topologies are more conducive to consensus efficiency, ii) some degree of partisanship enhances consensus efficiency even under confirmation bias and iii) open-mindedness (i.e. when partisans agree to exchange opinions with opposing partisans) might inhibit efficiency in some cases.

Keywords: Social Networks, Social Learning, Misinformation, Confirmation Bias

JEL Codes: C11, D83, D85

CONFIRMATION BIAS IN SOCIAL NETWORKS

MARCOS R. FERNANDES[1]

This version: January 2023

Abstract. In this study, I propose a theoretical social learning model to investigate how confirmation bias affects opinions when agents exchange information over a social network. Hence, besides exchanging opinions with friends, agents observe a public sequence of potentially ambiguous signals and interpret it according to a rule that includes confirmation bias. First, this study shows that regardless of level of ambiguity both for people or networked society, only two types of opinions can be formed, and both are biased. However, one opinion type is less biased than the other depending on the state of the world. The size of both biases depends on the ambiguity level and relative magnitude of the state and confirmation biases. Hence, long-run learning is not attained even when people impartially interpret ambiguity. Finally, analytically confirming the probability of emergence of the less-biased consensus when people are connected and have different priors is difficult. Hence, I used simulations to analyze its determinants and found three main results: i) some network topologies are more conducive to consensus efficiency, ii) some degree of partisanship enhances consensus efficiency even under confirmation bias and iii) open-mindedness (i.e. when partisans agree to exchange opinions with opposing partisans) might inhibit efficiency in some cases.

JEL Classification: C11, D83, D85

Keywords: Social Networks, Social Learning, Misinformation, Confirmation Bias.

[1] University of São Paulo, Department of Economics, e-mail: mrf.ross@gmail.com.

I thank Sandro Brusco, Pradeep Dubey, Ben Golub, Laura Karpuska, Ehud Lehrer, Ting Liu, Mihai Manea, Alejandro Melo, Ron Peretz, Max Silva, Michael Sinkey, Eilon Solan, Troy Tassier and Yair Tauman for helpful comments and suggestions. I also thank all participants of the 29th International Conference on Game Theory (Stony Brook), 23rd Coalition Theory Network workshop (Maastricht), 4th Annual Conference on Network Science and Economics (Nashville), 44th Eastern Economic Association conference (Boston) and the seminar participants at UFABC, Fordham University, Stony Brook Center for Behavioral Political Economy, INSPER and University of São Paulo. I gratefully acknowledge the Department of Economics of Stony Brook University (Juan Carlos Conesa), Stony Brook Graduate Students Organization and The Network Science in Economics conference series (Myrna Wooders) for financial support.

1. Introduction

People form opinions on various economic, political, social, and health issues based on information from both the media and people they trust (e.g. friends, coworkers, family, experts, etc). This information acquisition process usually occurs when the issue discussed has no clear-cut *right/wrong* or *true/false* distinction or when the available information cannot be easily understood. Consulting people's opinions, in this case, is an appealing and easy way to gather information. For many people, social networks then become primary tool to stay informed. Thus, understanding how beliefs depend on how agents perceive and process information is vital. In this study, I examine how opinions are affected by confirmation bias in a networked environment.

In psychology, confirmation bias denotes the interpretation of evidence in ways consistent with existing beliefs (Nickerson (1998), Molden and Higgins (2008)). This can be done in different ways, like restricting attention to favored hypothesis, disregarding evidence that could falsify the current worldview or overvaluing positive confirmatory instances. In all cases, people restrict attention to a single hypothesis and fail to carefully consider alternatives.

In social psychology, people interpret evidence when they are ambiguous (i.e. when evidence is conflicting). People may misinterpret scientists and experts after ambiguous announcements. Simonovic and Taber (2022) highlighted that when WHO declared the COVID-19 outbreak a global pandemic in 2020, experts did not precisely understand the extent and nature of the health risks or how disease transmission can be prevented. Hence, WHO provided *conflicting recommendations* to the public on whether wearing a mask was necessary. Other medical authorities also provided conflicting recommendations to the public regarding medicines and vaccines' efficacy. Conflicting evidences may have even contributed to people making their own assessment about the problem.

While friends may help people to aggregate information in some cases, in other cases, people may expose themselves to others who that rely on their own worldview to derive information from ambiguous evidence. In these cases, efficient aggregation of information is not guaranteed, and I investigate how opinions are influenced by people's biases.

To analyze this phenomenon, I consider a society where agents are interested to learn the underlying state $\theta \in \Theta = [0, 1]$. For instance, the underlying state θ might represent the efficacy of a new vaccine (e.g. from 0 to 1). All agents have prior beliefs about the vaccine's efficacy and observe a sequence of public signals, one at each date t. Public signals may be (i) informative or (ii) ambiguous. Informative signals are binary variables indicated as 1 if state on the right side of the 0-1 spectrum are more likely (i.e. if vaccine's efficacy is high) and 0 if the states on the left side of the 0-1 spectrum are more likely (i.e. if vaccine's efficacy is low). Hence, as signals realization does not convey full information on the underlying state, agents can only learn the true state (vaccine's efficacy) asymptotically. This is in the spirit of ongoing learning, where information accumulates through experience. In the case of ambiguous signals, agents are allowed to interpret these signals using a fairly general randomization rule proposed by Fryer Jr, Harms, and Jackson (2019) that accounts for confirmation bias. Hence, the interpretation of the ambiguous signal received at time t is influenced, to a greater or lesser extent, by the likelihood of 0 and 1 at time $t-1$ (see more details below). This captures situations wherein people feel impelled to explain ambiguous evidence about a particular issue.

As in Jadbabaie, Molavi, Sandroni, and Tahbaz-Salehi (2012), besides learning from public signals, agents exchange information through a social network. At the beginning of every period t, the public signal is realized. Thus, each agent first *interprets* signals (if ambiguous) using the randomization rule, *stores* the signal and *computes* the Bayesian posterior (opinion and precision). Every agent then sets their *final* opinions and precisions to be a linear combination of the Bayesian posterior opinions and precisions computed with the interpreted signal and opinions and precisions of friends (e.g., formal definition of neighbors in subsection 3.1) they met in the period before. Social connectivity among agents remains fixed over time and strong connectivity is assumed (i.e., all agents are exposed to all other agents either through a directed or undirected path in the social networks).

Hence, despite the level of ambiguity and both in the case of a single individual or a connected society, only two types of opinions can emerge, and both are biased: left- and right-biased opinions. However, one type of opinion is less biased than the other depending on the underlying state. Less-biased opinion is only guaranteed to emerge under a favorable combination of sufficiently low ambiguity and sufficiently pronounced states. If this condition holds, I show that the less-biased opinion is attained with probability 1. Moreover, long-run learning is not attained even if people are impartial when they interpret ambiguous signals (i.e., when interpreting evidence uniformly at random instead of using their own opinions). Those results contrast with those by Rabin and Schrag (1999) and Fryer Jr et al. (2019), who suggest that long-run learning occurs with a positive probability and that impartiality helps in learning the state. Furthermore, both the network effect presented here and signals realization, reinforce the interpreting dispute (*tug-of-war*) as people may have their own interpretation biases reinforced or attenuated by other agents.

Finally, confirming the probability of emergence of the less-biased consensus analytically is difficult, and I use Monte Carlo simulations to show its determinants. The presence of partisan agents (i.e., agents with skewed initial priors) in societies suffering from confirmatory bias have two main effects. (i) When the degree of partisanship is low, partisanship helps to counter the realization of initial misleading signals (e.g., realization of a 0 when $\theta \geq 0.5$). Thus, low partisanship increases the odds of reaching less-biased consensus. (ii) When the degree of partisanship is high, partisans exacerbate misinterpretation of signals. Thus, high partisanship reduces the odds of reaching less-biased consensus. Moreover, I also show that open-mindedness of partisan agents (i.e., when partisans agree to exchange opinions with partisans with polar opposite beliefs) might reduce the odds of reaching less-biased consensus in some network structures.

While this work does not generalize theoretical results for other conjugate families and numerical results for other network structures, both methods and cases explored are sufficiently general to capture important aspects of real-world networks. In every period, public signals realized and observed by all agents may represent information reported by sources including media outlets and international organizations. The level of ambiguity of the informational content reported by them, measured by a parameter $\mu \in (0, 1)$, represents the fraction of instances where a signal simultaneously conveys two conflicting meanings and agents feel impelled to interpret them. Parameters of the signal interpretation function dictate the interpretation behavior of every agent.

This work is structured as follows. Section 2 provides a brief literature review and highlights contributions. Section 3 describes a framework for updating beliefs when agents communicate over social networks with ambiguous signals and present main theoretical results. Section 4 describes a simulation exercise when priors heterogeneity (partisanship) is assumed. Section 5 concludes the study. Moreover, six appendices are available. Appendices A and B contain primitives of the Beta-Bernoulli conjugate family employed in this work. Appendix C contains proofs of auxiliary results, while Appendix D presents proofs of main results. Appendices E and F show simulation statistics and present regression robustness.

2. Literature review and contribution

Considerable empirical evidence on social psychology supports the idea that confirmation bias is extensive and appears in many ways. Most studies in the field confirm the human tendency of casting doubt on information that conflicts with preexisting beliefs and confirming preexisting beliefs when exposed to ambiguous information (see Nickerson (1998)). However, this selectivity in the acquisition and use of evidence occurs without intending to treat evidence in a biased way. Molden and Higgins (2004, 2008) note that both *vagueness* (when evidence is weak) and *ambiguity* (when evidence is conflicting and open to interpretations) induce interpretation. Conversely, Furnham and Ribchester (1995) and Furnham and Marks (2013) review literature on the subject and report evidence that the way people perceive and process information about ambiguous situations is related to their degree of ambiguity tolerance (i.e., individual differences in cognitive reaction to stimuli considered ambiguous). Therefore, ambiguity tolerance refers to underlying psychological differences that impel people to process, interpret, and react differently to ambiguous information.[1]

Thus, confirmation bias may oppose standard Bayesian updating processes as agents scrutinize signals in line with their worldviews. Some examples of decision-making models that account for Bayesian updating deviation are Hellman and Cover (1970), Rabin and Schrag (1999), Wilson (2014), Fryer Jr et al. (2019), Sikder, Smith, Vivo, and Livan (2020) and Buechel, Klößner, Meng, and Nassar (2022).

Studies by Rabin and Schrag (1999), Fryer Jr et al. (2019), Sikder et al. (2020) and Buechel et al. (2022) are the closest references to this work, in both spirit and results. Rabin and Schrag (1999), showed that signals believed as less likely are misinterpreted with an exogenous probability. Fryer Jr et al. (2019) stated that ambiguous signals, in its simplest version with binary states, are produced with a certain probability, and agents interpret those before conducting the Bayesian update. Interpreting these signals requires agents to use three methods that differ in intensity with which agents conform their interpretation with their current worldview. However, both works do not consider network communication.

Sikder et al. (2020) employ a slightly modified version of Rabin and Schrag (1999) to a networked environment (mostly focused on regular networks), where agents synchronously share the full set of signals with their neighbors. However, biased agents reject information incongruent with their preexisting beliefs, reduce the weight they place on other agents, and place the remaining weight on an external positively oriented "ghost" node, creating a polarization of unbiased agents in the steady state. In this study, I assume a general (connected) network structure among agents and allow them to set their final beliefs to be a linear combination of the Bayesian posterior and opinions of their neighbors as in Jadbabaie et al. (2012), regardless of their biases and signals received. A key difference relative to Sikder et al. (2020) is that my modeling strategy allows me to discuss the relative importance of the learning parameters, network structure, and connections heterophily (open-mindedness of partisans) in determining the probability of reaching the less-biased consensus.

Buechel et al. (2022) allow different types of signals to have different transmission capacities (i.e., asymmetric decay factor applied to positive and negative signals) when signals are shared in different networks, and show that for a society to aggregate information efficiently, different asymmetries must be balanced and that an agent's ratio of centralities between the two networks must be moderate compared to the ratio of centrality concentration in the two networks. In my

[1] More recently, scholars have considered the concept of tolerance of ambiguity as a reflection of the contemporary definition of ambiguity proposed by Ellsberg (1961). For a good coverage of the classic literature on ambiguity aversion, see Gilboa and Schmeidler (1989), Gilboa and Schmeidler (1993), and Epstein and Schneider (2007).

model, sharing asymmetries and misinformation is not considered even when partisan agents remain equally balanced and central. However, this work shares an interesting feature with my model relative to how equality of centralities is critical for reducing misinformation.

This work is also related to the literature of *bounded confidence* in networks. Overall, this literature focuses on models of social learning wherein agents overvalue the opinion of friends with similar beliefs. Hegselmann, Krause et al. (2002), Hegselmann and Krause (2005), Dandekar, Goel, and Lee (2013), Mao, Bolouki, and Akyol (2018), and Gallo and Langtry (2020) provide examples of this phenomenon. While bounded confidence involves the tendency to *conform* with the majority or leading people, confirmation bias is a failure in the Bayesian updating process. From this perspective, modeling confirmatory bias as either a bounded confidence or a failure in the Bayesian update has different consequences. On the one hand, bounded confidence presumes that the connections between agents are broken (or temporarily interrupted) according to opinions distance. Hence, nontrivial changes are implied in the network topology. In this literature, long-run polarization naturally occurs under bounded confidence. Polarization, hence, is a natural product of the initial heterophily of opinions in the system and eventual deletion of links. On the other hand, modeling confirmatory bias as a Bayesian update failure is inconsequential to the network topology and under the strong connectivity assumption leads to a bias (misinformation) that can be analytically studied.

Finally, numerous works on *social learning* assumed bounded and full rationality. Bayesian social learning literature (fully rational agents) mainly focuses on formulating stylized games with incomplete information and characterizing its equilibria. Specifically, rather than considering complex and repeated interactions, most works focus on environments where agents are myopic or interact only once (Banerjee (1992), Bala and Goyal (1998), Bala and Goyal (2001), Banerjee and Fudenberg (2004), Acemoglu, Dahleh, Lobel, and Ozdaglar (2011)).[2]

Non-bayesian learning (bounded rational agents) literature focuses on studying generalizations of the seminal DeGroot (1974) model. DeMarzo, Vayanos, and Zwiebel (2003) show that consensus result does not rely on the social weighting matrix being a stationary matrix. Acemoglu, Ozdaglar, and ParandehGheibi (2010) consider a random meeting (Poisson) model and characterize how the presence of forceful agents (i.e., agents who influence others disproportionately and hardly revise their beliefs) prevents information aggregation. Conversely, Golub and Jackson (2010) show that convergence holds if (and only if) the influence of the most influential agent vanishes as society grows unboundedly. Jadbabaie et al. (2012) is the first study to consider the possibility

[2] For an overview of recent research on learning in social networks, see Acemoglu and Ozdaglar (2011); Golub and Sadler (2017); Grabisch and Rusinowska (2020).

of constant arrival of informative signals in every period in networked environments. In their study, the update rule that sets the final belief as a linear combination of the Bayesian posterior and the neighbors' opinions is an efficient alternative to the complicated task of implementing Bayesian update in networks. Finally, similar to this work in modeling strategy, Azzimonti and Fernandes (2022) investigate how the structure of social networks and the presence of fake news affect the degree of polarization and misinformation. In their study, i) their model considers the presence of *bots* whose sole purpose is to deceive other agents, and that ii) connectivity among all agents evolves stochastically. Those two features combined are main drivers of misinformation and polarization cycles. Herein, the main source of bias derives from confirmation bias and connectivity among agents is assumed fixed. Thus, this study focuses on understanding how misinformation depends on both network structure and how agents interpret ambiguous signals.

