Hypocrisy in a Simple Social Interaction Model

Mikhail Anufriev [a,*] Kirill Borissov [b,†] Mikhail Pakhnin [b,‡]

This draft: DECEMBER 2022

[a] *Economics Discipline Group, Business School, University of Technology Sydney*
[b] *Department of Economics, European University at St. Petersburg*

Abstract

We propose a simple model of social interaction, where agents are embedded in a multi-layer network and repeatedly make their statements and form their beliefs. Agents tune statements to their peers from the audience network and update their beliefs by listening to their influencers from the influence network. We study the dynamics of the model and show that in the long run agents' statements may not coincide with their beliefs, that is, agents could be hypocritical. We characterize hypocrisy in terms of the audience and influence networks, and provide necessary and sufficient conditions for the presence and absence of hypocrisy in the long run. Our model is able to explain various social phenomena, ranging from effects of political propaganda to implications of the spiral of silence theory.

Keywords: Social networks; Hypocrisy; DeGroot learning; Dissonance minimization; Political propaganda; Spiral of silence

JEL Classification: D83, D85, D91, Z13

[*] *E-mail:* Mikhail.Anufriev@uts.edu.au
[†] *E-mail:* kirill@eu.spb.ru
[‡] *E-mail:* mpakhnin@eu.spb.ru

> The keyword here is BLACKWHITE. Like so many Newspeak words, this word has two mutually contradictory meanings. Applied to an opponent, it means the habit of impudently claiming that black is white, in contradiction of the plain facts. Applied to a Party member, it means a loyal willingness to say that black is white when Party discipline demands this. But it means also the ability to BELIEVE that black is white, and more, to KNOW that black is white, and to forget that one has ever believed the contrary.
>
> Orwell, George (1949). Nineteen Eighty-Four.

> Stepan Arkadyevitch took in and read a liberal paper, not an extreme one, but one advocating the views held by the majority. And in spite of the fact that science, art, and politics had no special interest for him, he firmly held those views on all these subjects which were held by the majority and by his paper, and he only changed them when the majority changed them — or, more strictly speaking, he did not change them, but they imperceptibly changed of themselves within him.
>
> Tolstoy, Leo (1878). Anna Karenina.

1 Introduction

Belief formation is a complicated process that attracts attention of economists, sociologists, political scientists, psychologists. This process is central in social sciences as it affects all aspects of individual choice and collective behavior, ranging from economic decisions within organizations and governments to the results of elections and political campaigns. Studying belief formation is even more important nowadays, when technological advances of Internet and social networks allow people to get quick opinions on any issue from a large and heterogeneous pool of sources. At the same time, the issues of validity of information and its impact on information aggregation become crucial with the appearance of "internet trolls" and "Kremlin bots" spreading political propaganda and fake news (see, e.g., Stukal et al., 2022).[1]

[1] According to some estimates, the budget of the largest Russian troll farm was around $10 mln already in 2014 (see, e.g., Seddon, 2014).

Recent research emphasizes that the structure of society, including available sources of information, strongly affects the outcome of belief formation process.[2] Furthermore, belief is actually a state of mind which becomes known only as a statement through some kind of conversation. It can be argued that *beliefs* about various issues (objective facts, social norms, economic variables) are based not on evidence but on the *statements* that we learn from our peers such as family, friends or coworkers, as well as from influencers such as newsmakers or media trendsetters.

Social psychologists show that both individual statements and beliefs are formed under various social influences. In the famous Asch (1955) experiment, around 30% of the subjects made an obviously wrong judgement about the length of the line after hearing the same incorrect statement from the stooges who were perceived as neutral. After the experiment, most of the subjects noted that they did not believe in their apparently incorrect answers but felt certain dissonance from disagreement with the others, while some subjects admitted that they believed in what the others have (deliberately wrong) stated. Following this experiment, Deutsch and Gerard (1955) introduced two different types of social influence: *normative social influence* (making statements to conform to the statements of the others) and *informational social influence* (updating beliefs according to what the others have stated).

Normative social influence leads us to make statements which do not coincide with our own beliefs but are tailored to the statements of others. This phenomenon is closely associated with *audience tuning* by which people move their statements towards their peers (see Higgins, 1999; Echterhoff et al., 2009). Examples of audience tuning are ubiquitous. Faculty members may not tell their true opinions about functioning of the university when asked by a provost. As illustrated by the opening quote from Orwell, people in authoritarian countries may agree with their dictator that black is white. Also, people may not express their sincere views on sensitive topics in discussions with friends in order not to upset them. Informational social influence leads us to accept others' statements as evidence about reality even though there may be no motivation to agree with the others. For instance, CEOs may trust the Federal Reserve Chair about expected inflation, because they assume that the Chair is much better informed than they are. As demonstrated by the opening quote from Tolstoy, changing views under the influence of majority was, and still is, very common.

In this paper we propose a model of social interaction which captures the effects

[2]See Golub and Sadler (2016) for a review of the economic research on learning in social networks that focuses on the questions of efficient information aggregation.

of both types of social influence. The key feature of our model is that it explicitly distinguishes (i) statements vs. beliefs of agents; and (ii) audience vs. influence networks for each agent. We represent society as a set of agents connected by means of two different directed networks. The *audience network* reflects the idea of normative social influence: statements of others cause people to say not what they believe. Agents tune their own statements to the statements of their peers in this network, as in the faculty-provost example. The *influence network* reflects the idea of informational social influence: people form their beliefs by listening to others and adopting others' statements. Agents pay attention to what local leaders or other influencers state and tend to believe in what they hear, as in the CEO-Fed Chair example.

We model belief formation in each period as a two-stage process. First, agents are engaged in a round of conversations where they make statements by minimizing dissonance from disagreement with their peers in the audience network. Second, agents update their beliefs taking into account the statements of influencers in the influence network. We study the dynamics of beliefs and statements in the model and find that hypocrisy, when agents' beliefs differ from their statements, may prevail in the long run. That is, some or all agents in the society may perpetually not say what they believe. At the same time, other agents may change their beliefs over time but eventually avoid hypocrisy, and this may happen because of both the extreme peer pressure from the audience network and the learning from their influencers' views.

As it turns out, the belief dynamics in the model is described as an average-based updating process in the spirit of DeGroot learning.[3] To characterize the long-run beliefs, we partition agents based on the multi-layer network induced by both audience and influence networks. We divide agents into disjoint groups of *stubborn agents* and *pure communication classes*. Stubborn agents are those who have no influencers. These agents do not change their initial beliefs. Agents from a pure communication class are either peers or influencers or both to each other, and they do not have peers and influencers outside of their class. Agents from each pure communication class reach consensus in the long run. There may be *remaining agents* that complete the partition of the society. Remaining agents converge to a weighted average of long-run beliefs of other groups.

[3]DeGroot (1974) proposed a model of social learning in which a new belief of each agent is a weighted average of current beliefs of all others where the non-negative weights are exogenously given. Recent applications of the DeGroot model include DeMarzo et al. (2003); Golub and Jackson (2010, 2012); Olcina et al. (2017). As these and other studies, we take advantage of high tractability of DeGroot learning. Moreover, this boundedly rational learning has gained an empirical support when compared with rational learning, see Chandrasekhar et al. (2020).

The main contribution of our paper is that it develops a framework to explain and rationalize the phenomenon of hypocrisy, and fully characterizes conditions under which agents become hypocritical in the long-run. With the results describing the long-run beliefs, we are able to trace the emergence of hypocrisy to the primitives of the model, that is, to the audience and influence networks which summarize individual sensitivities to disagreement with others' statements. Our results imply that hypocrisy is a direct consequence of conversations and audience tuning: when agents have no peers or when their peers are the same people as their influencers, there is no hypocrisy. Further, agents from pure communication classes are never hypocritical, and in this sense hypocrisy is not a local but global property of connectivity across people in the society.

Several examples illustrate the role of hypocrisy in social interactions. We show that hypocrisy makes political propaganda more powerful and successful, as people may spread others' opinions even when they do not believe in them. We also discuss examples of the Overton window shift (change from a mainstream to an extreme opinion) and spiral of silence (formation and reinforcement of a perceived majority opinion), and argue that they are essentially the same social phenomena which can be explained by our model.

Three recent papers are close to ours in that they also make a distinction between statements and beliefs, and study the implications of audience tuning. Arifovic et al. (2015) simulate a model where agents minimize dissonance arising from disagreement with their peers. Anufriev et al. (2021) provide a full analytical solution to the model of Arifovic et al. (2015) and characterize the long-run outcome in terms of dissonance sensitivities. Buechel et al. (2015) incorporate conformity as a source of misrepresentation of beliefs into the DeGroot model and study efficiency of information aggregation. In all these papers, long-run statements of agents coincide with their beliefs. The present paper differs from this literature in that we explicitly consider two different networks as sources of normative and informational social influence. As a result, agents in our model could be hypocritical even after the belief dynamics converge.

