HAL open science

A Multi-agent Model for Polarization Under Confirmation Bias in Social Networks

Mário S. Alvim, Bernardo Amorim, Sophia Knight, Santiago Quintero, Frank Valencia

To cite this version:

Mário S. Alvim, Bernardo Amorim, Sophia Knight, Santiago Quintero, Frank Valencia. A Multi-agent Model for Polarization Under Confirmation Bias in Social Networks. 41th International Conference on Formal Techniques for Distributed Objects, Components, and Systems (FORTE), Jun 2021, Valletta, Malta. pp.22-41, 10.1007/978-3-030-78089-0_2 . hal-03740263

HAL Id: hal-03740263

https://inria.hal.science/hal-03740263v1

Submitted on 29 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

A Multi-Agent Model for Polarization under Confirmation Bias in Social Networks ⋆

Mário S. Alvim[1], Bernardo Amorim[1], Sophia Knight[2], Santiago Quintero[3], and Frank Valencia[4,5]

[1] Department of Computer Science, UFMG, Brazil
[2] Department of Computer Science, University of Minnesota Duluth, USA
[3] LIX, École Polytechnique de Paris, France
[4] CNRS-LIX, École Polytechnique de Paris, France
[5] Pontificia Universidad Javeriana Cali, Colombia

Abstract. We describe a model for polarization in multi-agent systems based on Esteban and Ray's standard measure of polarization from economics. Agents evolve by updating their beliefs (opinions) based on an underlying influence graph, as in the standard DeGroot model for social learning, but under a *confirmation bias*; i.e., a discounting of opinions of agents with dissimilar views. We show that even under this bias polarization eventually vanishes (converges to zero) if the influence graph is strongly-connected. If the influence graph is a regular symmetric circulation, we determine the unique belief value to which all agents converge. Our more insightful result establishes that, under some natural assumptions, if polarization does not eventually vanish then either there is a disconnected subgroup of agents, or some agent influences others more than she is influenced. We also show that polarization does not necessarily vanish in weakly-connected graphs under confirmation bias. We illustrate our model with a series of case studies and simulations, and show how it relates to the classic DeGroot model for social learning.

Keywords: Polarization · Confirmation bias · Multi-Agent Systems · Social Networks

1 Introduction

Distributed systems have changed substantially in the recent past with the advent of social networks. In the previous incarnation of distributed computing [22] the emphasis was on consistency, fault tolerance, resource management and related topics; these were all characterized by *interaction between processes*. What marks the new era of distributed systems is an emphasis on the flow of epistemic information (facts, beliefs, lies) and its impact on democracy and on society at large.

Indeed in social networks a group may shape their beliefs by attributing more value to the opinions of outside influential figures. This cognitive bias is known as *authority*

⋆ Mário S. Alvim and Bernardo Amorim were partially supported by CNPq, CAPES and FAPEMIG. Santiago Quintero and Frank Valencia were partially supported by the ECOS-NORD project FACTS (C19M03).

bias [32]. Furthermore, in a group with uniform views, users may become extreme by reinforcing one another's opinions, giving more value to opinions that confirm their own preexisting beliefs. This is another common cognitive bias known as *confirmation bias* [4]. As a result, social networks can cause their users to become radical and isolated in their own ideological circle causing dangerous splits in society [5] in a phenomenon known as *polarization* [4].

There is a growing interest in the development of models for the analysis of polarization and social influence in networks [6, 8, 9, 12, 14, 15, 19, 20, 28, 31, 34, 35, 37]. Since polarization involves non-terminating systems with *multiple agents* simultaneously exchanging information (opinions), concurrency models are a natural choice to capture the dynamics of polarization.

The Model. In fact, we developed a multi-agent model for polarization in [3], inspired by linear-time models of concurrency where the state of the system evolves in discrete time units (in particular [27, 33]). In each time unit, the agents *update* their beliefs about the proposition of interest taking into account the beliefs of their neighbors in an underlying weighted *influence graph*. The belief update gives more value to the opinion of agents with higher influence (*authority bias*) and to the opinion of agents with similar views (*confirmation bias*). Furthermore, the model is equipped with a *polarization measure* based on the seminal work in economics by Esteban and Ray [13]. The polarization is measured at each time unit and it is 0 if all agents' beliefs fall within an interval of agreement about the proposition. The contributions in [3] were of an experimental nature and aimed at exploring how the combination of influence graphs and cognitive biases in our model can lead to polarization.

In the current paper we prove claims made from experimental observations in [3] using techniques from calculus, graph theory, and flow networks. The main goal of this paper is identifying how networks and beliefs are structured, for agents subject to confirmation bias, when polarization *does not* disappear. Our results provide insight into the phenomenon of polarization, and are a step toward the design of robust computational models and simulation software for human cognitive and social processes.

The closest related work is that on DeGroot models [9]. These are the standard linear models for social learning whose analysis can be carried out by linear techniques from Markov chains. A novelty in our model is that its update function extends the classical update from DeGroot models with confirmation bias. As we shall elaborate in Section 5 the extension makes the model no longer linear and thus mathematical tools like Markov chains do not seem applicable. Our model incorporates a polarization measure in a model for social learning and extends classical convergence results of DeGroot models to the confirmation bias case.

Main Contributions. The following are the main theoretical results established in this paper. Assuming confirmation bias and some natural conditions about belief values: (1) If polarization does not disappear then either there is disconnected subgroup of agents, or some agent influences others more than she is influenced, or all the agents are initially radicalized (i.e., each individual holds the most extreme value either in favor or against of a given proposition). (2) Polarization eventually disappears (converges to zero) if the influence graph is strongly-connected. (3) If the influence graph is a regular symmetric circulation we determine the unique belief value all agents converge to.

Organization. In Section 2 we introduce the model and illustrate a series of examples and simulations, uncovering interesting new insights and complex characteristics of the believe evolution. The theoretical contributions (1-3) above are given in Sections 3 and 4. We discuss DeGroot and other related work in Sections 5 and 6. Full proofs can be found in the corresponding technical report [2]. An implementation of the model in Python and the simulations are available on Github [1].

2 The Model

Here we refine the polarization model introduced in [3], composed of static and dynamic elements. We presuppose basic knowledge of calculus and graph theory [11, 38].

Static Elements of the Model *Static elements* of the model represent a snapshot of a social network at a given point in time. They include the following components:

- A (finite) set $\mathcal{A} = \{0, 1, \ldots, n-1\}$ of $n \geq 1$ *agents*.
- A *proposition* p of interest, about which agents can hold beliefs.
- A *belief configuration* $B{:}\mathcal{A}{\to}[0,1]$ s.t. each value B_i is the instantaneous confidence of agent $i{\in}\mathcal{A}$ in the veracity of proposition p. Extreme values 0 and 1 represent a firm belief in, respectively, the falsehood or truth of p.
- A *polarization measure* $\rho{:}[0,1]^{\mathcal{A}}{\to}\mathbb{R}$ mapping belief configurations to real numbers. The value $\rho(B)$ indicates how polarized belief configuration B is.

There are several polarization measures described in the literature. In this work we adopt the influential measure proposed by Esteban and Ray [13].