3. The model

Notation: All vectors are considered column vectors, unless stated otherwise. Given a vector $v \in \mathbb{R}^n$, I denote by v_i its i-th entry. When $v_i \geq 0$ for all entries, I write $v \geq 0$. Moreover, I define $v^\top$ as the transpose of the vector x. Hence, the inner (scalar) product of two vectors $x, y \in \mathbb{R}^n$ is denoted by $x^\top y$. I denote by $\mathbf{1}$ the vector with all entries equal to 1. A matrix W is considered to have size $m \times n$ whenever W has exactly m rows and n columns. Moreover, whenever $m = n$, W is called a square matrix of size n. The identity matrix of size n is denoted by $\mathbb{I}$. For a matrix W, W_{ij} denotes the entry in the i-th row and j-th column. The notation W^k_{ij} is used to denote the entry in the i-th row and j-th column of the matrix W^k, i.e. the matrix W raised to the power k. Finally, a vector v is said to be a stochastic vector when $v \geq 0$ and $\sum_i v_i = 1$. A square matrix W is said to be a (row) stochastic matrix when each row of W is a stochastic vector.

3.1. **Network structure.** Connectivity among agents in a network is described by a directed graph $G = (N, g)$, where $N = \{1, 2, \dots, n\}$ is the set of agents, fixed over time, and g is a binary $n \times n$ adjacency (or incidence) matrix, also fixed over time. Each element g_{ij} in the directed-graph represents the connection between agents i and j. More precisely, $g_{ij} = 1$ if individual i is paying attention to (i.e. receiving information from) individual j, and 0 if otherwise. As the graph is directed, some agents pay attention to others who are not necessarily reciprocating (i.e., $g_{ij} \neq g_{ji}$). The out-neighborhood of any agent i is the set of agents that i is receiving information from, and is denoted by $N_i^{out} = \{j \,|\, g_{ij} = 1\}$. Similarly, the in-neighborhood of any agent i is denoted by $N_i^{in} = \{j \,|\, g_{ji} = 1\}$, represents the set of agents that are receiving information from i. In undirected networks, $N_i^{in} = N_i^{out} = d_i$, where d_i is the number of neighbors agent i has, also known as degree centrality of agent i. Thus, the term $\hat{g}_{ij} = \frac{g_{ij}}{|N_i^{out}|} \in [0, 1]$ represents the weight that agent i gives to information received from their out-neighbor j. A network is considered *regular* if every node has the same degree of centrality, and that a network is *complete* if every node is connected with all other nodes. Finally, a directed path in G from agent i to agent j is defined as a sequence of agents beginning with i and ending with j such that each agent is a neighbor of the next agent in the sequence. A social network is *strongly connected* if a directed path from each agent to any other agent exists.

3.2. **Signals, initial beliefs and opinions.**

Let $\Theta = [0, 1]$ to denote the set of possible states of the world. For instance, one may find useful to interpret Θ as the effectiveness of a new vaccine, such that a state close to 0 means that the vaccine has low efficacy, whereas a state close to 1 means that vaccine has high efficacy.

Conditional on the state of the world θ, every agent observes a sequence of public signals s_t, one at each date $t \in \{1, 2, \ldots\}$. Public signals lie in the set $S = \{1, 0, a\}$. Considering the example of the vaccine's efficacy given above, a signal 1 is evidence that the new vaccine can prevent people from severe illness, a signal 0 is evidence of no efficacy, and a signal a is *ambiguous* and open to idiosyncratic interpretation (Section 3.3 explains how agents deal with those signals). Signals are independent over time, conditional on the state. The probability that a signal is ambiguous is $\mu \in (0, 1)$. Hence, the signal conveys informational aspects that could lead one to interpret as either 1 or 0. With the remaining probability $(1 - \mu)$, the information provided by the signal is unambiguous. In any state $\theta \in \Theta$, the probability that an unambiguous signal is 1 is $\theta \in [0, 1]$ and 0 with probability $1 - \theta$. The signal structure is depicted in the Figure 1.

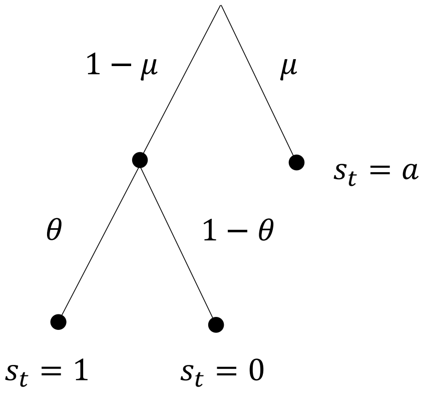


FIGURE 1. Signals structure

Each agent i in this society is assumed to start with an *initial belief* about an underlying state $f_{i,0}(\theta) \in \Delta\Theta$, represented by a Beta probability distribution over the set Θ with shape parameters $\alpha_{i,0}, \beta_{i,0} \geq 1$ and defined as follows:

$$f_{i,0}\left(\theta\right)=\begin{cases}\dfrac{\Gamma\left(\alpha_{i,0}+\beta_{i,0}\right)}{\Gamma\left(\alpha_{i,0}\right)\Gamma\left(\beta_{i,0}\right)}\,\theta^{\alpha_{i,0}-1}(1-\theta)^{\beta_{i,0}-1} & \text{, for } 0<\theta<1\\ 0 & \text{, otherwise,}\end{cases} \tag{1}$$

where $\Gamma(\cdot)$ is a Gamma function and the ratio of Gamma functions in the expression above is a normalization constant that ensures that the total probability integrates to 1.

Given prior beliefs and signals, *opinion* of agent i at time t is denoted by

$$y_{i,t}=\mathbb{E}\left[\theta|s_i^t\right]=\frac{\alpha_{i,t}}{\alpha_{i,t}+\beta_{i,t}},$$

where s_i^t is the history of signals received and interpreted by agent i up until time t.[3]

3.3. **Interpretation of ambiguous signals.** Although ambiguous signals are uninformative about the state and should be disregarded from a pure Bayesian perspective, agents are constrained to interpret ambiguous signals. This constraint captures the idea that, in some instances, people react to ambiguous pieces of information. They fail to perceive the lack of informational content of signals and end up using their prior worldview to derive meaning from them.

For the interpretation of ambiguous signals, I use a randomization rule proposed in Fryer Jr et al. (2019), adapted here for some technical idiosyncrasies. Hence, with probability $\gamma_i\in[\frac{1}{2},1]$ agent i conforms with his posterior at time $t-1$ and with probability $1-\gamma_i$ goes against it. Essentially, with probability

$$\psi_{i,t}=\gamma_i\,\mathbb{1}\{y_{i,t-1}\geq 0.5\}+(1-\gamma_i)\,\mathbb{1}\{y_{i,t-1}<0.5\} \tag{2}$$

agent i interprets the ambiguous signals as 1 and with the remaining probability $(1-\psi_{i,t})$ interprets the ambiguous signals as 0 at time t.[4]

Therefore, parameter γ_i represents the intensity of the confirmatory bias of an individual i. I only assume γ_i to be independent of opinion $y_{i,t}$ for any $i\in N$, history of opinions of all agents, and of all other parameters in this model. From this randomization rule, three cases of interest are available.

Definition 1. *An individual $i\in N$*

(1) *is **impartial** if $\gamma_i=\frac{1}{2}$,*

[3] Appendix A discusses the primitives of the Beta distribution and the Beta-Bernoulli conjugate family. For tractability, the opinion is intended as a real number that summarizes the entire belief. Hence, one can understand the opinion of an agent as the Bayesian estimator of θ that minimizes the mean squared error. One could also assume that the opinion of any agent i at time t could also be the Bayesian estimator of θ which minimizes the absolute error. As the mean, mode, and median of the Beta distribution are asymptotically equivalent, the functional form is irrelevant for the results.

[4] From Appendix B, note that, since mean and mode of the Beta distribution are very close for different choices of (α,β) and are asymptotically equivalent, using $y_{i,t-1}$ (the mean and note mode) to interpret public signals in Equation (2) is neutral to all results.

(2) *has* ***confirmatory tendency*** *if* $\frac{1}{2} < \gamma_i < 1$,

(3) *is* ***fully biased*** *(biased for short) if* $\gamma_i = 1$.

Hence, the signal interpretation functions, $s_t^{(0)}$ and $s_t^{(1)}$, for each individual at any point in time can be generally defined as follows:

$$s_{i,t}^{(0)} = \mathbb{1}\{s_t = 0\} + \mathbb{1}\{s_t = a\}\mathbb{1}\{u_t > \psi_{i,t}\} \tag{3}$$

$$s_{i,t}^{(1)} = \mathbb{1}\{s_t = 1\} + \mathbb{1}\{s_t = a\}\mathbb{1}\{u_t \leq \psi_{i,t}\}, \tag{4}$$

where $\psi_{i,t}$ is as defined in Equation (2), s_t is the publicly observed signal and u_t is the realization of a continuous $U\,[0,1]$ random variable at time t simply used to break the tie. Draws $\{u_t\}$ are independent across time and also independent of all other random variables in this model. Hence, the signal interpretation functions are basically transforming observed signals $\{s_t\}_{t=1}^{\infty}$ into binary interpretations. When the realized public signal is $s_t = 1$ ($s_t = 0$), all agents undoubtedly interpret it as 1 (as 0) and set $s_t^{(0)} = 0$ and $s_t^{(1)} = 1$ (set $s_t^{(0)} = 1$ and $s_t^{(1)} = 0$). However, when the realized public signal is ambiguous (i.e., $s_t = a$), agents use their prior information (summarized by $y_{i,t-1}$) to categorize the signal as either 0 or 1, as per Equation (2). Figure 2 shows a more detailed description of the signals interpretation scheme. Appendix A shows details on the signals likelihood function.

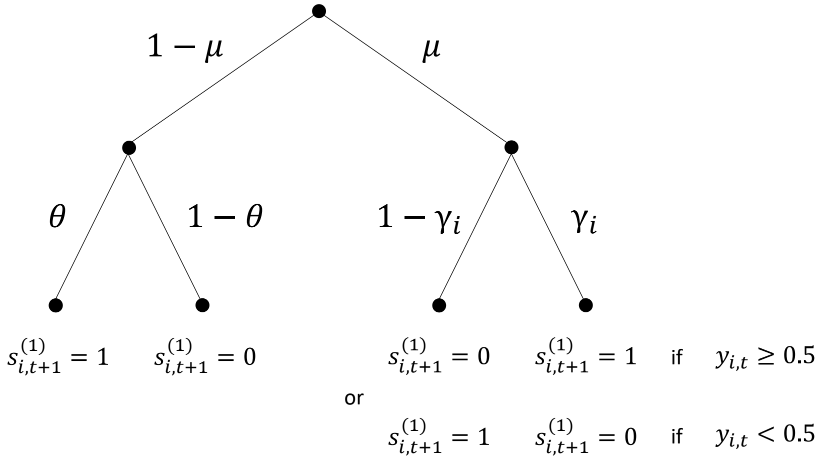


FIGURE 2. Signals interpretation by agent i upon receiving a public signal s_{t+1}

3.4. **Belief evolution.** Agents are assumed to update their beliefs based on public signals $s_t \in S = \{1, 0, a\}$ and on the influence of friends in their social network.

Hence, at the beginning of period t, a public signal is realized and signal s_t is observed by agent i. After observing the public signal s_t, agent i computes his posterior in a standard Bayesian fashion. Following Jadbabaie et al. (2012), I assume that the updated parameters α and β will be a convex combination between the parameters α and β of his Bayesian posterior and the weighted average of his neighbors' parameters.[5]

In mathematical terms, the update rule is as follows

$$\alpha_{i,t+1} = b\left[\alpha_{i,t} + s^{(1)}_{i,t+1}\right] + (1-b)\sum_j \hat{g}_{ij}\alpha_{j,t} \tag{5}$$

$$\beta_{i,t+1} = b\left[\beta_{i,t} + s^{(0)}_{i,t+1}\right] + (1-b)\sum_j \hat{g}_{ij}\beta_{j,t}, \tag{6}$$

where $b \in [0,1]$.

Notice that when $b = 1$, agents fully rely on the signals and behave like standard Bayesian agents. As b approaches zero, agents are more influenced by the network as more weight is given to their neighbors' opinions. Moreover, let $\alpha_t = (\alpha_{1,t}, \alpha_{2,t}, \dots, \alpha_{n,t})^\top$ and $\beta_t = (\beta_{1,t}, \beta_{2,t}, \dots, \beta_{n,t})^\top$ denote the column vectors of length n of agents beliefs parameters at time t, $\mathbb{I}$ be an identity matrix of dimension n and $B = \text{diag}(b, b, \dots, b)$ be the diagonal Bayesian (or self-reliance) matrix. We can rewrite Equation (5) as follows

$$\begin{aligned}\alpha_{t+1} &= B(\alpha_t + s^{(1)}_{t+1}) + (\mathbb{I} - B)\hat{g}\alpha_t \\ &= \left(B + (\mathbb{I} - B)\hat{g}\right)\alpha_t + Bs^{(1)}_{t+1} \\ &= W\alpha_t + Bs^{(1)}_{t+1}, \end{aligned} \tag{7}$$

and equation (6) as follows

$$\beta_{t+1} = W\beta_t + Bs^{(0)}_{t+1}, \tag{8}$$

where, $W = B + (\mathbb{I} - B)\hat{g}$ is a homogeneous row-stochastic matrix. Notice that as the graph G induced by the adjacency matrix g is assumed to be strongly connected, the graph induced by W is trivially strongly connected as well.

4. Theoretical results

4.1. **Single individual case.** Before illustrating the network effects over the opinions when agents are exposed to ambiguous signals, we first focus on explaining what happens in the case of a single individual. In this regard, the following result shows that only two types of opinions may emerge when an agent interprets ambiguity under confirmatory bias.

[5] One may alternatively interpret agents to share opinions (mean) and precisions (variance) with each other rather than sharing distribution parameters. Those are equivalent modeling strategies, and we only need to use the relationships $y = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ to fully determine α and β. Algebraic manipulation yields $\alpha = -\frac{y(\sigma^2+y^2-y)}{\sigma^2}$ and $\beta = \frac{(\sigma^2+y^2-y)(y-1)}{\sigma^2}$.

Proposition 1 (characterization). *If a single individual i randomizes interpretation of ambiguous signals according to Equation* (2)*, while disregarding neighbors' opinions* ($b = 1$)*, then their opinion converges to either $y_r = (1-\mu)\theta + \mu\gamma_i$ or $y_l = (1-\mu)\theta + \mu(1-\gamma_i)$ almost surely, regardless of initial belief* $(\alpha_{i,0}, \beta_{i,0})$.

Both left and right biased opinions can be considered tail events and, therefore, may emerge with some positive probability, in the spirit of any classic zero-one law. Moreover, both opinions are biased as any limiting opinion is a weighted average that places weight μ on the confirmatory bias parameter γ_i and weight $(1-\mu)$ on the true state θ. Thus, both the fraction of ambiguous signals μ and confirmatory bias γ_i are key misinformation drivers.