The rest of the paper is organized as follows. In Section 2 we set the stage by formalizing the ideas of audience tuning in conversations and influencers' impact on belief formation. Section 3 describes the dynamics of the model within each period of time, while Section 4 characterizes the long-run outcome of the model. Section 5 discusses hypocrisy in our model. We conclude in Section 6 with a brief summary and several examples. The proofs of all the results are relegated to the Appendix.

2 The model

Society consists of a set $\mathcal{N} = \{1, \ldots, N\}$ of agents communicating and exchanging information about certain issue. Time is discrete. In each period t agent i makes a statement $s_i(t)$ about the issue in conversations with peers. Next, agent forms a belief $b_i(t)$ about the issue taking into account the statements of influencers.[4]

Each period t starts with a round of conversations. In conversations, an agent minimizes dissonance which arises when the statement differs from (i) agent's current belief, reflecting the cost of being inconsistent (private dissonance); and (ii) the statements of others, reflecting the cost of disagreement with peers (social dissonance). Formally, within period t conversations occur at a fast timescale τ. Given the initial belief $b_i(t-1)$, which is fixed in this timescale, at each date τ agent i chooses statement $s_i^\tau(t)$ to minimize the dissonance with current belief $b_i(t-1)$ and previous statements of peers $s_j^{\tau-1}(t)$:

$$\min_{s_i^\tau(t)} \left\{ (s_i^\tau(t) - b_i(t-1))^2 + \sum\nolimits_{j=1}^{N} d_{ij} (s_i^\tau(t) - s_j^{\tau-1}(t))^2 \right\}, \tag{1}$$

where $d_{ij} \geq 0$ are dissonance sensitivities. The sensitivities are summarized in the *audience dissonance matrix* $\mathbf{D} = \{d_{ij}\}_{i,j=1}^N$, with *zero diagonal*.[5] If $d_{ij} = 0$, i does not take into account the statement of j when making own statement. If $d_{ij} > 0$, disagreement in statements of i and j results in positive cost for i in (1).

Audience dissonance matrix $\mathbf{D}$ is the adjacency matrix of the *audience network*. There is a directed edge between i and j in this network (j is a *peer* of i) whenever $d_{ij} > 0$. Audience network captures the idea of normative social influence of Deutsch and Gerard (1955): in conversations people tend to experience dissonance from disagreement with the expressed opinions. However, this tendency is not symmetric. For example, people in authoritarian countries may feel a dissonance when their statements differ from those of their dictator, whereas the dictator may feel no dissonance in any conversation. Employees may feel a dissonance when their message differs from those of their supervisors, and supervisors may feel a dissonance when talking to the CEO, but not when talking to the employees. Therefore we do not impose any further restriction on matrix $\mathbf{D}$. This matrix, in general, characterizes both the peers of each agent (direction of the egdes) and heterogeneous dissonance sensitivities to each peer (weights of the edges).

It follows from the standard results (see, e.g., Proposition 2 in Anufriev et al., 2021) that agents' statements in the fast timescale converge to vector $\mathbf{s}(t)$ whose

[4]Beliefs can be interpreted in a broad sense – also as opinions, judgements, estimations, norms or values. Similarly, one can think of statements as verbal statements or observable actions.
[5]In optimization problem (1), the private dissonance sensitivity is normalized to 1.

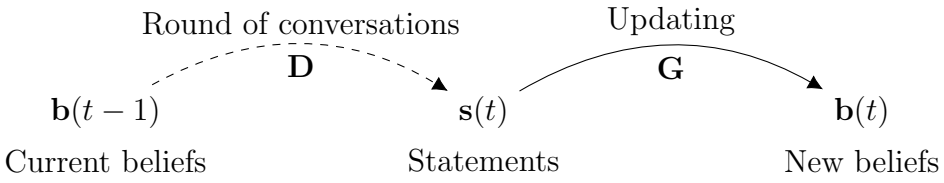


Figure 1: Timing of the model within period t. A round of conversations in a fast timescale is followed by the belief updating.

components simultaneously solve, for each agent, the problem

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

That is, for all i, $s_i^\tau(t) \xrightarrow[\tau\to\infty]{} s_i(t)$, given by

$$s_i(t) \;=\; \frac{1}{1+\delta_i} b_i(t-1) + \sum\nolimits_{j=1}^{N} \frac{d_{ij}}{1+\delta_i} s_j(t), \qquad i = 1, \ldots, N\,, \tag{2}$$

where $\delta_i = \sum_k d_{ik}$ denotes the i-th row sum of matrix $\mathbf{D}$. Statement of each agent is a weighted average of own current belief and the statements of peers. Dissonance minimization in (1) leads to a peer pressure, making agents' statements different from their beliefs and adjusted towards the statements of their peers. This is known as the audience tuning (Echterhoff et al., 2005).

After the round of conversations, agents update their beliefs, see Fig. 1. An agent minimizes costs which arise when the new belief differs from (i) agent's current belief, reflecting the cost of changing one's mind; and (ii) the statements of others, reflecting the cost of disagreement with influencers. Formally, given belief $b_i(t-1)$, agent i chooses new belief $b_i(t)$ by solving the following problem:

$$\min_{b_i(t)} \left\{ (b_i(t) - b_i(t-1))^2 + \sum\nolimits_{j=1}^{N} g_{ij}(b_i(t) - s_j(t))^2 \right\} , \tag{3}$$

where $g_{ij} \geq 0$ are influence sensitivities. These parameters form the *influence matrix* $\mathbf{G} = \{g_{ij}\}_{i,j=1}^{N}$, with *zero diagonal*.[6] If $g_{ij} = 0$, i does not take j's statement into account when forming a belief. If $g_{ij} > 0$, disagreement of i's belief with j's statement results in positive cost for i in (3).

Influence matrix $\mathbf{G}$ is the adjacency matrix of the *influence network*. There is a directed edge between i and j in this network (j is an *influencer* for i) whenever $g_{ij} > 0$. Influence network captures the idea of informational social influence of Deutsch and Gerard (1955): people base their beliefs on the statements of others.

[6]The private influence sensitivity in problem (3) is normalized to 1.

Influencers can be experts, scientists, opinion leaders, social media celebrities, and so on. Matrix $\mathbf{G}$ defines both the set of influencers for each agent (direction of the edges) and the sensitivity to each influencer (weights of the edges).

Due to strict convexity of the objective function in (3), a new belief of agent i is given by a weighted average of own current belief and the statements of influencers:

$$b_i(t) \;=\; \frac{1}{1+\gamma_i} b_i(t-1) + \sum\nolimits_{j=1}^{N} \frac{g_{ij}}{1+\gamma_i} s_j(t), \qquad i = 1, \ldots, N\,, \tag{4}$$

where $\gamma_i = \sum_k g_{ik}$ denotes the i-th row sum of matrix $\mathbf{G}$.

Audience dissonance matrix $\mathbf{D}$ and influence matrix $\mathbf{G}$ are primitives of the model: the model is defined by two arbitrary non-negative square matrices with zeros on the diagonal. It is important that influencers and peers of an agent are, in general, different individuals, so $\mathbf{D}$ and $\mathbf{G}$ are different. As mentioned in the Introduction, this is what distinguishes our framework from other related work.

3 Belief dynamics

Using (2) and (4), we can rewrite the joint dynamics of statements and beliefs in matrix form:

$$\begin{cases} \mathbf{b}(t-1) = (\mathbf{I} + \boldsymbol{\Delta} - \mathbf{D})\, \mathbf{s}(t)\,, & (5) \\ \mathbf{b}(t-1) = (\mathbf{I} + \boldsymbol{\Gamma})\, \mathbf{b}(t) - \mathbf{G}\mathbf{s}(t)\,, & (6) \end{cases}$$

where $\mathbf{b}(t)$ is the (column) vector of agents' beliefs, $\mathbf{I}$ is the identity matrix of size N, $\boldsymbol{\Delta}$ is the diagonal matrix whose i-entry is δ_i and $\boldsymbol{\Gamma}$ is the diagonal matrix whose i-entry is γ_i. Initial beliefs $\mathbf{b}(0)$ are given.