Definition 1 (Esteban-Ray Polarization). *Consider a set* $\mathcal{Y}{=}\{y_0, y_1, \ldots, y_{k-1}\}$ *of size* k*, s.t. each* $y_i{\in}\mathbb{R}$*. Let* $(\pi, y){=}(\pi_0, \pi_1, \ldots, \pi_{k-1}, y_0, y_1, \ldots, y_{k-1})$ *be a* distribution *on* $\mathcal{Y}$ *s.t.* π_i *is the frequency of value* $y_i{\in}\mathcal{Y}$ *in the distribution.* [6] *The* Esteban-Ray (ER) polarization measure *is defined as* $\rho_{ER}(\pi, y) = K \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \pi_i^{1+\alpha} \pi_j |y_i - y_j|$*, where* $K{>}0$ *is a constant, and typically* $\alpha{\approx}1.6$*.*

The higher the value of $\rho_{ER}(\pi, y)$, the more polarized distribution (π, y) is. The measure captures the intuition that polarization is accentuated by both intra-group homogeneity and inter-group heterogeneity. Moreover, it assumes that the total polarization is the sum of the effects of individual agents on one another. The measure can be derived from a set of intuitively reasonable axioms [13].

Note that ρ_{ER} is defined on a discrete distribution, whereas in our model a general polarization metric is defined on a belief configuration $B{:}\mathcal{A}{\to}[0,1]$. To apply ρ_{ER} to our setup we convert the belief configuration B into an appropriate distribution (π, y).

Definition 2 (k-bin polarization). *Let* D_k *be a discretization of the interval* $[0,1]$ *into* $k{>}0$ *consecutive non-overlapping, non-empty intervals (*bins*)* $I_0, I_1, \ldots, I_{k-1}$*. We use the term* borderline points *of* D_k *to refer to the end-points of* $I_0, I_1, \ldots, I_{k-1}$ *different from 0 and 1. We assume an underlying discretization* D_k *throughout the paper.*

[6] W.l.o.g. we can assume the values of π_i are all non-zero and add up to 1.

Given D_k and a belief configuration B, define the distribution (π, y) as follows. Let $\mathcal{Y}=\{y_0, y_1, \ldots, y_{k-1}\}$ where each y_i is the mid-point of I_i, and let π_i be the fraction of agents having their belief in I_i. The polarization measure ρ of B is $\rho(B) = \rho_{ER}(\pi, y)$.

Notice that when there is consensus about the proposition p of interest, i.e., when all agents in belief configuration B hold the same belief value, we have $\rho(B)=0$. This happens exactly when all agents' beliefs fall within the same bin of the underlying discretization D_k. The following property is an easy consequence from Def. 1 and Def. 2.

Proposition 1 (Zero Polarization). *Let $D_k=I_0, I_1, \ldots, I_{k-1}$ be the discretization of $[0,1]$ in Def. 2. Then $\rho(B)=0$ iff there exists $m\in\{0, \ldots, k-1\}$ s.t. for all $i\in\mathcal{A}$, $B_i\in I_m$.*

Dynamic Elements of the Model *Dynamic elements* formalize the evolution of agents' beliefs as they interact over time and are exposed to different opinions. They include:

- A *time frame* $\mathcal{T}=\{0, 1, 2, \ldots\}$ representing the discrete passage of time.
- A *family of belief configurations* $\{B^t{:}\mathcal{A}\to[0,1]\}_{t\in\mathcal{T}}$ s.t. each B^t is the belief configuration of agents in $\mathcal{A}$ w.r.t. proposition p at time step $t\in\mathcal{T}$.
- A *weighted directed graph* $\mathcal{I}{:}\mathcal{A}\times\mathcal{A}\to[0,1]$. The value $\mathcal{I}(i,j)$, written $\mathcal{I}_{i,j}$, represents the *direct influence* that agent i has on agent j, or the *weight* i carries with j. A higher value means stronger weight. Conversely, $\mathcal{I}_{i,j}$ can also be viewed as the *trust* or *confidence* that j has on i. We assume that $\mathcal{I}_{i,i}=1$, meaning that agents are self-confident. We shall often refer to $\mathcal{I}$ simply as the *influence* (graph) $\mathcal{I}$.
  We distinguish, however, the direct influence $\mathcal{I}_{i,j}$ that i has on j from the *overall effect* of i in j's belief. This effect is a combination of various factors, including direct influence, their current opinions, the topology of the influence graph, and how agents reason. This overall effect is captured by the update function below.
- An *update function* $\mu{:}(B^t, \mathcal{I})\mapsto B^{t+1}$ mapping belief configuration B^t at time t and influence graph $\mathcal{I}$ to new belief configuration B^{t+1} at time $t+1$. This function models the evolution of agents' beliefs over time. We adopt the following premises.

(i) **Agents present some Bayesian reasoning**: Agents' beliefs are updated in every time step by combining their current belief with a *correction term* that incorporates the new evidence they are exposed to in that step –i.e., other agents' opinions. More precisely, when agent j interacts with agent i, the former affects the latter moving i's belief towards j's, proportionally to the difference $B^t_j - B^t_i$ in their beliefs. The intensity of the move is proportional to the influence $\mathcal{I}_{j,i}$ that j carries with i. The update function produces an overall correction term for each agent as the average of all other agents' effects on that agent, and then incorporates this term into the agent's current belief. [7] The factor $\mathcal{I}_{j,i}$ allows the model to capture *authority bias* [32], by which agents' influences on each other may have different intensities (by, e.g., giving higher weight to an authority's opinion).

[7] Note that this assumption implies that an agent has an influence on himself, and hence cannot be used as a "puppet" who immediately assumes another's agent's belief.

(ii) **Agents may be prone to confirmation bias**: Agents may give more weight to evidence supporting their current beliefs while discounting evidence contradicting them, independently from its source. This behavior in known in the psychology literature as *confirmation bias* [4], and is captured in our model as follows. When agent j interacts with agent i, the update function moves agent i's belief toward that of agent j, proportionally to the influence $\mathcal{I}_{j,i}$ of j on i, but with a caveat: the move is stronger when j's belief is similar to i's than when it is dissimilar.

The premises above are formally captured in the following update-function.

Definition 3 (Confirmation-bias). *Let B^t be a belief configuration at time $t{\in}\mathcal{T}$, and $\mathcal{I}$ be an influence graph. The* confirmation-bias update-function *is the map $\mu^{CB}{:}(B^t,\mathcal{I}) \mapsto B^{t+1}$ with B^{t+1} given by $B_i^{t+1} = B_i^t + {}^1\!/_{|\mathcal{A}_i|} \sum_{j\in\mathcal{A}_i} \beta_{i,j}^t\, \mathcal{I}_{j,i}\, (B_j^t - B_i^t)$, for every agent $i{\in}\mathcal{A}$, where $\mathcal{A}_i = \{j{\in}\mathcal{A} \mid \mathcal{I}_{j,i}{>}0\}$ is the set of* neighbors *of i and $\beta_{i,j}^t{=}1{-}|B_j^t{-}B_i^t|$ is the* confirmation-bias factor *of i w.r.t. j given their beliefs at time t.*

The expression ${}^1\!/_{|\mathcal{A}_i|} \sum_{j\in\mathcal{A}_i} \beta_{i,j}^t\, \mathcal{I}_{j,i}\, (B_j^t - B_i^t)$ in Def. 3 is a *correction term* incorporated into agent i's original belief B_i^t at time t. The correction is the average of the effect of each neighbor $j{\in}\mathcal{A}_i$ on agent i's belief at that time step. The value B_i^{t+1} is the resulting updated belief of agent i at time $t{+}1$.

The confirmation-bias factor $\beta_{i,j}^t$ lies in the interval $[0,1]$, and the lower its value, the more agent i discounts the opinion provided by agent j when incorporating it. It is maximum when agents' beliefs are identical, and minimum they are extreme opposites.

