

**CES ifo**

**12700  
2026**

May 2026

# Working Papers

## Hypocrisy in a Simple Social Interaction Model

Mikhail Anufriev, Kirill Borissov, Mikhail Pakhnin

CES ifo

Imprint:

**CESifo Working Papers**

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion  
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany  
Telephone +49 (0)89 2180-2740

Email [office@cesifo.de](mailto:office@cesifo.de)  
<https://www.cesifo.org>

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge

- from the CESifo website: [www.ifo.de/en/cesifo/publications/cesifo-working-papers](http://www.ifo.de/en/cesifo/publications/cesifo-working-papers)
- from the SSRN website: [www.ssrn.com/index.cfm/en/cesifo/](http://www.ssrn.com/index.cfm/en/cesifo/)
- from the RePEc website: <https://ideas.repec.org/s/ces/ceswps.html>

# Hypocrisy in a Simple Social Interaction Model\*

Mikhail Anufriev<sup>a,b,†</sup>   Kirill Borissov<sup>c,‡</sup>   Mikhail Pakhnin<sup>d,§</sup>

<sup>a</sup> *Economics Discipline Group, Business School, University of Technology Sydney*

<sup>b</sup> *Department of Finance, VŠB - Technical University of Ostrava, Czechia*

<sup>c</sup> *School of Computational Social Sciences, Non-government*

*Educational Institution “European University at St. Petersburg”*

<sup>d</sup> *Department of Applied Economics, University of the Balearic Islands*

## Abstract

We develop a model of social interaction in which agents hold private beliefs but communicate public statements. We distinguish between two networks: an audience network shaping statements through social pressure, and an influence network governing belief updating. This separation drives a wedge between public statements and underlying beliefs, so that agents may persistently communicate opinions they do not hold. We provide necessary and sufficient network conditions for the emergence of such hypocrisy. The model generates rich patterns of opinion propagation, including the amplification of messages that agents themselves do not believe, and sheds light on phenomena such as propaganda diffusion, the spiral of silence, and political correctness.

**Keywords:** Social learning; Network theory; Hypocrisy; Political correctness; Dissonance minimization; Propaganda

**JEL Classification:** D83, D85, D91, Z13

---

\*We thank Sebastiano Della Lena, John Duffy, Evgeniya Goryacheva, Alan Kirman, Dunia López-Pintado, Paolo Pin, Mikhail Sokolov, Philip Ushchev, and Yves Zenou for useful and stimulating discussions on earlier drafts of the paper. Earlier versions also benefited greatly from comments by participants at WEHIA 2022 (Catania), the Workshop on Computational and Experimental Economics in memory of Jasmina Arifovic (Vancouver), CEF 2023 (Nice), and seminars at the University of Technology Sydney, the University of Adelaide and Pablo de Olavide University of Sevilla. Mikhail Anufriev acknowledges financial support from the Australian Research Council through Discovery Project DP170100429 and VSB-TU Ostrava under the SGS project SP2026/003. Mikhail Pakhnin acknowledges support from the Projects PID2023-152029OB-I00 and PID2024-160690NB-I00 financed by MICIU / AEI / 10.13039 / 501100011033 / FEDER, EU.

<sup>†</sup>*E-mail:* Mikhail.Anufriev@uts.edu.au

<sup>‡</sup>*E-mail:* kirill@eu.spb.ru

<sup>§</sup>*E-mail:* m.pakhnin@uib.es

# 1 Introduction

People often publicly express views they do not privately hold. In political debate, online communication, and organizational life, individuals may repeat prevailing narratives despite privately disagreeing with them. We refer to such persistent discrepancies between public statements and private beliefs as *hypocrisy*. While hypocrisy is commonly associated with strategic deception or manipulation, these discrepancies may also arise without any intent to mislead.

This paper studies how hypocrisy emerges, persists, and affects opinion propagation as a consequence of the network structure. We develop a tractable model of social interaction in which agents form private beliefs but communicate public statements. The key idea is that these two objects stem from distinct processes: statements are shaped by social pressure, while beliefs evolve through learning from observed statements. Belief formation therefore relies on a communication layer that may be systematically distorted relative to the distribution of private opinions. We characterize the conditions under which such distortions persist in the long run and illustrate how public narratives and aggregate outcomes can be shaped by statements that do not reflect agents' private beliefs.

Belief formation in economics is typically studied through two broad traditions. The first is Bayesian social learning, in which agents use Bayes' rule to infer information from signals, actions, or choices observed within a social structure. The second is DeGroot-style dynamics, in which agents repeatedly update their opinions by taking weighted averages of others' views.<sup>1</sup> Bayesian models provide a rational benchmark, while DeGroot models offer a behaviorally simple and analytically tractable framework for opinion dynamics.<sup>2</sup>

We retain the tractability of averaging-based opinion dynamics, but introduce two departures from the standard setting. First, we distinguish between privately held beliefs and publicly communicated statements. Second, we distinguish between the *audience network*, through which social pressure shapes statements, and the *influence network*, through which observed statements shape belief updating. These departures allow public communication to affect belief dynamics in ways that are absent from the standard DeGroot learning.

---

<sup>1</sup>Canonical contributions to Bayesian social learning include Banerjee (1992) and Bikhchandani et al. (1992), while network-based models include, among others, Bala and Goyal (1998) and Acemoglu et al. (2011). The original averaging model is due to DeGroot (1974). Important early network-based extensions include DeMarzo et al. (2003); Acemoglu et al. (2010); Golub and Jackson (2010, 2012). See also Golub and Sadler (2016) for a comprehensive survey of Bayesian and non-Bayesian models of social learning in networks.

<sup>2</sup>Experimental evidence comparing the two approaches suggests that individual behavior in networked learning environments is often better described by simple averaging heuristics than by fully rational Bayesian updating (e.g., Chandrasekhar et al., 2020; Grimm and Mengel, 2020).

These distinctions are motivated by evidence from social psychology showing that private beliefs and public statements are shaped by different social forces. In the classic conformity experiments of Asch (1955), subjects often reported judgments that contradicted their own perceptions after being exposed to unanimous but incorrect statements of others. To explain these observations, Deutsch and Gerard (1955) distinguished between *normative social influence*, which induces conformity in public expression, and *informational social influence*, which affects the updating of private beliefs. This distinction maps naturally into our framework: the audience network captures normative pressure, while the influence network captures informational influence.

The main contribution of the paper is a network-theoretic characterization of the origins of hypocrisy. We show that hypocrisy is a structural property of the multi-layer network induced by audience and influence relations. The distinction between these two layers can generate persistent public-private divergence even under simple averaging dynamics. In particular, whether an agent is hypocritical or not depends on how distinct long-run sources of influence enter belief updating and statement formation through the two interaction channels. We derive necessary and sufficient conditions for hypocrisy and show that an agent is generically hypocritical unless all relevant information either originates from a single independent source or loops back to the agent through an echo chamber in the audience network or influence network.

Our results imply that hypocrisy does not arise in standard DeGroot settings in which communication and influence coincide, but emerges naturally, though not always, once these channels are separated. More broadly, the analysis shows that persistent discrepancies between what agents believe and what they say can be generated by the topology of social interaction itself, without strategic manipulation, reputational signaling, or biased belief updating.

Beyond its characterization of hypocrisy, the model has implications for the propagation of opinions in networks. Because agents tune their statements to their audience, they may transmit narratives that differ from their own beliefs. As a result, public messages can spread without being privately adopted by the agents who pass them on. We show that this mechanism can amplify the reach of particular viewpoints: agents who do not believe a message may nevertheless communicate it, allowing it to influence individuals far from its original source. This provides a simple theoretical explanation for phenomena such as propaganda diffusion, the spiral of silence, and political correctness. More generally, the framework suggests that public opinion, as reflected in observed narratives, may shape aggregate outcomes while systematically differing from private opinion.

An important implication of the model is that hypocrisy can arise even in a non-strategic environment. Agents do not attempt to manipulate others or influence future beliefs, yet they may systematically communicate statements they do not hold, even after beliefs have converged. By isolating a minimal mechanism through which hypocrisy emerges and shapes the propagation of opinions, we show that failures of information aggregation may arise not from biased updating, but from distortions in the communication layer itself. Abstracting from strategic or forward-looking behavior allows the model to highlight the role of multi-layer network structure in sustaining discrepancies between private beliefs and public narratives. We discuss our contributions to the literature in detail below.

**Related literature.** Belief formation in networks has been extensively studied in the literature on non-Bayesian social learning and opinion dynamics, often through DeGroot-style averaging dynamics (see, e.g., Grabisch and Rusinowska, 2020, for a survey). Recent contributions have extended this framework to incorporate heterogeneous initial information (Banerjee et al., 2021), individual learning (Della Lena, 2024), endogenous network formation (Bolletta and Pin, 2025), endogenous information acquisition (Risco, 2025), and signed or group-based interactions (Della Lena et al., 2026). These and many other papers assume that agents communicate their beliefs truthfully.

Closer to our approach are the papers by Buechel et al. (2015), Arifovic et al. (2015) and Anufriev et al. (2023), which relax this assumption and explicitly study the role of conformity motives in opinion dynamics, belief aggregation, and convergence. Similarly to these papers, we distinguish between privately held beliefs and publicly communicated statements. Unlike them, however, we model beliefs and statements as outcomes of separate interaction processes operating through distinct networks. This multi-layer structure gives rise to persistent discrepancies between private beliefs and public statements, a phenomenon that we call *hypocrisy* and that cannot emerge in those models.

By establishing conditions of such hypocrisy, our paper also relates to the literature on conformity and social image. Classical contributions such as Akerlof (1980), Bernheim (1994), and Bénabou and Tirole (2006) study how social norms, reputational concerns, status incentives, and image considerations shape behavior. In this literature, discrepancies between private preferences and publicly observed behavior arise as equilibrium responses to social incentives. In our model, agents also face a trade-off between expressing views close to their privately held beliefs and conforming to social pressure. Our mechanism, however, is different. Unlike in this literature, private beliefs in our framework evolve endogenously through social

learning from publicly observed statements, so that public expression feeds back into future beliefs. While agents optimally resolve the static trade-off between conformity and expressing their privately held beliefs, they do not internalize the dynamic consequences of their current statements for future belief evolution.

In this respect, our paper is closer to models of norm conformity and assimilation dynamics in social networks, in which agents respond to social pressure or peer influence. For example, Buechel et al. (2014) studies norm conformity through the distinction between intrinsic and displayed traits, while Olcina et al. (2024) analyzes assimilation dynamics in social networks. Although our setting and focus differ, the distinction between privately held and publicly displayed characteristics in these papers parallels our distinction between beliefs and statements.

Our paper is also related to the literature on preference falsification, political correctness, and distorted public communication. Kuran (1987, 1995) study environments in which individuals publicly misrepresent privately held views because of social or political pressure. In Kuran’s framework, public expression affects beliefs about the distribution of private opinions in society and may generate pluralistic ignorance, as well as sudden shifts in public opinion once hidden views become publicly revealed.<sup>3</sup> Relatedly, Morris (2001) studies how concerns about political correctness distort strategic communication and information aggregation when agents account for the social consequences of their statements. While Kuran emphasizes active misrepresentation of views, the spiral of silence theory emphasizes the suppression or self-censorship of publicly expressed minority opinions because of perceived social pressure (Noelle-Neumann, 1974, 1984). As in this literature, agents in our model face a trade-off between expressing views close to their privately held beliefs and adapting statements to social pressure.

Our contribution is to embed this mechanism into a network model of social learning. In contrast to Kuran (1987) and Morris (2001), agents in our framework do not infer hidden private opinions from observed statements. Instead, publicly communicated statements directly enter the process of belief updating, so that distorted communication propagates through the network and affects long-run belief formation. We demonstrate that this mechanism can generate qualitatively related phenomena, including the amplification of perceived majority views, suppression of minority opinions, and propagation of narratives through indirect communication channels. These findings are broadly consistent with emerging empirical evidence showing that social pressure substantially affects publicly expressed opinions (Bursztyn et al., 2020; Ho and Huang, 2026), while media influence and

---

<sup>3</sup>Our current model does not feature such phenomena, since agents do not form higher-order beliefs about the private opinions underlying observed statements.

interpersonal communication may generate large network spillovers (DellaVigna and Kaplan, 2007; Bond et al., 2012; Yanagizawa-Drott, 2014).

The paper is organized as follows. Section 2 introduces the model and formalizes the roles of audience tuning and influence in belief formation. Section 3 solves the model and characterizes its long-run outcomes. Section 4 studies the emergence of hypocrisy and provides conditions for its occurrence. Section 5 discusses the implications of the model and illustrates its mechanisms through applications. Section 6 concludes. The proofs of all results are relegated to the Appendix.

## 2 The model

Society consists of a set of agents,  $\mathcal{N} = \{1, \dots, N\}$ , who communicate and exchange information regarding a given subject. Time is discrete. In each period  $t > 0$ , agent  $i$  makes a public statement  $s_i(t)$  about the subject and subsequently forms a belief  $b_i(t)$  based on the statements of others.<sup>4</sup>

Each period  $t$  begins with a round of communication. During this stage, each agent chooses a statement to minimize dissonance, which arises when their statement deviates from

- (i) their own previous belief, reflecting the cost of personal inconsistency (private dissonance); and
- (ii) the statements observed in their audience, reflecting the cost of social disagreement (social dissonance).

Formally, communication occurs on a faster timescale  $\tau$  within period  $t$ . Given the previous belief  $b_i(t-1)$ , which remains fixed at this timescale, agent  $i$  at each instance  $\tau$  chooses a statement  $s_i^\tau(t)$  to minimize dissonance with their previous belief and the preceding statements of others:

$$\min_{s_i^\tau(t)} \left\{ (s_i^\tau(t) - b_i(t-1))^2 + \sum_{j=1}^N d_{ij} (s_i^\tau(t) - s_j^{\tau-1}(t))^2 \right\}. \quad (1)$$

The parameters  $d_{ij} \geq 0$  capture the sensitivity of agent  $i$  to dissonance with agent  $j$  and are summarized in the *audience dissonance matrix*  $\mathbf{D} = \{d_{ij}\}_{i,j=1}^N$ , which has a zero diagonal. If  $d_{ij} = 0$ , agent  $i$  does not take  $j$ 's statement into account when choosing a statement, whereas if  $d_{ij} > 0$ , disagreement between  $i$  and  $j$  generates a cost.

---

<sup>4</sup>In our context, beliefs are broadly interpreted as private opinions, judgments, estimations, norms, or values, while statements correspond to publicly expressed opinions, narratives, or observable actions. We use the term *belief* to emphasize that our results apply even when agents aim to estimate an objective truth.

The matrix  $\mathbf{D}$  defines the *audience network*: there is a directed edge from  $i$  to  $j$  whenever  $d_{ij} > 0$ , indicating that agent  $j$  belongs to the audience relevant for agent  $i$ 's statement choice. This network captures normative social influence (Deutsch and Gerard, 1955), whereby individuals experience dissonance when their public statements diverge from those observed in their audience. Importantly, this influence need not be symmetric. For example, in authoritarian regimes, individuals may experience dissonance when their statements contradict those of a dictator, whereas the dictator may feel no dissonance in return. Similarly, employees may experience dissonance when their messages diverge from those of their supervisors, while supervisors may not face the same pressure when addressing subordinates. Note that the matrix  $\mathbf{D}$  captures both the direction and intensity of social pressure.

It follows from standard results (e.g., Lemma 2 in Anufriev et al., 2023) that statements converge in the fast timescale to a vector  $\mathbf{s}(t)$  whose components solve, for each  $i$ ,

$$\min_{s_i(t)} \left\{ (s_i(t) - b_i(t-1))^2 + \sum_{j=1}^N d_{ij} (s_i(t) - s_j(t))^2 \right\}. \quad (2)$$

The resulting statements satisfy

$$s_i(t) = \frac{1}{1 + \delta_i} b_i(t-1) + \sum_{j=1}^N \frac{d_{ij}}{1 + \delta_i} s_j(t), \quad i = 1, \dots, N, \quad (3)$$

where  $\delta_i = \sum_j d_{ij}$ . Thus, each agent's statement is a weighted average of their own previous belief and the statements observed in their audience. This process, known as *audience tuning* (Echterhoff et al., 2005), reflects the role of normative social influence in shaping public expression.