Another way to view this is by rewriting the expressions y_l and y_r as the true underlying state plus its biases

$$y_l = \theta + \underbrace{\mu\big((1-\gamma_i) - \theta\big)}_{\text{bias}}, \text{ and}$$
$$y_r = \theta + \underbrace{\mu\big(\gamma_i - \theta\big)}_{\text{bias}}.$$

Considering the vaccine efficacy example, say the underlying efficacy is $\theta = 0.8$ (i.e., people who got the vaccine were at 80% lower risk of contracting the disease), the confirmatory bias is $\gamma_i = 0.6$ and the fraction of ambiguous signals is 0.3. As per the y_r expression, the bias is $0.3 \times (0.6 - 0.8) = -0.06$, meaning that individual i believes the efficacy is $0.8 - 0.06 = 0.74$. As per the y_l expression, the bias is $0.3 \times \big((1-0.6) - 0.8\big) = -0.12$, meaning that individual i believes the efficacy is $0.8 - 0.12 = 0.68$. Hence, both opinions are biased and both μ and γ_i are important misinformation drivers. From this perspective, in the absence of ambiguity ($\mu = 0$), there would be no bias and no misinformation. Conversely, if $\mu > 0$, then misinformation could be fully mitigated if $\gamma_i = \theta$ (i.e. agents randomizing interpretation between 0 and 1 according to the true proportion θ) if $\theta > 0.5$ or if $\gamma_i = 1 - \theta$ if $\theta < 0.5$.

Based on this example, a first result of interest stemming from Proposition 1 is that, for any individual with confirmatory tendency, one opinion type is less biased than the other depending on the state θ. This is generalized as follows.

Corollary 1 (asymmetric bias). *For any individual with confirmatory tendency and for any ambiguity level, y_r (y_l) is less biased than y_l (y_r) if $\theta > \frac{1}{2}$ $\left(\theta < \frac{1}{2}\right)$. Conversely, both y_r and y_l are equally biased when $\theta = \frac{1}{2}$.*

Generally, these two opinions are not equally distant from θ (see Example 1). This is because the bias of each one depends on the relative size of θ and γ_i. As we are restricting attention to the case in which the agent has confirmatory tendency (i.e., $\gamma_i \geq \frac{1}{2}$), it is the case that agents make less mistakes when they are in the correct side of the spectrum. Therefore, ambiguity has to be low enough to not mislead agents and the state has to be high (or low) enough to nudge agents' opinions to the correct side.

Example 1. *Suppose that a biased individual ($\gamma_i = 1$) faces 20% of ambiguous signals ($\mu = 0.20$) and consider three particular values for the underlying state θ, say: low ($\theta_L = 0.1$), medium ($\theta_M = 0.5$) and high ($\theta_H = 0.9$).*[6] *In this case, Proposition 1 and Corollary 1 show that under state*

$$\theta_L, \begin{cases} y_r = (1-0.2)\times 0.1 + 0.2 \times 1 = 0.28 \text{ is formed with probability } 0, \text{ and} \\ y_l = (1-0.2)\times 0.1 + 0.2 \times (1-1) = 0.08 \text{ is formed with probability } 1. \end{cases}$$

Under state

$$\theta_M, \begin{cases} y_r = (1-0.2)\times 0.5 + 0.2 \times 1 = 0.60 \text{ is formed with some probability } p \in (0,1), \text{ and} \\ y_l = (1-0.2)\times 0.5 + 0.2 \times (1-1) = 0.40 \text{ is formed with probability } 1-p. \end{cases}$$

Finally, under state

$$\theta_H, \begin{cases} y_r = (1-0.2)\times 0.9 + 0.2 \times 1 = 0.92 \text{ is formed with probability } 1, \text{ and} \\ y_l = (1-0.2)\times 0.9 + 0.2 \times (1-1) = 0.72 \text{ is formed with probability } 0. \end{cases}$$

Although y_l and y_r are not equidistant from θ, the distance between y_l and y_r does not depend on θ and is equal to $\mu(2\gamma_i - 1)$. Hence, for any given μ, the distance between opinions y_l and y_r is maximal when $\gamma_i = 1$. In this case, the distance becomes μ. In the Example 1, note that as $\gamma_i = 1$, opinions distance is always $\mu = 0.2$, regardless of state θ.

Moreover, depending on the pair (θ, μ), we can show which opinion will be reached. Hence, we just need to determine which combinations of θ and μ are sufficient to allow both types of opinion to fall in the same side of the 0-1 spectrum (i.e. both y_l and y_r above 0.5 or both y_l and y_r below 0.5) and which combinations lead opinions to diverge in location (i.e. $y_r \geq 0.5$ and $y_l < 0.5$). These conditions lead to different regions of space $\Theta \times M = [0,1]^2$ (unit square): region L, characterized by both low state θ and low ambiguity level μ; region R, characterized by both high state and low ambiguity; whereas region $\mathcal{W}$ is the complement of the union of L and R. In mathematical terms,

[6] The arbitrary values chosen for θ in the Example 1 only mean to illustrate the results of Proposition 1 and Corollary 1 for a broad range of θ. For all purposes, $\theta \in \Theta = [0,1]$.

those partitions are characterized by the following:

$$R = \left\{ (\theta, \mu) | \frac{1}{2} < \theta \leq 1 \text{ and } 0 \leq \mu < \frac{\theta - 0.5}{\gamma_i + \theta - 1} \right\},$$

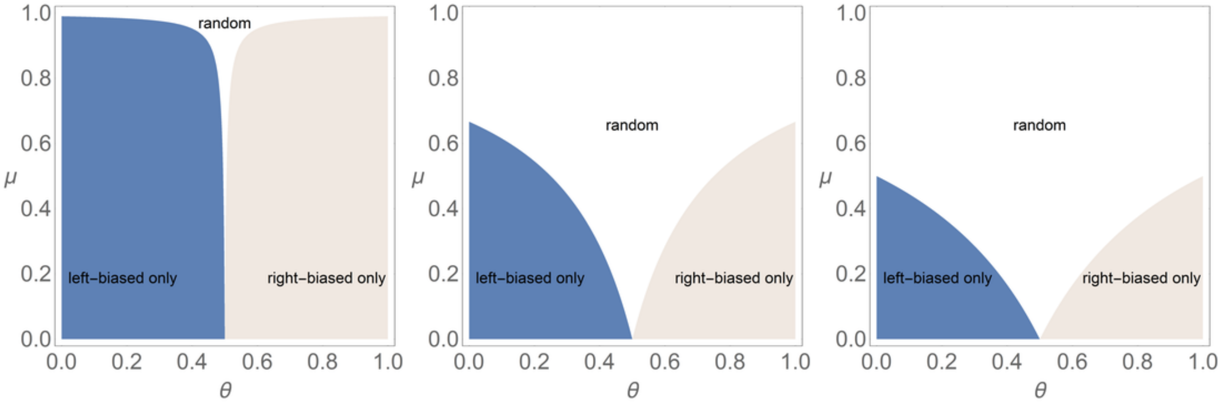
$$L = \left\{ (\theta, \mu) | 0 \leq \theta < \frac{1}{2} \text{ and } 0 \leq \mu < \frac{\theta - 0.5}{\theta - \gamma_i} \right\},$$

$$\mathcal{W} = [0,1]^2 \setminus \{R \cup L\}.$$

We can state the following result regarding the general conditions for the emergence of each opinion type.

Proposition 2 (opinion-type emergence)**.** *For any individual with confirmatory tendency, if $(\theta, \mu) \in R$, then the limiting opinion is the right-biased one with probability 1. If $(\theta, \mu) \in L$, then the limiting opinion is the left-biased one with probability 1. If $(\theta, \mu) \in \mathcal{W}$, then the limiting opinion is a random variable whose possible values are y_l and y_r.*

This result holds regardless of initial beliefs and observed sequence of signals. For three cases of confirmation bias, Figure 3 depicts the idea of Proposition 2. In case 1, when the agent is roughly impartial, case 2 when the agent has an intermediary level of confirmatory bias, and case 3 when agent is biased.



(A) case 1: $\gamma_i = 0.505$ (B) case 2: $\gamma_i = 0.750$ (C) case 3: $\gamma_i = 1.000$

FIGURE 3. Parameter space and emergence of different types of consensus

Lightly shaded areas on the right represent the set of parameters μ (vertical axis) and θ (horizontal axis) ensuring the emergence of a right-biased opinion. Conversely, darkly shaded areas represent the set of parameters ensuring the emergence of left-biased opinion. In both areas, for a given level of confirmatory bias, the left (right)-biased opinion emerge with probability 1 if there is both low frequency of ambiguous signals and low (high) state (i.e., below (above) 0.5).

On the other hand, the white area represents the combinations of θ and μ such that the opinion type becomes a random variable whose possible values are y_l and y_r (i.e., both opinion types may emerge with positive probability). Considering that when $(\theta, \mu) \in \mathcal{W}$ any of the two opinions may be formed, we define the probability p with which an individual reaches the less-biased opinion following the results from Proposition 1 and Corollary 1.

Definition 2. *For any given initial belief $(\alpha_{i,0}, \beta_{i,0})$, ambiguity level $\mu \in (0,1)$ and confirmation bias $\gamma_i \geq \frac{1}{2}$, the probability of less-biased opinion forming is*

$$p = \begin{cases} P\left(\lim\limits_{t\to\infty} y_{i,t} = y_l\right), & \text{when } \theta < 0.5, \text{ or} \\ P\left(\lim\limits_{t\to\infty} y_{i,t} = y_r\right), & \text{when } \theta > 0.5. \end{cases}$$

Another interesting case stemming from Proposition 1 is the one in which bias cannot be overcome even when an agent is impartial.

Corollary 2 (bias from impartiality)**.** *If an individual is impartial, then his limiting opinion is $(1-\mu)\,\theta + \mu\,\frac{1}{2}$ almost surely, regardless of his initial prior and the sequence of observed signals.*

Impartiality does not overcome bias because it forces agents to set a disproportionate probability mass in the center of the spectrum $(0,1)$. Hence, impartiality makes agents excessively centrist instead of making them neutral toward possible states. This is a direct consequence of the Beta-Bernoulli conjugate family employed here that would not occur in a binary state space (i.e., $\Theta = \{0,1\}$).

Moreover, under impartiality, for any mass of ambiguity $\mu > 0$, if true state is located in the left side of the 0-1 spectrum ($\theta < \frac{1}{2}$), then opinion has a positive bias and lies in $\left(\theta, \frac{1}{2}\right)$. Conversely, if $\theta > \frac{1}{2}$, then opinion has a negative bias and lies in $\left(\frac{1}{2}, \theta\right)$. The only instance when an individual learns the state is when $\theta = \frac{1}{2}$, a zero mass event if θ was drawn randomly from the interval $[0,1]$. The results presented so far both extend the intuition and contrast with Propositions 4 and 5 in Rabin and Schrag (1999) and with Propositions 2 and 3 in Fryer Jr et al. (2019). This extends the intuition to the case in which the state is continuously distributed over the interval 0-1 and contrasts because impartiality can no longer help an individual overcome bias, as per the result above.

Finally, at the other extreme, one could ask under what conditions an individual would reach an extreme opinion (i.e., either opinion 0 (extreme left) or opinion 1 (extreme right)). The next result shows that those cases can only be sustained under two extreme conditions: (i) the fraction of ambiguous signals is maximal ($\mu = 1$) and, (ii) individual is biased ($\gamma_i = 1$).

Corollary 3 (extreme opinions). *For any $\theta \in \Theta$ and any initial belief $(\alpha_{i,0}, \beta_{i,0})$, any individual $i \in N$ will form an extreme opinion (either 0 or 1) if they are biased ($\gamma = 1$) and the mass of ambiguity is maximal ($\mu = 1$).*

4.2. **Networked society.** Given the intuition of the single agent case described, one may ask what happens if agents also learned from their friends, besides learning from signals. This case imposes an extra challenge as the interpretation of ambiguous not only depends on the initial realization of signals but also on the influence of friends that potentially interpret ambiguity in different ways. "*Tug-of-war*" played between left and right biases has one extra driver: the network externalities.

Before discussing the implications of a network structure, we define the concept of consensus (Definition 3) and illustrate the social influence of agents, in terms of ergodicity of a Markov chain (Lemma 1), derived from the reliance weight matrix W. Appendix C contains the proof of the Lemma.

Definition 3 (consensus). *Society reaches a consensus almost surely for any initial beliefs if there is a y such that, for every $\epsilon > 0$ and $i \in N$,*

$$P\left(\lim_{t\to\infty} |y_{i,t} - y| < \epsilon\right) = 1.$$

Lemma 1 (strong connectivity). *The t-th power of matrix W, W^t, converges to a unique row-stochastic matrix with unit rank (all rows the same) as t tends to infinity, i.e.*

$$\lim_{t\to\infty} W^t = W^\infty = \mathbf{1}\pi^\top = \Pi,$$

where the invariant distribution π is the normalized left eigenvector of the matrix W associated to the unit eigenvalue, i.e. $\pi^\top W = \pi^\top$ and $\sum_i \pi_i = 1$.

A first case of interest is the limiting case in which individuals exclusively pay attention to friends. This represents the situation wherein agents disregard signals completely and are pure conformists. The consensus reached is slightly different from the classic DeGroot case, as the limiting opinion is not exactly a weighted average of the initial opinions, although is still very close to it. The discrepancy has to do with the fact that agents are exchanging opinions and precisions (parameters α and β). This is stated as follows.

Proposition 3 (DeGroot consensus). *If the social network $G = (N, g)$ is strongly connected, and agents disregard all public signals ($b = 0$), then society reaches consensus*

$$\bar{y} = \frac{\sum_j \Pi_{ij}\alpha_{j,0}}{\sum_j \Pi_{ij}\left(\alpha_{j,0} + \beta_{j,0}\right)},$$

for any $i \in N$ and where Π is the invariant distribution matrix.

Section 5 highlights the implications of this result wherein we explore the effects of priors heterogeneity on the probability of attaining the less-biased consensus. Next, we show that assuming strong connectivity, consensus is reached in this dynamic system and has a similar functional form of the individual limiting opinion in Proposition 1.