Remark 1 (Classical Update: Authority Non-Confirmatory Bias). In this paper we focus on confirmation-bias update and, unless otherwise stated, assume the underlying function is given by Def. 3. Nevertheless, in Sections 4 and 5 we will consider a *classical update* $\mu^C{:}(B^t,\mathcal{I}){\mapsto}B^{t+1}$ that captures non-confirmatory authority-bias and is obtained by replacing the confirmation-bias factor $\beta_{i,j}^t$ in Def. 3 with 1. That is, $B_i^{t+1}{=}B_i^t{+}{}^1\!/_{|\mathcal{A}_i|} \sum_{j\in\mathcal{A}_i} \mathcal{I}_{j,i}\, (B_j^t{-}B_i^t)$. (We refer to this function as *classical* because it is closely related to the standard update function of the DeGroot models for social learning from Economics [9]. This correspondence will be formalized in Section 5.)

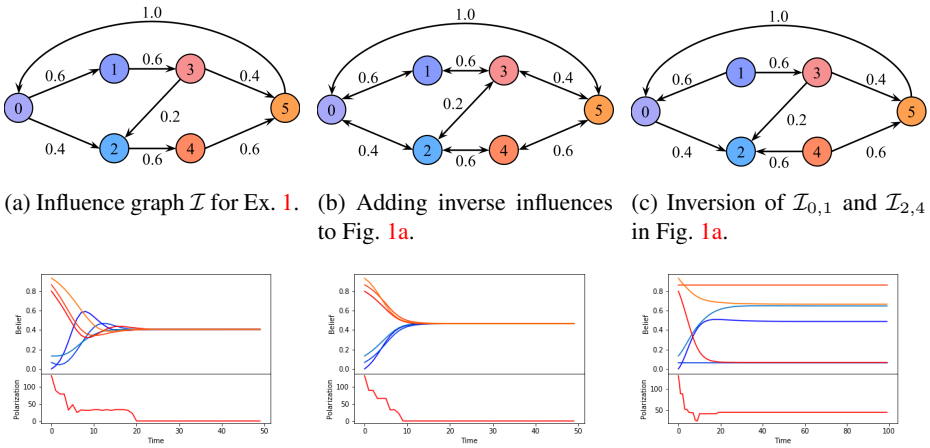
2.1 Running Example and Simulations

We now present a running example and several simulations that motivate our theoretical results. Recall that we assume $\mathcal{I}_{i,i}{=}1$ for every $i{\in}\mathcal{A}$. For simplicity, in all figures of influence graphs we omit self-loops.

In all cases we compute the polarization measure (Def. 2) using a discretization D_k of $[0,1]$ for $k{=}5$ bins, each representing a possible general position w.r.t. the veracity of the proposition p of interest: *strongly against*, $[0, 0.20)$; *fairly against*, $[0.20, 0.40)$; *neutral/unsure*, $[0.40, 0.60)$; *fairly in favour*, $[0.60, 0.80)$; and *strongly in favour*, $[0.80, 1]$.[8] We set parameters $\alpha{=}1.6$, as suggested by Esteban and Ray [13], and $K{=}1\,000$. In all definitions we let $\mathcal{A}{=}\{0, 1, \ldots, n{-}1\}$, and $i, j{\in}\mathcal{A}$ be generic agents.

As a running example we consider the following hypothetical situation.

[8] Recall from Def. 2 that our model allows arbitrary discretizations D_k –i.e., different number of bins, with not-necessarily uniform widths– depending on the scenario of interest.



(a) Influence graph $\mathcal{I}$ for Ex. 1. (b) Adding inverse influences to Fig. 1a. (c) Inversion of $\mathcal{I}_{0,1}$ and $\mathcal{I}_{2,4}$ in Fig. 1a.

(d) Beliefs and pol. for Fig. 1a. (e) Beliefs and pol. for Fig. 1b. (f) Belief and pol. for Fig. 1c.

Fig. 1: Influence graphs and evolution of beliefs and polarization for Ex. 1.

Example 1 (Vaccine Polarization). Consider the sentence "vaccines are safe" as the proposition p of interest. Assume a set $\mathcal{A}$ of 6 agents that is initially *extremely polarized* about p: agents 0 and 5 are absolutely confident, respectively, in the falsehood or truth of p, whereas the others are equally split into strongly in favour and strongly against p.

Consider first the situation described by the influence graph in Fig. 1a. Nodes 0, 1 and 2 represent anti-vaxxers, whereas the rest are pro-vaxxers. In particular, note that although initially in total disagreement about p, Agent 5 carries a lot of weight with Agent 0. In contrast, Agent 0's opinion is very close to that of Agents 1 and 2, even if they do not have any direct influence over him. Hence the evolution of Agent 0's beliefs will be mostly shaped by that of Agent 5. As can be observed in the evolution of agents' opinions in Fig. 1d, Agent 0 moves from being initially strongly against to being fairly in favour of p around time step 8. Moreover, polarization eventually vanishes (i.e., becomes zero) around time 20, as agents reach the consensus of being fairly against p.

Now consider the influence graph in Fig. 1b, which is similar to Fig. 1a, but with reciprocal influences (i.e., the influence of i over j is the same as the influence of j over i). Now Agents 1 and 2 do have direct influences over Agent 0, so the evolution of Agent 0's belief will be partly shaped by initially opposed agents: Agent 5 and the anti-vaxxers. But since Agent 0's opinion is very close to that of Agents 1 and 2, the confirmation-bias factor will help keeping Agent 0's opinion close to their opinion against p. In particular, in contrast to the situation in Fig. 1d, Agent 0 never becomes in favour of p. The evolution of the agents' opinions and their polarization is shown in Fig. 1e. Notice that polarization vanishes around time 8 as the agents reach consensus but this time they are more positive about (less against) p than in the first situation.

Finally, consider the situation in Fig. 1c obtained from Fig. 1a by inverting the influences of Agent 0 over Agent 1 and Agent 2 over Agent 4. Notice that Agents 1 and 4 are no longer influenced by anyone though they influence others. Thus, as shown in

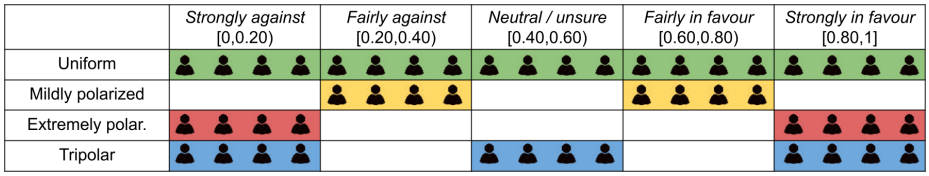
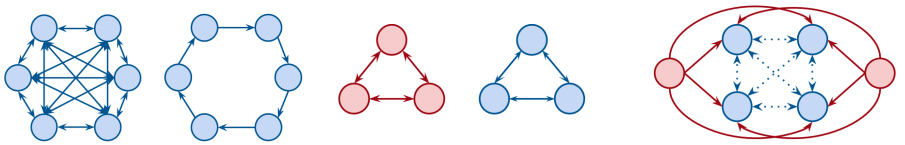


Fig. 2: Depiction of different initial belief configurations used in simulations.



(a) Clique (b) Circular (c) Disconnected groups (d) Unrelenting influencers

Fig. 3: The general shape of influence graphs used in simulations, for n=6 agents.