The statement formation process admits a game-theoretic interpretation (see also Bindel et al., 2015). Given previous beliefs  $\{b_i(t-1)\}_{i=1}^N$ , the profile of statements  $\{s_i(t)\}_{i=1}^N$  coincides with the unique Nash equilibrium of a static game in which agents simultaneously choose statements to minimize the quadratic loss in (2). Each agent trades off fidelity to their own belief against conformity with the statements of others in their audience. Uniqueness follows from strict convexity, and the resulting equilibrium is characterized by (3).

After communication, agents update their beliefs based on observed statements (see Fig. 1). An agent chooses a new belief  $b_i(t)$  by minimizing

$$\min_{b_i(t)} \left\{ (b_i(t) - b_i(t-1))^2 + \sum_{j=1}^N g_{ij} (b_i(t) - s_j(t))^2 \right\}. \quad (4)$$

The parameters  $g_{ij} \geq 0$  represent the sensitivity of agent  $i$  to the statements of

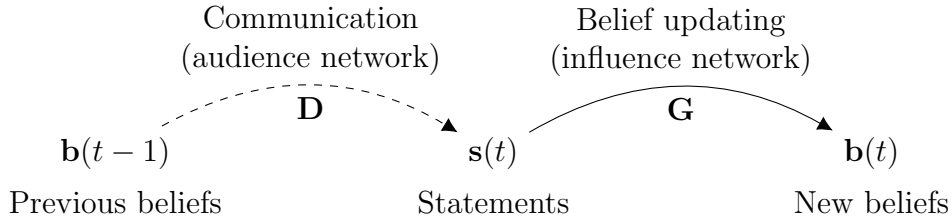


Figure 1: Timing of the model within period  $t$ . A round of communication is followed by belief updating.

agent  $j$  and define the *influence matrix*  $\mathbf{G} = \{g_{ij}\}_{i,j=1}^N$ , which also has a zero diagonal. If  $g_{ij} = 0$ , agent  $i$  ignores  $j$ 's statement when updating beliefs, whereas if  $g_{ij} > 0$ , disagreement generates a cost.

The matrix  $\mathbf{G}$  defines the *influence network*: there is a directed edge from  $i$  to  $j$  whenever  $g_{ij} > 0$ , meaning that agent  $j$  influences agent  $i$ 's belief. Thus,  $\mathbf{G}$  captures both the direction and intensity of informational social influence, whereby agents use others' statements as inputs for belief formation. Agents treat observed statements as primitives when updating their beliefs and do not attempt to infer the underlying beliefs that generated them. Accordingly, statements constitute the effective informational inputs in the learning process.

As in the case of the audience network,  $\mathbf{G}$  need not be symmetric: individuals may rely on media outlets, public figures, or official sources of information when forming beliefs, even though these sources are themselves unaffected by the beliefs of ordinary individuals. Similarly, employees may update their beliefs based on the views of supervisors, experts, or mentors, while these figures may place little or no weight on the public opinions of those they influence.

The solution to (4) is given by

$$b_i(t) = \frac{1}{1 + \gamma_i} b_i(t-1) + \sum_{j=1}^N \frac{g_{ij}}{1 + \gamma_i} s_j(t), \quad i = 1, \dots, N, \quad (5)$$

where  $\gamma_i = \sum_j g_{ij}$ . Similarly to (3), each agent's belief is a weighted average of their own previous belief and the statements from their influence network.

The model is thus fully characterized by two matrices,  $\mathbf{D}$  and  $\mathbf{G}$ , capturing normative and informational social influences, respectively. In general, these networks need not coincide, reflecting the fact that the agents who shape an individual's public statements may differ from those who influence their private beliefs. This distinction is central to our analysis and will be shown to generate systematic discrepancies between beliefs and statements.

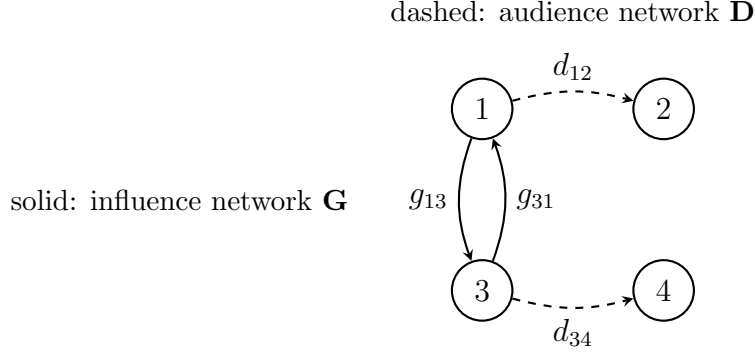


Figure 2: Illustrative four-agent example. Dashed arrows represent the audience network  $\mathbf{D}$  and solid arrows represent the influence network  $\mathbf{G}$ . Equivalently, the only nonzero entries are  $d_{12}$  and  $d_{34}$  in  $\mathbf{D}$ , and  $g_{13}$  and  $g_{31}$  in  $\mathbf{G}$ .

**Illustrative four-agent example.** Before turning to the formal analysis, it is instructive to consider the four-agent society in Fig. 2, which highlights the key ingredients of the model.

Suppose that in the audience network, agent 1 takes agent 2’s statement into account and agent 3 takes agent 4’s statement into account, so that  $d_{12}$  and  $d_{34}$  are positive and all other audience weights are zero. In the influence network, agents 1 and 3 listen to each other, so that  $g_{13}$  and  $g_{31}$  are positive and all other influence weights are zero. Thus, agents 2 and 4 affect statements but do not directly influence beliefs.<sup>5</sup>

At an intuitive level, this configuration already suggests the main mechanism of the model. Agents 1 and 3 update their beliefs based on each other’s statements, but these statements are themselves shaped by audience tuning toward agents 2 and 4. As a result, the beliefs of agents 2 and 4 indirectly affect the beliefs of agents 1 and 3, even though neither agent directly influences them in the influence network. Moreover, agents 1 and 3 may communicate statements that differ from their own beliefs, as they adjust their statements toward their respective audiences.

Thus, the example suggests both indirect opinion propagation through the audience network and discrepancies between beliefs and statements. These observations will be formalized in the next section.

<sup>5</sup>This configuration can be interpreted, for instance, in a political context in which agents 1 and 3 represent politicians or commentators who form their beliefs about the subject of interest through interaction with one another, while agents 2 and 4 represent their respective voter bases. These constituencies may not directly determine beliefs, but they shape how positions are publicly expressed. As a result, political discourse may reflect audience-driven narratives rather than underlying private beliefs.

### 3 Belief and statement dynamics

We now characterize the joint dynamics of statements and beliefs and study their long-run behavior. This clarifies how communication and learning jointly shape the relationship between private beliefs and public statements. A key implication is that these objects need not coincide in the long run.

#### 3.1 Statement formation

Using (3) and (5), the joint dynamics of statements and beliefs can be written as:

$$\mathbf{b}(t-1) = (\mathbf{I} + \mathbf{\Delta} - \mathbf{D}) \mathbf{s}(t), \quad \mathbf{b}(t-1) = (\mathbf{I} + \mathbf{\Gamma}) \mathbf{b}(t) - \mathbf{G} \mathbf{s}(t), \quad (6)$$

where  $\mathbf{b}(t)$  is the column vector of beliefs,  $\mathbf{s}(t)$  is the vector of statements,  $\mathbf{I}$  is the identity matrix of size  $N$ ,  $\mathbf{\Delta}$  is the diagonal matrix with entries  $\delta_i = \sum_j d_{ij}$ , and  $\mathbf{\Gamma}$  is the diagonal matrix with entries  $\gamma_i = \sum_j g_{ij}$ . Initial beliefs  $\mathbf{b}(0)$  are given.

We first recall several standard notions. Let  $\mathbf{A} = (a_{ij})$  be a square matrix with non-negative entries, i.e.,  $a_{ij} \geq 0$  for all  $i, j$  (denoted  $\mathbf{A} \geq 0$ ). A *path from  $i$  to  $j$  in  $\mathbf{A}$*  is a sequence  $i_1, \dots, i_J$  with  $i_1 = i$ ,  $i_J = j$ , and  $a_{i_k i_{k+1}} > 0$  for all  $k = 1, \dots, J-1$ . The matrix  $\mathbf{A}$  is *strongly connected* if for any  $i, j$  there exists a path from  $i$  to  $j$ . The matrix  $\mathbf{A}$  is *row-stochastic* if  $\sum_j a_{ij} = 1$  for all  $i$ .

The following result characterizes the formation of statements.

**Proposition 1.** (i) *The evolution of statements is given by  $\mathbf{s}(t) = \mathbf{P} \mathbf{b}(t-1)$ , where the row-stochastic matrix  $\mathbf{P}$  is defined as*

$$\mathbf{P} = (\mathbf{I} + \mathbf{\Delta} - \mathbf{D})^{-1}. \quad (7)$$

(ii) *For any  $i$ ,  $p_{ii} > 0$ . For any  $i$  and any  $j \neq i$ ,  $p_{ij} > 0$  if and only if there exists a path from  $i$  to  $j$  in  $\mathbf{D}$ .*

(iii)  *$\mathbf{P}$  is strongly connected if and only if  $\mathbf{D}$  is strongly connected.*

*Proof.* See Appendix A.1. ■

Proposition 1 shows that communication leads to statements that are weighted averages of previous beliefs. In particular, agent  $j$ 's belief may affect agent  $i$ 's statement not only directly, when  $d_{ij} > 0$ , but also through chains of audience interactions.

### 3.2 Belief dynamics

We now turn to belief formation. From (6), we obtain  $\mathbf{b}(t) = \mathbf{Q}\mathbf{s}(t)$ , where

$$\mathbf{Q} = (\mathbf{I} + \mathbf{\Gamma})^{-1} (\mathbf{I} + \mathbf{\Delta} + \mathbf{G} - \mathbf{D}).$$

Combining this with Proposition 1, we obtain  $\mathbf{b}(t) = \mathbf{T}\mathbf{b}(t-1)$ , where  $\mathbf{T} \equiv \mathbf{Q}\mathbf{P}$ .

**Proposition 2.** *The matrix  $\mathbf{T}$  defined as*

$$\mathbf{T} = (\mathbf{I} + \mathbf{\Gamma})^{-1} (\mathbf{I} + \mathbf{\Delta} + \mathbf{G} - \mathbf{D}) (\mathbf{I} + \mathbf{\Delta} - \mathbf{D})^{-1} \quad (8)$$

*is row-stochastic, and beliefs evolve according to*

$$\mathbf{b}(t) = \mathbf{T}\mathbf{b}(t-1).$$

*Proof.* See Appendix A.1. ■

Thus, belief updating follows an average-based process of DeGroot-style learning. Each agent’s new belief is a weighted average of previous beliefs, where the weights  $t_{ij}$  depend on both the audience network  $\mathbf{D}$  and the influence network  $\mathbf{G}$ .<sup>6</sup>

The interaction of the two networks suggests considering the multi-layer network induced by  $\mathbf{D}$  and  $\mathbf{G}$ , denoted by  $\mathbf{D} \vee \mathbf{G}$ . Its adjacency matrix is defined by

$$(\mathbf{D} \vee \mathbf{G})_{ij} = \begin{cases} 1 & \text{if } \max\{d_{ij}, g_{ij}\} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We say that agent  $i$  is *linked* to agent  $j$  if there exists a path from  $i$  to  $j$  in  $\mathbf{D} \vee \mathbf{G}$ .

The next result characterizes the structure of  $\mathbf{T}$  in terms of  $\mathbf{D}$  and  $\mathbf{G}$ .

**Proposition 3.** *Let  $\mathbf{T}$  be given by (8).*

- (i) *For any  $i$ ,  $t_{ii} > 0$ .*
- (ii) *For any  $i$  and any  $j \neq i$ ,  $t_{ij} > 0$  if and only if either  $g_{ij} > 0$  or there exists  $k$  such that  $g_{ik} > 0$  and there is a path from  $k$  to  $j$  in  $\mathbf{D}$ .*
- (iii)  *$t_{ii} = 1$  if and only if  $g_{ij} = 0$  for all  $j$ .*
- (iv) *Suppose that for each  $i$  there exists  $j$  such that  $g_{ij} > 0$ . Then  $\mathbf{T}$  is strongly connected if and only if  $\mathbf{D} \vee \mathbf{G}$  is strongly connected.*

---

<sup>6</sup>When  $\mathbf{G} = \mathbf{D}$ , we obtain  $\mathbf{Q} = \mathbf{I}$  and hence  $\mathbf{b}(t) = \mathbf{s}(t)$ , corresponding to the “saying is believing” effect (Higgins, 1999). This special case is studied in Anufriev et al. (2023).

*Proof.* See Appendix A.2. ■

First, Proposition 3 implies that each agent places a positive weight on their own previous belief. It is well known that this guarantees convergence of beliefs (see, e.g., Theorem 2 in Golub and Jackson, 2010).

Second, beliefs may be influenced *indirectly* through the audience network. If agent  $i$  listens to agent  $k$  in the influence network and the statement of  $k$  is itself shaped by agent  $j$  (possibly through a chain of audience links), then  $j$ 's belief enters the updating rule of agent  $i$  even when  $j$  is not connected with  $i$  in the influence network, and may not even be known to  $i$ . While indirect influence is a familiar feature of network learning models, here it arises *within a single period* and operates entirely through public statements shaped by audience tuning.

Third, agent  $i$  does not change their belief if and only if  $i$  is not influenced (i.e., if and only if the  $i$ th row of  $\mathbf{G}$  is zero). This motivates the following definition.

**Definition 1.** *An agent  $i$  with  $t_{ii} = 1$  is called stubborn.*

As we show below, stubborn agents play a key role in generating discrepancies between beliefs and statements.

Finally, when there are no stubborn agents, the matrix  $\mathbf{T}$  is strongly connected if and only if the multi-layer network  $\mathbf{D} \vee \mathbf{G}$  is strongly connected. This highlights that the relevant structure governing belief dynamics is the combination of the audience and influence networks.

**The illustrative example continued.** We illustrate the matrices for the four-agent example from Section 2; see Fig. 2. Agents 2 and 4 neither adjust their statements nor update their beliefs, and are therefore stubborn.

The matrix of statement formation is

$$\mathbf{P} = \begin{pmatrix} \frac{1}{1+d_{12}} & \frac{d_{12}}{1+d_{12}} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{1+d_{34}} & \frac{d_{34}}{1+d_{34}} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

so that agents 1 and 3 tune their statements toward agents 2 and 4, respectively.

The belief-updating matrix  $\mathbf{T}$  reflects this structure. Since agents 2 and 4 are stubborn, the second and fourth rows of  $\mathbf{T}$  have a 1 on the diagonal and 0 elsewhere. The first row of  $\mathbf{T}$  is

$$\left( \frac{1}{1+g_{13}}, \quad 0, \quad \frac{g_{13}}{1+g_{13}} \cdot \frac{1}{1+d_{34}}, \quad \frac{g_{13}}{1+g_{13}} \cdot \frac{d_{34}}{1+d_{34}} \right),$$

where the weights on agents 3 and 4 arise because agent 1 listens to agent 3, whose statement is shaped by agent 4. A similar decomposition holds for agent 3.

Thus, agent 1 assigns positive weight to agent 4, and agent 3 assigns positive weight to agent 2, even though agents 2 and 4 are neither directly observed nor influential. Audience tuning therefore enlarges the set of agents whose beliefs affect long-run outcomes.

### 3.3 Long-run behavior

Finally, we study the long-run properties of the model. From the belief dynamics, we have  $\mathbf{b}(t) = \mathbf{T}^t \mathbf{b}(0)$ . Since  $t_{ii} > 0$  for all  $i$ , beliefs converge: there exists

$$\mathbf{b}^* \equiv \lim_{t \rightarrow \infty} \mathbf{T}^t \mathbf{b}(0).$$

Since  $\mathbf{s}(t) = \mathbf{P} \mathbf{b}(t - 1)$ , statements also converge, and we denote

$$\mathbf{s}^* \equiv \lim_{t \rightarrow \infty} \mathbf{s}(t).$$

The following simple example illustrates that long-run beliefs need not coincide with long-run statements.