Proposition 4 (Network externality). *With network externalities ($0 < b < 1$), sequences $\{y_{i,t}\}_{t=1}^{\infty}$ generated by the update rule converge almost surely to either right-biased consensus $\bar{y}_r = (1-\mu)\theta + \mu\bar{\gamma}$ or left-biased consensus $\bar{y}_l = (1-\mu)\theta + \mu(1-\bar{\gamma})$ for all $i \in N$ and where Π is the invariant distribution matrix, and $\bar{\gamma} = \sum_j \Pi_{ij}\gamma_j$ for any $i \in N$.*

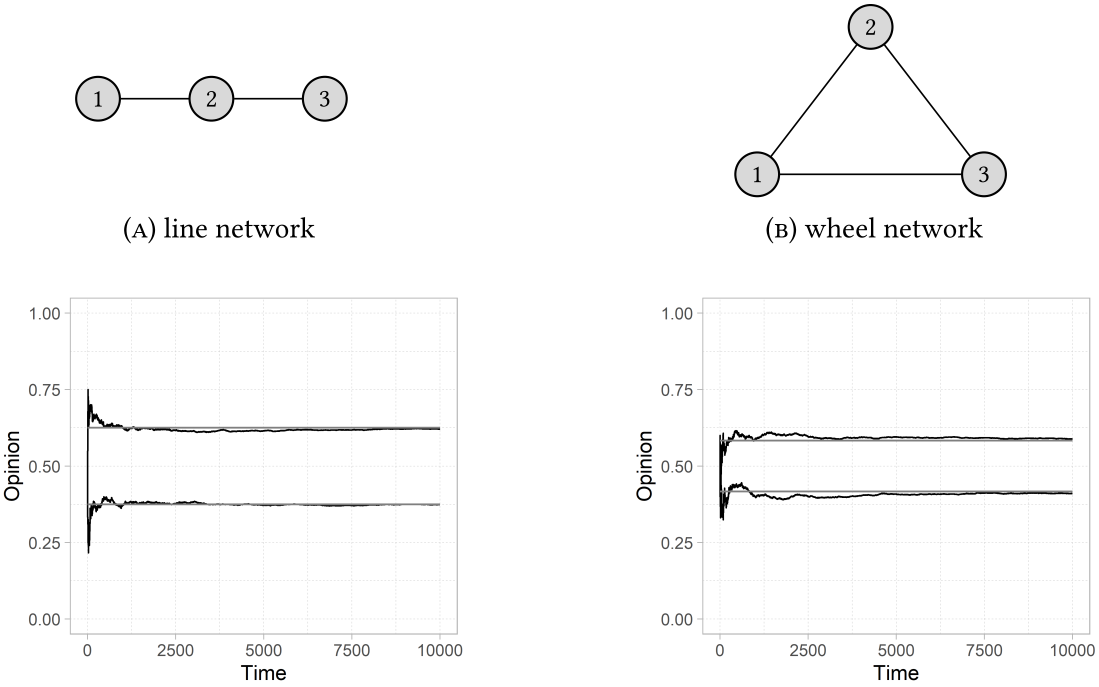
Therefore, in a networked society, consensus is a *weighted average* between the true state θ (with weight $1-\mu$) and the weighted average of confirmatory biases $\bar{\gamma}$ (with weight μ). Society aggregates information efficientrly and no bias exists if $\mu = 0$ or $\bar{\gamma} = \theta$ when $\theta > 0.5$ or if $\bar{\gamma} = 1 - \theta$ when $\theta < 0.5$. Additionally, parameter b impacts the vector of social influence through the invariant distribution π of the matrix W (see Lemma 1) and therefore does impact consensus. Note that when $b = 1$, social connection is lost, and polarization emerges with each individual opinion being a function of individual confirmation bias as in Proposition 1.

Moreover, the above results show that consensus type in this dynamic system is also a tail event (i.e., right-biased consensus will either almost surely emerges as the stable equilibrium or almost surely not emerge). If this does not emerge as an equilibrium of this system, then the left-biased consensus has truly emerged as the equilibrium. Figures 4c and 4d show the typical opinion sample paths (different simulations) of any agents' opinions in the line and wheel networks, respectively, and convergence to different consensus types (horizontal lines).

In terms of social efficiency, the next Section numerically characterizes the probability of emergence of efficient consensus. As this exercise is not trivial, we rely on simulations of the learning process described in Section (3) for selected classic network topologies and for different sets of parameters of interest and a Probit regression model to explore the variability of simulated data.

5. Monte Carlo simulations: determining consensus type in classic networks

Ascertaining p (Definition 2) analytically for a networked society is not trivial as several recursions are involved in the opinion formation process. Initial prior distributions can be disproportionately skewed (partisan agents) and, as a consequence of this, agents may be prone to



(A) line network

(B) wheel network

(C) two simulated opinion paths (black) in a line network and theoretical consensuses (gray)

(D) two simulated opinion paths (black) in a wheel network and theoretical consensuses (gray)

FIGURE 4. four simulations with parameters $T = 10,000$, $n = 3$, $\mu = \theta = b = 0.5$, $\alpha_{i,0} = \beta_{i,0} = 1$ for all $i \in N$ (so $y_{i,0} = 0.5$ for any i) and $\gamma = (\gamma_1, \gamma_2, \gamma_3) = (0.8, 1, 0.2)$.

interpret ambiguous signals differently. Besides, priors heterogeneity induces centrality heterogeneity (i.e., partisan agents might be located at more or less central nodes). This is challenging because partisans might influence other agents disproportionately and amplify underlying interpreting disputes in the network. Agents might also differ in the intensity of their confirmatory bias they suffer (i.e. agents with polar opposite bias might be directly connected or not). This might influence how much this heterogeneity affects the interpreting conflict. This task becomes particularly complex if one allows different partisan agents to have different confirmatory biases. The examples show the challenging nature of computing p analytically, as signals interpretations depend not only on the stream of public signals observed by agents but also on other agents' beliefs and their location.

5.1. **Initial beliefs.** This exercise reduces the dimension of initial beliefs into a single parameter $\tau \in \mathbb{R}_{\geq 0}$ that comprises both *common* and *heterogeneous* priors.

(a) Heterogeneous priors. Refer to the situation in which there are three types of agents at time $t = 0$ with different initial prior distributions: centrists ($\mathcal{C}_0$), leftists ($\mathcal{L}_0$) and rightists ($\mathcal{R}_0$). To distinguish the agents, consider two parameters that intend to measure the degree of *partisanship* of such agents $\tau_l, \tau_r \in \mathbb{N}_+$. Hence, such groups are defined as follows: centrists, $\mathcal{C}_0 = \{i \in N \mid \alpha_{i,0} = 1 \text{ and } \beta_{i,0} = 1\}$, left-partisan, $\mathcal{L}_0 = \{i \in N \mid \alpha_{i,0} = 1 \text{ and } \beta_{i,0} = 1 + \tau_l\}$, and right-partisan, $\mathcal{R}_0 = \{i \in N \mid \alpha_{i,0} = 1 + \tau_r \text{ and } \beta_{i,0} = 1\}$. Notice that the definition implies that initial opinions and precisions $y_{i,0} = \alpha_{i,0}(\alpha_{i,0}+\beta_{i,0})^{-1}$ and $\sigma_{i,0}^{-2} = (\alpha_{i,0}\beta_{i,0})^{-1}(\alpha_{i,0}+\beta_{i,0})^2(\alpha_{i,0}+\beta_{i,0}+1)$, respectively,

$$y_{i,0} = \begin{cases} \dfrac{1}{2+\tau_l}, & \text{if } i \in \mathcal{L}_0 \\ \dfrac{1}{2}, & \text{if } i \in \mathcal{C}_0 \\ \dfrac{1+\tau_r}{2+\tau_r}, & \text{if } i \in \mathcal{R}_0 \end{cases}$$

and

$$\sigma_{i,0}^{-2} = \begin{cases} \dfrac{6+5\tau_l+\tau_l^2}{1+\tau_l}, & \text{if } i \in \mathcal{L}_0 \\ 12, & \text{if } i \in \mathcal{C}_0 \\ \dfrac{6+5\tau_r+\tau_r^2}{1+\tau_r}, & \text{if } i \in \mathcal{R}_0. \end{cases}$$

Notice that $\lim_{\tau\to\infty} y_{i,0}$ is 0, $\frac{1}{2}$ and 1, whereas $\lim_{\tau\to\infty} \sigma_{i,0}^{-2}$ is $+\infty$, 12 and $+\infty$ for left-partisan, centrists and right-partisan, respectively.

(b) Common prior. Refer to the situation in which priors parameters are identical across agents (i.e., $\alpha_{i,0} = \alpha \in \mathbb{R}_+$ and $\beta_{i,0} = \beta \in \mathbb{R}_+$ for all $i \in N$). Particularly, when $\alpha = \beta = 1$, all agents hold a uniform common prior over the unit interval. For any other value, say $\alpha = \beta = k > 1$, agents hold a symmetric bell-shaped common prior over the unit interval, centered at 0.5. Moreover, as $k \to \infty$, the bell-shaped priors collapse to the point 0.5 (i.e., the precision of the prior diverges) and all opinions are $y_{i,0} = 0.5$. These cases represent the situation wherein agents begin as *centrists*, and the subsequent asymmetry of interpretation stems from the signals realizations.[7] Conversely, when $\alpha_{i,0} = \alpha$ and $\beta_{i,0} = \beta$ for all $i \in N$ and $\alpha > \beta$ ($\beta > \alpha$), the society holds a rightist (leftist) common prior (i.e., $y_{i,0} = y > 0.5$ ($y_{i,0} = y < 0.5$) and as $\frac{\alpha}{\beta} \to \infty$ ($\to 0$)),

[7] To be more precise, interpretation neutrality does not exist as there is a non-neutral tie-break rule in Equation (2). Thus, if the realization of the first public signal is a, then this signal will be interpreted as 1 by all agents, as per the tie-break rule. This is without loss of generality for the results presented in this work. The tie-break rule could have been defined in a way that the initial interpretation would be 0 and intuition and conclusions would remain the same. Finally, if we established no tie-break rules, a more intricate update rule would be needed to maintain the prior when facing an ambiguous signal and opinions were exactly 0.5. Hence agents would keep opinions unchanged until some non-ambiguous realization occurs. In this case, the results would not also change as the neutral tie-break would promote some of the states, and the nature of the problem remains unchanged.

the bell-shaped priors collapse to the point 1 (0) (i.e., the precision of the prior diverges and all opinions become extreme).

5.2. **Monte Carlo simulation.** To compute the empirical frequency of the emergence of the less-biased consensus ($\hat{p}$) for states and ambiguity level in $\mathcal{W}$, we simulate the learning process in Section (3) in selected classic networks (G) (Figure 5). The number of simulations is described by $S \in \mathbb{N}_+$ and interaction time is $t \in \mathbb{N}_+$.

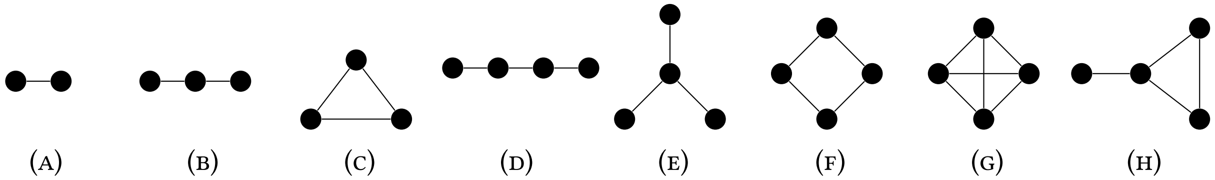


FIGURE 5. Classic networks – G

Each simulation allows some parameters to vary (see details below), so changes in $\hat{p}$ can be captured owing to changes in such parameters. However, in each simulation, parameter choices are identical for all networks. Hence, we can properly isolate effects on $\hat{p}$ owing to parameter variability. For any given simulation S, the realization of the public signals is also identical across networks. Therefore, given the choice of parameters in each simulation S, the simulated frequency of the less-biased consensus in a network G follows Definition 2, and is computed as follows:

$$\hat{p}_G = \frac{1}{S}\sum_{S} \mathbb{1}\Bigg\{ \left|\lim_{t\to\infty} y_{i,t}^{S,G} - ((1-\mu_S)\theta_S + \mu_S\gamma_S)\right| < \epsilon \text{ and } \theta_S > 0.5 \text{ , or} \\ \left|\lim_{t\to\infty} y_{i,t}^{S,G} - ((1-\mu_S)\theta_S + \mu_S(1-\gamma_S))\right| < \epsilon \text{ and } \theta_S < 0.5\Bigg\}, \tag{9}$$

for a small $\epsilon > 0$.

Partisanship effect. This exercise mainly aims to understand how degree of partisanship (τ) affects the probability of less-biased consensus emerging. Hence, both partisans are placed in the available nodes uniformly at random in each simulation. Thus, despite being in equal number, their level of centrality may differ in each simulation. Regarding the degree of partisanship, when $\tau_l = \tau_r = \tau = 0$, agents have a common uninformative prior (uniform distribution over the unit interval), and no partisan agents are found. Conversely, when $\tau_l = \tau_r = \tau > 0$, we have heterogeneous priors in which the degree of partisanship of both partisans is equally balanced.

Moreover, to avoid an extra layer of heterogeneity, we allow all agents to be biased (i.e., $\gamma_i = \gamma_S = \gamma = 1$ for all i and S). The benefit of fixing $\gamma = 1$ across agents and simulations is that we know how each node is interpreting ambiguous signals and, hence, this allows us to study the effect of partisans centrality. Thus, this simulation assumes the following configuration:

- **Fixed parameters** (for all simulations) are as follows:
  - **Learning**: $\gamma_i = \gamma = 1$ for all i,
  - **Bayesian**: $b = 0.5$,
  - **Duration**: $t = 700$,
  - **Priors**: $(\alpha_{i,0}, \beta_{i,0})$ = $(1, 1)$ for all $i \in N$, $|\mathcal{L}_0| = |\mathcal{R}_0| = 1$,
  - **Information**: $\mu = 0.6$.
- **Variable parameters** (for each simulation S) are
  - **Information**: $\theta_S \in \{0.2, 0.8\}$ and
  - **Initial prior**: $\tau_l = \tau_r = \tau_S$ such that $\tau_S \in \{0, 1, 10, 30\}$,
  - **Partisans location**: partisans are placed uniformly at random in the nodes of each network.

Table 1 presents the summary statistics of the simulated data. Besides reporting the simulated $\hat{p}$ for each network, the table shows other variables that present variability across simulations S, as follows:

(1) *degree centrality* of both partisans in networks whose degree centrality variance is positive (i.e., networks (B), (D), (E) and (H)),
(2) *open-mindedness* dummy variable valued at 1 when partisans with opposite beliefs are connected, and 0 if otherwise[8],
(3) a dummy variable named *first impression* (in reference to the work of Rabin and Schrag (1999)) that takes on value 1 when the realization of the very first public signal nudges the society toward the true state θ in that particular simulation and network[9],
(4) a dummy variable for when $\theta_S = 0.8$,
(5) and dummy variables for all levels of partisanship τ.

Besides analyzing the statistical relation established between $\hat{p}$ and different levels of τ, Table 2 presents and discusses the results of a Probit regression model wherein the dependent variable is a dummy variable that takes on the value 1 when less-biased consensus is formed in simulation S after t periods of agents interaction and the independent variables are (i) a dummy variable called

[8] For any given network induced by some adjacency matrix g, a partisan agent $i \in N$ is considered open-minded if for some other partisan agent $j \in N$ with opposite belief, we have that $j \in N_i^{out}(g)$. Conversely, i is narrow-minded if $j \notin N_i^{out}(g)$. Open-mindedness is defined quite similarly to *heterophily* already established in social and economic networks literature. Both reflect the tendency of different people to connect with each other.