To proceed, the following notions are useful. Consider an arbitrary square matrix $\mathbf{A}$ with non-negative elements. A *path in* $\mathbf{A}$ *from* i *to* j is a sequence $i_1, \ldots, i_J$ such that $i_1 = i$, $i_J = j$ and $a_{i_k i_{k+1}} > 0$ for each $k = 1, \ldots, J-1$. A network and its adjacency matrix $\mathbf{A}$ are *strongly connected* if for any i and j there is a path in $\mathbf{A}$ from i to j. Matrix $\mathbf{A}$ is *row-stochastic* if $\sum_j a_{ij} = 1$ for all i.

The next result of Anufriev et al. (2021) describes the formation of statements.

Proposition 1. *(i) Evolution of statements is given by*

$$\mathbf{s}(t) \;=\; \mathbf{P}\mathbf{b}(t-1)\,,$$

with row-stochastic matrix $\mathbf{P}$ *defined as*

$$\mathbf{P} \;=\; (\mathbf{I} + \boldsymbol{\Delta} - \mathbf{D})^{-1}\,. \tag{7}$$

(ii) For any i, $p_{ii} > 0$. For any i, j, $p_{ij} > 0$ iff there is a path in $\mathbf{D}$ from i to j.

(iii) $\mathbf{P}$ is strongly connected iff $\mathbf{D}$ is strongly connected.

Proposition 1 establishes that rounds of conversations in our model result in statements being the average of agents' beliefs and characterizes the properties of the corresponding matrix $\mathbf{P}$ in terms of dissonance sensitivities. For each agent i, the self-weight p_{ii} is always strictly positive. Moreover, belief of some agent j affects the statement of agent i ($p_{ij} > 0$) not only when j is a peer of i ($d_{ij} > 0$) but also when there is a chain of successive peers leading from i to j.

Consider now the formation of beliefs. It follows from (6) that

$$\mathbf{b}(t) \;=\; \mathbf{Q}\mathbf{s}(t)\,,$$

where

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

It then follows from Proposition 1 that evolution of beliefs is determined by the matrix $\mathbf{T} \equiv \mathbf{QP}$. Our next result implies that $\mathbf{T}$ is row-stochastic.

Proposition 2. *Evolution of beliefs is given by*

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

where the matrix $\mathbf{T}$ defined as

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

is row-stochastic.

Proof. See Appendix A.1. ■

Agent's interaction with peers and influencers gives rise to average-based updating of the beliefs in a society in the spirit of DeGroot learning. At each date, agent i's new belief is the weighted average of current beliefs of all agents. The weights of beliefs, t_{ij}, are determined as a combination of dissonance and influence sensitivities, and depend both on the structure of audience network reflected by $\mathbf{D}$ and on the structure of influence network reflected by $\mathbf{G}$.[7]

[7]The special case where audience and influence networks coincide is studied by Anufriev et al. (2021). Formally, when $\mathbf{G} = \mathbf{D}$, we have $\mathbf{Q} = \mathbf{I}$, so that agents are exposed to the "saying is believing" effect broadly discussed in cognitive psychology (Higgins, 1999). In this case, $\mathbf{b}(t) = \mathbf{s}(t)$, and agents' new beliefs coincide with their just made statements. This paper, however, shows that for this effect to arise, the sets of peers and influencers for any agent should coincide. In some situations (such as students' clubs or reading groups) this may be a plausible assumption, but in general it is not.

The interdependence of peers and influencers suggests that the relevant network in our model is a multi-layer network induced by both $\mathbf{D}$ and $\mathbf{G}$ which we denote by $\mathbf{D} \star \mathbf{G}$. It is convenient to define this network via the adjacency matrix $\mathbf{D} \star \mathbf{G}$ of size N as follows. The (i, j)-entry in $\mathbf{D} \star \mathbf{G}$ is equal to 1 when $\max\{d_{ij}, g_{ij}\} > 0$ (either $d_{ij} > 0$ or $g_{ij} > 0$ or both) and is equal to 0 otherwise. We say that agent i *is linked* to agent j in $\mathbf{D} \star \mathbf{G}$ if there is a path in $\mathbf{D} \star \mathbf{G}$ from i to j.

The next result characterizes the properties of $\mathbf{T}$ in terms of the primitives $\mathbf{D}$ and $\mathbf{G}$ and is crucial to establish the dynamic properties of our model.

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

(i) For any i*, the self-weight* $t_{ii} > 0$.

(ii) The weight $t_{ij} > 0$ *iff either* $g_{ij} > 0$ *or there is* k *such that* $g_{ik} > 0$ *and there is a path in* $\mathbf{D}$ *from* k *to* j.

(iii) The weight $t_{ii} = 1$ *iff* $g_{ij} = 0$ *for all* j.

(iv) Suppose that for all i, $g_{ij} > 0$ *at least for some* j. *Then* $\mathbf{T}$ *is strongly connected iff* $\mathbf{D} \star \mathbf{G}$ *is strongly connected.*

Proof. See Appendix A.1. ■

First, Proposition 3 implies that each agent always attaches a positive weight to own current belief. It is well-known (see, e.g., Theorem 2 in Golub and Jackson, 2010), that this property excludes the possibility of cycles and guarantees the convergence of beliefs.

Second, apart from a natural channel when influencer j affects agent i's belief directly, there are indirect channels to affect i's belief. Even if j is not an influencer of i, j's belief will affect i via any chain of peers leading from j to some other i's influencer, say agent k. The mechanism is clear. If i takes statements made by k into account when updating the belief, and j is, for example, a peer of k, then k tunes own statement to the statement of j, and in this way agent j's belief will have non-zero weight for agent i. In general, due to rounds of conversations, j should not even be the peer of k, any chain of peers from j to k will suffice. To appreciate this property, note that i may not even be aware of j's existence and yet be affected by j's belief.

Third, agent i never changes belief iff i has no influencers. This motivates the following definition.

Definition 1. *Agent* i *with* $t_{ii} = 1$ *in matrix* $\mathbf{T}$ *from* (8) *is a* stubborn *agent.*

If agent i is stubborn, then matrix $\mathbf{G}$ has zero row i. As we shall see further, stubborn agents play a special role in the emergence of hypocrisy.

Finally, when there are no stubborn agents in society, matrices $\mathbf{T}$ and $\mathbf{D} \star \mathbf{G}$ are strongly connected (or not) at the same time. Recall that the elements of $\mathbf{D} \star \mathbf{G}$ are only 0 or 1 and that the associated multi-layer network effectively concatenates the audience and influence networks.

4 Long-run beliefs and statements

We now study the long-run properties of our model. By iterating the dynamics of beliefs, we get $\mathbf{b}(t) = \mathbf{T}\mathbf{b}(t-1) = \ldots = \mathbf{T}^t\mathbf{b}(0)$. As mentioned above, since $t_{ii} > 0$ for all i, beliefs of each agent converge: there exists $\mathbf{b}^* \equiv \lim_{t\to\infty} \mathbf{T}^t\mathbf{b}(0)$. Then the statements of each agent also converge: there exists $\mathbf{s}^* \equiv \lim_{t\to\infty} \mathbf{s}(t)$.

As the next example demonstrates, the limiting beliefs may be different among agents and do not need to coincide with the agents' limiting statements.

Example 1. Consider the society in Fig. 2a (the dashed line indicates the edge in the audience network, see the caption). There are two agents, both have no influencers. Agent 2 is a peer of agent 1 whose dissonance sensitivity is d_1. As both agents are stubborn, they do not change their beliefs. Agent 2 has no peers and 2's statement always coincides with belief, $s_2^* = b_2^* = b_2(0)$. In contrast, agent 1 who believes in $b_1^* = b_1(0)$, tunes the message. Taking the limit in (2), we find that agent 1 states $s_1^* = \frac{1}{1+d_1} b_1(0) + \frac{d_1}{1+d_1} b_2(0)$, a weighted average of own belief and agent 2's belief. □

In this example, the society does not reach consensus (except for the knife-edge case of equal initial beliefs). Moreover, agent 1 does not state a true belief even in the long-run. This motivates the following definition.

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

We will study the phenomenon of hypocrisy in the next section. Before that, we need to characterize the limiting beliefs and statements in the model. For this purpose, we introduce the following notions.

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

A communication class is a set of agents who are peers and/or influencers only to each other. It follows that no agent can belong to two different communication classes. However, a stubborn agent can belong to some communication class: such agent does not have influencers, and yet may have any number of peers. Hereafter, we refer to stubborn agents and pure communication classes as *independent classes* in $\mathbf{D} \star \mathbf{G}$.

The set of all agents $\mathcal{N}$ can be partitioned into a set $\mathcal{S}$ of stubborn agents, a number of pure communication classes $\mathcal{C}_m$ indexed by $m = 1, \ldots, M$, and a set $\mathcal{R}$ of all remaining agents. This partition implies a block structure of matrix $\mathbf{T}$ and fully determines the long-run beliefs which we briefly describe next. For the details, see Appendix A.2.