**Two-agent example.** Consider a two-agent society shown in Fig. 3. There are no influence links, so both agents are stubborn and  $b_i^* = b_i(0)$  for  $i = 1, 2$ . Agent 1 accounts for agent 2’s statement with dissonance sensitivity  $d_{12} > 0$ , while agent 2 does not adjust their statement. Thus, agent 2 states their belief,  $s_2^* = b_2(0)$ , whereas agent 1’s statement is a weighted average of the two beliefs:

$$s_1^* = \frac{1}{1 + d_{12}} b_1(0) + \frac{d_{12}}{1 + d_{12}} b_2(0).$$

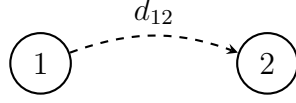
Hence, agent 1’s limiting statement differs from their limiting belief unless  $b_1(0) = b_2(0)$ , a condition that holds only on a zero-measure set of initial beliefs and is therefore non-generic.

This simple example highlights the possibility of persistent discrepancies between beliefs and statements, which we refer to as *hypocrisy*.

**Definition 2.** Agent  $i$  is hypocritical if  $b_i^* \neq s_i^*$ .

The possibility of *long-run hypocrisy* is a distinctive feature of our framework. In previous models of social interaction incorporating conformity or dissonance effects (Arifovic et al., 2015; Buechel et al., 2015; Anufriev et al., 2023), discrepancies

dashed: audience network  $\mathbf{D}$



no influence links ( $\mathbf{G} = \mathbf{0}$ )

Figure 3: Two-agent example. Agent 1 takes agent 2’s statement into account in the audience network, while both agents are stubborn. Agent 1 is generically hypocritical.

between beliefs and statements may arise within a single period but vanish in the long run as the dynamics converge. By contrast, in our model, such discrepancies may persist once the audience and influence networks are disentangled.

Understanding when hypocrisy emerges requires characterizing the model’s long-run outcomes as a function of the network structure. In our framework, the belief-updating matrix  $\mathbf{T}$  is generated by the interaction of two networks,  $\mathbf{D}$  and  $\mathbf{G}$ . To express the relevant structure directly in terms of the multi-layer network  $\mathbf{D} \vee \mathbf{G}$ , we recall and extend the standard notion of *closed communication classes*.

**Definition 3.** A set  $\mathcal{C} \subset \mathcal{N}$  is a closed communication class (CCC) in  $\mathbf{D} \vee \mathbf{G}$  if for all  $i, j \in \mathcal{C}$  there exist paths from  $i$  to  $j$  and from  $j$  to  $i$  (i.e.,  $\mathcal{C}$  is strongly connected), and there are no paths from any  $i \in \mathcal{C}$  to any  $k \notin \mathcal{C}$ . A closed communication class without stubborn agents is called a pure communication class.

CCCs are maximal strongly connected subsets of agents with no outgoing links to the rest of the network. They play a central role in shaping long-run outcomes in standard models (see, e.g., Chapter 8 of Jackson, 2008).

A key difference in our framework is that CCCs in  $\mathbf{D} \vee \mathbf{G}$  do *not* generally coincide with those in  $\mathbf{T}$ , which govern long-run beliefs. In particular, such a class may contain stubborn agents (through the audience network), whereas under  $\mathbf{T}$  each stubborn agent forms a singleton class. This observation motivates the notion of *pure* communication classes. Lemma A.1 in Appendix A.3 shows that the CCCs in  $\mathbf{T}$  consist precisely of singleton stubborn agents and the pure communication classes in  $\mathbf{D} \vee \mathbf{G}$ .

We refer to pure communication classes and singleton stubborn agents as *independent classes*. Accordingly, with respect to  $\mathbf{D} \vee \mathbf{G}$ , the agent set  $\mathcal{N}$  can be partitioned into (i) the pure communication classes  $\mathcal{C}_m$ ,  $m = 1, \dots, M$ , (ii) the set  $\mathcal{S}$  of stubborn agents, and (iii) the set  $\mathcal{R}$  of remaining agents. Under this partition,  $\mathbf{T}$  admits a canonical block form, which determines the limiting beliefs. We characterize the long-run outcome below (see Proposition A.1 in Appendix A.3 for the formal analysis).

**Limiting beliefs.** First, all agents in a pure communication class  $\mathcal{C}_m$  converge to the same long-run belief

$$b^{*m} = \sum_{i \in \mathcal{C}_m} \pi_i^m b_i(0),$$

where the weights  $\{\pi_i^m\}_{i \in \mathcal{C}_m}$  form the unique positive left eigenvector of the submatrix of  $\mathbf{T}$  restricted to  $\mathcal{C}_m$  associated with eigenvalue 1, normalized to sum to one. Therefore,  $b^{*m}$  is a weighted average of the initial beliefs of agents in  $\mathcal{C}_m$ , where the weight  $\pi_i^m$  corresponds to agent  $i$ 's *eigenvector centrality* within the class. Generically (i.e., except on a zero-measure set of initial beliefs and parameter values in  $\mathbf{D}$  and  $\mathbf{G}$ ), different pure communication classes have different limiting beliefs.<sup>7</sup>

Eigenvector centralities within  $\mathcal{C}_m$  satisfy the following system of equations:

$$\pi_i^m = \frac{1 + \gamma_i}{\gamma_i} \frac{1}{1 + \delta_i} \sum_{h \in \mathcal{C}_m} \left( d_{hi} + \frac{g_{hi}}{\gamma_h} \right) \frac{\gamma_h}{1 + \gamma_h} \pi_h^m. \quad (9)$$

Thus, eigenvector centralities reflect both normative and informational social influence: an agent's centrality increases when others tune their messages toward that agent or form their beliefs under that agent's influence. Conversely, an agent's centrality, and hence their impact on the limiting belief within the class, decreases when the agent strongly tunes their statements to others or is highly influenced by others' views.<sup>8</sup>

Second, stubborn agents never update their beliefs:

$$b_s^* = b_s(0), \quad s \in \mathcal{S}.$$

Third, each remaining agent  $r \in \mathcal{R}$  converges to a weighted average of the long-run beliefs of the independent classes that influence  $r$ . In particular, if  $r$  is linked to a single independent class in  $\mathbf{D} \vee \mathbf{G}$ , then  $r$ 's belief converges to the long-run belief of this class.

**Limiting statements.** Long-run statements satisfy  $\mathbf{s}^* = \mathbf{P}\mathbf{b}^*$ . For a non-stubborn agent  $i$ , taking limits in (5) yields

$$b_i^* = \sum_j \frac{g_{ij}}{\gamma_i} s_j^*. \quad (10)$$

---

<sup>7</sup>In the language of the social learning literature, each pure communication class reaches consensus. However, society as a whole fails to reach consensus whenever there are at least two independent classes.

<sup>8</sup>Formally, the eigenvector centrality of agent  $i$  within  $\mathcal{C}_m$  increases with  $d_{hi}$  and  $g_{hi}$  and decreases with  $d_{ij}$  and  $g_{ij}$  for any  $h, j \in \mathcal{C}_m$  (see Proposition A.2 in Appendix A.4).

Similarly, taking limits in (3) gives

$$s_i^* = \frac{1}{1 + \delta_i} b_i^* + \sum_j \frac{d_{ij}}{1 + \delta_i} s_j^*. \quad (11)$$

These relations reflect the two-stage structure of the model: beliefs are updated based on observed statements through the influence network, while statements are formed based on the audience network. Together, (10) and (11) define a system of linear equations that jointly determines the long-run beliefs and statements.

**The illustrative example completed.** We now complete the analysis of the four-agent network introduced in Section 2; see Fig. 2. Agents 2 and 4 neither adjust their statements nor update their beliefs. Hence, they are stubborn and non-hypocritical:  $b_2^* = s_2^* = b_2(0)$  and  $b_4^* = s_4^* = b_4(0)$ .

Agents 1 and 3, by contrast, are non-stubborn. The set  $\{1, 3\}$  is not a CCC in  $\mathbf{D} \vee \mathbf{G}$ : although the two agents are mutually reachable through the influence network, the set is not closed, as there are paths from 1 to 2 and from 3 to 4. Hence, the only independent classes are the singleton stubborn agents  $\{2\}$  and  $\{4\}$ , while agents 1 and 3 belong to the set of remaining agents.

For agents 1 and 3, it follows from (10) that

$$b_1^* = s_3^*, \quad b_3^* = s_1^*,$$

while (11) implies

$$s_1^* = \frac{b_1^* + d_{12}b_2(0)}{1 + d_{12}}, \quad s_3^* = \frac{b_3^* + d_{34}b_4(0)}{1 + d_{34}},$$

so that we obtain

$$b_1^* = \frac{d_{12}b_2(0) + d_{34}(1 + d_{12})b_4(0)}{d_{12} + d_{34} + d_{12}d_{34}}, \quad b_3^* = \frac{d_{12}(1 + d_{34})b_2(0) + d_{34}b_4(0)}{d_{12} + d_{34} + d_{12}d_{34}}.$$

While both long-run beliefs are weighted averages of  $b_2(0)$  and  $b_4(0)$ , they generally differ, so both agents are hypocritical. In particular,

$$s_1^* - b_1^* = \kappa(b_2(0) - b_4(0)), \quad s_3^* - b_3^* = -\kappa(b_2(0) - b_4(0)),$$

where  $\kappa \equiv d_{12}d_{34}/(d_{12} + d_{34} + d_{12}d_{34})$ . Thus, the two agents exhibit hypocrisy of equal magnitude but in opposite directions. Each agent's statement is pulled toward their respective audience, agent 1 toward agent 2, and agent 3 toward agent 4, while their beliefs are shaped by each other's statements. Differences in

the audiences' beliefs therefore generate a systematic wedge between statements and beliefs, with magnitude governed by audience weights.

This example highlights three features of the model. First, separating the audience and influence networks can generate persistent discrepancies between beliefs and statements. Second, even when agents influence each other's beliefs, they may fail to reach consensus because statements are tuned to different audiences. Third, agents who do not directly influence beliefs (such as agents 2 and 4) may still affect them indirectly through the statements of intermediaries.

## 4 Hypocrisy

A central contribution of the model is that it rationalizes *hypocrisy*, defined as a discrepancy between agent's limiting beliefs and statements. Importantly, this phenomenon arises in a simple, non-strategic environment: agents do not attempt to manipulate others or influence future beliefs, but instead update their beliefs based on observed statements, which may not fully reflect underlying beliefs due to dissonance minimization. While discrepancies between beliefs and statements naturally arise within a period, the key question is whether they persist in the long run and how they depend on the structure of the multi-layer network.

The two examples discussed above (see Figs. 2 and 3) may suggest that separating the audience and influence networks is sufficient to generate long-run hypocrisy. However, this intuition is incomplete: separation alone is not sufficient. We present two examples in which hypocrisy is absent, showing that persistence depends on how agents are connected to different independent classes in the multi-layer network  $\mathbf{D} \vee \mathbf{G}$ . We then provide formal results.

**Illustrative cases of no hypocrisy.** Fig. 4 presents two modifications of the four-agent example introduced earlier. In both cases, the same set of agents is considered, but the structure of audience and influence links differs.

In the left panel, agents 1 and 3 influence each other, while only agent 1 adjusts their statement toward agent 2. Agent 4 is stubborn and no longer influences agent 3. Thus, within the connected component consisting of agents 1, 2, and 3, the singleton set  $\{2\}$  is the only pure communication class. The distortion in agent 1's statement toward agent 2 feeds into agent 3's belief updating and then back to agent 1. As a result, both agents' beliefs shift toward agent 2's initial belief. In the long run, these three agents share the same belief, and statements coincide with beliefs, so there is no hypocrisy.



Figure 4: Two variations of the four-agent example without hypocrisy. Dashed arrows denote the audience network  $\mathbf{D}$ , while solid arrows denote the influence network  $\mathbf{G}$ .

In the right panel, the audience structure from the baseline example is preserved, but agents 2 and 4 now influence each other and form a pure communication class. They converge to a common belief, which propagates to agents 1 and 3. In the long run, all agents share the same belief (a weighted average of the initial beliefs of agents 2 and 4), and there is no hypocrisy.

These examples show that separating the audience and influence networks is not sufficient to generate long-run hypocrisy; rather, the emergence of hypocrisy depends on the structure of the multi-layer network  $\mathbf{D} \vee \mathbf{G}$ .

#### 4.1 Sufficient conditions for no hypocrisy

In general, the presence or absence of hypocrisy depends on initial beliefs and on the weighted networks  $\mathbf{D}$  and  $\mathbf{G}$ . The multi-layer network  $\mathbf{D} \vee \mathbf{G}$  determines which independent classes can affect each agent, while the weights determine how these classes enter long-run beliefs and statements. Our objective is to relate hypocrisy as directly as possible to the structure of  $\mathbf{D} \vee \mathbf{G}$ .

We begin with a structural sufficient condition for the absence of hypocrisy.

**Proposition 4** (No hypocrisy within pure communication classes). *If agent  $i$  belongs to a pure communication class in  $\mathbf{D} \vee \mathbf{G}$ , then  $b_i^* = s_i^*$ .*

*Proof.* If agent  $i$  belongs to a pure communication class  $\mathcal{C}_m$ , then all agents in  $\mathcal{C}_m$  share the same long-run belief  $b^{*m}$ . If  $p_{ij} > 0$ , then, by Proposition 1, there exists a path from  $i$  to  $j$  in  $\mathbf{D}$ , and hence also in  $\mathbf{D} \vee \mathbf{G}$ . Since  $\mathcal{C}_m$  is closed in  $\mathbf{D} \vee \mathbf{G}$ , it follows that  $j \in \mathcal{C}_m$ . Using  $\mathbf{s}^* = \mathbf{P}\mathbf{b}^*$ , we obtain  $s_i^* = \sum_j p_{ij} b_j^* = b^{*m} = b_i^*$ . ■

Proposition 4 reflects the fact that agents within a pure communication class reach consensus: their beliefs converge to a common value. Since statements are weighted averages of beliefs, they coincide with this common belief. Thus, if the

entire society forms a single pure communication class, hypocrisy cannot emerge. More broadly, discrepancies between beliefs and statements can arise only outside pure communication classes, that is, among stubborn or remaining agents.

We next turn to a complementary condition based on interaction weights.

**Proposition 5** (No hypocrisy under aligned interaction weights). *If, for agent  $i$ , either  $\delta_i = 0$  or  $\delta_i > 0$ ,  $\gamma_i > 0$ , and  $d_{ij}/\delta_i = g_{ij}/\gamma_i$  for all  $j$ , then  $b_i^* = s_i^*$ .*

*Proof.* If  $\delta_i = 0$  (i.e.,  $d_{ij} = 0$  for all  $j$ ), then  $s_i(t) = b_i(t-1)$  for all  $t$ , hence  $s_i^* = b_i^*$ . Otherwise, letting  $\alpha_{ij} = d_{ij}/\delta_i = g_{ij}/\gamma_i$ , Eqs. (10) and (11) imply  $b_i^* = \sum_j \alpha_{ij} s_j^*$  and thus  $s_i^* = (b_i^* + \delta_i \sum_j \alpha_{ij} s_j^*)/(1 + \delta_i) = (b_i^* + \delta_i b_i^*)/(1 + \delta_i) = b_i^*$ . ■

Proposition 5 implies that hypocrisy is driven by audience tuning. Agents are non-hypocritical if they either do not adjust their statements to the audience ( $\delta_i = 0$ ) or their audience and influence weights coincide up to normalization. In particular, there is no hypocrisy when  $\mathbf{D} = \mathbf{0}$  (standard DeGroot learning) or when  $\mathbf{D} = \mathbf{G}$  (the “saying is believing” case studied in Anufriev et al., 2023).

## 4.2 Characterization of generic hypocrisy

The previous results identify two reasons why hypocrisy may fail to arise. First, as illustrated by the two-agent example (see Fig. 3), discrepancies between beliefs and statements may vanish for particular initial beliefs. Second, Proposition 5 shows that hypocrisy may also be absent when the relative weights governing an agent’s statement formation coincide exactly with those governing belief updating, even though the two sets of weights come from distinct interaction networks.