[9] In mathematical terms, the first impression in simulation S and network G is defined as

$$\text{FI}^{S,G} = \mathbb{1}\{s_1^{S,G} = a \text{ or } s_1^{S,G} = 1, \text{ if } \theta_S \geq 0.5\} + \mathbb{1}\{s_1^{S,G} = 0, \text{ if } \theta_S < 0.5\}.$$

partisan centrality advantage (PCA) that takes on value 1 if the rightist (leftist) is more central than the leftist (rightist) when $\theta \geq 0.5$ ($\theta < 0.5$), (ii) open-mindedness dummy (OM), (iii) the first impression dummy, (iv) the dummy variable for when $\theta_S = 0.8$, and (v) all dummy variables for all different levels of partisanship τ. The exercise aims at exploring the variability of these variables to assess their relative importance in terms of increasing the odds of the less biased consensus to be reached.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
$\hat{p}_{(A)}$	21,040	0.809	0.393	0	1	1	1	1
$\hat{p}_{(B)}$	21,040	0.679	0.467	0	0	1	1	1
$\hat{p}_{(C)}$	21,040	0.810	0.392	0	1	1	1	1
$\hat{p}_{(D)}$	21,040	0.735	0.441	0	0	1	1	1
$\hat{p}_{(E)}$	21,040	0.698	0.459	0	0	1	1	1
$\hat{p}_{(F)}$	21,040	0.826	0.379	0	1	1	1	1
$\hat{p}_{(G)}$	21,040	0.787	0.409	0	1	1	1	1
$\hat{p}_{(H)}$	21,040	0.700	0.458	0	0	1	1	1
$\mathcal{R}_0$ degree in (B)	21,040	1.332	0.471	1	1	1	2	2
$\mathcal{R}_0$ degree in (D)	21,040	1.503	0.500	1	1	2	2	2
$\mathcal{R}_0$ degree in (E)	21,040	1.516	0.875	1	1	1	3	3
$\mathcal{R}_0$ degree in (H)	21,040	2.010	0.704	1	2	2	3	3
$\mathcal{L}_0$ degree in (B)	21,040	1.333	0.471	1	1	1	2	2
$\mathcal{L}_0$ degree in (D)	21,040	1.503	0.500	1	1	2	2	2
$\mathcal{L}_0$ degree in (E)	21,040	1.501	0.866	1	1	1	3	3
$\mathcal{L}_0$ degree in (H)	21,040	1.996	0.707	1	1	2	2	3
Open mind (OM) in (B)	21,040	0.666	0.472	0	0	1	1	1
Open mind (OM) in (D)	21,040	0.501	0.500	0	0	1	1	1
Open mind (OM) in (E)	21,040	0.508	0.500	0	0	1	1	1
Open mind (OM) in (H)	21,040	0.667	0.471	0	0	1	1	1
First impression (FI)	21,040	0.619	0.486	0	0	1	1	1
$\mathbb{1}\{\theta = 0.8\}$	21,040	0.500	0.500	0	0	0.5	1	1
$\mathbb{1}\{\tau = 0\}$	21,040	0.250	0.433	0	0	0	0.2	1
$\mathbb{1}\{\tau = 1\}$	21,040	0.250	0.433	0	0	0	0.2	1
$\mathbb{1}\{\tau = 10\}$	21,040	0.250	0.433	0	0	0	0.2	1
$\mathbb{1}\{\tau = 30\}$	21,040	0.250	0.433	0	0	0	0.2	1

TABLE 1. Summary statistics - simulated $\hat{p}$ and parameters

Moreover, as the main goal is to understand the effect of partisanship on $\hat{p}$, Table 3 shows $\hat{p}$ for different levels of τ.

Based on statistical and regression analyses of simulation data, we present results that hold for classic network structures presented above and for the case one draws a single pair in $\mathcal{W}$ uniformly at random (i.e., under no knowledge of θ or μ). Although simulation results are not generalized to a broader range of network topologies, the structures analyzed are sufficiently general and have similar characteristics of real-world networks. Hence, as per the data from

	Dep. Variable: probability of emergence of less biased consensus ($\hat{p}_G$)							
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Partisan centrality advantage (PCA)		1.86^{***}		0.87^{***}	1.91^{***}			1.62^{***}
		(0.04)		(0.04)	(0.05)			(0.05)
Open mind (OM)		-1.06^{***}		0.003	-0.96^{***}	0.50^{***}		0.45^{***}
		(0.04)		(0.03)	(0.04)	(0.03)		(0.04)
First impression (FI)	1.64^{***}	2.13^{***}	1.93^{***}	1.13^{***}	2.24^{***}	2.32^{***}	2.16^{***}	1.36^{***}
	(0.03)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.04)	(0.05)
PCA × FI		0.83^{***}		0.91^{***}	0.77^{***}			0.61^{***}
		(0.08)		(0.08)	(0.10)			(0.07)
PCA × OM				-0.05				-0.39^{***}
				(0.06)				(0.06)
OM × FI		-1.10^{***}		-0.29^{***}	-1.18^{***}	-0.61^{***}		-0.10^{*}
		(0.06)		(0.04)	(0.06)	(0.06)		(0.05)
$\mathbb{1}\{\tau = 1\}$	0.54^{***}	0.48^{***}	0.56^{***}	0.58^{***}	0.34^{***}	0.52^{***}	0.33^{***}	0.42^{***}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
$\mathbb{1}\{\tau = 10\}$	0.54^{***}	-0.41^{***}	0.66^{***}	0.13^{***}	-0.37^{***}	0.87^{***}	0.38^{***}	-0.08^{*}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
$\mathbb{1}\{\tau = 30\}$	0.54^{***}	-0.41^{***}	0.66^{***}	-0.16^{***}	-0.38^{***}	0.98^{***}	0.39^{***}	-0.65^{***}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
$\mathbb{1}\{\theta = 0.8\}$	-0.33^{***}	-0.02	-0.26^{***}	-0.05	0.04	-0.15^{***}	-0.19^{***}	-0.06^{*}
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)
Constant	-0.15^{***}	0.03	-0.31^{***}	-0.27^{***}	-0.08^{**}	-0.68^{***}	-0.30^{***}	-0.85^{***}
	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Observations	21,040	21,040	21,040	21,040	21,040	21,040	21,040	21,040
Log Likelihood	-7,873.93	-7,648.32	-7,053.26	-9,546.68	-7,350.89	-6,348.13	-6,941.64	-8,660.61
Akaike Inf. Crit.	15,759.90	15,316.60	14,118.50	19,115.40	14,721.80	12,712.30	13,895.30	17,343.20

Note: *p<0.05; **p<0.01; ***p<0.001

TABLE 2. Regression results: Probit

simulations with common prior ($\tau = 0$, no partisanship), we can see that network structure has limited effect over $\hat{p}$. This evidence is stated as the following result.

Result 1 (topology neutrality)**.** *If society is biased ($\gamma = 1$) and have common prior (i.e., $\tau = 0$), then network topology has no significant impact on $\hat{p}$.*

The intuition of this result relies on the fact that as signals are public and all agents share the same bias intensity $\gamma_i = 1$, no interpretation diversity exists regardless of signals realization. If agents begin observing signal 1, then all agents will become more rightists and network externalities cannot countervail this effect anyhow. The same argument applies to all other signals, including the ambiguous one. Hence, this is identical to the case of a single individual learning from signals. Moreover, based on the data from simulations with a common prior ($\tau = 0$, no partisanship) and low priors heterogeneity ($\tau = 1$, low partisanship), partisanship seems to have a non-negative effect on consensus efficiency.

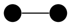
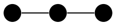
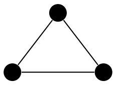
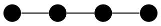
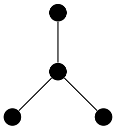
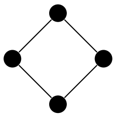
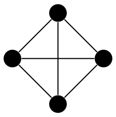
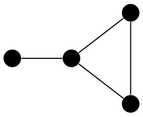
Size	Network Topology	Type	Label	$\hat{p}$ $(\tau{=}0)$	$\hat{p}$ $(\tau{=}1)$	$\hat{p}$ $(\tau{=}10)$	$\hat{p}$ $(\tau{=}30)$
$n = 1$		single agent	(SA)	0.702	-	-	-
$n = 2$		line (complete)	(A)	0.702	0.844	0.844	0.844
$n = 3$		line	(B)	0.702	0.802	0.603	0.610
		wheel (complete)	(C)	0.702	0.834	0.852	0.852
$n = 4$		line	(D)	0.702	0.841	0.739	0.659
		star	(E)	0.721	0.787	0.643	0.641
		wheel	(F)	0.702	0.818	0.884	0.899
		complete	(G)	0.727	0.801	0.810	0.812
		paw	(H)	0.720	0.814	0.706	0.558
S				21,040	21,040	21,040	21,040

TABLE 3. Simulated frequency of the emergence of less biased consensus $\hat{p}$.

Result 2 (low partisanship effect)**.** *In expected terms, a biased society with low degree of partisanship (*$\tau = 1$*) can reach the less-biased consensus as the same biased society with no partisanship at all (*$\tau = 0$*).*

This can be seen in two ways: (i) there is a statistically significant difference between proportions in Table 3 under $\tau = 0$ and $\tau = 1$, and (ii) coefficients of the dummy variable $\mathbb{1}\{\tau = 1\}$ are all positive and significant.[10]

Partisanship acts to counter the effect of initial ambiguous signals. Under no partisan influence, agents' interpretations depend exclusively on the signals. The realization of the initial signal is crucial to determine what bias opinions will have and, hence, it is determinant to consensus efficiency. Conversely, when some partisan agents are present, priors parameters α's and β's are shifted up, by right- and left-partisans, respectively, which makes opinions more robust to initial signal realization. However, some "optimal" level of partisanship exists as high partisanship, for

[10] Note that the coefficients are with respect to the omitted dummy variable $\mathbb{1}\{\tau = 0\}$. The omission is needed so there is no perfect colinearization.

most topologies, has a nonmonotonic effect over the probability of emergence of the less-biased consensus. This result is generalized as follows.

Result 3 (high partisanship effect)**.** *In expected terms, a biased society with low partisanship ($\tau = 1$) can reach the efficient consensus as the same biased society with high partisanship ($\tau = 30$). Exceptions include the wheel and complete networks (i.e. (C), (F) and (G)) in which $\hat{p}$ is non-decreasing with partisanship.*

This can be seen in two ways: (i) there is a statistically significant difference between the proportions in Table 3 under $\tau = 1$ and $\tau = 30$, and (ii) the coefficients of the dummy variable $\mathbb{1}\{\tau = 30\}$ are higher than the coefficients of the dummy variable $\mathbb{1}\{\tau = 1\}$ for the referred networks.

Moreover, if an imbalance in partisanship exists, then partisan agents can unbalance opinions in the same way realization of the first signals do. More explicitly, a partisan agent with high degree of partisanship will almost never interpret ambiguous evidence in a way that disagrees with his beliefs and a similar effect applies to his neighbors. However, partisan agents might be more or less connected and even connected to each other.

Naturally, in networks (A), (C), and (G), partisan agents are invariably open-minded as those networks are complete (i.e., all agents are connected with every other agent in the network, regardless of their types). Hence, analyzing the effect of OM in networks (B), (D), (E), (F), and (H) as those agents are not always connected. Table 4 reports simulated probability $\hat{p}$ in those cases, and the next result is stated immediately.

Result 4 (open-minded partisans)**.** *In expected terms, for biased agents, open-mindedness of partisans increases odds of less-biased consensus formation in networks (F) and (H). Conversely, narrow-mindedness of partisans increases the odds of the less-biased consensus formation in networks (B) and (E).*

This can be seen in two ways: (i) a statistically significant (positive) difference between the proportions in Table 4 under the open-minded (OM) and narrow-minded (NM) cases. That is, OM increases the odds relative to the NM case for networks (F) and (H), whereas NM increases the odds with respect to the OM case in networks (B) and (E). (ii) Coefficients of the dummy variable OM in the Probit regression are only positive for networks (F) and (H) and negative for the (B) and (E) networks.

The intuition relies on Results 2 and 3. In networks (B) and (E), OM would imply that one partisan is disproportionately more influential than the other (i.e., there would be a partisan

Network	Partisans	$\hat{p}$ $(\tau{=}0)$	$\hat{p}$ $(\tau{=}1)$	$\hat{p}$ $(\tau{=}10)$	$\hat{p}$ $(\tau{=}30)$
(B)	pooled	0.702	0.802	0.603	0.61
	open-minded	0.706	0.806	0.493	0.502
	norrow-minded	0.693	0.794	0.821	0.826
(D)	pooled	0.702	0.841	0.739	0.659
	open-minded	0.693	0.846	0.668	0.649
	norrow-minded	0.711	0.837	0.812	0.668
(E)	pooled	0.721	0.787	0.643	0.641
	open-minded	0.724	0.797	0.507	0.498
	norrow-minded	0.719	0.776	0.782	0.784
(F)	pooled	0.702	0.818	0.884	0.899
	open-minded	0.713	0.829	0.918	0.936
	norrow-minded	0.681	0.796	0.812	0.825
(H)	pooled	0.72	0.814	0.706	0.558
	open-minded	0.719	0.82	0.711	0.579
	norrow-minded	0.723	0.803	0.695	0.517

TABLE 4. Open and norrow-minded partisans - Result 4

centrality imbalance). In expected terms, the benefit of having a central partisan that induces the underlying true state is offset by the costs of having the opposite situation of the polar opposite partisan inducing misinformation. In network (F), OM means that partisans will moderate quickly as they are connected directly to each other but can still shield society from initial misleading signals, as discussed. Finally, in network (H), OM avoids centrality imbalance.

Another case of interest is the one in which agents are connected through a line.

Result 5 (line networks)**.** *In expected terms, for biased agents connected through any sufficiently long line network ($n \geq 3$), high partisanship ($\tau > 1$) reduces the odds of reaching the less biased consensus. Moreover, for any given level of partisanship $\tau > 0$, a longer line (higher n) increases the odds of reaching the less biased consensus.*

This can be seen in two ways: (i) there is a statistically significant difference between proportions in Table 3 in both networks (B) and (D) for the proportions when $\tau > 1$ compared to the ones under $\tau \leq 1$, and (ii) the coefficients of the dummy variable $\mathbb{1}\{\tau = 30\}$ are negative for the referred networks.

A final result is related to the higher odds of reaching the less biased consensus when nodes are equally central and partisanship is high.

Result 6 (regular networks)**.** *In expected terms, for biased agents connected through any regular network, higher levels of partisanship increase the odds of reaching the less biased consensus.*

Networks (A), (C), (F), and (G) are all regular. Table 3 shows that for any of these networks, $\hat{p}$ increases as τ increases. Moreover, coefficients of the dummy variables for the levels $\tau = 1$ to $\tau = 30$ are in increasing order. This is because no partisan agent outweighs the opposing partisan easily in terms of influence. Thus, partisanship moderates the interpreting dispute by keeping the society close to the center of the 0-1 spectrum long enough, so informative signals accumulate and nudge the society toward the less-biased consensus.

6. Conclusions

Confirmation bias is one of the most notorious cognitive biases documented and, as it is a systematic deviation from rationality, have a significant influence in the process of belief formation. In this sense, as social networks appear as a primary tool for many people to get informed and debate their worldviews, one could expect confirmatory bias to have some influence on the opinion formation. However, to date, how such phenomenon influences opinions in a networked environment has not been understood. To explore this topic, we consider a social learning model in which a fraction of signals external to the social network is ambiguous and open to idiosyncratic interpretation. The interpretation of these signals is affected by people's confirmatory biases. Moreover, we also allow agents to be influenced by their friends and set their beliefs to be a linear combination of the (biased) Bayesian posterior and the (also biased) friends' posteriors.

My model shows that biased agents connected through social networks can only reach two types of consensus and both are biased, one to the left and the other to the right. However, one consensus type is less-biased than the other depending on the state. Moreover, I demonstrate that long-run learning is not attained even if agents are impartial when interpreting ambiguous signals. Those results contradict Rabin and Schrag (1999) and Fryer Jr et al. (2019) in which long-run learning takes place with a positive probability, and impartiality helps learning the state. Furthermore, the network effect presented, together with signal realizations, reinforces the interpreting "*tug-of-war*" as agents might have their own biases confirmed (or mitigated) by other agents.