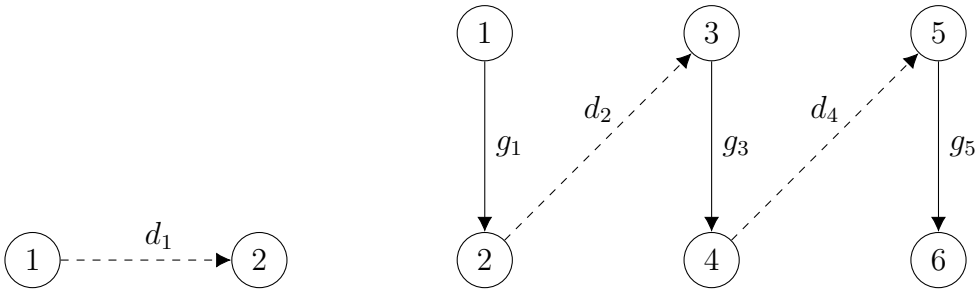
First, stubborn agents do not change their beliefs. Long-run belief of a stubborn agent always coincides with initial belief: $b_s^* = b_s(0)$, for all $s \in \mathcal{S}$.

Second, each pure communication class reaches a consensus: all agents from $\mathcal{C}_m$ converge to the same long-run belief b^{*m} given by a weighted average of their initial beliefs, $b^{*m} = \boldsymbol{\pi}^m \mathbf{b}(0)|_{\mathcal{C}_m}$. The weights $\boldsymbol{\pi}^m$ that agents have on the consensus belief of their pure communication class are agents' eigenvector centralities within the class. Centralities are given by the left unit eigenvector (corresponding to eigenvalue 1) of the respective block in matrix $\mathbf{T}$: $\boldsymbol{\pi}^m \mathbf{T}|_{\mathcal{C}_m} = \boldsymbol{\pi}^m$, where $\mathbf{T}|_{\mathcal{C}_m}$ is $\mathbf{T}$ restricted to the set $\mathcal{C}_m$. Clearly, the long-run beliefs of different independent classes do not coincide, unless initial beliefs are chosen in a very special way. In what follows, *we consider a generic case where long-run beliefs of different independent classes are different, ignoring a zero-measure set of initial beliefs.* Hence when there are at least two independent classes, society as a whole generically does not reach consensus.

Third, long-run belief of a remaining agent is a weighted average of the long-run beliefs of the agents from independent classes to whom this agent is linked in $\mathbf{D} \star \mathbf{G}$. Thus, initial beliefs of agents from $\mathcal{R}$ do not matter: their long-run beliefs are fully determined by initial beliefs of stubborn agents and agents from pure communication classes. In particular, if agent $r \in \mathcal{R}$ is linked only to a single independent class, then r's belief converges to the long-run belief of this class.

As for the statements, note that $\mathbf{s}^* = \mathbf{P}\mathbf{b}^*$, and hence the long-run statements are weighted averages of the long-run beliefs. Suppose that agent i is not stubborn. Then, taking the limit in (4), we find that i's long-run belief is a weighted average of the statements of i's influencers:

$$b_i^* \;=\; \sum\nolimits_j \frac{g_{ij}}{\gamma_i} s_j^* \,. \tag{9}$$



(a) Stubborn agent 1 is hypocritical while stubborn agent 2 is not.

(b) Propagation of propaganda: 2 and 4 do not believe 6 but convey 6's belief.

Figure 2: Audience network (dashed lines) and influence network (solid lines) in Example 1 (left panel) and Example 2 (right panel).

In particular, when agent i has only one influencer, i's long-run belief coincides with the statement of this influencer. It then follows from (2) that i's long-run statement is a weighted average of the statements of i's influencers and peers:

$$s_i^* \;=\; \frac{1}{1+\delta_i}\left(b_i^* + \sum\nolimits_j d_{ij} s_j^*\right) \;=\; \frac{1}{1+\delta_i}\sum\nolimits_j \left(\frac{g_{ij}}{\gamma_i} + d_{ij}\right) s_j^* \,. \tag{10}$$

We conclude this section with an example that illustrates the formation of long-run beliefs and the interaction of audience and influence networks. This example shows how influence of a single agent (e.g., a Kremlin bot) can propagate due to audience tuning, and why our beliefs may be affected by the beliefs of those who are far from us in terms of social distance.

Example 2. Consider the "chain" society of six agents (see Fig. 2b). Agent 6 is a propagandist who has no influencers and peers, so $b_6^* = s_6^* = b_6(0)$. Agent 5 is influenced by 6 and has no peers, so 5 converges to 6's statement (which is 6's initial belief): $b_5^* = s_5^* = b_6(0)$. Agent 4 is stubborn, $b_4^* = b_4(0)$, but tunes a statement to that of agent 5, so $s_4^* = \frac{1}{1+d_4} b_4(0) + \frac{d_4}{1+d_4} b_6(0)$, as in Example 1. Agent 3 is influenced by 4 and has no peers, so both belief and statement of 3 converge to 4's statement. Hence long-run belief of agent 3 is a weighted average of initial beliefs of agents 4 and 6: $b_3^* = s_3^* = \frac{1}{1+d_4} b_4(0) + \frac{d_4}{1+d_4} b_6(0)$. Agent 2 is stubborn but tunes a statement to that of agent 3, so $b_2^* = b_2(0)$ and $s_2^* = \frac{1}{1+d_2} b_2(0) + \frac{d_2}{1+d_2} s_3^*$.

As a result, agent 1, who is influenced only by agent 2, converges to 2's statement, that is, to a weighted average of initial beliefs of agents 2, 4 and 6:

$$b_1^* = s_1^* = \frac{1}{1+d_2} b_2(0) + \frac{d_2}{(1+d_2)(1+d_4)} b_4(0) + \frac{d_2 d_4}{(1+d_2)(1+d_4)} b_6(0) \,.$$

Note that agent 1 is separated from the propagandist 6 by all other agents. Never-

theless, since agents 2 and 4 tune their messages to their peers, 6's belief propagates through the multi-layer network and affects agent 1. Furthermore, the weights in the long-run belief of agent 1 depend only on the dissonance sensitivities of agents 2 and 4, and do not depend on influence sensitivities of 1, 3 or 5. This example shows that hypocrisy makes political propaganda much more far-reaching than it may seem: even though agents 2 and 4 do not believe in anything stated by the propagandist 6, they still convey 6's belief further. □

5 Hypocrisy

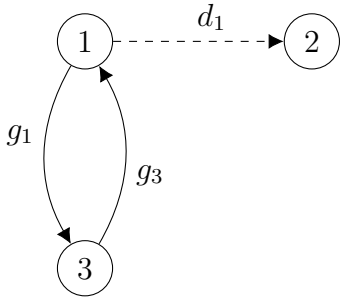
A central contribution of our model is that it is able to rationalize hypocrisy, which is a phenomenon where agents' beliefs and statements do not coincide in the long run: $\mathbf{b}^* \neq \mathbf{s}^*$. Note that in the short run our model implies that, due to audience tuning, agents generally do not say what they believe, and, due to impact of influencers, agents generally do not believe in what they say. Therefore, the relevant questions are whether these effects remain in the long run and how hypocrisy is related to the structure of the multi-layer network.

5.1 Examples

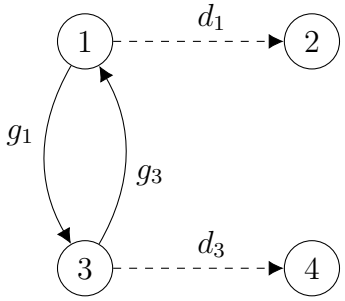
We illustrate the absence and presence of hypocrisy with several examples. As we have seen in Example 1, a stubborn agent 2, who is the only cause of 1's hypocrisy, is not hypocritical. Thus, while stubborn agents can be hypocritical or not, it is important that they can be the cause of hypocrisy for other agents. However, the following example suggests that stubborn agent as a peer does not necessarily cause hypocrisy.