To abstract from such knife-edge cases, we use the term *generic* with respect to both initial beliefs and parameters, while preserving the zero pattern of  $\mathbf{D} \vee \mathbf{G}$ .<sup>9</sup>

Proposition 4 implies that only stubborn and remaining agents can be hypocritical. We characterize these agents separately. The following two statements are the main results of the paper: they provide graph-theoretic necessary and sufficient conditions for hypocrisy among stubborn and remaining agents, respectively.

**Proposition 6** (Generic hypocrisy of stubborn agents). *A stubborn agent  $s \in \mathcal{S}$  is generically hypocritical if and only if there exists an agent  $j$  with  $d_{sj} > 0$  such that  $j$  belongs to, or is linked to, an independent class in  $\mathbf{D} \vee \mathbf{G}$  other than  $\{s\}$ .*

*Proof.* See Appendix A.6. ■

<sup>9</sup>Formally, agent  $i$  is *generically hypocritical* with respect to initial beliefs and parameters if  $b_i^* \neq s_i^*$  for all initial beliefs outside a proper linear subspace of  $\mathbb{R}^N$  and for all values of the nonzero parameters in  $\mathbf{D}$  and  $\mathbf{G}$ , except on a zero-measure subset, conditional on a fixed  $\mathbf{D} \vee \mathbf{G}$ .

The condition for stubborn agents has a simple network interpretation in terms of the audience network  $\mathbf{D}$ . A stubborn agent becomes hypocritical whenever their public statement is tuned to agents whose long-run beliefs are influenced by a different independent class. In that case, the limiting belief of another class reaches the agent through the statement-formation process, so the limiting statement no longer coincides with the agent's own belief.

**Proposition 7** (Generic hypocrisy of remaining agents). *A remaining agent  $r \in \mathcal{R}$  is generically hypocritical unless at least one of the following conditions holds:*

- (i) *agent  $r$  is linked only to a single independent class in  $\mathbf{D} \vee \mathbf{G}$ ;*
- (ii) *all agents reachable from  $r$  in the audience network  $\mathbf{D}$  are remaining and are linked only to  $r$  in  $\mathbf{D} \vee \mathbf{G}$ ;*
- (iii) *all agents reachable from  $r$  in the influence network  $\mathbf{G}$  are remaining and are linked only to  $r$  in  $\mathbf{D} \vee \mathbf{G}$ .*

*Proof.* See Appendix A.6. ■

This result admits an interpretation parallel to Proposition 6, based on the mismatch between the audience and influence networks. A remaining agent becomes hypocritical whenever the flow of information through the audience network, which affects statement formation, differs from the flow of information through the influence network, which affects belief updating. In that case, distinct independent classes influence the agent through different channels, so the agent's limiting statement does not coincide with the limiting belief.

Proposition 7 identifies three exceptional network topologies under which this mismatch disappears. Condition (i) corresponds to a single source of influence: all relevant information originates from the same independent class. Conditions (ii) and (iii) correspond to *echo chambers* in the audience and influence networks, respectively. In these cases, the agent effectively interacts only with agents whose public and private opinions ultimately trace back to that same agent, so statements and beliefs coincide in the long run.

Thus, a remaining agent is hypocritical if and only if these conditions fail. More broadly, hypocrisy generically arises whenever agents either learn from individuals toward whom they do not tune their statements, or tune their statements toward individuals from whom they do not learn.

Taken together, these results imply a key insight: the presence of multiple independent classes is necessary but not sufficient for generic hypocrisy. In the presence of echo chambers, when the audience or sources of influence merely repeat

the same discourse, beliefs and statements coincide. By contrast, discrepancies arise when different classes affect an agent through distinct channels (the audience and influence networks), and are therefore weighted differently in belief updating and statement formation.

A simple sufficient condition for such discrepancies, and hence for hypocrisy, follows from Proposition 7. Suppose there exists an agent  $m$  reachable from  $r$  in the audience network  $\mathbf{D}$  whose long-run belief places positive weight to one independent class, and an agent  $n$  reachable from  $r$  in the influence network  $\mathbf{G}$  whose long-run belief places positive weight to another independent class. Then  $r$  is generically hypocritical. Conversely, if all such agents place unit weight on the same independent class, then  $r$  is non-hypocritical.

This sufficient condition is satisfied in the four-agent example in Fig. 2 for agent  $r = 1$ . Agent  $m = 2$  belongs to the audience of agent 1 and assigns full weight to its own independent class  $\{2\}$ . Agent  $n = 3$  influences the belief of agent 1 and tunes the statement to another independent class  $\{4\}$ . This distortion feeds into agent 1's belief updating and propagates back to agent 3 (cf. the discussion of the left panel of Fig. 4). Thus, agent 1's long-run belief is influenced by both independent classes  $\{2\}$  and  $\{4\}$  (through agent 3), but audience tuning shifts agent 1's statement more toward  $\{2\}$ , making agent 1 generically hypocritical. The converse case arises in the variation of Fig. 4 (right panel). Here, agents 2 and 4 form a single independent class  $\mathcal{C} = \{2, 4\}$ . All agents assign unit weight to this class, and there is no hypocrisy.

An important implication of Proposition 7 is that these examples admit a natural generalization. Agents 2 and 4 can be interpreted not as single individuals, but as representatives of broader groups of agents (recall the interpretation in footnote 5). Formally, each of them can be replaced by a collection of agents forming a pure communication class. In this interpretation, an independent class aggregates a set of agents whose views converge to a common long-run belief and therefore acts as a single source of influence.

Thus, hypocrisy arises from the interaction between heterogeneous information sources and distinct social channels. It emerges generically when society is sufficiently heterogeneous (there are multiple independent classes) and sufficiently interconnected (different classes affect agents through different pathways). Conversely, hypocrisy disappears in benchmark cases of consensus or echo chambers.

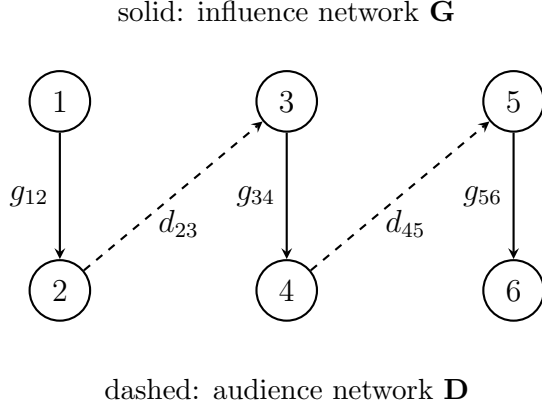


Figure 5: Propagation through audience tuning. Agents 2 and 4 do not observe agent 6, but transmit statements affected by agent 6, allowing beliefs to propagate beyond the influence network.

## 5 Discussion and applications

Our results show that discrepancies between beliefs and statements are not merely a short-run phenomenon but can persist in the long run as a consequence of the structure of social interactions. A key implication is that belief formation operates on a distorted layer of communication: agents update their beliefs based on statements that are themselves shaped by social pressure. When the networks governing communication and learning differ, the aggregation of information through statements does not coincide with the aggregation of underlying beliefs.

A useful interpretation of the model is that beliefs represent *private opinions*, while statements represent *publicly expressed opinions* or *narratives*. Agents do not observe private beliefs, but instead respond to narratives, which may systematically differ from the underlying beliefs. We illustrate the implications of this mechanism through various examples.

**Propagation through audience tuning.** The first example illustrates how audience tuning amplifies the spread of beliefs beyond the influence network. Consider the six-agent chain shown in Fig. 5. Agents 2, 4, and 6 are stubborn and are the only ones who influence agents 1, 3, and 5, respectively. In the audience network, agent 2 tunes statements toward agent 3, and agent 4 toward agent 5.<sup>10</sup>

We have  $b_6^* = s_6^* = b_6(0)$  and  $b_5^* = s_5^* = b_6(0)$ , so neither agent 6 nor agent 5

<sup>10</sup>One interpretation is inspired by reactions to official media influence in authoritarian regimes, where individuals may repeat state-sponsored narratives even while explicitly stating that they do not rely on state-controlled media. Agent 6 may represent a state TV channel, and agent 5 a viewer. Agent 4, who holds a different view, is a close acquaintance of agent 5 and moderates their public statements (e.g., on social media) to avoid dissonance. Agent 3 listens only to agent 4. Agents 2 and 1 play analogous roles with respect to agents 4 and 3.

is hypocritical. Agent 4, however, tunes their statements toward agent 5, so  $s_4^*$  is a weighted average of  $b_4^* = b_4(0)$  and  $b_5^* = b_6(0)$ . Agent 3 listens only to agent 4; hence  $b_3^* = s_3^* = s_4^*$ . Agent 2 tunes their statements toward agent 3, so  $s_2^*$  is a weighted average of  $b_2(0)$  and  $s_3^*$ . Agent 1 listens only to agent 2, so that

$$b_1^* = s_1^* = \frac{1}{1 + d_{23}} b_2(0) + \frac{d_{23}}{(1 + d_{23})(1 + d_{45})} b_4(0) + \frac{d_{23}d_{45}}{(1 + d_{23})(1 + d_{45})} b_6(0).$$

Thus, agent 1 is affected by agent 6 despite not being connected to them in the influence network. The key mechanism is that intermediate agents transmit statements that are shaped by their audiences rather than by their own beliefs.

More generally, audience tuning creates *indirect channels of influence*: beliefs propagate along paths that combine influence links with audience links. As a result, the effective reach of an information source is determined not only by the influence network  $\mathbf{G}$ , but by the multi-layer network  $\mathbf{D} \vee \mathbf{G}$ . In longer chains, this mechanism allows influence to travel arbitrarily far, even when the influence network itself is locally sparse.

This mechanism provides a simple foundation for the spread of propaganda. A source can influence agents far removed from it, not because they directly observe or trust it, but because others relay its message while adapting their statements to their audiences. Importantly, these intermediaries need not believe the message they transmit, so narratives can diffuse even when they are not widely believed.

At the aggregate level, this mechanism can also distort how information is weighted in society. In the absence of audience tuning, beliefs would reflect the independent sources of information (here, all three stubborn agents). With audience tuning, however, agents may place disproportionate weight on sources that are amplified through communication chains, so that certain narratives may dominate collective outcomes even when they originate from a small part of the network.<sup>11</sup>

The next two examples show how the *direction* of audience pressure shapes which beliefs are amplified in the population.

---

<sup>11</sup>While direct causal evidence on the micro-level mechanisms of propaganda transmission is limited, a growing empirical literature shows that its effects extend beyond directly exposed individuals. Yanagizawa-Drott (2014) shows that exposure to radio propaganda during the Rwandan genocide increased participation in violence not only in areas with direct signal coverage but also in neighboring regions, consistent with spillovers through interpersonal communication. Related evidence shows that social influence can propagate through networks at scale (Bond et al., 2012), and that media exposure has substantial persuasive effects (DellaVigna and Kaplan, 2007). In parallel, work on social pressure in opinion expression shows that individuals may adjust their publicly expressed views in response to their social environment (Bursztn et al., 2020). Taken together, these findings are consistent with a setting in which narratives propagate through audience network and need not coincide with underlying private beliefs.

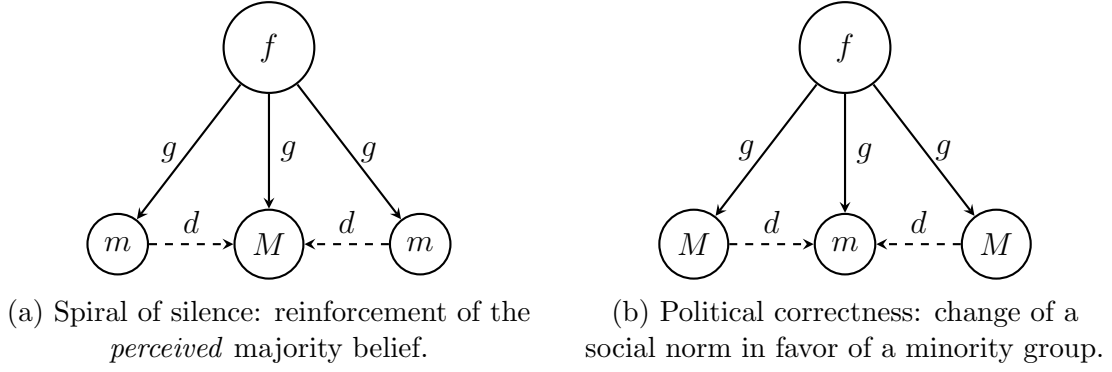


Figure 6: Audience (dashed lines) and influence (solid lines) networks in the spiral of silence (left panel) and political correctness (right panel) examples.

**Spiral of silence.** This example is inspired by the spiral of silence theory (Noelle-Neumann, 1974, 1984), according to which individuals who perceive their views to be socially unpopular may refrain from expressing them due to social pressure. Recent work provides empirical support for this mechanism, emphasizing the role of self-censorship and limited attention to silence in shaping beliefs and public discourse. Ho and Huang (2026) show experimentally that individuals holding socially inappropriate views are significantly more likely to self-censor, and that inattention to silence leads others to overestimate the prevalence of socially appropriate views. This generates a self-reinforcing dynamic in which distorted expression and biased inference amplify each other over time. In our framework, this mechanism is captured through audience tuning, which distorts public expression without affecting underlying beliefs.

Consider the four-agent network in Fig. 6a. Agents labeled  $m$  and  $M$  are stubborn, with  $b_m(0) \neq b_M(0)$ . Although there are two agents with belief  $m$ , they perceive their views to be in the minority and adjust their statements toward those of agent  $M$  through audience pressure. The parameter  $d$  captures the strength of this pressure. Agent  $f$  is a follower who forms beliefs based on the statements of the other agents.

While the private beliefs of the  $m$  agents remain unchanged, their statements are shifted toward  $M$ . As a result, the follower observes a distribution of statements that overrepresents the  $M$  position. The long-run belief of agent  $f$  is

$$b_f^* = b_M(0) + \frac{2}{3(1+d)}(b_m(0) - b_M(0)). \quad (12)$$

As  $d$  increases, the influence of the  $m$  agents diminishes, and the follower's belief tends to  $b_M(0)$ .<sup>12</sup>

<sup>12</sup>If there are  $k$  agents with belief  $m$ , each subject to dissonance  $d$  toward a single  $M$  agent, then

Thus, even without any change in private beliefs, social pressure can generate a misperception of consensus in favor of  $M$ . Although the  $m$  agents are in the actual majority, their views are underrepresented in public discourse, so the follower observes  $M$  as the dominant statement. The dominant view is therefore shaped not by the distribution of private beliefs, but by the direction of audience pressure, which suppresses the expression of perceived minority views.<sup>13</sup>

**Political correctness.** We now consider a closely related mechanism underlying political correctness, which reverses the direction of perceived social pressure. In contrast to the spiral of silence, where agents behave as if their views are in the minority, here agents behave as if expressing the majority view is socially discouraged. As a result, social pressure operates in the opposite direction: agents holding the majority belief adjust their statements toward a minority position.

Consider the network in Fig. 6b. Agents labeled  $M$  and  $m$  are stubborn, with  $b_M(0) \neq b_m(0)$ . Although agents with belief  $M$  are in the majority, they perceive society as discouraging the open expression of this view and consequently adjust their statements toward those of agent  $m$  in response to audience pressure. The parameter  $d$  captures the strength of this pressure. As before, agent  $f$  is a follower who learns from the statements of the other agents.

While the private beliefs of the  $M$  agents are not updated, their statements are shifted toward  $m$ . As a result, the follower observes a distribution of statements that overrepresents the minority position. The long-run belief of agent  $f$  is

$$b_f^* = b_m(0) + \frac{2}{3(1+d)}(b_M(0) - b_m(0)).$$

As  $d$  increases, the influence of the  $M$  agents falls, and the follower's belief tends to  $b_m(0)$ . Thus, even without any change in private beliefs, social pressure can generate a distortion in favor of a minority position.