Fig. 1f, their beliefs do not change over time, which means that the group does not reach consensus and polarization never disappears though it is considerably reduced. □

The above example illustrates complex non-monotonic, overlapping, convergent, and non-convergent evolution of agent beliefs and polarization even in a small case with n=6 agents. Next we present simulations for several influence graph topologies with n=1 000 agents, which illustrate more of this complex behavior emerging from confirmation-bias interaction among agents. Our theoretical results in the next sections bring insight into the evolution of beliefs and polarization depending on graph topologies.

In all simulations we limit execution to T time steps varying according to the experiment. A detailed mathematical specification of simulations can be found in the corresponding technical report [2].

We consider the following initial belief configurations, depicted in Fig. 2: a *uniform* belief configuration with a set of agents whose beliefs are as varied as possible, all equally spaced in the interval $[0, 1]$; a *mildly polarized* belief configuration with agents evenly split into two groups with moderately dissimilar inter-group beliefs compared to intra-group beliefs; an *extremely polarized* belief configuration representing a situation in which half of the agents strongly believe the proposition, whereas half strongly disbelieve it; and a *tripolar* configuration with agents divided into three groups.

As for influence graphs, we consider the following ones, depicted in Fig. 3:

- A C-*clique* influence graph $\mathcal{I}^{clique}$ in which each agent influences every other with constant value C=0.5. This represents a social network in which all agents interact among themselves, and are all immune to authority bias.
- A *circular* influence graph $\mathcal{I}^{circ}$ representing a social network in which agents can be organized in a circle in such a way each agent is only influenced by its predecessor and only influences its successor. This is a simple instance of a balanced graph (in which each agent's influence on others is as high as the influence received, as in Def. 9 ahead), which is a pattern commonly encountered in some sub-networks.

- A *disconnected* influence graph $\mathcal{I}^{disc}$ representing a social network sharply divided into two groups in such a way that agents within the same group can considerably influence each other, but not at all the agents in the other group.
- An *unrelenting influencers* influence graph $\mathcal{I}^{unrel}$ representing a scenario in which two agents exert significantly stronger influence on every other agent than these other agents have among themselves. This could represent, e.g., a social network in which two totalitarian media companies dominate the news market, both with similarly high levels of influence on all agents. The networks have clear agendas to push forward, and are not influenced in a meaningful way by other agents.

We simulated the evolution of agents' beliefs and the corresponding polarization of the network for all combinations of initial belief configurations and influence graphs presented above. The results, depicted in Figure 4, will be used throughout this paper to illustrate some of our formal results. Both the Python implementation of the model and the Jupyter Notebook containing the simulations are available on Github [1].

3 Belief and Polarization Convergence

Polarization tends to diminish as agents approximate a *consensus*, i.e., as they (asymptotically) agree upon a common belief value for the proposition of interest. Here and in Section 4 we consider meaningful families of influence graphs that guarantee consensus *under confirmation bias*. We also identify fundamental properties of agents, and the value of convergence. Importantly, we relate influence with the notion of *flow* in flow networks, and use it to identify necessary conditions for polarization not converging to zero.

3.1 Polarization at the limit

Prop. 1 states that our polarization measure on a belief configuration (Def. 2) is zero exactly when all belief values in it lie within the same bin of the underlying discretization $D_k{=}I_0 \dots I_{k-1}$ of $[0,1]$. In our model polarization converges to zero if all agents' beliefs converge to a same non-borderline value. More precisely:

Lemma 1 (Zero Limit Polarization). *Let v be a non-borderline point of D_k such that for every $i{\in}\mathcal{A}$, $\lim_{t\to\infty} B_i^t{=}v$. Then $\lim_{t\to\infty} \rho(B^t){=}0$.*

To see why we exclude the $k{-}1$ borderline values of D_k in the above lemma, assume $v{\in}I_m$ is a borderline value. Suppose that there are two agents i and j whose beliefs converge to v, but with the belief of i staying always within I_m whereas the belief of j remains outside of I_m. Under these conditions one can verify, using Def. 1 and Def. 2, that ρ will not converge to 0. This situation is illustrated in Fig. 5b assuming a discretization $D_2 = [0, 1/2), [1/2, 1]$ whose only borderline is $1/2$. Agents' beliefs converge to value $v{=}1/2$, but polarization does not converge to 0. In contrast, Fig.5c illustrates Lem.1 for $D_3 = [0, 1/3), [1/3, 2/3), [2/3, 1]$. [9]

[9] It is worthwhile to note that this discontinuity at borderline points matches real scenarios where each bin represents a sharp action an agent takes based on his current belief value. Even when

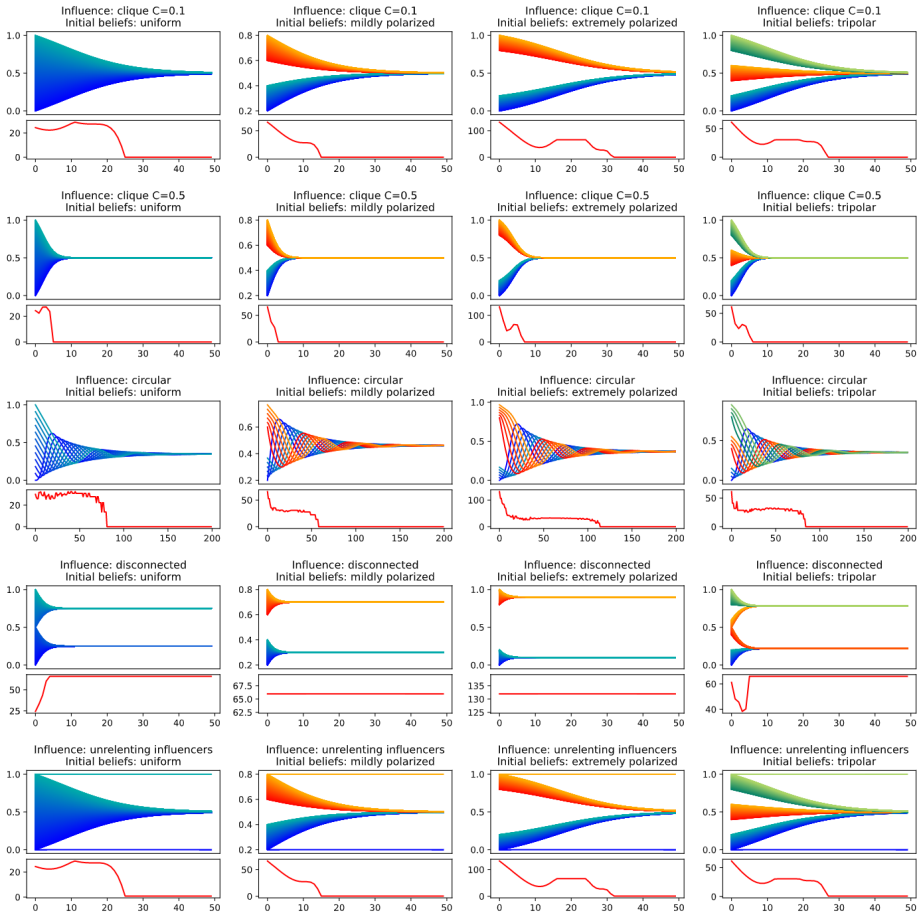
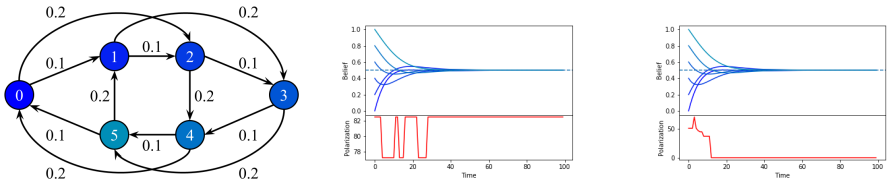


Fig. 4: Evolution of belief and polarization under confirmation bias. Horizontal axes represent time. Each row contains all graphs with the same influence graph, and each column all graphs with the same initial belief configuration. Simulations of circular influences used n=12 agents, the rest used n=1 000 agents.