Example 3. Add agent 3 to the network from Example 1 (see Fig. 3a). Agents 1 and 3 are influenced by each other. At each date, 1 makes a statement which differs from belief and is closer to the statement (and initial belief) of 2. Agent 3 is influenced by 1's statement and inadvertently takes into account 2's belief when forming new belief. Since 1 is in turn influenced by 3, 1's belief also shifts towards the belief of agent 2. Eventually society reaches a consensus: beliefs of all agents converge to the initial belief of 2. Therefore, there is no hypocrisy in this society: in the long run all agents state what they believe, $s_i^* = b_i^* = b_2(0)$, for all i. □

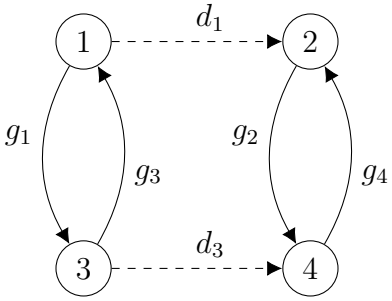
In Example 3, agent 1 ceases to be hypocritical because agent 3 shifts 1's belief precisely towards 1's statement which is always tuned to a stubborn agent 2. The



(a) No hypocrisy: 1 and 3 influence each other to believe 2.



(b) Two stubborn agents as sources of hypocrisy.



(c) No hypocrisy: 2 and 4 form a pure communication class.

Figure 3: Audience (dashed lines) and influence (solid lines) networks in Example 3 (up left), Example 4 (up right) and Example 5 (bottom).

following example shows that when agent 3 also tunes the statement to some other stubborn agent, 1 and 3 shift each other's beliefs differently, so two mutual influencers with different peers are hypocritical.

Example 4. Add agent 4 to the network from Example 3 (see Fig. 3b). Agent 4 is stubborn and is a peer of 3 (whose dissonance sensitivity is d_3). Long-run beliefs and statements of stubborn agents 2 and 4 coincide with their initial beliefs: $b_2^* = s_2^* = b_2(0)$ and $b_4^* = s_4^* = b_4(0)$.

Agent 1 is influenced by 3 who tunes statement to 4. At the same time, 3 is influenced by 1 who tunes statement to 2. Therefore, long-run beliefs of 1 and 3 are a weighted average of the initial beliefs of stubborn agents 2 and 4. It can be directly checked that $b_1^* = \frac{d_1}{d_1+d_3+d_1d_3} b_2(0) + \frac{d_3(1+d_1)}{d_1+d_3+d_1d_3} b_4(0)$, and $b_3^* = \frac{d_1(1+d_3)}{d_1+d_3+d_1d_3} b_2(0) + \frac{d_3}{d_1+d_3+d_1d_3} b_4(0)$. Note that the weights depend only on dissonance sensitivities d_1 and d_3 and are different for agents 1 and 3. The weight of $b_2(0)$ is always higher for 3 than for 1: while 2 is closer to 1 in terms of social distance, 1 is not directly influenced by 2's statement unlike 3. Moreover, the higher is d_1, the higher is the role of $b_2(0)$ in the long-run beliefs of both 1 and 3.

Since 3 is the only influencer for 1, 1 believes in what 3 says, and hence $s_3^* = b_1^*$. Similarly, 3 believes in what 1 says, so $s_1^* = b_3^*$. Therefore, both agents 1 and 3 are

hypocritical. Since 1 is tuning statements to 2, 1's long-run statement is closer to the belief of 2 and further from the belief of 4 than 1's own long-run belief. □

Furthermore, the following example confirms that it is two different independent classes, and not two different peers, that cause hypocrisy for remaining agents.

Example 5. Suppose that in the network from Example 4 agents 2 and 4 influence each other with influence sensitivities g_2 and g_4 (see Fig. 3c). Then 2 and 4 converge to the same long-run belief, and hence 1 and 3 also converge to that belief. In this case society reaches a consensus. The long-run belief of all agents is a weighted average of the initial beliefs of agents 2 and 4: for all i, $b_i^* = \frac{g_4(1+g_2)}{g_2+g_4+2g_2g_4}b_2(0) + \frac{g_2(1+g_4)}{g_2+g_4+2g_2g_4}b_4(0)$. The weights depend only on the influence sensitivities g_2 and g_4. The higher is g_2, the higher is the role of $b_4(0)$ in the long-run beliefs of all agents. Clearly, there is no hypocrisy, as the long-run statements of all agents are also the same and coincide with their long-run belief. □

5.2 Formal results

Next, we provide specific conditions for the absence and presence of hypocrisy. In our model hypocrisy is determined by three factors: initial beliefs $\mathbf{b}(0)$ of all agents, dissonance and influence sensitivities (weights summarized in matrices $\mathbf{D}$ and $\mathbf{G}$), and the structure of the network (links described by a $0,1$-matrix $\mathbf{D} \star \mathbf{G}$). Since it is the latter factor that is the most fundamental, our goal is to relate hypocrisy to the structure of the network $\mathbf{D} \star \mathbf{G}$.

We begin by noting that, irrespective of initial beliefs and sensitivities, there is no hypocrisy when agents reach consensus. If agents have the same long-run belief, then their statements, being weighted averages of the beliefs, are also the same and coincide with their common belief. Since this argument holds for any pure communication class in the long run, we arrive at the following result.

Proposition 4. *If agent i belongs to a pure communication class in $\mathbf{D} \star \mathbf{G}$, then $b_i^* = s_i^*$.*

Proposition 4 implies that if society consists only of pure communication classes in $\mathbf{D} \star \mathbf{G}$, then there is no hypocrisy. This result can be interpreted as follows: in the long run all agents say what they believe if society is either strongly connected (every agent is linked with every other agent and there are no stubborn agents) or perfectly segregated (that is, segmented into multiple parts which do not interact with each other).

It follows from Proposition 4 that only stubborn agents and remaining agents can be hypocritical. By comparing (9) and (10), we obtain a common sufficient condition for the absence of hypocrisy for both types of agents stated in terms of dissonance and influence sensitivities.

Proposition 5. *If for agent i $d_{ij} = 0$ for all j or $\frac{d_{ij}}{\delta_i} = \frac{g_{ij}}{\gamma_i}$ for all j, then $b_i^* = s_i^*$.*

Proposition 5 emphasizes that hypocrisy is a direct consequence of conversations and audience tuning. Agent i is not hypocritical when i has no peers and is not engaged in audience tuning or when i has the same peers and influencers with the same sensitivities up to a constant factor. In particular, there is no hypocrisy in society if $\mathbf{D} = \mathbf{0}$ (as in the standard DeGroot model) or $\mathbf{G} = \mathbf{D}$ (as in the model studied in Anufriev et al., 2021).

However, the conditions of Proposition 5 are not necessary. It is clear that agents may not be hypocritical also "accidentially", when initial beliefs or the weights in matrices $\mathbf{D}$ and $\mathbf{G}$ are picked up in a very special way. Ignoring knife-edge cases, we call agent i *generically hypocritical* when $b_i^* \neq s_i^*$ for all initial beliefs $\mathbf{b}(0)$ and for all positive weights in matrices $\mathbf{D}$ and $\mathbf{G}$ except for a zero-measure subsets. In other words, generic hypocrisy is a phenomenon which depends only on the structure of the network $\mathbf{D} \star \mathbf{G}$. Then for each stubborn agent and each remaining agent we can provide a corresponding necessary and sufficient condition for the presence of generic hypocrisy stated in terms of paths in network $\mathbf{D} \star \mathbf{G}$.

Proposition 6. *(i) A stubborn agent $s \in \mathcal{S}$ is generically hypocritical iff s is linked in $\mathbf{D} \star \mathbf{G}$ to some other independent class.*

(ii) A remaining agent $r \in \mathcal{R}$ is generically hypocritical iff r has a peer who belongs or is linked to some independent class in $\mathbf{D} \star \mathbf{G}$, and r has an influencer who belongs or is linked to another independent class in $\mathbf{D} \star \mathbf{G}$.

Proof. See Appendix A.3. ■

Part (i) of Proposition 6 is based on the following observation: when conditions of Proposition 5 are not met, a stubborn agent s is not generically hypocritical iff all s's peers are only remaining agents who are linked in $\mathbf{D} \star \mathbf{G}$ only to s. More precisely, if s is hypocritical, then s has at least one peer who belongs or is linked to another independent class in $\mathbf{D} \star \mathbf{G}$. Unless initial beliefs and values of sensitivities are chosen in a very special way, this is also a sufficient condition for hypocrisy of a stubborn agent (see also Example 1).

Part (ii) of Proposition 6 follows from the fact that when conditions of Proposition 5 are not met, a remaining agent r is not generically hypocritical iff r is linked

in $\mathbf{D} \star \mathbf{G}$ to a single independent class or all remaining agents reach consensus. If r is hypocritical, then r has at least one peer who belongs or is linked to some independent class in $\mathbf{D} \star \mathbf{G}$, and r's influencer belongs or is linked in $\mathbf{D} \star \mathbf{G}$ to another independent class. Again, unless initial beliefs and values of sensitivities are chosen in a very special way, these conditions are also sufficient (see also Example 4). Loosely speaking, the presence of hypocrisy is a rather typical phenomenon, while the absence of hypocrisy is a very special case.