Together with the previous example, a common conclusion emerges: the dominant view is shaped not by the distribution of private beliefs, but by the direction

---

the coefficient in Eq. (12) becomes  $k/(k+1)/(1+d)$ . Thus, regardless of how numerous the agents with minority beliefs are, for sufficiently large  $d$  the follower's belief is arbitrarily close to that of the  $M$  agent. If some of the  $m$  agents remain silent, this effectively removes their influence on  $f$  (i.e., the corresponding influence links), further accelerating convergence toward  $b_M(0)$ . The same conclusion holds if agent  $M$  is replaced by a pure communication class.

<sup>13</sup>This example is a static snapshot of a dynamic spiral. In a dynamic setting, agents like  $f$  may update their beliefs and adjust their own statements, reinforcing the perceived dominance of  $M$ . As in Ho and Huang (2026), the interaction between self-censorship and inference can generate self-reinforcing dynamics, in which distorted expression leads to biased beliefs, which in turn further suppress dissent. Our example captures the cross-sectional distortion induced by this mechanism, while a dynamic extension would generate the full spiral of silence.

of audience pressure. The same underlying mechanism, audience-driven distortion of statements, can amplify either minority or majority positions, depending on the direction of perceived social pressure induced by the audience network.

The above two examples suggest that communication distortions can affect not only individual beliefs but also aggregate outcomes. We now illustrate this point by extending these settings to a simple voting framework.

**Collective decisions under distorted communication.** Consider a binary state of the world,  $\mu \in \{0, 1\}$ , and suppose that agents interpret their beliefs as subjective probabilities that  $\mu = 1$ . After beliefs have converged, each agent casts a vote according to  $v_i = \mathbb{1}\{b_i^* \geq 1/2\}$ , that is, an agent votes for alternative 1 whenever their belief assigns probability at least  $1/2$  to  $\mu = 1$ , and votes for 0 otherwise.<sup>14</sup> The winning alternative is determined by majority rule. The analysis in Section 3 implies that the collective outcome depends on how the beliefs of independent classes are transmitted to the rest of the population.

To make this explicit for the spiral-of-silence and political-correctness examples, suppose the population consists of three groups: a share  $\lambda_M$  of stubborn agents with belief  $b_M(0)$ , a share  $\lambda_m$  of stubborn agents with different belief  $b_m(0)$ , and a share  $\lambda_f$  of followers, who place equal weight on all stubborn agents in their influence network, with  $\lambda_M + \lambda_m + \lambda_f = 1$ .

The long-run belief of the followers is

$$b_f^* = b_T(0) + \frac{\lambda_O}{(\lambda_T + \lambda_O)(1 + d)}(b_O(0) - b_T(0)),$$

where  $T = M$  and  $O = m$  in the spiral-of-silence example, and  $T = m$ ,  $O = M$  in the political-correctness example, so that  $b_T(0)$  and  $b_O(0)$  denote the belief of the position toward which statements are tuned and of the opposite side, respectively. When  $d$  is close to zero, the follower's belief reflects the weighted average of observed positions. As  $d$  increases, however, the follower's belief shifts toward  $b_T(0)$ , reflecting the growing distortion in communication.

The aggregate vote share for alternative 1 is then

$$V = \lambda_M \mathbb{1}\{b_M(0) \geq 1/2\} + \lambda_m \mathbb{1}\{b_m(0) \geq 1/2\} + \lambda_f \mathbb{1}\{b_f^* \geq 1/2\}.$$

This expression makes clear that communication distortions matter whenever followers are electorally relevant. Even if the long-run beliefs of the independent

---

<sup>14</sup>The function  $\mathbb{1}\{\cdot\}$  denotes the indicator function, which takes the value 1 if the condition inside the braces is satisfied and 0 otherwise.

classes themselves remain unchanged, audience tuning can move  $b_f^*$  beyond the threshold  $1/2$ . If  $\lambda_f$  is sufficiently large, this can reverse the majority outcome.

In the spiral-of-silence example, followers' beliefs move toward the position represented by  $M$ , even though agents with belief  $m$  are more numerous. If followers form a sufficiently large share of the population, stronger audience pressure can shift the collective decision toward the perceived majority position. This, in turn, can reinforce the perceived dominance of  $M$ , further discouraging the expression of opposing views and amplifying the distortion over time. In the political-correctness example, the same logic works in the opposite direction: stronger audience pressure can shift followers' beliefs toward the minority position and thereby tilt the aggregate outcome in its favor.

Thus, our framework shows that the social consequences of hypocrisy arise through distorted aggregation: society may act on beliefs that reflect the structure of communication rather than the underlying distribution of private opinions.

**Wisdom of crowds.** While the preceding discussion of collective decisions assumes a binary state of the world, the impact of communication distortions on information aggregation is particularly salient when the state is a continuous variable. A standard measure of the efficiency of social learning is the proximity of agents' views to the true state in the long run, commonly referred to as *wisdom* (see, e.g., Golub and Jackson, 2010; Buechel et al., 2015). Since our framework highlights the discrepancy between private beliefs and public statements, we can compare the efficiency of information aggregation along both dimensions.

Suppose that there is a true state of the world  $\mu \in \mathbb{R}$ , which is unknown to the agents. Initial beliefs are independent and unbiased signals of the true state: each  $b_i(0)$  is a random variable with expected value  $\mu$  and variance  $\sigma^2 > 0$ , and the random variables  $\{b_i(0)\}_{i=1}^N$  are pairwise uncorrelated.

The dynamics of the model imply that the limiting beliefs and statements are also random variables. We introduce *belief wisdom* (respectively, *statement wisdom*) as the mean squared error of the limiting belief (respectively, statement) as an estimator of the true state. Formally, for each agent  $i$ , belief wisdom  $\mathcal{E}_{b_i}$  and statement wisdom  $\mathcal{E}_{s_i}$  are defined as

$$\mathcal{E}_{b_i} = \mathbb{E} [(b_i^* - \mu)^2], \quad \text{and} \quad \mathcal{E}_{s_i} = \mathbb{E} [(s_i^* - \mu)^2].$$

Wisdom depends on the partition of agents into closed communication classes in  $\mathbf{D} \vee \mathbf{G}$ . We discuss belief and statement wisdom for different agents below (see Proposition A.3 in Appendix A.5 for the formal analysis).

Since agents in a pure communication class  $\mathcal{C}_m$  reach consensus, they are non-hypocritical, and hence their belief and statement wisdom coincide,  $\mathcal{E}_{s_m} = \mathcal{E}_{b_m}$ . Nevertheless, since the consensus belief is a weighted average of initial beliefs within the class, agents in  $\mathcal{C}_m$  become collectively wiser in the long run than they were initially. Unless the belief of  $\mathcal{C}_m$  is fully determined by a single influential agent, the limiting belief is closer to the truth than the initial beliefs. Intuitively, averaging independent signals causes idiosyncratic errors to partially cancel out, reducing the variance of the aggregate belief.

Since stubborn agents  $s \in \mathcal{S}$  never update beliefs, their belief wisdom equals the variance of the initial belief,  $\mathcal{E}_{b_s} = \sigma^2$ . At the same time, hypocrisy improves the informational quality of public communication. Although a hypocritical stubborn agent retains a noisy private belief, their public statement aggregates the long-run belief of a different independent class (see Proposition 6). Thus, for stubborn agents, audience dissonance makes statements more informative than beliefs and the public statement becomes a more accurate estimator of the truth:  $\mathcal{E}_{s_s} < \mathcal{E}_{b_s}$ .

Using stubborn agents as an example, we can illustrate the two effects that push wisdom in opposite directions. First, the *information diversification effect*: when stubborn agents have identical dissonance sensitivities to their audiences, statement wisdom improves with the number of independent classes to which they tune their statements. In a homogeneous society, the more independent sources of information are incorporated into the statement of a hypocritical stubborn agent, the closer that statement is to the truth.

Second, the *influence concentration effect*: statement wisdom deteriorates as the distribution of audience weights becomes more unequal. If a hypocritical stubborn agent experiences disproportionately strong audience dissonance toward some independent class, the accuracy of their statement is lower than when the weights are distributed equally, minimizing the variance of a weighted average. The interplay between these two effects, which depend on both the direction and intensity of social pressure, determines the statement wisdom of stubborn agents.

By similar logic, a remaining agent  $r \in \mathcal{R}$  also becomes wiser in the long run than initially, unless  $r$  is linked to a single stubborn agent in  $\mathbf{D} \vee \mathbf{G}$  or to a single pure communication class whose belief is determined by a single agent. Generically, agent  $r$ 's belief and statement wisdom coincide if and only if agent  $r$  is non-hypocritical; otherwise, there is no systematic ordering between the two. Intuitively, beliefs and statements of hypocritical remaining agents are affected by different independent classes (reflecting the information diversification effect) with different audience and influence weights (reflecting the influence concentration effect). Whether agent  $r$ 's belief is more accurate than their statement or vice

versa depends on the network structure and the parameter values in  $\mathbf{D}$  and  $\mathbf{G}$ , making the comparison between belief and statement wisdom ambiguous.

**Polarization and echo chambers.** Finally, we consider how the relationship between beliefs and statements depends on the structure of social interactions in polarized environments. A natural interpretation of polarization is that the influence network becomes segmented into groups, so that agents primarily learn from others within their own group. In our framework, such groups correspond to independent classes with distinct limiting beliefs.

In the extreme case of echo chambers, both the influence and audience networks are aligned and segregated across groups. Each group then forms an independent class, and beliefs and statements coincide within groups. In this case, polarization is not associated with hypocrisy: agents express views consistent with their beliefs, but disagreement persists across groups.

A different pattern arises when beliefs are polarized but communication remains cross-cutting. Suppose that agents update their beliefs within their own group, but adjust their statements in response to a broader audience that includes members of other groups. As a result, the networks governing belief formation and communication are misaligned. As shown in Propositions 6 and 7, in this case agents face pressure to adapt their statements to audiences whose beliefs differ from their own, generating discrepancies between beliefs and statements.

Thus, polarization does not necessarily increase hypocrisy. Rather, hypocrisy arises when polarized belief formation coexists with cross-group communication. When communication is also segregated, polarization leads to divergence in beliefs without distorting their public expression.

**Discussion.** The above examples highlight a common mechanism: statements aggregate information through the audience network, while beliefs aggregate information through the influence network. When these networks differ, belief formation operates on a distorted layer of communication, so that agents respond to narratives rather than directly to private opinions.

This distortion can amplify particular views, extend the reach of influence, and generate persistent discrepancies between beliefs and statements. It can also affect aggregate outcomes, as society may act on beliefs shaped by communication rather than by the underlying distribution of opinions. More generally, the model implies that the structure of social interactions determines both how information spreads and how it is ultimately aggregated in society.

## 6 Conclusion

In this paper, we develop a model of social interaction that distinguishes between audience and influence networks and explicitly separates privately held beliefs from publicly expressed statements. We show that when audience and influence networks differ, discrepancies between beliefs and statements may persist in the long run. Thus, audience tuning in conversations may lead agents to communicate opinions they do not hold, even after private beliefs have converged. We refer to this phenomenon as *hypocrisy*. Our main contribution is to provide a network-based characterization of conditions under which such hypocrisy arises.

We demonstrate how the structure of the multi-layer network combining audience and influence relations determines long-run beliefs and statements. We show that hypocrisy is not an exceptional phenomenon but emerges naturally when these two channels of interaction differ. In particular, hypocrisy arises when agents rely on different sources for belief formation and public expression. There is no discrepancy between private beliefs and public statements whenever agents ultimately rely on a single independent source of information or when the sources involved in belief updating and statement formation form an echo chamber that reinforces previous opinions.

Our model provides a unified perspective on several social phenomena. Since agents adjust their statements to their audience, opinions may spread *independently* of belief adoption: agents may transmit messages they do not believe, and these messages may shape the beliefs of others, potentially over long distances in the network. The same mechanism also helps interpret the spiral of silence and political correctness. In both cases, social pressure distorts public statements, which then feed into belief formation, so that collective outcomes may reflect perceived or socially acceptable narratives rather than underlying private beliefs.

The analysis abstracts from several important considerations. First, the model does not incorporate reputational concerns, signaling motives, or strategic manipulation of information, which are central in many environments. At the same time, demonstrating the emergence of hypocrisy in such a simple setting is meaningful, as it shows that these phenomena do not require sophisticated strategic behavior. Second, when updating their beliefs, agents take observed statements as primitives and do not attempt to infer the underlying beliefs or audience pressures that may have generated them. In many real environments, however, individuals do try to “read others’ minds” and discount messages that appear strategically or socially distorted. Third, we take the network structure as given, whereas in practice audience and influence networks may evolve endogenously.

These limitations suggest several directions for future research. One extension would be to embed audience tuning into a richer informational setting in which agents make inferences about the motives and beliefs behind others' statements. This would allow one to study how far strategic sophistication reduces or amplifies individual hypocrisy. Our results provide a natural foundation for such an analysis. Another promising direction is to endogenize network formation; see, for example, Anufriev et al. (2023) and Bolletta and Pin (2025) for recent work in related settings. It would also be valuable to study settings with private signals and misinformation, along the lines of Della Lena (2024), where individual and social learning interact, and better private information need not improve aggregate outcomes, due to audience tuning. Such extensions would facilitate a closer connection between the model and empirical settings, for instance by allowing one to quantify the extent of belief vs. statement divergence in real-world networks.

## Appendix

### A.1 Proof of Propositions 1 and 2

Proposition 1 follows from Propositions 1 and 2 in Anufriev et al. (2023). To prove Proposition 2, recall that  $\mathbf{s}(t) = \mathbf{P}\mathbf{b}(t-1)$ , where  $\mathbf{P} = (\mathbf{I} + \mathbf{\Delta} - \mathbf{D})^{-1}$ . By (6),

$$\mathbf{b}(t) = (\mathbf{I} + \mathbf{\Gamma})^{-1}(\mathbf{b}(t-1) + \mathbf{G}\mathbf{s}(t)) = (\mathbf{I} + \mathbf{\Gamma})^{-1}(\mathbf{I} + \mathbf{G}\mathbf{P})\mathbf{b}(t-1).$$

Therefore  $\mathbf{b}(t) = \mathbf{T}\mathbf{b}(t-1)$ , where  $\mathbf{T} = (\mathbf{I} + \mathbf{\Gamma})^{-1}(\mathbf{I} + \mathbf{G}\mathbf{P})$ . Since

$$(\mathbf{I} + \mathbf{\Delta} + \mathbf{G} - \mathbf{D})\mathbf{P} = (\mathbf{I} + \mathbf{\Delta} - \mathbf{D})\mathbf{P} + \mathbf{G}\mathbf{P} = \mathbf{I} + \mathbf{G}\mathbf{P},$$

this expression for  $\mathbf{T}$  coincides with (8).

It remains to show that  $\mathbf{T}$  is row-stochastic. Clearly  $\mathbf{T} \geq 0$  since  $\mathbf{P} \geq 0$  and  $\mathbf{G} \geq 0$ . Moreover,  $\mathbf{P}\mathbf{1} = \mathbf{1}$ , where  $\mathbf{1}$  denotes the vector of ones, and  $\mathbf{G}\mathbf{1} = \mathbf{\Gamma}\mathbf{1}$ , so

$$\mathbf{T}\mathbf{1} = (\mathbf{I} + \mathbf{\Gamma})^{-1}(\mathbf{I} + \mathbf{G}\mathbf{P})\mathbf{1} = (\mathbf{I} + \mathbf{\Gamma})^{-1}(\mathbf{1} + \mathbf{G}\mathbf{1}) = (\mathbf{I} + \mathbf{\Gamma})^{-1}(\mathbf{I} + \mathbf{\Gamma})\mathbf{1} = \mathbf{1}.$$

Thus  $\mathbf{T}$  is row-stochastic.

## A.2 Proof of Proposition 3

Since  $\mathbf{T} = (\mathbf{I} + \mathbf{\Gamma})^{-1}(\mathbf{I} + \mathbf{GP})$ , we have

$$t_{ij} = \frac{\mathbf{1}_{\{i=j\}} + \sum_k g_{ik} p_{kj}}{1 + \gamma_i}, \quad (\text{A.1})$$

where  $\mathbf{1}_{\{i=j\}}$  is the indicator function, equal to 1 if  $i = j$  and 0 otherwise.