(a) Influence graph.

(b) Beliefs and polarization, 2 bins, for graph in Fig. 5a.

(c) Beliefs and polarization, 3 bins, for graph in Fig. 5a.

Fig. 5: Belief convergence to borderline value 1/2. Polarization does not converge to 0 with equal-length 2 bins (Fig. 5b) and but it does with 3 equal-length bins (Fig. 5c).

3.2 Convergence under Confirmation Bias in Strongly Connected Influence

We now introduce the family of *strongly-connected* influence graphs, which includes cliques, that describes scenarios where each agent has an influence over all others. Such influence is not necessarily *direct* in the sense defined next, or the same for all agents, as in the more specific cases of cliques.

Definition 4 (Influence Paths). *Let* $C \in (0, 1]$. *We say that* i *has a* direct influence C *over* j, *written* $i \xrightarrow{C} j$, *if* $\mathcal{I}_{i,j} = C$.

An influence path *is a finite sequence of* distinct *agents from* $\mathcal{A}$ *where each agent in the sequence has a direct influence over the next one. Let* p *be an influence path* $i_0 i_1 \ldots i_n$. *The* size *of* p *is* $|p|{=}n$. *We also use* $i_0 \xrightarrow{C_1} i_1 \xrightarrow{C_2} \ldots \xrightarrow{C_n} i_n$ *to denote* p *with the direct influences along this path. We write* $i_0 \overset{C}{\rightsquigarrow}_p i_n$ *to indicate that the* product influence *of* i_0 *over* i_n *along* p *is* $C{=}C_1{\times}\ldots{\times}C_n$.

We often omit influence or path indices from the above arrow notations when they are unimportant or clear from the context. We say that i *has an* influence *over* j *if* $i{\rightsquigarrow}j$.

The next definition is akin to the graph-theoretical notion of strong connectivity.

Definition 5 (Strongly Connected Influence). *We say that an influence graph* $\mathcal{I}$ *is* strongly connected *if for all* $i, j{\in}\mathcal{A}$ *such that* $i{\neq}j$, $i{\rightsquigarrow}j$.

Remark 2. For technical reasons we assume that, *initially*, there are no two agents $i, j{\in}\mathcal{A}$ such that $B_i^0{=}0$ and $B_j^0{=}1$. This implies that for every $i, j{\in}\mathcal{A}$: $\beta_{i,j}^0{>}0$ where $\beta_{i,j}^0$ is the confirmation bias of i towards j at time 0 (See Def. 3). Nevertheless, at the end of this section we will address the cases in which this condition does not hold.

We shall use the notion of maximum and minimum belief values at a given time t.

---

two agents' beliefs are asymptotically converging to a same borderline value from different sides, their discrete decisions will remain distinct. E.g., in the vaccine case of Ex. 1, even agents that are asymptotically converging to a common belief value of 0.5 will take different decisions on whether or not to vaccinate, depending on which side of 0.5 their belief falls. In this sense, although there is convergence in the underlying belief values, there remains polarization w.r.t. real-world actions taken by agents.

Definition 6 (Extreme Beliefs). *Define* $max^t = \max_{i\in\mathcal{A}} B_i^t$ *and* $min^t = \max_{i\in\mathcal{A}} B_i^t$.

It is worth noticing that *extreme agents* –i.e., those holding extreme beliefs– do not necessarily remain the same across time steps. Fig. 1d illustrates this point: Agent 0 goes from being the one most against the proposition of interest at time $t{=}0$ to being the one most in favour of it around $t{=}8$. Also, the third row of Fig. 4 shows simulations for a circular graph under several initial belief configurations. Note that under all initial belief configurations different agents alternate as maximal and minimal belief holders.

Nevertheless, in what follows will show that the beliefs of all agents, under strongly-connected influence and confirmation bias, converge to the same value since the difference between min^t and max^t goes to 0 as t approaches infinity. We begin with a lemma stating a property of the confirmation-bias update: *The belief value of any agent at any time is bounded by those from extreme agents in the previous time unit*.

Lemma 2 (Belief Extremal Bounds). *For every* $i \in \mathcal{A}$, $min^t \leq B_i^{t+1} \leq max^t$.

The next corollary follows from the assumption in Rmk. 2 and Lemma 2.

Corollary 1. *For every* $i, j{\in}\mathcal{A}$, $t{\geq}0$*:* $\beta_{i,j}^t{>}0$.

Note that monotonicity does not necessarily hold for belief evolution. This is illustrated by Agent 0's behavior in Fig. 1d. However, it follows immediately from Lemma 2 that $min^{\cdot}$ and $max^{\cdot}$ are monotonically increasing and decreasing functions of t.

Corollary 2 (Monotonicity of Extreme Beliefs). $max^{t+1}{\leq}max^t$ *and* $min^{t+1}{\geq}min^t$ *for all* $t{\in}\mathbb{N}$.

Monotonicity and the bounding of $max^{\cdot}$, $min^{\cdot}$ within $[0, 1]$ lead us, via the Monotonic Convergence Theorem [38], to the existence of *limits for beliefs of extreme agents*.

Theorem 1 (Limits of Extreme Beliefs). *There are* $U, L{\in}[0, 1]$ *s.t.* $\lim_{t\to\infty} max^t{=}U$ *and* $\lim_{t\to\infty} min^t{=}L$.

We still need to show that U and L are the same value. For this we prove a distinctive property of agents under strongly connected influence graphs: the belief of any agent at time t will influence every other agent by the time $t{+}|\mathcal{A}|{-}1$. This is precisely formalized below in Lemma 3. First, however, we introduce some bounds for confirmation-bias, influence as well as notation for the limits in Th.1.

Definition 7 (Min Factors). *Define* $\beta_{min}{=}\min_{i,j\in\mathcal{A}} \beta_{i,j}^0$ *as the minimal confirmation bias factor at* $t{=}0$*. Also let* $\mathcal{I}_{min}$ *be the smallest positive influence in* $\mathcal{I}$*. Furthermore, let* $L{=}\lim_{t\to\infty} min^t$ *and* $U{=}\lim_{t\to\infty} max^t$.

Notice that since min^t and max^t do not get further apart as the time t increases (Cor. 2), $\min_{i,j\in\mathcal{A}} \beta_{i,j}^t$ is a non-decreasing function of t. Therefore β_{min} acts as a lower bound for the confirmation-bias factor in every time step.

Proposition 2. $\beta_{min} = \min_{i,j\in\mathcal{A}} \beta_{i,j}^t$ *for every* $t > 0$.

The factor β_{min} is used in the next result to establish that the belief of agent i at time t, the minimum confirmation-bias factor, and the maximum belief at t act as bound of the belief of j at $t{+}|p|$, where p is an influence path from i and j.

Lemma 3 (Path bound). *If $\mathcal{I}$ is strongly connected:*

1. *Let p be an arbitrary path $i \overset{C}{\rightsquigarrow}_p j$. Then $B_j^{t+|p|} \leq max^t{+}C\beta_{min}^{|p|}/|\mathcal{A}|^{|p|}(B_i^t{-}max^t)$.*
2. *Let $\mathtt{m}^t{\in}\mathcal{A}$ be an agent holding the least belief value at time t and p be a path such that $\mathtt{m}^t \rightsquigarrow_p i$. Then $B_i^{t+|p|} \leq max^t{-}\delta$, with $\delta = (\mathcal{I}_{min}\beta_{min}/|\mathcal{A}|)^{|p|}(U{-}L)$.*

Next we establish that all beliefs at time $t{+}|\mathcal{A}|{-}1$ are smaller than the maximal belief at t by a factor of at least ϵ depending on the minimal confirmation bias, minimal influence and the limit values L and U.