We conclude this section by using Propositions 4 and 6 to characterize hypocrisy in arbitrary communication class $\mathcal{C}$. First, suppose that there are *no stubborn agents* in $\mathcal{C}$, that is, $\mathcal{C}$ is a pure communication class. Then agents from $\mathcal{C}$ reach consensus and have the same long-run belief. Moreover, all agents say what they believe and are not hypocritical.

Second, suppose that $\mathcal{C}$ contains *a single stubborn agent* s. Then all other members of $\mathcal{C}$ are remaining agents who are linked to a single independent class. It follows that all agents from $\mathcal{C}$ converge to s's belief and are not hypocritical: eventually, s says what s believes, and all others believe in what s says. While in both cases members of $\mathcal{C}$ are not hypocritical and reach consensus, the long-run outcome is different. Without a stubborn agent, agents' common long-run belief is a weighted average of their initial beliefs, while in the presence of a single stubborn agent only the initial belief of a stubborn agent matters.

Third, suppose that there are *two or more stubborn agents* in $\mathcal{C}$. Then two cases are possible. First, all the communication with stubborn agents is carried out by a single remaining agent r, while all other remaining agents in $\mathcal{C}$ are linked only to each other (and to this agent r). Then all the remaining agents reach consensus and are not hypocritical. Second, there is no such "bottleneck" remaining agent who is the only connection of class $\mathcal{C}$ with stubborn agents. Then due to audience tuning all remaining agents are generically hypocritical. In both cases, all stubborn agents in $\mathcal{C}$ are necessarily hypocritical unless they are peers only to each other. Thus, our results suggest that there is hypocrisy iff society is sufficiently segregated (there are no less than two independent classes with different initial beliefs), and at the same time society is sufficiently connected (different agents have peers from different independent classes). It can be argued that hypocrisy is a very natural and prevalent phenomenon in almost any society.

6 Discussion and conclusion

In this paper we make a distinction between audience and influence networks in the model of social interaction, and show that audience tuning in conversations

can lead to hypocrisy even in the long run. As a final remark, we discuss the possible implications of our model. In particular, our model highlights the fact that the spiral of silence (formation and reinforcement of the perceived majority belief) and Overton window shift (change from a mainstream to an extreme belief) can be regarded as the same social phenomenon.

First, consider the spiral of silence example. According to the spiral of silence theory (see Noelle-Neumann, 1974), the threat of isolation forces the perceived minority to stay silent about their true beliefs. Let us show that hypocrisy plays an important role in the process of formation and reinforcement of majority belief.

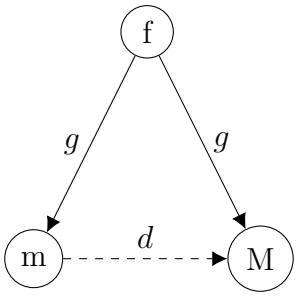
Suppose that a society consists of three agents with different initial beliefs (see Fig. 4a). Stubborn agents m and M are interpreted as holders of perceived minority and majority beliefs respectively. They need not even be actual minority and majority in society. A perceived minority faces a fear of isolation from society which forces them not to express their true beliefs but to make statements close to that of perceived majority. Dissonance sensitivity d reflects the subjective level of fear that society will neglect or ostracize its members on account of their beliefs. Agent f is a follower who is equally influenced (that is, has the same influence sensitivity g) by both influencers in forming a belief.[8]

Long-run beliefs of m and M coincide with their initial beliefs. However, due to audience tuning, perceived minority is hypocritical: m's statement is shifted from m's true belief to the statement (and the true belief) of M. As a result, the follower, who forms beliefs by paying equal attention to the statements of majority and minority, is increasingly influenced by the perceived majority belief. The long-run belief of f is a weighted average of the beliefs of both influencers: $b_f^* = b_M(0) + \frac{1}{2(1+d)}(b_m(0) - b_M(0))$. The weights of initial beliefs of M and m in f's long-run belief do not depend on influence sensitivity g and are fully determined by m's dissonance sensitivity d. The higher is d, the closer is the long-run belief of the follower to the perceived majority belief.

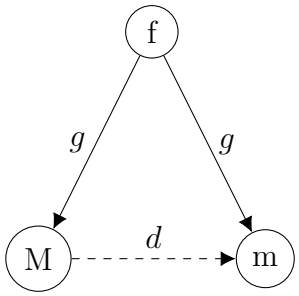
In the long run, even though the perceived minority m keeps the initial belief, sufficiently high social pressure (reflected in m's dissonance sensitivity d) pulls the belief of the follower arbitrarily close to the perceived majority belief. This conclusion is still valid even when f has different influence sensitivities with respect to m and M. Thus, as a self-fulfilling prophecy, the perceived majority belief indeed becomes the majority belief (recall also the opening quote from Tolstoy).

Second, consider the Overton window example. According to the Overton window model (see, e.g., Lehman, 2010), at each moment there is a "window of

[8]It is crucial that agent m is stubborn, while agents M and f can be replaced by arbitrary independent classes in $\mathbf{D} \star \mathbf{G}$.



(a) Spiral of silence: reinforcement of the perceived majority belief.



(b) Overton window shift: change from a mainstream to an extreme belief.

Figure 4: Audience (dashed lines) and influence (solid lines) networks in the spiral of silence (left panel) and Overton window (right panel) examples.

political possibility" which defines a set of mainstream opinions, and any views outside of this set are politically unacceptable. Let us show that dissonance and hypocrisy are social forces that may shift the Overton window and make more extreme views politically possible. This example is similar to the spiral of silence example, but in this case agents m and M change places (see Fig. 4b). Stubborn agent M is interpreted as a holder of the majority (mainstream) belief, while stubborn agent m is interpreted as a holder of minority (extreme) belief.

A majority does not want to offend a minority, and feels a pressure to make a statement closer to the statement of minority. Dissonance sensitivity d in this case reflects the tolerance to other views. Though M holds a mainstream belief, a sufficiently high level of tolerance leads the follower, who is equally influenced by majority and minority, to converge arbitrarily close to the extreme belief of minority m: $b_f^* = b_M(0) + \frac{1}{2(1+d)}(b_m(0) - b_M(0))$. Thus, by manipulating the level of tolerance, a minority may shift the Overton window to allow more extreme views into public discourse.

Appendix. Proofs

A.1 Characterization of matrix T

First, we show that elements of $\mathbf{T}$ are non-negative. Recall that $\mathbf{T} = \mathbf{QP}$, where $q_{ii} = \frac{1+\delta_i}{1+\gamma_i}$ and $q_{ij} = \frac{g_{ij}-d_{ij}}{1+\delta_i}$. It also follows from (7) that $(1+\delta_i)p_{ii} = 1+\sum_k d_{ik}p_{ki}$, and $(1+\delta_i)p_{ij} = \sum_k d_{ik}p_{kj}$. Then the self-weight t_{ii} can be written as

$$t_{ii} = \frac{(1+\delta_i)p_{ii} + \sum_k g_{ik}p_{ki} - \sum_k d_{ik}p_{ki}}{1+\gamma_i} = \frac{1+\sum_k g_{ik}p_{ki}}{1+\gamma_i}\,. \tag{A.1}$$

Similarly, the weight t_{ij} for $j \neq i$ can be written as

$$t_{ij} = \sum_k q_{ik} p_{kj} = \frac{(1+\delta_i)p_{ij} + \sum_k g_{ik}p_{kj} - \sum_k d_{ik}p_{kj}}{1+\gamma_i} = \frac{\sum_k g_{ik}p_{kj}}{1+\gamma_i}. \qquad \text{(A.2)}$$

Clearly, $0 < t_{ii} \leq 1$, and $0 \leq t_{ij} < 1$.

Second, we show that $\mathbf{T}$ is row-stochastic. Since $\mathbf{P}$ is row-stochastic, we have $\mathbf{P}^{-1}\mathbb{1} = \mathbb{1}$, where $\mathbb{1}$ is the N-vector of ones. Using the definition of $\mathbf{Q}$, we have

$$\mathbf{T}\mathbb{1} = \mathbf{Q}\mathbb{1} = (\mathbf{I}+\mathbf{\Gamma})^{-1}\left(\mathbf{P}^{-1}+\mathbf{G}\right)\mathbb{1} = (\mathbf{I}+\mathbf{\Gamma})^{-1}(\mathbf{I}+\mathbf{\Gamma})\,\mathbb{1},$$

and hence $\mathbf{T}\mathbb{1} = \mathbb{1}$, which proves Proposition 2.