Finally, as deriving the probability of emergence of the less-biased consensus is challenging, we relied on Monte Carlo simulations to show its determinants. We show that the presence of partisan agents in societies who suffer from confirmatory bias have two main effects on the expected consensus efficiency: (i) it helps countervail the misinterpretation of initial signals when there degree of partisanship is low and for that it increases expected efficiency; and (ii) exacerbates misinterpretation of signals when the degree of partisanship is high, reducing expected consensus efficiency. Moreover, we also show that open-mindedness of partisan agents, i.e., when partisans agree to exchange opinions with partisans with polar opposite beliefs, might reduce expected consensus efficiency in some social topologies.

These results suggest that policies designed to mitigate partisanship and confirmatory bias effects in social networks have to consider also the positive and negative network externalities induced by them in different settings.

References

Acemoglu, D., K. Bimpikis, and A. Ozdaglar (2014): "Dynamics of information exchange in endogenous social networks," *Theoretical Economics*, 9, 41–97.

Acemoglu, D., M. A. Dahleh, I. Lobel, and A. Ozdaglar (2011): "Bayesian learning in social networks," *The Review of Economic Studies*, 78, 1201–1236.

Acemoglu, D. and A. Ozdaglar (2011): "Opinion dynamics and learning in social networks," *Dynamic Games and Applications*, 1, 3–49.

Acemoglu, D., A. Ozdaglar, and A. ParandehGheibi (2010): "Spread of (mis) information in social networks," *Games and Economic Behavior*, 70, 194–227.

Allahverdyan, A. E. and A. Galstyan (2014): "Opinion dynamics with confirmation bias," *PloS one*, 9, e99557.

Andreoni, J. and T. Mylovanov (2012): "Diverging opinions," *American Economic Journal: Microeconomics*, 4, 209–32.

Andrews, R. J., T. D. Logan, and M. J. Sinkey (2018): "Identifying confirmatory bias in the field: Evidence from a poll of experts," *Journal of Sports Economics*, 19, 50–81.

Azzimonti, M. and M. Fernandes (2022): "Social media networks, fake news, and polarization," *European Journal of Political Economy*, 102256.

Bala, V. and S. Goyal (1998): "Learning from neighbours," *The review of economic studies*, 65, 595–621.

——— (2001): "Conformism and diversity under social learning," *Economic theory*, 17, 101–120.

Baliga, S., E. Hanany, and P. Klibanoff (2013): "Polarization and ambiguity," *American Economic Review*, 103, 3071–83.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2014): "Gossip: Identifying central individuals in a social network," Tech. rep., National Bureau of Economic Research.

Banerjee, A. and D. Fudenberg (2004): "Word-of-mouth learning," *Games and economic behavior*, 46, 1–22.

Banerjee, A. V. (1992): "A simple model of herd behavior," *The quarterly journal of economics*, 107, 797–817.

——— (1993): "The economics of rumours," *The Review of Economic Studies*, 60, 309–327.

Bowen, R., D. Dmitriev, and S. Galperti (2021): "Learning from shared news: when abundant information leads to belief polarization," Tech. rep., National Bureau of Economic Research.

Buechel, B., S. Klössner, F. Meng, and A. Nassar (2022): "Misinformation due to asymmetric information sharing," *Université de Fribourg Working Papers SES, N.528, Vl. 2022.*

Dandekar, P., A. Goel, and D. T. Lee (2013): "Biased assimilation, homophily, and the dynamics of polarization," *Proceedings of the National Academy of Sciences*, 110, 5791–5796.

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

DeGroot, M. H. and M. J. Schervish (2012): *Probability and statistics*, Pearson Education.

DeMarzo, P. M., D. Vayanos, and J. Zwiebel (2003): "Persuasion bias, social influence, and unidimensional opinions," *The Quarterly journal of economics*, 118, 909–968.

Ellison, G. and D. Fudenberg (1993): "Rules of thumb for social learning," *Journal of political Economy*, 101, 612–643.

Ellsberg, D. (1961): "Risk, ambiguity, and the Savage axioms," *The quarterly journal of economics*, 643–669.

Epstein, L. G., J. Noor, A. Sandroni, et al. (2010): "Non-bayesian learning," *The BE Journal of Theoretical Economics*, 10, 1–20.

Epstein, L. G. and M. Schneider (2007): "Learning under ambiguity," *The Review of Economic Studies*, 74, 1275–1303.

Fryer, R., M. O. Jackson, et al. (2008): "A categorical model of cognition and biased decision-making," *BE Journal of Theoretical Economics*, 8, 1–42.

Fryer Jr, R. G., P. Harms, and M. O. Jackson (2019): "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization," *Journal of the European Economic Association*, 17, 1470–1501.

Furnham, A. and J. Marks (2013): "Tolerance of ambiguity: A review of the recent literature," *Psychology*, 4, 717–728.

Furnham, A. and T. Ribchester (1995): "Tolerance of ambiguity: A review of the concept, its measurement and applications," *Current psychology*, 14, 179–199.

Gale, D. and S. Kariv (2003): "Bayesian learning in social networks," *Games and economic behavior*, 45, 329–346.

Gallo, E. and A. Langtry (2020): "Social networks, confirmation bias and shock elections," *arXiv preprint arXiv:2011.00520.*

Gennaioli, N. and A. Shleifer (2010): "What comes to mind," *The Quarterly journal of economics*, 125, 1399–1433.

Gilboa, I. and D. Schmeidler (1989): "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 18, 141–153.

——— (1993): "Updating ambiguous beliefs," *Journal of economic theory*, 59, 33–49.

Glaeser, E. L. and C. R. Sunstein (2013): "Why does balanced news produce unbalanced views?" Tech. rep., National Bureau of Economic Research.

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

Golub, B. and E. Sadler (2017): "Learning in social networks," *Available at SSRN 2919146*.

Grabisch, M. and A. Rusinowska (2020): "A survey on nonstrategic models of opinion dynamics," *Games*, 11, 65.

Han, P. K., B. J. Zikmund-Fisher, C. W. Duarte, M. Knaus, A. Black, A. M. Scherer, and A. Fagerlin (2018): "Communication of scientific uncertainty about a novel pandemic health threat: Ambiguity aversion and its mechanisms," *Journal of health communication*, 23, 435–444.

Hegselmann, R. and U. Krause (2005): "Opinion dynamics driven by various ways of averaging," *Computational Economics*, 25, 381–405.

Hegselmann, R., U. Krause, et al. (2002): "Opinion dynamics and bounded confidence models, analysis, and simulation," *Journal of artificial societies and social simulation*, 5.

Hellman, M. E. and T. M. Cover (1970): "Learning with Finite Memory," *The Annals of Mathematical Statistics*, 41, 765 – 782.

Jackson, M. O., E. Kalai, and R. Smorodinsky (1999): "Bayesian representation of stochastic processes under learning: de Finetti revisited," *Econometrica*, 67, 875–893.

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

Kalai, E. and E. Lehrer (1994): "Weak and strong merging of opinions," *Journal of Mathematical Economics*, 23, 73–86.

Lord, C. G., L. Ross, and M. R. Lepper (1979): "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." *Journal of personality and social psychology*, 37, 2098.

Mahler, R. P. (1995): "Combining ambiguous evidence with respect to ambiguous a priori knowledge. Part II: Fuzzy Logic," *Fuzzy Sets and Systems*, 75, 319–354.

Mao, Y., S. Bolouki, and E. Akyol (2018): "Spread of information with confirmation bias in cyber-social networks," *IEEE Transactions on Network Science and Engineering*, 7, 688–700.

Mercier, H. and D. Sperber (2011): "Why do humans reason? Arguments for an argumentative theory." *Behavioral and brain sciences*, 34, 57–74.

Meyer, C. D. (2000): *Matrix analysis and applied linear algebra*, vol. 71, Siam.

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

Molden, D. C. and E. T. Higgins (2004): "Categorization under uncertainty: Resolving vagueness and ambiguity with eager versus vigilant strategies," *Social Cognition*, 22, 248–277.

——— (2008): "How preferences for eager versus vigilant judgment strategies affect self-serving conclusions," *Journal of Experimental Social Psychology*, 44, 1219–1228.

Mullainathan, S. (2002): "A memory-based model of bounded rationality," *The Quarterly Journal of Economics*, 117, 735–774.

Nickerson, R. S. (1998): "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, 2, 175–220.

Rabin, M. and J. L. Schrag (1999): "First impressions matter: A model of confirmatory bias," *The quarterly journal of economics*, 114, 37–82.

Sherman, D. K. and G. L. Cohen (2006): "The psychology of self-defense: Self-affirmation theory," *Advances in experimental social psychology*, 38, 183–242.

Siegrist, K. (2021): "Probability, Mathematical Statistics, Stochastic Processes," *Available online at:* `https://stats.libretexts.org/Bookshelves/Probability_Theory` *(Accessed 07.01.2022).*

Sikder, O., R. E. Smith, P. Vivo, and G. Livan (2020): "A minimalistic model of bias, polarization and misinformation in social networks," *Scientific reports*, 10, 1–11.

Simonovic, N. and J. M. Taber (2022): "Psychological impact of ambiguous health messages about COVID-19," *Journal of Behavioral Medicine*, 45, 159–171.

Sinkey, M. (2015): "How do experts update beliefs? Lessons from a non-market environment," *Journal of Behavioral and Experimental Economics*, 57, 55–63.

Wilson, A. (2014): "Bounded memory and biases in information processing," *Econometrica*, 82, 2257–2294.

Appendix A. Beta-Bernoulli model and likelihood function of interpreted signals

At any time t, the belief of agent i is represented by the Beta probability distribution with parameters $\alpha_{i,t}$ and $\beta_{i,t}$

$$f_{i,t}\left(\theta\right) = \begin{cases} \dfrac{\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}\,\theta^{\alpha_{i,t}-1}(1-\theta)^{\beta_{i,t}-1} & \text{, for } 0<\theta<1 \\ 0 & \text{, otherwise,} \end{cases} \tag{10}$$

where $\Gamma(\cdot)$ is a Gamma function and the ratio of Gamma functions in the expression above is a normalization constant that ensures that the total probability integrates to 1. In this sense,

$$f_{i,t}\left(\theta\right) \propto \theta^{\alpha_{i,t}-1}(1-\theta)^{\beta_{i,t}-1}.$$

The idiosyncratic likelihood induced by the agent i's interpretation of the public signal s_{t+1} is

$$\ell_i(s_{t+1}|\theta) = \theta^{s^{(1)}_{i,t+1}}(1-\theta)^{s^{(0)}_{i,t+1}}$$

and, therefore, the standard Bayesian posterior is computed as

$$f_{i,t+1}(\theta|s_{t+1}) = \frac{\ell_i(s_{t+1}|\theta)\,f_{i,t}\left(\theta\right)}{\displaystyle\int_{\Theta}\ell_i(s_{t+1}|\theta)\,f_{i,t}\left(\theta\right)d\theta}.$$

Since the denominator of the expression above is just a normalizing constant, the posterior distribution is said to be proportional to the product of the prior distribution and the likelihood function as

$$\begin{aligned} f_{i,t+1}(\theta|s_{t+1}) &\propto \ell_i(s_{t+1}|\theta)\,f_{i,t}\left(\theta\right) \\ &\propto \theta^{\alpha_{i,t}+s^{(1)}_{i,t+1}-1}\left(1-\theta\right)^{\beta_{i,t}+s^{(0)}_{i,t+1}-1}. \end{aligned}$$

Therefore, the posterior distribution is

$$f_{i,t+1}\left(\theta\right) = \begin{cases} \dfrac{\Gamma\left(\alpha_{i,t+1}+\beta_{i,t+1}\right)}{\Gamma\left(\alpha_{i,t+1}\right)\Gamma\left(\beta_{i,t+1}\right)}\,\theta^{\alpha_{i,t+1}-1}(1-\theta)^{\beta_{i,t+1}-1} & \text{, for } 0<\theta<1 \\ 0 & \text{, otherwise,} \end{cases}$$

where $\alpha_{i,t+1} = \alpha_{i,t} + s^{(1)}_{i,t+1}$ and $\beta_{i,t+1} = \beta_{i,t} + s^{(0)}_{i,t+1}$.

Appendix B. Beta Distribution: Mode, Mean, Median

Mode. The mode of a random variable beta-distributed is the value that appears most often. It is the value θ at which its probability density function takes its maximum value. As per Equation (10), the mode $\theta^{mod}_{i,t}$, for any i at any point in time t, is the $\arg\max_\theta f_{i,t}(\theta)$. Computed as

$$\frac{df_{i,t}}{d\theta}=\frac{\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}\left[(\alpha_{i,t}-1)\theta^{\alpha_{i,t}-2}(1-\theta)^{\beta_{i,t}-1}-\theta^{\alpha_{i,t}-1}(\beta_{i,t}-1)(1-\theta)^{\beta_{i,t}-2}\right]=0.$$

Implying that

$$(\alpha_{i,t}-1)\theta^{\alpha_{i,t}-2}(1-\theta)^{\beta_{i,t}-1}-\theta^{\alpha_{i,t}-1}(\beta_{i,t}-1)(1-\theta)^{\beta_{i,t}-2}=0,$$

and therefore

$$\theta^{mod}_{i,t}=\begin{cases}\dfrac{\alpha_{i,t}-1}{\alpha_{i,t}+\beta_{i,t}-2} & \text{, for } \alpha_{i,t},\beta_{i,t}>1\\ 0 & \text{, for } \alpha_{i,t}=1,\beta_{i,t}>1\\ 1 & \text{, for } \alpha_{i,t}>1,\beta_{i,t}=1\\ \text{any value in } (0,1) & \text{, for } \alpha_{i,t},\beta_{i,t}=1\end{cases} \tag{11}$$

Mean. The mean of a random variable Beta-distributed, denoted by $\theta^{mean}_{i,t}$ for any i and t, is computed as follows

$$\begin{aligned}\theta^{mean}_{i,t}&=\int_0^1\theta\frac{\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}\theta^{\alpha_{i,t}-1}(1-\theta)^{\beta_{i,t}-1}d\theta\\&=\frac{\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}\int_0^1\theta^{(\alpha_{i,t}+1)-1}(1-\theta)^{\beta_{i,t}-1}d\theta\\&=\frac{\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}\frac{\Gamma\left(\alpha_{i,t}+1\right)\Gamma\left(\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}+\beta_{i,t}+1\right)}\\&=\frac{\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}\frac{\alpha_{i,t}\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}{(\alpha_{i,t}+\beta_{i,t})\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}=\frac{\alpha_{i,t}}{\alpha_{i,t}+\beta_{i,t}}.\end{aligned} \tag{12}$$

Median. There is no general closed formula for the median of the beta distribution for arbitrary values of the parameter $\alpha_{i,t}$ and $\beta_{i,t}$. The median, denoted by $\theta^{med}_{i,t}$, is the function that satisfies

$$\frac{\Gamma\left(\alpha_{i,t}+\beta_{i,t}\right)}{\Gamma\left(\alpha_{i,t}\right)\Gamma\left(\beta_{i,t}\right)}\int_0^{\theta^{med}_{i,t}}\theta^{\alpha_{i,t}-1}(1-\theta)^{\beta_{i,t}-1}=\frac{1}{2}.$$

An accurate approximation of the value of the median of the beta distribution, for both $\alpha_{i,t},\beta_{i,t}\geq 1$, is given by

$$\theta^{med}_{i,t}=\frac{\alpha_{i,t}-\frac{1}{3}}{\alpha_{i,t}+\beta_{i,t}-\frac{2}{3}}.^{11} \tag{13}$$

Therefore, if $1<\alpha_{i,t}<\beta_{i,t}$, then $\theta^{mod}_{i,t}<\theta^{med}_{i,t}<\theta^{mean}_{i,t}$. If $1<\beta_{i,t}<\alpha_{i,t}$, then the order of the inequalities is reversed. Finally, it is trivial to see that those three statistical measures are asymptotically equal as $\alpha_{i,t},\beta_{i,t}\to\infty$.