Lemma 4. *Suppose that $\mathcal{I}$ is strongly-connected.*

1. *If $B_i^{t+n} \leq max^t - \gamma$ and $\gamma \geq 0$ then $B_i^{t+n+1} \leq max^t - \gamma/|\mathcal{A}|$.*
2. *$B_i^{t+|\mathcal{A}|-1} \leq max^t - \epsilon$, where ϵ is equal to $(\mathcal{I}_{min}\beta_{min}/|\mathcal{A}|)^{|\mathcal{A}|-1}(U - L)$.*

Lem. 4(2) states that $\max^{\cdot}$ decreases by at least ϵ after $|A|{-}1$ steps. Therefore, after $m(|A| - 1)$ steps it should decrease by at least $m\epsilon$.

Corollary 3. *If $\mathcal{I}$ is strongly connected, $max^{t+m(|\mathcal{A}|-1)}{\leq}max^t{-}m\epsilon$ for ϵ in Lem. 4.*

We can now state that in strongly connected influence graphs extreme beliefs eventually converge to the same value. The proof uses Cor. 1 and Cor. 3 above.

Theorem 2. *If $\mathcal{I}$ is strongly connected then $\lim_{t\to\infty} max^t = \lim_{t\to\infty} min^t$.*

Combining Th. 2, the assumption in Rmk. 2 and the Squeeze Theorem, we conclude that for strongly-connected graphs, all agents' beliefs converge to the same value.

Corollary 4. *If $\mathcal{I}$ is strongly connected then for all $i, j{\in}\mathcal{A}$, $\lim_{t\to\infty} B_i^t{=}\lim_{t\to\infty} B_j^t$.*

The Extreme Cases. We assumed in Rmk. 2 that there were no two agents i, j s.t. $B_i^t{=}0$ and $B_j^t{=}1$. Th. 3 below addresses the situation in which this does not happen. More precisely, it establishes that under confirmation-bias update, in any strongly-connected, non-radical society, agents' beliefs eventually converge to the same value.

Definition 8 (Radical Beliefs). *An agent $i{\in}\mathcal{A}$ is called* radical *if $B_i{=}0$ or $B_i{=}1$. A belief configuration B is* radical *if every $i{\in}\mathcal{A}$ is radical.*

Theorem 3 (Confirmation-Bias Belief Convergence). *In a strongly connected influence graph and under the confirmation-bias update-function, if B^0 is not radical then for all $i, j{\in}\mathcal{A}$, $\lim_{t\to\infty} B_i^t{=}\lim_{t\to\infty} B_j^t$. Otherwise for every $i{\in}\mathcal{A}$, $B_i^t{=}B_i^{t+1}{\in}\{0, 1\}$.*

We conclude this section by emphasizing that belief convergence is not guaranteed in non strongly-connected graphs. Fig. 1c from the vaccine example shows such a graph where neither belief convergence nor zero-polarization is obtained.

4 Conditions for Polarization

We now use concepts from flow networks to identify insightful necessary conditions for polarization never disappearing. Understanding the conditions when polarization *does not* disappear under confirmation bias is one of the main contributions of this paper.

Balanced Influence: Circulations The following notion is inspired by the *circulation problem* for directed graphs (or flow network) [11]. Given a graph $G = (V, E)$ and a function $c{:}E{\to}\mathbb{R}$ (called *capacity*), the problem involves finding a function $f{:}E{\to}\mathbb{R}$ (called *flow*) such that: (1) $f(e){\leq}c(e)$ for each $e{\in}E$; and (2) $\sum_{(v,w)\in E} f(v,w) = \sum_{(w,v)\in E} f(w,v)$ for all $v{\in}V$. If such an f exists it is called a *circulation* for G and c.

Thinking of flow as influence, the second condition, called *flow conservation*, corresponds to requiring that each agent influences others as much as is influenced by them.

Definition 9 (Balanced Influence). *We say that $\mathcal{I}$ is* balanced *(or a* circulation*) if every $i \in \mathcal{A}$ satisfies the constraint $\sum_{j\in\mathcal{A}} \mathcal{I}_{i,j}{=}\sum_{j\in\mathcal{A}} \mathcal{I}_{j,i}$.*

Cliques and circular graphs, where all (non-self) influence values are equal, are balanced (see Fig. 3b). The graph of our vaccine example (Fig. 1) is a circulation that it is neither a clique nor a circular graph. Clearly, influence graph $\mathcal{I}$ is balanced if it is a solution to a circulation problem for some $G{=}(\mathcal{A}, \mathcal{A}{\times}\mathcal{A})$ with capacity $c{:}\mathcal{A}{\times}\mathcal{A}{\to}[0, 1]$.

Next we use a fundamental property from flow networks describing flow conservation for graph cuts [11]. Interpreted in our case it says that any group of agents $A{\subseteq}\mathcal{A}$ influences other groups as much as they influence A.

Proposition 3 (Group Influence Conservation). *Let $\mathcal{I}$ be balanced and $\{A, B\}$ be a partition of $\mathcal{A}$. Then $\sum_{i\in A}\sum_{j\in B} \mathcal{I}_{i,j} = \sum_{i\in A}\sum_{j\in B} \mathcal{I}_{j,i}$.*

We now define *weakly connected influence*. Recall that an undirected graph is *connected* if there is path between each pair of nodes.

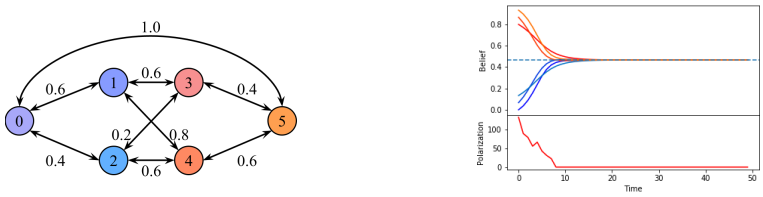
Definition 10 (Weakly Connected Influence). *Given an influence graph $\mathcal{I}$, define the undirected graph $G_{\mathcal{I}}{=}(\mathcal{A}, E)$ where $\{i, j\}{\in}E$ if and only if $\mathcal{I}_{i,j}{>}0$ or $\mathcal{I}_{j,i}{>}0$. An influence graph $\mathcal{I}$ is called* weakly connected *if the undirected graph $G_{\mathcal{I}}$ is connected.*

Weakly connected influence relaxes its strongly connected counterpart. However, every balanced, weakly connected influence is strongly connected as implied by the next lemma. Intuitively, circulation flows never leaves strongly connected components.

Lemma 5. *If $\mathcal{I}$ is balanced and $\mathcal{I}_{i,j}{>}0$ then $j{\rightsquigarrow}i$.*

Conditions for Polarization We have now all elements to identify conditions for permanent polarization. The convergence for strongly connected graphs (Th. 3), the polarization at the limit lemma (Lem. 1), and Lem. 5 yield the following noteworthy result.