Third, we prove Proposition 3. Parts (i)–(iii) follow directly from (A.1)–(A.2). To prove part (iv), suppose that there is a path in $\mathbf{T}$ from i to j, i.e., $t_{ik}t_{kl}\cdots t_{nj} > 0$. It follows from part (ii) that there are paths in $\mathbf{D}\star\mathbf{G}$ from i to k, from k to l, ..., and from n to j. A path in $\mathbf{D}\star\mathbf{G}$ from i to j can be constructed by combining these paths. Suppose that $\mathbf{G}$ does not have zero rows and $\mathbf{D}\star\mathbf{G}$ is strongly connected. Consider a path in $\mathbf{D}\star\mathbf{G}$ from i to j, i.e., $\max\{g_{ik}, d_{ik}\}\cdot\max\{g_{kl}, d_{kl}\}\cdots\max\{g_{nj}, d_{nj}\} > 0$. Two cases are possible. First, $g_{ik} > 0$. Then it follows from part (ii) that $t_{ik} > 0$. Further, if $g_{kl} > 0$, then again $t_{kl} > 0$, while if $g_{kl} = 0$, then $d_{kl} > 0$ and hence $t_{il} > 0$. Repeating the argument, we obtain a path in $\mathbf{T}$ from i to j. Second, $g_{ik} = 0$. Under our assumptions, there is h such that $g_{ih} > 0$, and there is a path in $\mathbf{D}\star\mathbf{G}$ from h to j. Therefore, we are in the conditions of the previous case, and hence there is a path in $\mathbf{T}$ from i to j.

A.2 Characterization of long-run beliefs

Without loss of generality, the set of all agents $\mathcal{N}$ can be partitioned into a set $\mathcal{S}$ of stubborn agents, M pure communication classes $\mathcal{C}_1$, ..., $\mathcal{C}_M$ in $\mathbf{D}\star\mathbf{G}$, and a set of all remaining agents $\mathcal{R}$. Let the numbers of agents in these sets be S, C_1, ..., C_M, and R, respectively. The following proposition characterizes the long-run beliefs of different sets of agents in our model.

Proposition A.1. *Let $\mathbf{b}(0)$ be an arbitrary vector of the initial beliefs with $\mathbf{b}(0)|_{\mathcal{S}}$ and $\mathbf{b}(0)|_{\mathcal{C}_m}$ denoting its restrictions to the sets $\mathcal{S}$ and $\mathcal{C}_m$, respectively. There exists $\mathbf{b}^* \equiv \lim_{t\to\infty}\mathbf{b}(t)$, and the vector of long-run beliefs is given by $\mathbf{b}^* = \{\mathbf{b}^*_{\mathcal{S}}, \mathbf{b}^*_1, \ldots, \mathbf{b}^*_M, \mathbf{b}^*_{\mathcal{R}}\}^T$. Here $\mathbf{b}^*_{\mathcal{S}} = \mathbf{b}(0)|_{\mathcal{S}}$. For each $m = 1, \ldots, M$, vector $\mathbf{b}^*_m$ has identical entries b^{*m} defined as*

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

where $\boldsymbol{\pi}^m$ is the unique positive left eigenvector whose entries sum to 1 of the matrix $\mathbf{T}$ restricted to the set $\mathcal{C}_m$ (corresponding to eigenvalue 1). The components of the vector $\mathbf{b}^*_{\mathcal{R}}$ are given by

$$b^*_r = \sum\nolimits_{s=1}^{S} \bar{\omega}_{rs} b_s(0) + \sum\nolimits_{m=1}^{M} \omega_{rm} b^{*m}\,, \qquad r \in \mathcal{R}\,, \tag{A.3}$$

where the unique non-negative weights $\left(\{\bar{\omega}_{rs}\}_{r,s=1}^{R,S} \;\; \{\omega_{rm}\}_{r,m=1}^{R,M}\right)$ sum to one: for each $r \in \mathcal{R}$, $\sum_{s=1}^{S} \bar{\omega}_{rs} + \sum_{m=1}^{M} \omega_{rm} = 1$.

Proof. Since $t_{ii} > 0$ for all i, matrix $\mathbf{T}$ is convergent (see, e.g., Theorem 2 in Golub and Jackson, 2010). That is, for any $\mathbf{b}(0)$ there exists a limit $\mathbf{b}^* \equiv \lim_{t\to\infty} \mathbf{T}^t \mathbf{b}(0)$. Note that $\mathbf{T}\mathbf{b}^* = \lim_{t\to\infty} \mathbf{T}^{t+1}\mathbf{b}(0) = \mathbf{b}^*$, so $\mathbf{b}^*$ is a right eigenvector of $\mathbf{T}$ corresponding to the eigenvalue 1. In general, $\mathbf{b}^*$ is not unique and depends on $\mathbf{b}(0)$.

The partition of the set of all agents implies that matrix $\mathbf{T}$ has the following block structure:

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

where $\mathbf{I}_S$ is the identity matrix of size S corresponding to the set of stubborn agents; for $m = 1, \dots, M$, matrix $\mathbf{T}_m = \mathbf{T}|_{\mathcal{C}_m}$ is $\mathbf{T}$ restricted to the pure communication class $\mathcal{C}_m$; $\mathbf{T}_{\mathcal{R}\mathcal{R}}$ represents the interaction of remaining agents with themselves; and matrices $\mathbf{T}_{\mathcal{R}\mathcal{S}}$ and $\mathbf{T}_{\mathcal{R}m}$ represent the interaction of remaining agents with stubborn agents and agents from pure communication class $\mathcal{C}_m$ respectively.

We have

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

where

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

Using the structure of $\mathbf{T}^t$, we can characterize the long-run beliefs indepen-

dently in each block. First, stubborn agents never change their beliefs: in the long-run $b_s^* = b_s(0)$, for all $s = 1, \ldots, S$.

Second, each pure communication class $\mathcal{C}_m$ in $\mathbf{D} \star \mathbf{G}$ is by construction strongly connected and does not contain stubborn agents. By Proposition 3, $\mathbf{T}_m$ is row-stochastic and strongly connected. Since $t_{ii} > 0$ for all i, $\mathbf{T}_m$ is convergent. By the Perron–Frobenius theorem, $\mathbf{T}_m$ has a unique positive right eigenvector corresponding to eigenvalue 1, which is a vector with identical components. Therefore, all agents from $\mathcal{C}_m$ converge to the same belief b^{*m}: for all initial beliefs $\mathbf{b}(0)|_{\mathcal{C}_m}$,

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

where $\mathbb{1}_{C_m}$ is the C_m-vector of ones.

All the rows of the limiting matrix $\mathbf{T}_m^\infty = \lim_{t\to\infty} \mathbf{T}_m^t$ are identical: for any $\mathbf{b}(0)|_{\mathcal{C}_m}$, $(\mathbf{T}_m^\infty \mathbf{b}(0)|_{\mathcal{C}_m})_i = \boldsymbol{\pi}^m \mathbf{b}(0)|_{\mathcal{C}_m}$. By construction, a (row) vector $\boldsymbol{\pi}^m$ is a left unit eigenvector corresponding to eigenvalue 1 of $\mathbf{T}_m$: $\boldsymbol{\pi}^m \mathbf{T}_m = \boldsymbol{\pi}^m$. By the Perron–Frobenius theorem, there is a unique such positive eigenvector. Thus, the consensus belief b^{*m} is a weighted average of initial beliefs of agents from the class $\mathcal{C}_m$: $b^{*m} = \boldsymbol{\pi}^m \mathbf{b}(0)|_{\mathcal{C}_m}$.