[11] With relative error of less than 4%, rapidly decreasing to zero as both shape parameters increase.

Appendix C. Auxiliary Definitions and Lemmas

Proof of Lemma 1. In order to see how W^t behaves as t grows large, I rewrite W using its diagonal decomposition. In particular, let v be the squared matrix of left-hand eigenvectors of W and $D = (d_1, d_2, \dots, d_n)^\top$ the eigenvector of size n associated to the unity eigenvalue $\lambda_1 = 1$. Without loss of generality, we assume the following normalization $\mathbf{1}^\top D = 1$. Therefore, $W = v^{-1}\Lambda v$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is the squared matrix with eigenvalues on its diagonal, ranked in terms of absolute values, i.e. $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. More generally, for any time t we write

$$W^t = v^{-1}\Lambda^t v.$$

Since v^{-1} has ones in all entries of its first column, it follows that

$$W^t_{ij} = d_j + \sum_r \lambda_r^t v_{ir}^{-1} v_{rj},$$

for each r, where λ_r is the r-th largest eigenvalue of W. Therefore, $\lim_{t\to\infty} W^t_{ij} = D\mathbf{1}^\top$, i.e. each row of W^t for all $t \geq \bar{t}$ converge to D, which coincides with the stationary distribution. Moreover, if the eigenvalues are ordered the way we have assumed, then $\|W^t - D\mathbf{1}^\top\| = o(|\lambda_2|^t)$, i.e. the convergence rate will be dictated by the second largest eigenvalue, as the others converge to zero more quickly as t grows. ■

Lemma 2. *The opinion of every agent i in any point in time t, $y_{i,t}$, can be written as*

$$y_{i,t} = \frac{\sum_{j=1}^n W^t_{ij}\alpha_{j,0} + bK(i,t)}{\sum_{j=1}^n W^t_{ij}\left(\alpha_{j,0} + \beta_{j,0}\right) + bL(i,t)},$$

where $K(i,t) = \sum_{k=0}^{t-1}\sum_{j=1}^{n} W^k_{ij} s^{(1)}_{j,t-k}$ *and* $L(i,t) = \sum_{k=0}^{t-1}\sum_{j=1}^{n} W^k_{ij}\left(s^{(0)}_{j,t-k} + s^{(1)}_{j,t-k}\right)$.

Proof. The update process of both parameters described by the equations (7) and (8) can be solved iteratively for any period t as

$$\alpha_t = W^t\alpha_0 + \sum_{k=0}^{t-1} W^k B s^{(1)}_{t-k} \tag{14}$$

$$\beta_t = W^t\beta_0 + \sum_{k=0}^{t-1} W^k B s^{(0)}_{t-k}. \tag{15}$$

In agebraic formulation, we have that each entry of the vector in equation (14) can be written as

$$\begin{aligned}\alpha_{i,t} &= \sum_{j=1}^{n} W_{ij}^{t}\alpha_{j,0} + \sum_{k=0}^{t-1}\sum_{j=1}^{n} W_{ij}^{k} b s_{j,t-k}^{(1)} \\ &= \sum_{j=1}^{n} W_{ij}^{t}\alpha_{j,0} + b\sum_{k=0}^{t-1}\sum_{j=1}^{n} W_{ij}^{k} s_{j,t-k}^{(1)} \\ &= \sum_{j=1}^{n} W_{ij}^{t}\alpha_{j,0} + b\,\sum_{k=0}^{t-1}\sum_{j=1}^{n} W_{ij}^{k} s_{j,t-k}^{(1)} \\ &= \sum_{j=1}^{n} W_{ij}^{t}\alpha_{j,0} + b\,K(i,t). \end{aligned} \tag{16}$$

Similarly for the expression $\alpha_{i,t} + \beta_{i,t}$ using both equations (14) and (15) as follows

$$\begin{aligned}\alpha_{i,t} + \beta_{i,t} &= \sum_{j=1}^{n} W_{ij}^{t}\left(\alpha_{j,0} + \beta_{j,0}\right) + b\,\sum_{k=0}^{t-1}\sum_{j=1}^{n} W_{ij}^{k}\left(s_{j,t-k}^{(0)} + s_{j,t-k}^{(1)}\right) \\ &= \sum_{j=1}^{n} W_{ij}^{t}\left(\alpha_{j,0} + \beta_{j,0}\right) + b\,L(i,t). \end{aligned} \tag{17}$$

Therefore, from the definition of opinion we have that $y_{i,t} = \frac{\alpha_{i,t}}{\alpha_{i,t}+\beta_{i,t}}$ and the statement is proven. ■

Lemma 3. *Let $k \in [0,1]$, $X_1, X_2, \ldots, X_t$ be a sequence of i.n.i.d. random variables such that $\mathbb{P}(X_t \geq x) = p$ and $u_1, u_2, \ldots, u_t$ be i.i.d. $U[0,1]$ random variables. Moreover, assume that the pair (X_t, u_t) is independent, for any t. In this case, the expressions $\mathbb{E}\left[\mathbb{1}\{u_t \leq \mathbb{1}\{X_t \geq x\}k\}\right]$ and $\mathbb{E}\left[\mathbb{1}\{u_t \leq \mathbb{E}\left[\mathbb{1}\{X_t \geq x\}\right]k\}\right]$ are equal.*

Proof. The first expression can be written as

$$\mathbb{E}\left[\mathbb{1}\{u_t \leq \mathbb{1}\{X_t \geq x\}k\}\right] = (1-p)\mathbb{E}\left[\mathbb{1}\{u_t \leq 0\}\right] + p\mathbb{E}\left[\mathbb{1}\{u_t \leq k\}\right] = pF_u(k) = pk.$$

The second expression simplifies to

$$\mathbb{E}\left[\mathbb{1}\{u_t \leq \mathbb{E}\left[\mathbb{1}\{X_t \geq x\}\right]k\}\right] = \mathbb{E}\left[\mathbb{1}\{u_t \leq (1-p)0 + pk\}\right] = \mathbb{E}\left[\mathbb{1}\{u_t \leq pk\}\right] = pk.$$

■

Lemma 4 (Convergence). *The sequences $\{\{y_{i,t}\}_{i=1}^{n}\}_{t=1}^{\infty}$ generated by the update rule converge almost surely as $t \to \infty$.*

Proof. For the individual case, Siegrist (2021) (Section 12.8.5) shows that there is an equivalence between the Beta-Bernoulli process and the Pólya's urn process. In the Pólya's urn proccess the sequence of random variables (drawn balls' colors) is not independent, but is exchangeable. Thus, the joint distribution of the interpreted signals (colors) is invariant under a permutation. Thus, the sequence of the proportion of signals interpreted as 1 is a martingale, and standard martingale convergence theorems ensure the convergence of this process.

For the networked case, Lemmas (1) and (2) in Jadbabaie et al. (2012) prove convergence of this process for a general case based on the same assumption that the social interaction matrix W is strongly connected and, for that, it always has at least one eigenvalue equal to 1 and that there exists a non-negative left eigenvector v corresponding to this eigenvalue. As a result, they show that $\sum_{i=1}^{n} v_i f_{i,t}(\theta^*)$ is a submartingale with respect to the filtration $\mathcal{F}_t$ (interpreted signals).

■

Appendix D. Proofs of main propositions and corollaries

Proof of Proposition 1.

$$
\begin{aligned}
\lim_{t\to\infty} y_{i,t} &= \lim_{t\to\infty} \frac{\alpha_{i,0} + \sum_{k=1}^{t} s_{i,k}^{(1)}}{\alpha_{i,0} + \sum_{k=1}^{t} s_{i,k}^{(1)} + \beta_{i,0} + \sum_{k=1}^{t} s_{i,k}^{(0)}} \\
&= \lim_{t\to\infty} \frac{\alpha_{i,0} + \sum_{k=1}^{t} \left(\mathbb{1}\{s_k = 1\} + \mathbb{1}\{s_k = a\}\mathbb{1}\{u_k \le \psi_{i,k}\}\right)}{\alpha_{i,0} + \beta_{i,0} + \sum_{k=1}^{t} \left(\mathbb{1}\{s_k = 1\} + \mathbb{1}\{s_k = 0\} + \mathbb{1}\{s_k = a\}\right)} \\
&= \lim_{t\to\infty} \frac{\frac{\alpha_{i,0}}{t} + \frac{1}{t}\sum_{k=1}^{t} \left(\mathbb{1}\{s_k = 1\} + \mathbb{1}\{s_k = a\}\mathbb{1}\{u_k \le \psi_{i,k}\}\right)}{\frac{\alpha_{i,0}+\beta_{i,0}}{t} + \frac{1}{t}\sum_{k=1}^{t} \left(\mathbb{1}\{s_k = 1\} + \mathbb{1}\{s_k = 0\} + \mathbb{1}\{s_k = a\}\right)} \\
&= \frac{\mathbb{E}_t\left[\mathbb{1}\{s_t = 1\}\right] + \mathbb{E}_t\left[\mathbb{1}\{s_t = a\}\right] \lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} \left(\mathbb{1}\{u_k \le \psi_{i,k}\}\right)}{\mathbb{E}_t\left[(\mathbb{1}\{s_t = 1\}\right] + \mathbb{E}_t\left[\mathbb{1}\{s_t = 0\}\right] + \mathbb{E}_t\left[\mathbb{1}\{s_t = a\}\right])} \\
&= (1-\mu)\theta + \mu \lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} \left(\mathbb{1}\{u_k \le \psi_{i,k}\}\right) \\
&= (1-\mu)\theta + \mu \lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} \left(\mathbb{1}\{u_k \le \gamma_{i,k}\, \mathbb{1}\{y_{i,k-1} \ge 0.5\} + (1-\gamma_{i,k})\, \mathbb{1}\{y_{i,k-1} < 0.5\}\}\right) \\
&= (1-\mu)\theta + \mu \mathbb{E}_t\left[\mathbb{1}\{u_t \le \mathbb{E}_t\left[\gamma_i\, \mathbb{1}\{y_{i,t-1} \ge 0.5\} + (1-\gamma_i)\, \mathbb{1}\{y_{i,t-1} < 0.5\}\}\right]\right] \\
&= (1-\mu)\theta + \mu \mathbb{E}_t\left[\mathbb{1}\{u_t \le \mathbb{E}_t\left[\mathbb{1}\{y_{i,t-1} \ge 0.5\}\right](2\gamma_i - 1) + 1 - \gamma_i\}\right]
\end{aligned}
$$

According to Lemma 4, convergence ensures that $\mathbb{E}_t\left[\mathbb{1}\{y_{i,t-1} \ge 0.5\} = \mathbb{P}\left(y_{i,\infty} \ge 0.5\right)\right]$ either takes on value 1 or 0. For simplicity, say the first case is denoted by A, and the second by B.

Therefore,

$$\begin{aligned}\lim_{t\to\infty} y_{i,t} &= \begin{cases}(1-\mu)\theta+\mu\mathbb{E}_t\left[\mathbb{1}\{u_t\leq\gamma_i\}\right] & \text{, if } A\\ (1-\mu)\theta+\mu\mathbb{E}_t\left[\mathbb{1}\{u_t\leq 1-\gamma_i\}\right] & \text{, if } B\end{cases}\\ &= \begin{cases}(1-\mu)\theta+\mu F_u\left(\gamma_i\right) & \text{, if } A\\ (1-\mu)\theta+\mu F_u\left(1-\gamma_i\right) & \text{, if } B\end{cases}\\ &= \begin{cases}(1-\mu)\theta+\mu\gamma_i & \text{, if } A\\ (1-\mu)\theta+\mu\left(1-\gamma_i\right) & \text{, if } B\end{cases}\end{aligned} \tag{18}$$

■

Proof of Proposition 2. The claim is supported by the solution of two systems of inequalities S_1 (for right-biased opinion) and S_2 (for left-biased opinion) below.

$$S_1=\begin{cases}(1-\mu)\theta+\mu\gamma_i>\frac{1}{2}\\ (1-\mu)\theta+\mu(1-\gamma_i)>\frac{1}{2}\\ 0<\mu\leq 1\\ 0\leq\theta\leq 1\\ \frac{1}{2}<\gamma_i\leq 1\end{cases} \qquad S_2=\begin{cases}(1-\mu)\theta+\mu\gamma_i<\frac{1}{2}\\ (1-\mu)\theta+\mu(1-\gamma_i)<\frac{1}{2}\\ 0<\mu\leq 1\\ 0\leq\theta\leq 1\\ \frac{1}{2}<\gamma_i\leq 1\end{cases}$$

The solution of those systems, together with the equation (18) in Proof of proposition 1 ensure the uniqueness of opinion types in the parameter spaces defined in the statement. ■

Proof of Corollary 1. From Proposition 1, we can write both right-biased and left-biased opinions as $\theta+\mu(\gamma_i-\theta)$ and $\theta+\mu(1-\gamma_i-\theta)$, respectively, where the second term in each expression represents their respective biases. From those expressions, we can see that both sign and magnitude of those biases naturally depend on the relative size of θ and γ_i. For both biases to be positive, we need $\theta<\min\{\gamma_i,1-\gamma_i\}=1-\gamma_i$, since $\gamma_i>\frac{1}{2}$. For both biases to be negative, we need $\theta>\max\{\gamma_i,1-\gamma_i\}=\gamma_i$, since $\gamma_i>\frac{1}{2}$. For the right-bias to be positive and the left-bias to be negative, we need $1-\gamma_i<\theta<\gamma_i$ to hold. The case in which the right bias is negative while the right-bias is positive never holds, since we assume $\gamma_i>\frac{1}{2}$. Therefore, we have the following summary.

(1) if $\theta<1-\gamma_i$, then both biases are strictly positive
(2) if $1-\gamma_i<\theta<\gamma_i$, then right-bias is strictly positive and left-bias is strictly negative
(3) if $\theta>\gamma_i$, then both biases are strictly negative.

In the case (1) listed above, we say that the right-bias is less than the left bias whenever $\mu(\gamma_i - \theta) < \mu(1 - \gamma_i - \theta)$, meaning that $\gamma_i < \frac{1}{2}$. However, this contradicts the assumption that individual is confirmatory and we can conclude that whenever $\theta < 1 - \gamma_i$, the left-biased opinion is less biased than the right-biased one. In the case (3), we say that the right-bias is less than the left bias whenever $\mu(\theta - \gamma_i) < \mu(\gamma_i + \theta - 1)$, meaning that the statement is true if $\gamma_i > \frac{1}{2}$. Therefore, if $\theta > \gamma_i$, the right-biased opinion is less biased than the left-biased one. Finally, in the case (2), we say that the right-bias is less than the left bias whenever $\mu(\gamma_i - \theta) < \mu(\gamma_i + \theta - 1)$, meaning that it can only be true when $\theta > \frac{1}{2}$. These three arguments together prove the statement and we conclude that the right-bias is less than the left bias whenever $\theta > \frac{1}{2}$ (and vice-versa).