*Part (i).* From (A.1),

$$t_{ii} = \frac{1 + \sum_k g_{ik} p_{ki}}{1 + \gamma_i} > 0.$$

*Part (ii).* Let  $i \neq j$ . By (A.1),

$$t_{ij} > 0 \iff \exists k \text{ such that } g_{ik} > 0 \text{ and } p_{kj} > 0.$$

By Proposition 1,  $p_{jj} > 0$ , and for  $k \neq j$ ,  $p_{kj} > 0$  iff there is a path from  $k$  to  $j$  in  $\mathbf{D}$ . Hence  $t_{ij} > 0$  iff there exists  $k$  such that  $g_{ik} > 0$  and either  $k = j$  or there is a path from  $k$  to  $j$  in  $\mathbf{D}$ .

*Part (iii).* If  $g_{ij} = 0$  for all  $j$ , then  $\gamma_i = 0$  and (A.1) gives  $t_{ii} = 1$ . Conversely, if  $g_{ik} > 0$  for some  $k \neq i$ , then by Proposition 1,  $p_{kk} > 0$ , so  $t_{ik} \geq g_{ik} p_{kk} / (1 + \gamma_i) > 0$ . Thus row  $i$  of  $\mathbf{T}$  has a positive off-diagonal entry, so  $t_{ii} < 1$ .

*Part (iv).* Suppose  $\mathbf{T}$  is strongly connected. Let  $i$  and  $j$  be arbitrary, and let  $i = v_0, \dots, v_m = j$  be a path in  $\mathbf{T}$ . By part (ii), each edge  $v_\ell \rightarrow v_{\ell+1}$  induces a path in  $\mathbf{D} \vee \mathbf{G}$ : either a  $\mathbf{G}$ -edge or a  $\mathbf{G}$ -edge followed by a path in  $\mathbf{D}$ . Concatenating these paths yields a path from  $i$  to  $j$  in  $\mathbf{D} \vee \mathbf{G}$ , so  $\mathbf{D} \vee \mathbf{G}$  is strongly connected.

Conversely, suppose each row of  $\mathbf{G}$  is non-zero and  $\mathbf{D} \vee \mathbf{G}$  is strongly connected. Fix  $i, j$ , and choose  $k$  with  $g_{ik} > 0$ . Since there is a path from  $k$  to  $j$  in  $\mathbf{D} \vee \mathbf{G}$ , appending  $i \rightarrow k$  yields a path from  $i$  to  $j$  whose first edge lies in  $\mathbf{G}$ . Partition this path into blocks, each consisting of one  $\mathbf{G}$ -edge followed by  $\mathbf{D}$ -edges. By part (ii), each block induces a positive edge in  $\mathbf{T}$  between its endpoints. These endpoints therefore form a path in  $\mathbf{T}$  from  $i$  to  $j$ , so  $\mathbf{T}$  is strongly connected.

## A.3 Characterization of long-run beliefs

We first characterize the closed communication classes of  $\mathbf{T}$ .

**Lemma A.1.** *The closed communication classes of  $\mathbf{T}$  are exactly:*

- (i) *the pure communication classes in  $\mathbf{D} \vee \mathbf{G}$ ; and*
- (ii) *the singleton sets  $\{s\}$  for stubborn agents  $s$ .*

*Proof.* Let  $\mathcal{C}$  be a pure communication class in  $\mathbf{D} \vee \mathbf{G}$ . Fix  $i, j \in \mathcal{C}$ . Since  $i$  is non-stubborn, there exists  $k$  with  $g_{ik} > 0$ . As  $\mathcal{C}$  is closed in  $\mathbf{D} \vee \mathbf{G}$ ,  $k \in \mathcal{C}$ . Since  $\mathcal{C}$  is strongly connected, there is a path from  $k$  to  $j$  in  $\mathbf{D} \vee \mathbf{G}$  within  $\mathcal{C}$ . Prepending  $i \rightarrow k$ , we obtain a path from  $i$  to  $j$  starting with a  $\mathbf{G}$ -edge. By Proposition 3(ii), this induces a path in  $\mathbf{T}$  contained in  $\mathcal{C}$ , so  $\mathbf{T}|_{\mathcal{C}}$  is strongly connected.

To show closedness, let  $i \in \mathcal{C}$  and  $t_{i\ell} > 0$ . By Proposition 3(ii), there exists  $k$  with  $g_{ik} > 0$  and either  $k = \ell$  or a path from  $k$  to  $\ell$  in  $\mathbf{D}$ . Since  $i \rightarrow k$  is in  $\mathbf{G}$  and  $\mathcal{C}$  is closed,  $k \in \mathcal{C}$ , and thus  $\ell \in \mathcal{C}$ . Hence  $\mathcal{C}$  is closed in  $\mathbf{T}$ .

Conversely, let  $\mathcal{C}$  be a closed class of  $\mathbf{T}$ . If  $\mathcal{C} = \{i\}$ , then  $t_{ii} = 1$ , so  $i$  is stubborn by Proposition 3(iii). If  $\mathcal{C}$  contains more than one agent, then it contains no stubborn agents, since each stubborn agent forms a singleton closed class of  $\mathbf{T}$ .

We first show that  $\mathcal{C}$  is strongly connected in  $\mathbf{D} \vee \mathbf{G}$ . If  $i, j \in \mathcal{C}$ , there is a path from  $i$  to  $j$  in  $\mathbf{T}$ . Each positive edge of  $\mathbf{T}$  induces a path in  $\mathbf{D} \vee \mathbf{G}$  by Proposition 3(ii), so concatenating these paths yields a path from  $i$  to  $j$  in  $\mathbf{D} \vee \mathbf{G}$ .

It remains to show that  $\mathcal{C}$  is closed in  $\mathbf{D} \vee \mathbf{G}$ . Fix  $i \in \mathcal{C}$ . If  $g_{i\ell} > 0$ , then  $t_{i\ell} > 0$ , so  $\ell \in \mathcal{C}$  by the closedness of  $\mathcal{C}$  in  $\mathbf{T}$ . If  $d_{i\ell} > 0$ , choose any  $v \in \mathcal{C} \setminus \{i\}$ . Since  $\mathcal{C}$  is a communication class of  $\mathbf{T}$ , there exists a path from  $v$  to  $i$  in  $\mathbf{T}$ . Let  $u$  be the predecessor of  $i$  on that path. Then  $u \in \mathcal{C}$  and  $t_{ui} > 0$ . By Proposition 3(ii), there exists  $k$  such that  $g_{uk} > 0$  and either  $k = i$  or there is a path from  $k$  to  $i$  in  $\mathbf{D}$ . Appending the edge  $i \rightarrow \ell$  of  $\mathbf{D}$ , we obtain a path from  $k$  to  $\ell$  in  $\mathbf{D}$ . Proposition 3(ii) then implies  $t_{u\ell} > 0$ , so  $\ell \in \mathcal{C}$ . Thus  $\mathcal{C}$  is a pure communication class in  $\mathbf{D} \vee \mathbf{G}$ .

Let  $s$  be stubborn. By Proposition 3(iii), row  $s$  of  $\mathbf{T}$  has a single nonzero entry equal to 1 on the diagonal, so  $\{s\}$  is a closed class.  $\blacksquare$

We can now partition  $\mathcal{N}$  into pure communication classes  $\mathcal{C}_1, \dots, \mathcal{C}_M$  in  $\mathbf{D} \vee \mathbf{G}$ , stubborn agents  $\mathcal{S}$ , and remaining agents  $\mathcal{R}$ . Let their sizes be  $C_1, \dots, C_M$ ,  $S$ , and  $R$ , respectively.

**Proposition A.1.** *For any initial beliefs  $\mathbf{b}(0)$ , the limit  $\mathbf{b}^* = \lim_{t \rightarrow \infty} \mathbf{b}(t)$  exists and can be written as*

$$\mathbf{b}^* = \{\mathbf{b}_1^*, \dots, \mathbf{b}_M^*, \mathbf{b}_S^*, \mathbf{b}_R^*\}^T.$$

For each  $m$ ,

$$\mathbf{b}_m^* = b^{*m} \mathbf{1}_{C_m} = [\boldsymbol{\pi}^m \mathbf{b}(0)|_{\mathcal{C}_m}] \mathbf{1}_{C_m},$$

where  $\mathbf{1}_{C_m}$  is the  $C_m$ -dimensional vector of ones and  $\boldsymbol{\pi}^m$  satisfies  $\boldsymbol{\pi}^m = \boldsymbol{\pi}^m \mathbf{T}|_{\mathcal{C}_m}$ , normalized so that its entries sum to one. For  $s \in \mathcal{S}$ ,  $b_s^* = b_s(0)$ , and for  $r \in \mathcal{R}$ ,

$$b_r^* = \sum_{m=1}^M \omega_{rm} b^{*m} + \sum_{s=1}^S \bar{\omega}_{rs} b_s(0),$$

where the weights  $\{\omega_{rm}\}_{m=1}^M$  and  $\{\bar{\omega}_{rs}\}_{s=1}^S$  are non-negative, sum to one, and are positive only for classes reachable from  $r$  in  $\mathbf{T}$ .

*Proof.* Since  $t_{ii} > 0$  for all  $i$ , the matrix  $\mathbf{T}$  is convergent (see, e.g., Theorem 2 in Golub and Jackson, 2010). Thus, for any initial beliefs  $\mathbf{b}(0)$ , the limit

$$\mathbf{b}^* \equiv \lim_{t \rightarrow \infty} \mathbf{T}^t \mathbf{b}(0)$$

exists and satisfies  $\mathbf{T}\mathbf{b}^* = \mathbf{b}^*$ .

By Lemma A.1, after a permutation of indices the matrix  $\mathbf{T}$  admits the block form

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_1 & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{T}_M & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_S & \mathbf{0} \\ \mathbf{T}_{\mathcal{R}1} & \cdots & \mathbf{T}_{\mathcal{R}M} & \mathbf{T}_{\mathcal{R}S} & \mathbf{T}_{\mathcal{R}\mathcal{R}} \end{pmatrix},$$

where the blocks  $\mathbf{T}_m = \mathbf{T}|_{\mathcal{C}_m}$  correspond to pure communication classes, the identity block corresponds to stubborn agents, and  $\mathcal{R}$  indexes the remaining agents.

It follows that for each  $t \geq 1$ ,

$$\mathbf{T}^t = \begin{pmatrix} \mathbf{T}_1^t & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{T}_M^t & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_S & \mathbf{0} \\ \mathbf{T}_{\mathcal{R}1}(t) & \cdots & \mathbf{T}_{\mathcal{R}M}(t) & \mathbf{T}_{\mathcal{R}S}(t) & \mathbf{T}_{\mathcal{R}\mathcal{R}}^t \end{pmatrix},$$

where  $\mathbf{T}_{\mathcal{R}m}(t)$  and  $\mathbf{T}_{\mathcal{R}S}(t)$  denote the corresponding blocks of  $\mathbf{T}^t$ , capturing the influence of class  $\mathcal{C}_m$  and stubborn agents on the remaining agents after  $t$  steps:

$$\mathbf{T}_{\mathcal{R}m}(t) = \sum_{\tau=0}^{t-1} \mathbf{T}_{\mathcal{R}\mathcal{R}}^\tau \mathbf{T}_{\mathcal{R}m} \mathbf{T}_m^{t-1-\tau}, \quad \text{and} \quad \mathbf{T}_{\mathcal{R}S}(t) = \sum_{\tau=0}^{t-1} \mathbf{T}_{\mathcal{R}\mathcal{R}}^\tau \mathbf{T}_{\mathcal{R}S}.$$

We use this block form to characterize the long-run beliefs of different agents.

Fix  $m \in \{1, \dots, M\}$ . By Proposition 2,  $\mathbf{T}_m$  is row-stochastic, and by Lemma A.1, the set  $\mathcal{C}_m$  is strongly connected. Moreover, Proposition 3(i) implies that all diagonal entries of  $\mathbf{T}_m$  are positive. Hence, by the Perron–Frobenius theorem,  $\lim_{t \rightarrow \infty} \mathbf{T}_m^t = \mathbf{1}_{C_m} \boldsymbol{\pi}^m$ , where  $\boldsymbol{\pi}^m$  is the unique positive left eigenvector of  $\mathbf{T}_m$  associated with eigenvalue 1, normalized so that  $\boldsymbol{\pi}^m \mathbf{1}_{C_m} = 1$ . Therefore,

$$\mathbf{b}_m^* = \lim_{t \rightarrow \infty} \mathbf{T}_m^t \mathbf{b}(0)|_{\mathcal{C}_m} = \mathbf{1}_{C_m} \boldsymbol{\pi}^m \mathbf{b}(0)|_{\mathcal{C}_m} = b^{*m} \mathbf{1}_{C_m},$$

with  $b^{*m} = \boldsymbol{\pi}^m \mathbf{b}(0)|_{\mathcal{C}_m}$ .

For each stubborn agent  $s$ , row  $s$  of  $\mathbf{T}$  has a single nonzero entry equal to 1 on the diagonal. Hence  $b_s(t) = b_s(0)$  for all  $t$ , and  $\mathbf{b}_s^* = \mathbf{b}(0)|_s$ .

Finally, consider remaining agents. Since no agent in  $\mathcal{R}$  belongs to a closed class of  $\mathbf{T}$ , it follows (see, e.g., Theorem 10 in DeMarzo et al., 2003) that

$$\lim_{t \rightarrow \infty} \mathbf{T}_{\mathcal{R}\mathcal{R}}^t = \mathbf{0}, \quad \text{and} \quad \sum_{\tau=0}^{\infty} \mathbf{T}_{\mathcal{R}\mathcal{R}}^{\tau} = (\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}})^{-1}.$$

Using the block form of  $\mathbf{T}^t$ , we obtain

$$\lim_{t \rightarrow \infty} \mathbf{T}_{\mathcal{R}m}(t) = (\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}})^{-1} \mathbf{T}_{\mathcal{R}m} \mathbf{1}_{C_m} \boldsymbol{\pi}^m, \quad \text{and} \quad \lim_{t \rightarrow \infty} \mathbf{T}_{\mathcal{R}s}(t) = (\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}})^{-1} \mathbf{T}_{\mathcal{R}s}.$$

Therefore, the long-run beliefs of remaining agents are weighted averages of the long-run beliefs of stubborn agents and pure communication classes:

$$\mathbf{b}_{\mathcal{R}}^* = (\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}})^{-1} \left( \sum_{m=1}^M \mathbf{T}_{\mathcal{R}m} \mathbf{1}_{C_m} b^{*m} + \mathbf{T}_{\mathcal{R}s} \mathbf{b}_s^* \right).$$

Define  $\boldsymbol{\Omega} \equiv (\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}})^{-1} (\mathbf{T}_{\mathcal{R}1} \mathbf{1}_{C_1} \cdots \mathbf{T}_{\mathcal{R}M} \mathbf{1}_{C_M} \mathbf{T}_{\mathcal{R}s})$ , with entries  $\{\omega_{rm}\}_{r,m=1}^{R,M}$  and  $\{\bar{\omega}_{rs}\}_{r,s=1}^{R,S}$ . All these coefficients are non-negative. Since  $\mathbf{T}$  is row-stochastic,

$$\sum_{m=1}^M \mathbf{T}_{\mathcal{R}m} \mathbf{1}_{C_m} + \mathbf{T}_{\mathcal{R}s} \mathbf{1}_S + \mathbf{T}_{\mathcal{R}\mathcal{R}} \mathbf{1}_R = \mathbf{1}_R,$$

so multiplying by  $(\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}})^{-1}$  yields  $\boldsymbol{\Omega} \mathbf{1} = \mathbf{1}_R$ . Thus each row of  $\boldsymbol{\Omega}$  sums to one.