Theorem 4 (Conditions for Polarization). *Suppose that $\lim_{t\to\infty} \rho(B^t){\neq}0$. Then either: (1) $\mathcal{I}$ is not balanced; (2) $\mathcal{I}$ is not weakly connected; (3) B^0 is radical; or (4) for some borderline value v, $\lim_{t\to\infty} B_i^t{=}v$ for each $i{\in}\mathcal{A}$.*



(a) Regular and reciprocal influence.

(b) Beliefs and pol. for Fig. 6a.

Fig. 6: Influence and evolution of beliefs and polar.

Hence, at least one of the four conditions is necessary for the persistence of polarization. If (1) then there must be at least one agent that influences more than what he is influenced (or vice versa). This is illustrated in Fig. 1c from the vaccine example, where Agent 2 is such an agent. If (2) then there must be isolated subgroups of agents; e.g., two isolated strongly-connected components the members of the same component will achieve consensus but the consensus values of the two components may be very different. This is illustrated in the fourth row of Fig. 4. Condition (3) can be ruled out if there is an agent that is not radical, like in all of our examples and simulations. As already discussed, (4) depends on the underlying discretization D_k (e.g., assuming equal-length bins if v is borderline in D_k it is not borderline in D_{k+1}, see Fig. 5.).

Reciprocal and Regular Circulations The notion of circulation allowed us to identify potential causes of polarization. In this section we will also use it to identify meaningful topologies whose symmetry can help us predict the exact belief value of convergence.

A *reciprocal* influence graph is a circulation where the influence of i over j is the same as that of j over i, i.e, $\mathcal{I}_{i,j}{=}\mathcal{I}_{j,i}$. Also a graph is (*in-degree*) *regular* if the in-degree of each nodes is the same; i.e., for all $i, j{\in}\mathcal{A}$, $|\mathcal{A}_i|{=}|\mathcal{A}_j|$.

As examples of regular and reciprocal graphs, consider a graph $\mathcal{I}$ where all (non-self) influence values are equal. If $\mathcal{I}$ is *circular* then it is a regular circulation, and if $\mathcal{I}$ is a *clique* then it is a reciprocal regular circulation. Also we can modify slightly our vaccine example to obtain a regular reciprocal circulation as shown in Fig. 6.

The importance of regularity and reciprocity of influence graphs is that their symmetry is sufficient to the determine the exact value all the agents converge to under confirmation bias: *the average of initial beliefs*. Furthermore, under classical update (see Rmk. 1), we can drop reciprocity and obtain the same result. The result is proven using Lem. 5, Th. 3, Cor. 5, the squeeze theorem and by showing that $\sum_{i\in\mathcal{A}} B_i^t{=}\sum_{i\in\mathcal{A}} B_i^{t+1}$ using symmetries derived from reciprocity, regularity, and the fact that $\beta_{i,j}^t{=}\beta_{j,i}^t$.

Theorem 5 (Consensus Value). *Suppose that $\mathcal{I}$ is regular and weakly connected. If $\mathcal{I}$ is reciprocal and the belief update is confirmation-bias, or if the influence graph $\mathcal{I}$ is a circulation and the belief update is classical, then* $\lim_{t\to\infty} B_i^t = {}^1\!/_{|\mathcal{A}|} \sum_{j\in\mathcal{A}} B_j^0$ *for every $i{\in}\mathcal{A}$.*

5 Comparison to DeGroot's model

DeGroot proposed a very influential model, closely related to our work, to reason about learning and consensus in multi-agent systems [9], in which beliefs are updated by a constant stochastic matrix at each time step. More specifically, consider a group $\{1, 2, \ldots, k\}$ of k agents, s.t. each agent i holds an initial (real-valued) opinion F_i^0 on a given proposition of interest. Let $T_{i,j}$ be a non-negative weight that agent i gives to agent j's opinion, s.t. $\sum_{j=1}^{k} T_{i,j}{=}1$. DeGroot's model posits that an agent i's opinion F_i^t at any time $t{\geq}1$ is updated as $F_i^t{=}\sum_{j=1}^{k} T_{i,j} F_i^{t-1}$. Letting F^t be a vector containing all agents' opinions at time t, the overall update can be computed as $F^{t+1}{=}TF^t$, where $T{=}\{T_{i,j}\}$ is a stochastic matrix. This means that the t-th configuration (for $t{\geq}1$) is related to the initial one by $F^t{=}T^t F^0$, which is a property thoroughly used to derive results in the model.

When we use classical update (as in Remark 1), our model reduces to DeGroot's via the transformation $F_i^0{=}B_i^0$, and $T_{i,j}{=}1/_{|\mathcal{A}_i|}\, \mathcal{I}_{j,i}$ if $i{\neq}j$, or $T_{i,j}{=}1{-}1/_{|\mathcal{A}_i|} \sum_{j\in\mathcal{A}_i} \mathcal{I}_{j,i}$ otherwise. Notice that $T_{i,j}{\leq}1$ for all i and j, and, by construction, $\sum_{j=1}^{k} T_{i,j}{=}1$ for all i. The following result is an immediate consequence of this reduction.

Corollary 5. *In a strongly connected influence graph $\mathcal{I}$, and under the classical update function, for all $i, j{\in}\mathcal{A}$, $\lim_{t\to\infty} B_i^t{=}\lim_{t\to\infty} B_j^t$.*

Unlike its classical counterpart, however, the confirmation-bias update (Def. 3) does not have an immediate correspondence with DeGroot's model. Indeed, this update is not linear due the confirmation-bias factor $\beta_{i,j}^t{=}1{-}|B_j^t{-}B_i^t|$. This means that in our model there is no immediate analogue of the relation among arbitrary configurations and the initial one as the relation in DeGroot's model (i.e., $F^t{=}T^t F^0$). Therefore, proof techniques usually used in DeGroot's model (e.g., based on Markov properties) are not immediately applicable to our model. In this sense our model is an extension of DeGroot's, and we need to employ different proof techniques to obtain our results.

6 Conclusions and Other Related Work

We proposed a model for polarization and belief evolution for multi-agent systems under confirmation-bias. We showed that whenever all agents can directly or indirectly influence each other, their beliefs always converge, and so does polarization as long as the convergence value is not a borderline point. We also identified necessary conditions for polarization not to disappear, and the convergence value for some important network topologies. As future work we intend to extend our model to model evolution of beliefs and measure polarization in situations in which agents hold opinions about multiple propositions of interest.

Related Work. As mentioned in the introduction and discussed in detail in Section 5, the closest related work is on DeGroot models for social learning [9]. We summarize some other relevant approaches put into perspective the novelty of our approach.

Polarization Polarization was originally studied as a psychological phenomenon in [26], and was first rigorously and quantitatively defined by economists Esteban and

Ray [13]. Their measure of polarization, discussed in Section 2, is influential, and we adopt it in this paper. Li et al.[20], and later Proskurnikov et al. [31] modeled consensus and polarization in social networks. Like much other work, they treat polarization simply as the lack of consensus and focus on when and under what conditions a population reaches consensus.Elder's work [12] focuses on methods to avoid polarization, without using a quantitative definition of polarization. [6] measures polarization but purely as a function of network topology, rather than taking agents' quantitative beliefs and opinions into account, in agreement with some of our results.

Formal Models Sîrbu et al. [37] use a model that updates probabilistically to investigate the effects of algorithmic bias on polarization by counting the number of opinion clusters, interpreting a single opinion cluster as consensus. Leskovec et al. [14] simulate social networks and observe group formation over time.