Finally, when $\theta = \frac{1}{2}$, the biases are equal since $|\gamma_i - \frac{1}{2}| = |\frac{1}{2} - \gamma_i|$ for any γ_i. ■

Proof of Corollary 2. When an individual j is always impartial, we have that

$$\begin{aligned}\psi_{j,t} &= \frac{1}{2}\,\mathbb{1}\{y_{j,t-1} \geq 0.5\} + \frac{1}{2}\,\mathbb{1}\{y_{j,t-1} < 0.5\} \\ &= \frac{1}{2}\,\mathbb{1}\{y_{j,t-1} \geq 0.5\} + \frac{1}{2}\,(1 - \mathbb{1}\{y_{j,t-1} \geq 0.5\}) \\ &= \frac{1}{2}, \end{aligned} \tag{19}$$

for all t. Since u_t is a continuous $U\,[0,1]$ random variable in every period t, we have that

$$\mathbb{E}_t\left[\mathbb{1}\left\{u_t \leq \frac{1}{2}\right\}\right] = \mathbb{P}\left(u_t \leq \frac{1}{2}\right) = F_u\left(\frac{1}{2}\right) = \frac{\frac{1}{2} - 0}{1 - 0} = \frac{1}{2}, \tag{20}$$

where $F_u(\cdot)$ is the cumulative distribution function of $U\,[0,1]$. Thus, equations (18) and (20) together prove the statement when agents are impartial (both always impartial and moderately impartial). ■

Proof of Corollary 3. Say extreme opinion 1 (i.e. $y_{i,\infty} = 1$) is formed, then as per Propositions 1 and 2 we know this is the right-biased opinion and therefore it should be the case that $(1-\mu)\theta + \mu\gamma_i = 1$. Conversely, say extreme opinion 0 (i.e. $y_{i,\infty} = 0$) is formed. Then, we know this is the left-biased opinion and it should be that $(1-\mu)\theta + \mu(1-\gamma_i) = 0$. These two conditions together imply that $\mu(2\gamma_i - 1) = 1$. If we generally consider that $0 \leq \mu \leq 1$ and $\frac{1}{2} \leq \gamma_i \leq 1$, then the relation $\mu(2\gamma_i - 1) = 1$ is only met when $\mu = \gamma_i = 1$. ■

Proof of Proposition 4. As per Lemma 2 in the Appendix C, the limiting opinion of any agent i can be written as

$$\lim_{t\to\infty} y_{i,t} = \lim_{t\to\infty} \frac{\frac{1}{t}\sum_{j=1}^{n} W_{ij}^{t}\alpha_{j,0} + b\frac{1}{t}K(i,t)}{\frac{1}{t}\sum_{j=1}^{n} W_{ij}^{t}\left(\alpha_{j,0}+\beta_{j,0}\right) + b\frac{1}{t}L(i,t)}$$

$$= \lim_{t\to\infty} \frac{\dfrac{1}{t}\displaystyle\sum_{k=0}^{t-1}\sum_{j=1}^{n} W_{ij}^{k} s_{j,t-k}^{(1)}}{\dfrac{1}{t}\displaystyle\sum_{k=0}^{t-1}\sum_{j=1}^{n} W_{ij}^{k}\left(s_{j,t-k}^{(0)} + s_{j,t-k}^{(1)}\right)}.$$

By Lemma 1 we can split both series in the numerator and denominator in two parts

$$\lim_{t\to\infty} y_{i,t} = \lim_{t\to\infty} \frac{\dfrac{1}{t}\left(\displaystyle\sum_{k=0}^{t_{\text{mix}}}\sum_{j=1}^{n} W_{ij}^{k} s_{j,t-k}^{(1)} + \sum_{k=t_{\text{mix}}+1}^{t-1}\sum_{j=1}^{n} W_{ij}^{k} s_{j,t-k}^{(1)}\right)}{\dfrac{1}{t}\left(\displaystyle\sum_{k=0}^{t_{\text{mix}}}\sum_{j=1}^{n} W_{ij}^{k}\left(s_{j,t-k}^{(0)} + s_{j,t-k}^{(1)}\right) + \sum_{k=t_{\text{mix}}+1}^{t-1}\sum_{j=1}^{n} W_{ij}^{k}\left(s_{j,t-k}^{(0)} + s_{j,t-k}^{(1)}\right)\right)}$$

$$= \lim_{t\to\infty} \frac{\dfrac{1}{t}\displaystyle\sum_{k=t_{\text{mix}}+1}^{t-1}\sum_{j=1}^{n} W_{ij}^{k} s_{j,t-k}^{(1)}}{\dfrac{1}{t}\displaystyle\sum_{k=t_{\text{mix}}+1}^{t-1}\sum_{j=1}^{n} W_{ij}^{k}\left(s_{j,t-k}^{(0)} + s_{j,t-k}^{(1)}\right)}.$$

Since the subindex k spans from t_{mix} onwards (i.e. when the chain is already mixed), we can use the invariant distribution matrix in the previous expression. Therefore the limiting opinion becomes

$$
\begin{aligned}
\lim_{t\to\infty} y_{i,t} &= \lim_{t\to\infty} \frac{\sum_{j=1}^{n} \Pi_{ij} \frac{1}{t} \sum_{k=t_{\text{mix}}+1}^{t-1} s^{(1)}_{j,t-k}}{\sum_{j=1}^{n} \Pi_{ij} \frac{1}{t} \sum_{k=t_{\text{mix}}+1}^{t-1} \left(s^{(0)}_{j,t-k} + s^{(1)}_{j,t-k}\right)} \\
&= \frac{\sum_{j=1}^{n} \Pi_{ij} \lim_{t\to\infty} \frac{t-1-t_{\text{mix}}}{t} \frac{1}{t-1-t_{\text{mix}}} \sum_{k=t_{\text{mix}}+1}^{t-1} s^{(1)}_{j,t-k}}{\sum_{j=1}^{n} \Pi_{ij} \lim_{t\to\infty} \frac{t-1-t_{\text{mix}}}{t} \frac{1}{t-1-t_{\text{mix}}} \sum_{k=t_{\text{mix}}+1}^{t-1} \left(s^{(0)}_{j,t-k} + s^{(1)}_{j,t-k}\right)} \\
&= \frac{\sum_{j=1}^{n} \Pi_{ij} \lim_{t\to\infty} \frac{1}{t-1-t_{\text{mix}}} \sum_{k=t_{\text{mix}}+1}^{t-1} \left(\mathbb{1}\{s_{t-k}=1\} + \mathbb{1}\{s_{t-k}=a\}\mathbb{1}\{u_{t-k} \le \psi_{j,t-k}\}\right)}{\sum_{j=1}^{n} \Pi_{ij} \lim_{t\to\infty} \frac{1}{t-1-t_{\text{mix}}} \sum_{k=t_{\text{mix}}+1}^{t-1} \left(\mathbb{1}\{s_{t-k}=0\} + \mathbb{1}\{s_{t-k}=1\} + \mathbb{1}\{s_{t-k}=a\}\right)} \\
&= \frac{\sum_j \Pi_{ij} \mathbb{E}_t \left[\mathbb{1}\{s_t=1\} + \mathbb{1}\{s_t=a\}\mathbb{1}\{u_t \le \psi_{j,t}\}\right]}{\sum_j \Pi_{ij} \mathbb{E}_t \left[\mathbb{1}\{s_t=0\} + \mathbb{1}\{s_t=1\} + \mathbb{1}\{s_t=a\}\right]} \\
&= (1-\mu)\theta + \mu \sum_j \Pi_{ij} \mathbb{E}_t \left[\mathbb{1}\{u_t \le \psi_{j,t}\}\right],
\end{aligned}
$$

where the term $\mathbb{E}_t\left[\mathbb{1}\{u_t \le \psi_{j,t}\}\right]$ is as in Proposition 1, implying that the limiting consensus is

$$
\lim_{t\to\infty} y_{i,t} = \begin{cases} (1-\mu)\theta + \mu \sum_j \Pi_{ij}\gamma_j & \text{, if } A \\ (1-\mu)\theta + \mu \sum_j \Pi_{ij}(1-\gamma_j) & \text{, if } B \end{cases}
$$

■

Proof of Proposition 3. From Equation (16) in Appendix C, we know that $\alpha_{i,t}$, for any i, can be iterated forwardly as

$$
\alpha_{i,t} = \sum_{j=1}^{n} W^t_{ij}\alpha_{j,0} + b \sum_{k=0}^{t-1} \sum_{j=1}^{n} W^k_{ij} s^{(1)}_{j,t-k}.
$$

.

Similarly, the expression $\alpha_{i,t} + \beta_{i,t}$ in Equation (17) can be written as

$$
\alpha_{i,t} + \beta_{i,t} = \sum_{j=1}^{n} W^t_{ij}\left(\alpha_{j,0} + \beta_{j,0}\right) + b \sum_{k=0}^{t-1} \sum_{j=1}^{n} W^k_{ij} \left(s^{(0)}_{j,t-k} + s^{(1)}_{j,t-k}\right).
$$

Thus, if $b = 0$, the opinion of any agent $i \in N$ at any time t boils down to

$$y_{i,t} = \frac{\sum_{j=1}^{n} W_{ij}^{t} \alpha_{j,0}}{\sum_{j=1}^{n} W_{ij}^{t} \left(\alpha_{j,0} + \beta_{j,0}\right)}$$

and therefore

$$\lim_{t \to \infty} y_{i,t} = y = \frac{\sum_{j=1}^{n} \Pi_{ij} \alpha_{j,0}}{\sum_{j=1}^{n} \Pi_{ij} \left(\alpha_{j,0} + \beta_{j,0}\right)}$$

for any i. In this case, the limiting opinion of any agent i can be written as in the case when $b = 0$ shown above. ■

Appendix E. Tests concerning differences among proportions

E.1. **Definition.** To decide whether observed differences among sample proportions are significant or whether they can be attributed to chance we must use tests concerning differences among proportions. For that, suppose that $x_1, x_2, \ldots, x_k$ are observed values of k independent random variables $X_1, X_2, ..., X_k$ having binomial distributions with the parameters n_1 and θ_1, n_2 and $\theta_2, \ldots, n_k$ and θ_k. If the sample sizes are sufficiently large, we can approximate the distributions of the independent random variables

$$Z_i = \frac{X_i - n_i\theta_i}{\sqrt{n_i\theta_i(1-\theta_i)}} \qquad \text{for } i = 1, 2, \ldots, k$$

with standard normal distributions. Therefore, we know that we can look upon the test-statistic

$$\chi^2 = \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\theta_i)^2}{n_i\theta_i(1-\theta_i)}$$

as a value of a random variable having chi-square distribution with k degrees of freedom. When the null hypothesis H_0 is $\theta_1 = \theta_2 = \cdots = \theta_k$ and the alternative hypothesis is that at least one of the θ's is different, we can use the *pooled estimate*

$$\hat{\theta} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} n_i}$$

and the test statistic becomes

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\hat{\theta})^2}{n_i\hat{\theta}(1-\hat{\theta})}$$

a random variable whose value has chi-square distribution with $k-1$ degrees of freedom because an estimate is substituted for the unknown parameter θ.

i	j	c_i	c_j	$\hat{p}_i(c_i)$	$\hat{p}_j(c_j)$	$CI_{5\%}$	$CI_{95\%}$	χ^2	p-value
(A)	(E)	$\tau = 0$	$\tau = 0$	0.702	0.721	-0.037	-0.002	4.915	0.027
(A)	(G)	$\tau = 0$	$\tau = 0$	0.702	0.727	-0.042	-0.007	7.871	0.005
(A)	(H)	$\tau = 0$	$\tau = 0$	0.702	0.72	-0.035	-0.001	4.174	0.041
(E)	(G)	$\tau = 0$	$\tau = 0$	0.721	0.727	-0.022	0.012	0.347	0.556
(E)	(H)	$\tau = 0$	$\tau = 0$	0.721	0.72	-0.016	0.019	0.03	0.862
(G)	(H)	$\tau = 0$	$\tau = 0$	0.727	0.72	-0.01	0.024	0.582	0.446

Table 5. Hypothesis Test for Proportions - Result 1

i	j	c_i	c_j	$\hat{p}_i(c_i)$	$\hat{p}_j(c_j)$	$CI_{5\%}$	$CI_{95\%}$	χ^2	p-value
(A)	(A)	$\tau = 0$	$\tau = 30$	0.688	0.678	-0.006	0.026	1.641	0.2
(B)	(B)	$\tau = 0$	$\tau = 30$	0.688	0.59	0.082	0.115	140.542	0
(D)	(D)	$\tau = 0$	$\tau = 30$	0.688	0.648	0.024	0.056	23.694	0
(B)	(D)	$\tau = 1$	$\tau = 1$	0.707	0.766	-0.077	-0.041	41.405	0
(B)	(D)	$\tau = 30$	$\tau = 30$	0.59	0.648	-0.078	-0.039	32.948	0

Table 6. Two Population Proportions - Result 5

Appendix F. Probit regression model - Robustness

	Dep. Variable: probability of emergence of less biased consensus ($\hat{p}_G$)		
	Pooled	Pooled ($\theta = 0.2$)	Pooled ($\theta = 0.8$)
Partisan centrality advantage (PCA)	1.05^{***} (0.03)	1.06^{***} (0.03)	1.23^{***} (0.08)
Open mind (OM)	-0.40^{***} (0.01)	-0.25^{***} (0.02)	-0.93^{***} (0.04)
First impression (FI)	1.71^{***} (0.02)	1.63^{***} (0.03)	1.85^{***} (0.04)
PCA × FI	0.37^{***} (0.04)	0.18^{**} (0.06)	0.37^{***} (0.07)
PCA × OM	0.52^{***} (0.03)	0.41^{***} (0.04)	0.76^{***} (0.08)
OM × FI	-0.29^{***} (0.02)	-0.27^{***} (0.03)	0.02 (0.05)
$\mathbb{1}\{\tau = 1\}$	0.46^{***} (0.01)	0.84^{***} (0.01)	-0.51^{***} (0.02)
$\mathbb{1}\{\tau = 10\}$	0.19^{***} (0.01)	0.75^{***} (0.01)	-1.06^{***} (0.02)
$\mathbb{1}\{\tau = 30\}$	0.07^{***} (0.01)	0.68^{***} (0.01)	-1.23^{***} (0.02)
$\mathbb{1}\{n = 2\}$	0.47^{***} (0.02)	-0.02 (0.02)	1.47^{***} (0.05)
$\mathbb{1}\{n = 3\}$	0.49^{***} (0.02)	-0.09^{***} (0.02)	1.68^{***} (0.05)
$\mathbb{1}\{n = 4\}$	0.40^{***} (0.02)	-0.25^{***} (0.02)	1.74^{***} (0.05)
$\mathbb{1}\{G = (B)\}$	-1.15^{***} (0.02)	-0.94^{***} (0.02)	-1.57^{***} (0.03)
$\mathbb{1}\{G = (D)\}$	-0.84^{***} (0.02)	-0.52^{***} (0.02)	-1.44^{***} (0.03)
$\mathbb{1}\{G = (E)\}$	-0.96^{***} (0.02)	-0.73^{***} (0.02)	-1.44^{***} (0.03)
$\mathbb{1}\{G = (F)\}$	0.02 (0.02)	0.13^{***} (0.02)	-0.13^{***} (0.03)
$\mathbb{1}\{G = (H)\}$	-1.02^{***} (0.02)	-0.76^{***} (0.02)	-1.54^{***} (0.03)
$\mathbb{1}\{\theta = 0.8\}$	-0.10^{***} (0.01)		
Observations	168,320	84,160	84,160
Log Likelihood	-67,244.40	-41,815.30	-21,296.40
Akaike Inf. Crit.	134,525.00	83,664.70	42,626.90
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 7. Probit regression with pooled data