Third, consider the remaining agents. It is easily checked (see, e.g., Theorem 10 in DeMarzo et al., 2003) that $\lim_{t\to\infty} \mathbf{T}_{\mathcal{R}\mathcal{R}}^t = \mathbf{0}$, while

$$\lim_{t\to\infty} \mathbf{T}_{\mathcal{R}\mathcal{S}}(t) \;=\; \sum_{\tau=0}^{\infty} \mathbf{T}_{\mathcal{R}\mathcal{R}}^\tau \mathbf{T}_{\mathcal{R}\mathcal{S}} = \left(\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}}\right)^{-1} \mathbf{T}_{\mathcal{R}\mathcal{S}} \,,$$

and

$$\lim_{t\to\infty} \mathbf{T}_{\mathcal{R}m}(t) \;=\; \sum_{\tau=0}^{\infty} \mathbf{T}_{\mathcal{R}\mathcal{R}}^\tau \mathbf{T}_{\mathcal{R}m} \mathbf{T}_m^\infty = \left(\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}}\right)^{-1} \mathbf{T}_{\mathcal{R}m} \mathbf{T}_m^\infty \,.$$

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

$$\mathbf{b}_{\mathcal{R}}^* = \left(\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}}\right)^{-1} \left(\mathbf{T}_{\mathcal{R}\mathcal{S}} \mathbf{b}_{\mathcal{S}}^* + \mathbf{T}_{\mathcal{R}1} \mathbb{1}_{C_1} b^{*1} + \ldots + \mathbf{T}_{\mathcal{R}M} \mathbb{1}_{C_M} b^{*M}\right) .$$

The weights of the agents from independent classes for all agents $r \in \mathcal{R}$ form the $R \times (S + M)$ matrix $\boldsymbol{\Omega} = \left(\{\bar{\omega}_{rs}\}_{r,s=1}^{R,S} \quad \{\omega_{rm}\}_{r,m=1}^{R,M}\right)$ given by $\boldsymbol{\Omega} = \left(\mathbf{I}_R - \mathbf{T}_{\mathcal{R}\mathcal{R}}\right)^{-1} \left(\mathbf{T}_{\mathcal{R}\mathcal{S}} \;\; \mathbf{T}_{\mathcal{R}1} \mathbb{1}_{C_1} \;\; \cdots \;\; \mathbf{T}_{\mathcal{R}M} \mathbb{1}_{C_M}\right)$. Since $\mathbf{T}$ is row-stochastic, $\boldsymbol{\Omega}$ is also row-stochastic. ■

A.3 Proof of Proposition 6

Since $\mathbf{s}^* = \mathbf{P}\mathbf{b}^*$, agent i is not hypocritical iff $b_i^* = \sum_n p_{in} b_n^*$. Without loss of generality, suppose that all independent classes are stubborn agents. Consider a stubborn agent i. Using (A.3), the above condition can be rewritten as

$$b_i(0) \;=\; \sum\nolimits_{s\in\mathcal{S}} p_{is} b_s(0) + \sum\nolimits_{r\in\mathcal{R}} p_{ir} \sum\nolimits_{s\in\mathcal{S}} \bar{\omega}_{rs} b_s(0)\,,$$

or

$$\Big(1 - p_{ii} - \sum\nolimits_{r\in\mathcal{R}} p_{ir}\bar{\omega}_{ri}\Big)\, b_i(0) \;=\; \sum\nolimits_{s\in\mathcal{S}, s\neq i} \Big(p_{is} + \sum\nolimits_{r\in\mathcal{R}} p_{ir}\bar{\omega}_{rs}\Big)\, b_s(0)\,. \quad \text{(A.4)}$$

Eq. (A.4) trivially holds when initial beliefs of independent classes belong to an $S-1$ dimensional subspace (for fixed matrices $\mathbf{D}$ and $\mathbf{G}$) or when positive elements in $\mathbf{P}$ are specifically chosen (for fixed initial beliefs). Ignoring these zero-measure cases, a stubborn agent i is not generically hypocritical iff $p_{ii} + \sum_{r\in\mathcal{R}} p_{ir}\bar{\omega}_{ri} = 1$ and $p_{is} + \sum_{r\in\mathcal{R}} p_{ir}\bar{\omega}_{rs} = 0$ for all $s \neq i$. Since $\mathbf{P}$ and $\mathbf{\Omega}$ are row-stochastic, this is possible iff $p_{ii} = 1$ (i has no peers) or $p_{is} = 0$ for all $s \neq i$ and $\bar{\omega}_{ri} = 1$ for all r with $p_{ir} > 0$ (i's peers are only those remaining agents who are linked in $\mathbf{D} \star \mathbf{G}$ only to i). This proves part (i).

Consider a remaining agent r. Using (A.3), we obtain

$$s_r^* \;=\; \sum\nolimits_{s\in\mathcal{S}} \Big(p_{rr}\bar{\omega}_{rs} + p_{rs} + \sum\nolimits_{n\in\mathcal{R}, n\neq r} p_{rn}\bar{\omega}_{ns}\Big)\, b_s(0)\,,$$

and hence a remaining agent r is not hypocritical iff

$$\sum\nolimits_{s\in\mathcal{S}} \bar{\omega}_{rs} b_s(0) \;=\; \sum\nolimits_{s\in\mathcal{S}} \Big(p_{rr}\bar{\omega}_{rs} + p_{rs} + \sum\nolimits_{n\in\mathcal{R}, n\neq r} p_{rn}\bar{\omega}_{ns}\Big)\, b_s(0)\,. \quad \text{(A.5)}$$

Unless initial beliefs of independent classes belong to an $S-1$ dimensional subspace defined by (A.5), a remaining agent r is not generically hypocritical iff for all $s \in \mathcal{S}$,

$$(1 - p_{rr})\bar{\omega}_{rs} \;=\; p_{rs} + \sum\nolimits_{n\in\mathcal{R}, n\neq r} p_{rn}\bar{\omega}_{ns}\,. \quad \text{(A.6)}$$

Note that in the considered case, for all $s \in \mathcal{S}$, we have

$$(1 - t_{rr})\bar{\omega}_{rs} \;=\; t_{rs} + \sum\nolimits_{n\in\mathcal{R}, n\neq r} t_{rn}\bar{\omega}_{ns}\,. \quad \text{(A.7)}$$

By comparing (A.6) and (A.7), we infer that (A.6) trivially holds when $p_{rr} = 1$ or $\frac{p_{rj}}{1-p_{rr}} = \frac{t_{rj}}{1-t_{rr}}$ for all j, which are the cases described in Proposition 5.

Acknowledgements

We thank Evgeniya Goryacheva, Alan Kirman, Paolo Pin, Mikhail Sokolov, Philip Ushchev, and Yves Zenou for useful and stimulating discussions. The earlier versions of the paper have greatly benefited from comments of participants of the 7th Workshop on Networks in Economics and Finance (September 2018, Lucca), CEPET Workshop (June 2019, Udine), and WEHIA 2022 (June 2022, Catania). Mikhail Anufriev acknowledges financial support from the Australian Research Council through Discovery Project DP170100429 and is thankful to hospitality of the European University at St. Petersburg.

References

Anufriev, M., Borissov, K., and Pakhnin, M. (2021). Dissonance Minimization and Conversation in Social Networks. CESifo Working Paper 9433.

Arifovic, J., Eaton, B. C., and Walker, G. (2015). The Coevolution of Beliefs and Networks. *Journal of Economic Behavior and Organization*, **120**, pp. 46–63.

Asch, S. E. (1955). Opinions and Social Pressure. *Scientific American*, **193** (5), pp. 31–35.

Buechel, B., Hellmann, T., and Klößner, S. (2015). Opinion Dynamics and Wisdom under Conformity. *Journal of Economic Dynamics and Control*, **52**, pp. 240–257.

Chandrasekhar, A. G., Larreguy, H., and Xandri, J. P. (2020). Testing Models of Social Learning on Networks: Evidence from Two Experiments. *Econometrica*, **88** (1), pp. 1–32.

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

DeMarzo, P. M., Vayanos, D., and Zwiebel, J. (2003). Persuasion Bias, Social Influence, and Unidimensional Opinions. *Quarterly Journal of Economics*, **118** (3), pp. 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), pp. 629–636.

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), pp. 257.

Echterhoff, G., Lang, S., Krämer, N., and Higgins, E. T. (2009). Audience-tuning Effects on Memory: The Role of Audience Status in Sharing Reality. *Social Psychology*, **40** (3), pp. 150–163.

Golub, B. and Jackson, M. O. (2010). Naive Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics*, **2** (1), pp. 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), pp. 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.

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, pp. 33–49. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Lehman, J. G. (2010). An Introduction to the Overton Window of Political Possibility. https://www.mackinac.org/12481.

Noelle-Neumann, E. (1974). The Spiral of Silence. A Theory of Public Opinion. *Journal of Communication*, **24** (2), pp. 43–51.

Olcina, G., Panebianco, F., and Zenou, Y. (2017). Conformism, Social Norms and the Dynamics of Assimilation. Discussion Paper 12166, Centre for Economic Policy Research.

Seddon, M. (2014). Documents Show How Russia's Troll Army Hit America. https://www.buzzfeednews.com/article/maxseddon/documents-show-how-russias-troll-army-hit-america.

Stukal, D., Sanovich, S., Bonneau, R., and Tucker, J. A. (2022). Why Botter: How Pro-Government Bots Fight Opposition in Russia. *American Political Science Review*, **116** (3), pp. 843–857.