The Degroot models developed in [9] and used in [15] are closest to ours. Rather than examining polarization and opinions, this work is concerned with the network topology conditions under which agents with noisy data about an objective fact converge to an accurate consensus, close to the true state of the world. As already discussed the basic DeGroot models do not include confirmation bias, however [7, 17, 23, 25, 36] all generalize DeGroot-like models to include functions that can be thought of as modelling confirmation bias in different ways, but with either no measure of polarization or a simpler measure than the one we use. [24] discusses DeGroot models where the influences change over time, and [16] presents results about generalizations of these models, concerned more with consensus than with polarization.

Logic-based approaches Liu et al. [21] use ideas from doxastic and dynamic epistemic logics to qualitatively model influence and belief change in social networks. Seligman et al. [34, 35] introduce a basic "Facebook logic." This logic is non-quantitative, but its interesting point is that an agent's possible worlds are different social networks. This is a promising approach to formal modeling of epistemic issues in social networks. Christoff [8] extends facebook logic and develops several non-quantitative logics for social networks, concerned with problems related to polarization, such as information cascades. Young Pederson et al. [28–30] develop a logic of polarization, in terms of positive and negeative links between agents, rather than in terms of their quantitative beliefs. Hunter [19] introduces a logic of belief updates over social networks where closer agents in the social network are more trusted and thus more influential. While beliefs in this logic are non-quantitative, there is a quantitative notion of influence between users.

Other related work The seminal paper Huberman et al. [18] is about determining which friends or followers in a user's network have the most influence on the user. Although this paper does not quantify influence between users, it does address an important question to our project. Similarly, [10] focuses on finding most influential agents. The work on highly influential agents is relevant to our finding that such agents can maintain a network's polarization over time.

Bibliography

[1] Alvim, M.S., Amorim, B., Knight, S., Quintero, S., Valencia, F.: (2020), https://github.com/Sirquini/Polarization

[2] Alvim, M.S., Amorim, B., Knight, S., Quintero, S., Valencia, F.: A multi-agent model for polarization under confirmation bias in social networks (Technical Report). arXiv preprint (2021)

[3] Alvim, M.S., Knight, S., Valencia, F.: Toward a formal model for group polarization in social networks. In: The Art of Modelling Computational Systems. Lecture Notes in Computer Science, vol. 11760, pp. 419–441. Springer (2019)

[4] Aronson, E., Wilson, T., Akert, R.: Social Psychology. Upper Saddle River, NJ : Prentice Hall, 7 edn. (2010)

[5] Bozdag, E.: Bias in algorithmic filtering and personalization. Ethics and Information Technology (09 2013)

[6] Calais Guerra, P., Meira Jr, W., Cardie, C., Kleinberg, R.: A measure of polarization on social media networks based on community boundaries. Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013 pp. 215–224 (01 2013)

[7] Cerreia-Vioglio, S., Corrao, R., Lanzani, G., et al.: Robust Opinion Aggregation and its Dynamics. IGIER, Università Bocconi (2020)

[8] Christoff, Z., et al.: Dynamic logics of networks: information flow and the spread of opinion. Ph.D. thesis, PhD Thesis, Institute for Logic, Language and Computation, University of Amsterdam (2016)

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

[10] DeMarzo, P.M., Vayanos, D., Zwiebel, J.: Persuasion bias, social influence, and unidimensional opinions. The Quarterly journal of economics **118**(3), 909–968 (2003)

[11] Diestel, R.: Graph Theory. Springer-Verlag, fifth ed edn. (2015)

[12] Elder, A.: The interpersonal is political: unfriending to promote civic discourse on social media. Ethics and Information Technology pp. 1–10 (2019)

[13] Esteban, J.M., Ray, D.: On the measurement of polarization. Econometrica **62**(4), 819–851 (1994)

[14] Gargiulo, F., Gandica, Y.: The role of homophily in the emergence of opinion controversies. arXiv preprint arXiv:1612.05483 (2016)

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

[16] Golub, B., Sadler, E.: Learning in social networks. Available at SSRN 2919146 (2017)

[17] Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence, models, analysis and simulation. Journal of Artificial Societies and Social Simulation **5**(3), 2 (2002)

[18] Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. arXiv preprint arXiv:0812.1045 (2008)

[19] Hunter, A.: Reasoning about trust and belief change on a social network: A formal approach. In: International Conference on Information Security Practice and Experience. pp. 783–801. Springer (2017)
[20] Li, L., Scaglione, A., Swami, A., Zhao, Q.: Consensus, polarization and clustering of opinions in social networks. IEEE Journal on Selected Areas in Communications **31**(6), 1072–1083 (2013)
[21] Liu, F., Seligman, J., Girard, P.: Logical dynamics of belief change in the community. Synthese **191**(11), 2403–2431 (Jul 2014)
[22] Lynch, N.A.: Distributed Algorithms. Morgan Kaufmann Publishers (1996)
[23] Mao, Y., Bolouki, S., Akyol, E.: Spread of information with confirmation bias in cyber-social networks. IEEE Transactions on Network Science and Engineering **7**(2), 688–700 (2020)
[24] Moreau, L.: Stability of multiagent systems with time-dependent communication links. IEEE Transactions on Automatic Control **50**(2), 169–182 (2005)
[25] Mueller-Frank, M.: Reaching Consensus in Social Networks. IESE Research Papers D/1116, IESE Business School (Feb 2015)
[26] Myers, D.G., Lamm, H.: The group polarization phenomenon. Psychological Bulletin (1976)
[27] Nielsen, M., Palamidessi, C., Valencia, F.D.: Temporal concurrent constraint programming: Denotation, logic and applications. Nord. J. Comput. **9**(1), 145–188 (2002)
[28] Pedersen, M.Y.: Polarization and echo chambers: A logical analysis of balance and triadic closure in social networks
[29] Pedersen, M.Y., Smets, S., Ågotnes, T.: Analyzing echo chambers: A logic of strong and weak ties. In: Blackburn, P., Lorini, E., Guo, M. (eds.) Logic, Rationality, and Interaction. pp. 183–198. Springer, Berlin, Heidelberg (2019)
[30] Pedersen, M.Y., Smets, S., Ågotnes, T.: Further steps towards a logic of polarization in social networks. In: Dastani, M., Dong, H., van der Torre, L. (eds.) Logic and Argumentation. pp. 324–345. Springer International Publishing, Cham (2020)
[31] Proskurnikov, A.V., Matveev, A.S., Cao, M.: Opinion dynamics in social networks with hostile camps: Consensus vs. polarization. IEEE Transactions on Automatic Control **61**(6), 1524–1536 (June 2016)
[32] Ramos, V.J.: Analyzing the Role of Cognitive Biases in the Decision-Making Process. IGI Global (2019)
[33] Saraswat, V.A., Jagadeesan, R., Gupta, V.: Foundations of timed concurrent constraint programming. In: LICS. pp. 71–80. IEEE Computer Society (1994)
[34] Seligman, J., Liu, F., Girard, P.: Logic in the community. In: Indian Conference on Logic and Its Applications. pp. 178–188. Springer (2011)
[35] Seligman, J., Liu, F., Girard, P.: Facebook and the epistemic logic of friendship. CoRR **abs/1310.6440** (2013)
[36] Sikder, O., Smith, R., Vivo, P., Livan, G.: A minimalistic model of bias, polarization and misinformation in social networks. Scientific Reports **10** (03 2020)
[37] Sîrbu, A., Pedreschi, D., Giannotti, F., Kertész, J.: Algorithmic bias amplifies opinion polarization: A bounded confidence model. arXiv preprint arXiv:1803.02111 (2018)
[38] Sohrab, H.H.: Basic Real Analysis. Birkhauser Basel, 2nd ed edn. (2014)