Finally, from the series representation of  $\boldsymbol{\Omega}$ , a weight  $\omega_{rm}$  or  $\bar{\omega}_{rs}$  is positive only if there exists a path from  $r$  to the corresponding independent class in  $\mathbf{T}$ . ■

## A.4 Eigenvector centralities

To simplify notation, suppose that  $\mathbf{D} \vee \mathbf{G}$  is strongly connected and that there are no stubborn agents (otherwise, our analysis applies to a single pure communication class in  $\mathbf{D} \vee \mathbf{G}$ ). By Proposition 3(iv),  $\mathbf{T}$  is strongly connected. By Proposition A.1, agents reach consensus and their limiting belief is given by  $b^* = \boldsymbol{\pi} \mathbf{b}(0)$ , where  $\boldsymbol{\pi}$  satisfies  $\boldsymbol{\pi} = \boldsymbol{\pi} \mathbf{T}$  and consists of eigenvector centralities. Using Eq. (8), we have

$$\begin{aligned} \boldsymbol{\pi} &= \boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Gamma})^{-1} (\mathbf{I} + \boldsymbol{\Delta} + \mathbf{G} - \mathbf{D}) (\mathbf{I} + \boldsymbol{\Delta} - \mathbf{D})^{-1} && \Longleftrightarrow \\ \boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Delta} - \mathbf{D}) &= \boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Gamma})^{-1} (\mathbf{I} + \boldsymbol{\Delta} + \mathbf{G} - \mathbf{D}) && \Longleftrightarrow \\ \boldsymbol{\pi} (\mathbf{I} - (\mathbf{I} + \boldsymbol{\Gamma})^{-1}) (\mathbf{I} + \boldsymbol{\Delta} - \mathbf{D}) &= \boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Gamma})^{-1} \mathbf{G} && \Longleftrightarrow \\ \boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma} (\mathbf{I} + \boldsymbol{\Delta} - \mathbf{D}) &= \boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Gamma})^{-1} \mathbf{G}. \end{aligned}$$

The above equality can be written as

$$\boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma} (\mathbf{I} + \boldsymbol{\Delta}) = \boldsymbol{\pi} (\mathbf{I} + \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma} (\mathbf{D} + \boldsymbol{\Gamma}^{-1} \mathbf{G}),$$

which yields Eq. (9) componentwise.

The following result characterizes the comparative statics of eigenvector centralities with respect to audience and influence weights (elements of  $\mathbf{D}$  and  $\mathbf{G}$ ).

**Proposition A.2.** *Suppose that  $\mathbf{D} \vee \mathbf{G}$  is strongly connected and that there are no stubborn agents. Let  $\boldsymbol{\pi} = \{\pi_i\}_{i \in \mathcal{N}}$  denote the eigenvector centralities.*

- (i) *Suppose that  $\max\{d_{ij}, g_{ij}\} > 0$ . If either  $d_{ij}$  or  $g_{ij}$  increases, then  $\pi_i$  strictly decreases and  $\pi_j$  strictly increases.*
- (ii) *The relative change in eigenvector centrality of any other agent is no greater than that of agent  $j$  and no smaller than that of agent  $i$ .*

*Proof.* It follows from Eq. (9) that  $\boldsymbol{\pi}$  is the positive unit left eigenvector of the non-negative, strongly connected matrix  $\mathbf{X}$  with  $x_{ii} = 0$  and

$$x_{ij} = \frac{1 + \gamma_j}{\gamma_j} \frac{1}{1 + \delta_j} \left( d_{ij} + \frac{g_{ij}}{\gamma_i} \right) \frac{\gamma_i}{1 + \gamma_i}, \quad i \neq j.$$

Let  $\max\{d_{ij}, g_{ij}\} > 0$ . If  $d_{ij}$  increases to  $\tilde{d}_{ij} > d_{ij}$ , this perturbation increases the element  $x_{ij}$ , thereby increasing the  $j$ -th column of  $\mathbf{X}$ . Simultaneously, it increases the sum  $\delta_i$ . Because  $\delta_i$  appears in the denominator of every element in the  $i$ -th column, the entire  $i$ -th column strictly decreases. All other columns remain unchanged. Similarly, increasing  $g_{ij}$  to  $\tilde{g}_{ij} > g_{ij}$  directly increases  $x_{ij}$  (increasing column  $j$ ) and increases  $\gamma_i$ , which decreases the factor  $\frac{1+\gamma_i}{\gamma_i}$  present in  $x_{hi}$ , again strictly decreasing the entire  $i$ -th column of  $\mathbf{X}$ .

In both cases, the perturbed matrix  $\tilde{\mathbf{X}}$  is formed by strictly increasing elements in column  $j$  and strictly decreasing elements in column  $i$ . Let  $\tilde{\boldsymbol{\pi}}$  be the normalized left eigenvector of  $\tilde{\mathbf{X}}$ . Part (ii) follows from Lemma 1 in Dietzenbacher (1990), stating that for any agent  $k$ , the relative changes in eigenvector centrality satisfy

$$\frac{\tilde{\pi}_j}{\pi_j} \geq \frac{\tilde{\pi}_k}{\pi_k} \geq \frac{\tilde{\pi}_i}{\pi_i}. \quad (\text{A.2})$$

To prove part (i), note that since  $\mathbf{X}$  is strongly connected, the non-trivial column perturbations guarantee that  $\boldsymbol{\pi}$  and  $\tilde{\boldsymbol{\pi}}$  are not proportional; hence, the inequalities in (A.2) are strict for at least one  $k$ . Rearranging the first inequality yields  $\pi_k/\pi_j \geq \tilde{\pi}_k/\tilde{\pi}_j$ . Summing over all  $k$  and using  $\sum_k \pi_k = \sum_k \tilde{\pi}_k = 1$ , we

obtain:

$$\frac{1}{\pi_j} = \sum_k \frac{\pi_k}{\pi_j} > \sum_k \frac{\tilde{\pi}_k}{\tilde{\pi}_j} = \frac{1}{\tilde{\pi}_j},$$

which implies  $\tilde{\pi}_j > \pi_j$ . Applying the same logic to the second inequality in (A.2) yields  $\tilde{\pi}_i < \pi_i$ .  $\blacksquare$

## A.5 Wisdom of crowds

The following result compares belief and statement wisdom for pure communication classes, stubborn agents, and remaining agents.

**Proposition A.3.** *1. For agents in a pure communication class  $\mathcal{C}_m$ ,  $\mathcal{E}_{b_m} \leq \sigma^2$ , with equality if and only if  $\pi_i^m = 1$  for some  $i \in \mathcal{C}_m$ . Moreover,  $\mathcal{E}_{s_m} = \mathcal{E}_{b_m}$ .*

*2. For a stubborn agent  $s \in \mathcal{S}$ ,  $\mathcal{E}_{b_s} = \sigma^2$ . Moreover,  $\mathcal{E}_{s_s} \leq \mathcal{E}_{b_s}$ , with equality if and only if stubborn agent  $s$  is non-hypocritical.*

*3. For a remaining agent  $r \in \mathcal{R}$ ,  $\mathcal{E}_{b_r} \leq \sigma^2$  and  $\mathcal{E}_{s_r} \leq \sigma^2$ . Moreover, generically  $\mathcal{E}_{s_r} = \mathcal{E}_{b_r}$  if and only if agent  $r$  is non-hypocritical. If agent  $r$  is hypocritical, then either  $\mathcal{E}_{s_r} < \mathcal{E}_{b_r}$  or  $\mathcal{E}_{s_r} > \mathcal{E}_{b_r}$ , depending on the parameter values.*

*Proof.* Recall that  $\mathbf{b}^* = \mathbf{T}^\infty \mathbf{b}(0)$  and  $\mathbf{T}^\infty$  is row-stochastic. Moreover,  $\mathbf{s}^* = \mathbf{P} \mathbf{b}^*$  and  $\mathbf{P}$  is row-stochastic. Therefore, both limiting beliefs and statements are unbiased:  $\mathbb{E} \mathbf{b}^* = \mathbb{E} \mathbf{s}^* = \mu \mathbf{1}$ . Since the mean squared error of an estimator equals the sum of its squared bias and variance, belief wisdom of agent  $i$  equals the variance of the limiting belief:

$$\mathcal{E}_{b_i} = (\mathbb{E}[b_i^* - \mu])^2 + \text{Var}[b_i^*] = \text{Var}[b_i^*].$$

The same holds true for the statement wisdom:  $\mathcal{E}_{s_i} = \text{Var}[s_i^*]$ .

*Part (i).* All agents in a pure communication class  $\mathcal{C}_m$  converge to the same limiting belief  $b^{*m} = \sum_{i \in \mathcal{C}_m} \pi_i^m b_i(0)$ , where  $\pi_i^m$  is the eigenvector centrality of agent  $i$  within  $\mathcal{C}_m$ . Since the initial beliefs are pairwise uncorrelated, belief wisdom of each agent in  $\mathcal{C}_m$  is the same and is given by

$$\mathcal{E}_{b_m} = \text{Var} \left[ \sum_{i \in \mathcal{C}_m} \pi_i^m b_i(0) \right] = \sum_{i \in \mathcal{C}_m} \text{Var} [\pi_i^m b_i(0)] = \sum_{i \in \mathcal{C}_m} (\pi_i^m)^2 \sigma^2.$$

Since  $\pi_i^m \in [0, 1]$  and  $\sum_{i \in \mathcal{C}_m} \pi_i^m = 1$ , we have

$$\mathcal{E}_{b_m} = \sum_{i \in \mathcal{C}_m} (\pi_i^m)^2 \sigma^2 \leq \sum_{i \in \mathcal{C}_m} \pi_i^m \sigma^2 = \sigma^2,$$

with equality if only if  $\pi_i^m = 1$  for some  $i$ . Moreover, since agents in  $\mathcal{C}_m$  are non-hypocritical and share the same limiting statement, we have  $\mathcal{E}_{s_m} = \mathcal{E}_{b_m}$ .

*Part (ii).* Since for a stubborn agent  $s \in \mathcal{S}$ ,  $b_s^* = b_s(0)$ , their belief wisdom is given by  $\mathcal{E}_{b_s} = \sigma^2$ . Furthermore, their limiting statement can be written as

$$\begin{aligned} s_s^* &= \sum_j p_{sj} b_j^* = \sum_{\varsigma=1}^S p_{s\varsigma} b_\varsigma(0) + \sum_{m=1}^M \sum_{i \in \mathcal{C}_m} p_{si} \pi_i^m b_i(0) \\ &\quad + \sum_{r \in \mathcal{R}} p_{sr} \left( \sum_{m=1}^M \omega_{rm} \sum_{i \in \mathcal{C}_m} \pi_i^m b_i(0) + \sum_{\varsigma=1}^S \bar{\omega}_{r\varsigma} b_\varsigma(0) \right) \\ &= \sum_{\varsigma=1}^S \left( p_{s\varsigma} + \sum_{r \in \mathcal{R}} p_{sr} \bar{\omega}_{r\varsigma} \right) b_\varsigma(0) + \sum_{m=1}^M \sum_{i \in \mathcal{C}_m} \left( p_{si} + \sum_{r \in \mathcal{R}} p_{sr} \omega_{rm} \right) \pi_i^m b_i(0), \end{aligned}$$

Again, given that the initial beliefs are uncorrelated,

$$\mathcal{E}_{s_s} = \sum_{\varsigma=1}^S \left[ p_{s\varsigma} + \sum_{r \in \mathcal{R}} p_{sr} \bar{\omega}_{r\varsigma} \right]^2 \sigma^2 + \sum_{m=1}^M \left[ \sum_{i \in \mathcal{C}_m} \left( p_{si} + \sum_{r \in \mathcal{R}} p_{sr} \omega_{rm} \right) \pi_i^m \right]^2 \sigma^2.$$

Since the sum of weights is equal to 1, we have  $\mathcal{E}_{s_s} \leq \sigma^2$ . Moreover, since  $p_{ss} > 0$ , equality holds if and only if  $p_{ss} + \sum_{r \in \mathcal{R}} p_{sr} \bar{\omega}_{rs} = 1$ , that is, if and only if  $s_s^* = b_s^*$ .

*Part (iii).* Since the limiting belief of a remaining agent  $r \in \mathcal{R}$  is a weighted average of the limiting beliefs of the independent classes that influence agent  $r$  (see Proposition A.1), we have  $b_r^* = \sum_j \lambda_j^b b_j(0)$ , where  $\sum_j \lambda_j^b = 1$ . Moreover, since  $\mathbf{s}^* = \mathbf{P}\mathbf{b}^*$ , we have  $s_r^* = \sum_k \lambda_k^s b_k(0)$ , where  $\sum_k \lambda_k^s = 1$  as well. Hence,

$$\mathcal{E}_{b_r} = \sum_j (\lambda_j^b)^2 \sigma^2, \quad \text{and} \quad \mathcal{E}_{s_r} = \sum_k (\lambda_k^s)^2 \sigma^2.$$

Clearly,  $\mathcal{E}_{b_r} \leq \sigma^2$ , with equality if and only if either  $\lambda_s^b = 1$  for some  $s \in \mathcal{S}$  or  $\lambda_m^b = 1$  for some  $\mathcal{C}_m$  such that  $b^{*m} = b_i(0)$  for  $i \in \mathcal{C}_m$ . The same holds for  $\mathcal{E}_{s_r}$ .

We have  $\mathcal{E}_{b_r} = \mathcal{E}_{s_r}$  if and only if  $\sum_j (\lambda_j^b)^2 = \sum_k (\lambda_k^s)^2$ , which generically holds if and only if  $\lambda_j^b = \lambda_j^s$  for all  $j$ , that is, if and only if  $b_r^* = s_r^*$ . Otherwise,  $\mathcal{E}_{b_r} \leq \mathcal{E}_{s_r}$  if and only if  $\sum_j (\lambda_j^b)^2 \leq \sum_k (\lambda_k^s)^2$ , and which sum is larger depends on the network structure and the parameter values in  $\mathbf{D}$  and  $\mathbf{G}$ .  $\blacksquare$

## A.6 Proof of Propositions 6 and 7

Without loss of generality, and to simplify notation, suppose that all independent classes in  $\mathbf{D} \vee \mathbf{G}$  consist of stubborn agents. From the perspective of other agents, a pure communication class in the long run reduces to a single unified source of a fixed statement, behaving exactly like a stubborn agent.

Let  $\mathcal{S}$  be the set of stubborn agents of size  $S$ , and  $\mathcal{R}$  be the set of remaining agents of size  $R$ . Without loss of generality, it is sufficient to consider the case where  $\mathcal{R}$  is strongly connected within  $\mathcal{R}$ : for all  $i, j \in \mathcal{R}$  there exist paths in  $\mathbf{D} \vee \mathbf{G}$

from  $i$  to  $j$  and from  $j$  to  $i$ . Agents that are not reachable do not affect the long-run beliefs and statements of the remaining agents.

The relevant matrices  $\mathbf{D}$  and  $\mathbf{G}$  admit the following block form:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{\mathcal{SS}} & \mathbf{D}_{\mathcal{SR}} \\ \mathbf{D}_{\mathcal{RS}} & \mathbf{D}_{\mathcal{RR}} \end{pmatrix}, \quad \text{and} \quad \mathbf{G} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{G}_{\mathcal{RS}} & \mathbf{G}_{\mathcal{RR}} \end{pmatrix},$$

where  $R \times S$  matrices  $\mathbf{D}_{\mathcal{RS}}$  and  $\mathbf{G}_{\mathcal{RS}}$  represent the interaction of remaining agents with stubborn agents, and  $R \times R$  matrices  $\mathbf{D}_{\mathcal{RR}}$  and  $\mathbf{G}_{\mathcal{RR}}$  represent the interaction of remaining agents among themselves.

We begin by deriving the statements of remaining agents  $\mathbf{s}_{\mathcal{R}}^*$  via statements of stubborn agents  $\mathbf{s}_{\mathcal{S}}^*$ . In the matrix form, Eq. (10) reads

$$\mathbf{b}_{\mathcal{R}}^* = \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RS}} \mathbf{s}_{\mathcal{S}}^* + \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RR}} \mathbf{s}_{\mathcal{R}}^*,$$

where  $\mathbf{\Gamma}_R^{-1}$  is the  $R \times R$  diagonal matrix whose  $r$ -entry is  $1/\gamma_r$  corresponding to remaining agent  $r$ . Similarly, Eq. (11) in the matrix form is given by

$$\mathbf{s}_{\mathcal{R}}^* = \mathbf{H}_R \left( \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RS}} + \mathbf{D}_{\mathcal{RS}} \right) \mathbf{s}_{\mathcal{S}}^* + \mathbf{H}_R \left( \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RR}} + \mathbf{D}_{\mathcal{RR}} \right) \mathbf{s}_{\mathcal{R}}^*,$$

where  $\mathbf{H}_R \equiv (\mathbf{I} + \mathbf{\Delta}_R)^{-1}$  is the  $R \times R$  diagonal matrix whose  $r$ -entry is  $1/(1 + \delta_r)$ .

Multiplying both sides by  $\mathbf{H}_R^{-1}$  and rearranging the terms, we get

$$\left( \mathbf{I} - \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RR}} + \mathbf{\Delta}_R - \mathbf{D}_{\mathcal{RR}} \right) \mathbf{s}_{\mathcal{R}}^* = \left( \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RS}} + \mathbf{D}_{\mathcal{RS}} \right) \mathbf{s}_{\mathcal{S}}^*.$$

Since  $\mathcal{R}$  is strongly connected, the matrix  $\mathbf{I} - \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RR}} + \mathbf{\Delta}_R - \mathbf{D}_{\mathcal{RR}}$  is irreducible. Moreover, it is diagonally dominant, and hence invertible. Then  $\mathbf{s}_{\mathcal{R}}^* = \mathbf{\Psi} \mathbf{s}_{\mathcal{S}}^*$ , where the  $R \times S$  matrix  $\mathbf{\Psi}$  is given by

$$\mathbf{\Psi} = \left( \mathbf{I} - \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RR}} + \mathbf{\Delta}_R - \mathbf{D}_{\mathcal{RR}} \right)^{-1} \left( \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RS}} + \mathbf{D}_{\mathcal{RS}} \right).$$

Note that  $\mathbf{\Psi}$  is such that its  $(r, s)$ -entry  $\psi_{rs}$  is determined by all paths in  $\mathbf{D} \vee \mathbf{G}$  from remaining agent  $r$  to stubborn agent  $s$  of all lengths  $t = 1, 2, \dots$ . Moreover,  $\mathbf{\Psi}$  is row-stochastic, since  $(\mathbf{I} - \mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RR}} + \mathbf{\Delta}_R - \mathbf{D}_{\mathcal{RR}}) \mathbf{1}_R = (\mathbf{\Gamma}_R^{-1} \mathbf{G}_{\mathcal{RS}} + \mathbf{D}_{\mathcal{RS}}) \mathbf{1}_S$ .

Thus, *long-run statements of stubborn agents are generically independent*. Since  $\mathbf{b}^* = \mathbf{P}^{-1} \mathbf{s}^* = (\mathbf{I} + \mathbf{\Delta} - \mathbf{D}) \mathbf{s}^*$ , and  $\mathbf{D}$  has the block form, we have

$$\mathbf{b}_{\mathcal{S}}^* = (\mathbf{I} + \mathbf{\Delta}_S - \mathbf{D}_{\mathcal{SS}}) \mathbf{s}_{\mathcal{S}}^* - \mathbf{D}_{\mathcal{SR}} \mathbf{s}_{\mathcal{R}}^* = (\mathbf{I} + \mathbf{\Delta}_S - \mathbf{D}_{\mathcal{SS}} - \mathbf{D}_{\mathcal{SR}} \mathbf{\Psi}) \mathbf{s}_{\mathcal{S}}^*.$$

Since  $\mathbf{\Psi}$  is row-stochastic, the right-hand side matrix is strictly diagonally domi-

nant. Then it is invertible, the columns of its inverse form a basis in  $\mathbb{R}^S$ , and its inverse uniquely transforms (generically independent)  $\mathbf{b}_s^*$  into  $\mathbf{s}_s^*$ .

We now prove Proposition 6. A stubborn agent  $i$  is non-hypocritical if and only if  $b_i^* = s_i^*$ . Using Eq. (11), this condition takes the form

$$s_i^* = \sum_j \frac{d_{ij}}{\delta_i} s_j^* = \sum_{s \in \mathcal{S}} \frac{d_{is}}{\delta_i} s_s^* + \sum_{r \in \mathcal{R}} \frac{d_{ir}}{\delta_i} s_r^*.$$

Taking into account that  $s_r^* = \sum_s \psi_{rs} s_s^*$ , the above condition can be written as

$$\left( \delta_i - \sum_{r \in \mathcal{R}} d_{ir} \psi_{ri} \right) s_i^* - \sum_{s \in \mathcal{S}, s \neq i} \left( d_{is} + \sum_{r \in \mathcal{R}} d_{ir} \psi_{rs} \right) s_s^* = 0. \quad (\text{A.3})$$

Since long-run statements of stubborn agents  $\mathbf{s}_s^*$  are independent, (A.3) holds generically if and only if each coefficient is zero. Since  $\Psi$  is row-stochastic, this is possible if and only if either  $\delta_i = 0$  (agent  $i$  has no audience) or  $d_{is} = 0$  for all  $s \neq i$  and  $\psi_{ri} = 1$  for all  $r$  with  $d_{ir} > 0$  (that is, there is no path in  $\mathbf{D} \vee \mathbf{G}$  from  $i$  to any other independent class). In all other cases,  $i$  is generically hypocritical.

We now turn to Proposition 7. Using Eqs. (10) and (11), and the definition of  $\Psi$ , the condition  $b_i^* = s_i^*$  for a remaining agent  $i$  can be written as

$$\sum_j \frac{g_{ij}}{\gamma_i} s_j^* = \sum_j \frac{d_{ij}}{\delta_i} s_j^* \iff \sum_{s \in \mathcal{S}} \left( \frac{g_{is}}{\gamma_i} + \sum_{r \in \mathcal{R}} \frac{g_{ir}}{\gamma_i} \psi_{rs} \right) s_s^* = \sum_{s \in \mathcal{S}} \left( \frac{d_{is}}{\delta_i} + \sum_{r \in \mathcal{R}} \frac{d_{ir}}{\delta_i} \psi_{rs} \right) s_s^*.$$

Since  $\mathbf{s}_s^*$  are independent, agent  $i$  is generically non-hypocritical if and only if for all  $s$

$$\frac{g_{is}}{\gamma_i} + \sum_{r \in \mathcal{R}} \frac{g_{ir}}{\gamma_i} \psi_{rs} = \frac{d_{is}}{\delta_i} + \sum_{r \in \mathcal{R}} \frac{d_{ir}}{\delta_i} \psi_{rs}. \quad (\text{A.4})$$

Generically, (A.4) holds only in three cases. First, when agent  $i$  is linked only to a single independent class  $\{\bar{s}\}$  in  $\mathbf{D} \vee \mathbf{G}$ . In this case, for all reachable remaining agents  $r$ ,  $\psi_{r\bar{s}} = 1$  and  $\psi_{rs} = 0$  for all other stubborn agents  $s$ . By the row-stochasticity of  $\Psi$ , (A.4) reduces to  $1 = 1$  for  $\bar{s}$  and  $0 = 0$  for all other  $s$ .

Second, when agent  $i$ 's audience consists only of remaining agents who are linked exclusively to  $i$  in  $\mathbf{D} \vee \mathbf{G}$ . In this case, the statements of all agents to whom  $i$  tunes their message replicate agent  $i$ 's statement:  $s_j^* = s_i^*$  for all  $j$  such that  $d_{ij} > 0$ . Thus, agent  $i$ 's statement satisfies

$$s_i^* = \frac{1}{1 + \delta_i} b_i^* + \frac{\delta_i}{1 + \delta_i} s_i^*,$$

implying that  $s_i^* = b_i^*$ . It follows that the elements of  $\Psi$  satisfy  $\psi_{js} = \psi_{is}$  for all  $j$  such that  $d_{ij} > 0$ , and hence the right-hand side of (A.4) simplifies to  $\psi_{is}$  for each

s. Moreover, the left-hand side of (A.4), which represents the weight of agent  $s$ 's statement in agent  $i$ 's belief, also coincides with  $\psi_{is}$  as  $s_i^* = b_i^*$ .

Third, when agent  $i$ 's sources of influence are only remaining agents who are linked to  $i$  in  $\mathbf{D} \vee \mathbf{G}$ . This case can be analyzed similarly to the one above.

Suppose that conditions (i)–(iii) from Proposition 7 do not hold. Since condition (i) fails, there are at least two independent classes reachable from  $i$  in  $\mathbf{D} \vee \mathbf{G}$ . Then one of the following must hold. First, the set of independent classes affecting agent  $i$ 's statement differs from the set of classes affecting  $i$ 's belief. In this case, there exists a stubborn agent  $s$  whose statement receives positive weight on one side of (A.4) and zero weight on the other, implying that (A.4) fails for this  $s$ .

Second, the same set of independent classes (say,  $\{\bar{s}\} \cup \{\underline{s}\}$ ) affects both agent  $i$ 's belief and statement. Then the coefficients can be perturbed so that equality in (A.4) fails. Indeed, since conditions (ii)–(iii) fail, both sides of (A.4) for both  $s = \bar{s}$  and  $s = \underline{s}$  must be strictly between 0 and 1. Moreover, either  $0 < d_{i\bar{s}} < 1$  and  $0 < d_{i\underline{s}} < 1$ , or there exists a remaining agent  $r$ , reachable from  $i$  in  $\mathbf{D} \vee \mathbf{G}$ , who is linked to  $\{\bar{s}\}$  or  $\{\underline{s}\}$  not through  $i$ . In the former case, perturbing  $d_{i\bar{s}}$  changes the two sides of (A.4) differently. In the latter case, either  $\psi_{r\bar{s}}$  varies independently of  $d_{i\bar{s}}$  or  $\psi_{r\underline{s}}$  varies independently of  $d_{i\underline{s}}$ .

Thus, outside the cases described in conditions (i)–(iii), Eq. (A.4) may hold only on a zero-measure set of parameters, implying that agent  $i$  is generically hypocritical with respect to initial beliefs and the parameter values in  $\mathbf{D}$  and  $\mathbf{G}$ .

## References

- Acemoglu, D., Dahleh, M. A., Lobel, I., and Ozdaglar, A. (2011). Bayesian Learning in Social Networks. *Review of Economic Studies*, **78** (4), 1201–1236.
- Acemoglu, D., Ozdaglar, A., and ParandehGheibi, A. (2010). Spread of (Mis)information in Social Networks. *Games and Economic Behavior*, **70** (2), 194–227.
- Akerlof, G. A. (1980). A Theory of Social Custom, of Which Unemployment May Be One Consequence. *Quarterly Journal of Economics*, **94** (4), 749–775.
- Anufriev, M., Borissov, K., and Pakhnin, M. (2023). Dissonance Minimization and Conversation in Social Networks. *Journal of Economic Behavior and Organization*, **215**, 167–191.
- Arifovic, J., Eaton, B. C., and Walker, G. (2015). The Coevolution of Beliefs and Networks. *Journal of Economic Behavior and Organization*, **120**, 46–63.

- Asch, S. E. (1955). Opinions and Social Pressure. *Scientific American*, **193** (5), 31–35.
- Bala, V. and Goyal, S. (1998). Learning from Neighbours. *Review of Economic Studies*, **65** (3), 595–621.
- Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *Quarterly Journal of Economics*, **107** (3), 797–817.
- Banerjee, A. V., Breza, E., Chandrasekhar, A. G., and Mobius, M. (2021). Naive Learning with Uninformed Agents. *American Economic Review*, **111** (11), 3540–3574.
- Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, **96** (5), 1652–1678.
- Bernheim, B. D. (1994). A Theory of Conformity. *Journal of Political Economy*, **102** (5), 841–877.
- Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, **100** (5), 992–1026.
- Bindel, D., Kleinberg, J., and Oren, S. (2015). How Bad is Forming Your Own Opinion? *Games and Economic Behavior*, **92**, 248–265.
- Bolletta, U. and Pin, P. (2025). Dynamic Opinion Updating with Endogenous Networks. *European Economic Review*, **176**, 105045.
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person Experiment in Social Influence and Political Mobilization. *Nature*, **489** (7415), 295–298.
- Buechel, B., Hellmann, T., and Klößner, S. (2015). Opinion Dynamics and Wisdom under Conformity. *Journal of Economic Dynamics and Control*, **52**, 240–257.
- Buechel, B., Hellmann, T., and Pichler, M. M. (2014). The Dynamics of Continuous Cultural Traits in Social Networks. *Journal of Economic Theory*, **154**, 274–309.
- Bursztyn, L., Egorov, G., and Fiorin, S. (2020). From Extreme to Mainstream: The Erosion of Social Norms. *American Economic Review*, **110** (11), 3522–3548.

- Chandrasekhar, A. G., Larreguy, H., and Xandri, J. P. (2020). Testing Models of Social Learning on Networks: Evidence from Two Experiments. *Econometrica*, **88** (1), 1–32.
- DeGroot, M. H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*, **69** (345), 118–121.
- Della Lena, S. (2024). The Spread of Misinformation in Networks with Individual and Social Learning. *European Economic Review*, **168**, 104804.
- Della Lena, S., Merlino, L., and Zenou, Y. (2026). Affective Polarization, Media Outlets, and Opinion Dynamics. RFBerlin Discussion Paper 001/26.
- DellaVigna, S. and Kaplan, E. (2007). The Fox News Effect: Media Bias and Voting. *Quarterly Journal of Economics*, **122** (3), 1187–1234.
- DeMarzo, P. M., Vayanos, D., and Zwiebel, J. (2003). Persuasion Bias, Social Influence, and Unidimensional Opinions. *Quarterly Journal of Economics*, **118** (3), 909–968.
- Deutsch, M. and Gerard, H. B. (1955). A Study of Normative and Informational Social Influences upon Individual Judgment. *Journal of Abnormal and Social Psychology*, **51** (3), 629–636.
- Dietzenbacher, E. (1990). Perturbations of the Perron Vector: Applications to Finite Markov Chains and Demographic Population Models. *Environment and Planning A*, **22** (6), 747–761.
- Echterhoff, G., Higgins, E. T., and Groll, S. (2005). Audience-tuning Effects on Memory: The Role of Shared Reality. *Journal of Personality and Social Psychology*, **89** (3), 257.
- Golub, B. and Jackson, M. O. (2010). Naive Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics*, **2** (1), 112–149.
- Golub, B. and Jackson, M. O. (2012). How Homophily Affects the Speed of Learning and Best-response Dynamics. *Quarterly Journal of Economics*, **127** (3), 1287–1338.
- Golub, B. and Sadler, E. (2016). Learning in Social Networks. In Bramoulle, Y., Galeotti, A., and Rogers, B. W., editors, *The Oxford Handbook of the Economics of Networks*. Oxford University Press, Oxford.

- Grabisch, M. and Rusinowska, A. (2020). A Survey on Nonstrategic Models of Opinion Dynamics. *Games*, **11** (4), 1–29.
- Grimm, V. and Mengel, F. (2020). Experiments on Belief Formation in Networks. *Journal of the European Economic Association*, **18** (1), 49–82.
- Higgins, E. T. (1999). “Saying Is Believing” Effects: When Sharing Reality about Something Biases Knowledge and Evaluations. In Thompson, L. L., Levine, J. M., and Messick, D. M., editors, *Shared Cognition in Organizations: The Management of Knowledge*, volume 1, 33–49. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.
- Ho, Y. and Huang, Y. (2026). Breaking the Spiral of Silence. Mimeo.
- Jackson, M. O. (2008). *Social and Economic Networks*. Princeton University Press.
- Kuran, T. (1987). Preference Falsification, Policy Continuity and Collective Conservatism. *Economic Journal*, **97** (387), 642–665.
- Kuran, T. (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.
- Morris, S. (2001). Political Correctness. *Journal of Political Economy*, **109** (2), 231–265.
- Noelle-Neumann, E. (1974). The Spiral of Silence. A Theory of Public Opinion. *Journal of Communication*, **24** (2), 43–51.
- Noelle-Neumann, E. (1984). *The Spiral of Silence: Public Opinion — Our Social Skin*. University of Chicago Press.
- Olcina, G., Panebianco, F., and Zenou, Y. (2024). Conformism, Social Pressure, and the Dynamics of Integration. *Journal of Economic Behavior and Organization*, **220**, 279–304.
- Risco, M. (2025). Network Effects on Information Acquisition by DeGroot Updaters. *Economic Theory*, **79**, 201–234.
- Yanagizawa-Drott, D. (2014). Propaganda and Conflict: Evidence from the Rwandan Genocide. *Quarterly Journal of Economics*, **129** (4), 1947–1994.