NBER WORKING PAPER SERIES

SOCIAL MEDIA NETWORKS, FAKE NEWS, AND POLARIZATION

Marina Azzimonti
Marcos Fernandes

Working Paper 24462
http://www.nber.org/papers/w24462

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2018

We would like to thank the participants of the 2017 NBER Summer Institute on Political Economy (Boston), 22nd Coalition Theory NetworkWorkshop (Glasgow), 27th International Conference on Game Theory (Stony Brook), 43rd Eastern Economic Association Conference (New York) and seminar series at London Business School/Government (London), Warwick, Harris Public School (Chicago). In particular we are grateful for the valuable comments from Jesse Shapiro, Daron Acemoglu, Ernesto Dal Bo, Matthew Jackson, Yair Tauman and Helios Herrera. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Social Media Networks, Fake News, and Polarization
Marina Azzimonti and Marcos Fernandes
NBER Working Paper No. 24462
March 2018
JEL No. C45,C63,D72,D8,D83,D85,D91

ABSTRACT

We study how the structure of social media networks and the presence of fake news might affect the degree of misinformation and polarization in a society. For that, we analyze a dynamic model of opinion exchange in which individuals have imperfect information about the true state of the world and are partially bounded rational. Key to the analysis is the presence of internet bots: agents in the network that do not follow other agents and are seeded with a constant flow of biased information. We characterize how the flow of opinions evolves over time and evaluate the determinants of long-run disagreement among individuals in the network. To that end, we create a large set of heterogeneous random graphs and simulate a long information exchange process to quantify how the bots' ability to spread fake news and the number and degree of centrality of agents susceptible to them affect misinformation and polarization in the long-run.

Marina Azzimonti
Economics Department
Stony Brook University
100 Nicolls Road
Stony Brook, NY 11794
and NBER
marina.azzimonti@gmail.com

Marcos Fernandes
Economics Department
Stony Brook University
100 Nicolls Road
Stony Brook, NY 11794
marcos.fernandes@stonybrook.edu

1 Introduction

In the last few years, the United States has become more polarized than ever. A recent survey conducted by The Pew Research Center indicates that Republicans and Democrats are further apart ideologically than at any point since 1994 (see Figure (1)).

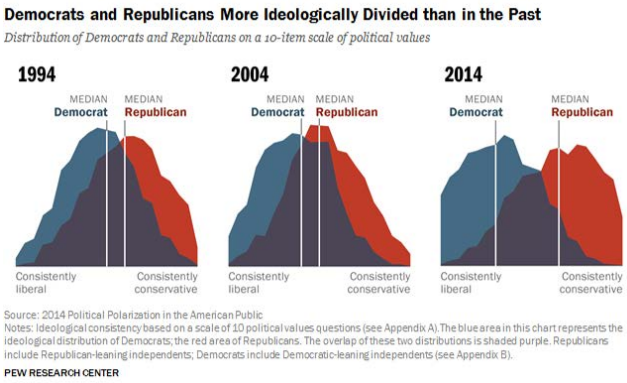


Figure 1: Political Polarization in the American Public (2014, Pew Research Center)

Traditional theories in economics and political science typically model disagreement as arising from one of two sources: (i) differences in preferences and (ii) informational frictions. In the first case, agents may disagree on the optimal level of a given policy because they benefit differently from it. This happens when their income or wealth levels are different (such as in the case of redistributive policies) or when they have different preferences over public goods (e.g. defense vs education or health-care, etc.). In the case of informational frictions, there may exist an optimal action, but society may not know exactly what it is. Examples are the need for environmental policy, mandatory vaccination, restrictions on certain groups of immigrants, unconventional monetary policy, or simply choosing one political candidate over another. Individuals may learn about the desirability of the policy by acquiring information, but to the extent that they are exposed to biased sources of information, their beliefs may differ at the time in which decisions must be taken.

There is a large literature trying to explain how slanted news and media bias may affect voters' opinions by generating misinformation and exacerbating polarization (see Della Vigna and Kaplan, 2007 or Martin and Yurukoglu, 2015). While this literature has been mostly focused on traditional media, such as newspapers, radio, and cable TV—broadly covered under the umbrella of 'broadcasting'—recent interest has shifted towards social media. There are several reasons for this shift. First, because individuals are increasingly obtaining information from social media

networks. According to a 2016 study by the Pew Research Center and the John S. and James L. Knight Foundation, 62% of adults get their news from social media (a sharp increase from the 49% observed in 2012).[1] Among these, two-thirds of Facebook users (66%) get news on the site, nearly six-in-ten Twitter users (59%) get news on Twitter, and seven-in-ten Reddit users get news on that platform.

Second, the technology of communication in social media is significantly different. In the world of broadcasting, agents are mostly consumers of information. There is a small number of news outlets that reach a large (and relatively passive) audience. In the world of Web 2.0, or 'social media,' individuals are not only consuming information, but they are also producing it. This technological change is less well understood. A key aspect of social media communication is that one given message can reach a large audience almost immediately. Another important change is that it is much more difficult for individuals to back out the reliability of a piece of information, as they observe a distilled signal from a friend in their network without necessarily knowing its source.

This is relevant when coupled with another phenomena that became prevalent particularly around 2016 presidential election: the massive spread of *fake news* (also referred to as disinformation campaigns, cyber propaganda, cognitive hacking, and information warfare) through the internet. As defined by Gu, Kropotov, and Yarochkin (2016), 'Fake news is the promotion and propagation of news articles via social media. These articles are promoted in such a way that they appear to be spread by other users, as opposed to being paid-for advertising. The news stories distributed are designed to influence or manipulate users' opinions on a certain topic towards certain objectives.' While the concept of propaganda is not new, social media apparently has made the spreading of ideas faster and more scalable, making it potentially easier for propaganda material to reach a wider set of people. Relative to more traditional ways of spreading propaganda, fake news are extremely difficult to detect posing a challenge for social media users, moderators, and governmental agencies trying control their dissemination. A December 2016 Pew Research Center study found that 'about two-in-three U.S. adults (64%) say fabricated news stories cause a great deal of confusion about the basic facts of current issues and events.' Moreover, 23% admit to having shared a made-up news story (knowingly or not) on social media. Understanding how fake news spread and affect opinions in a networked environment is at the core of our work.

With the dispersion of news through social media, and more generally the internet, and given

[1]The distribution of social media users is similar across education levels, race, party affiliation and age. About 22% of 18-29 year olds are social media users, 34% are aged 30-49, 26% are aged 50-64, and 19% 65 and older.

that a growing proportion of individuals, politicians, and media outlets are relying more intensively on this networked environment to get information and to spread their world-views, it is natural to ask whether and to what extent misinformation and polarization might be exacerbated by social media communication.

In this context, we study a dynamic model of opinion formation in which individuals who are connected through a social network have imperfect information about the true state of the world, denoted by θ. For instance, the true state of the world can be interpreted as the relative quality of two candidates competing for office, the optimality of a specific government policy or regulation, the degree of government intervention in specific markets, etc.

Individuals can obtain information about the true state of the world from unbiased sources external to the network, like scientific studies, unbiased news media, reports from non-partisan research centers such as the Congressional Budget Office, etc. This is modeled as an informative and unbiased private signal received by each agent. Due to limited observability of the structure of the network and the probability distribution of signals observed by others, individuals are assumed to be incapable of learning in a fully Bayesian way. Moreover, we assume that individuals are unable to process all the available information and for that they can also rely on the information from their social neighbors (i.e. individuals connected to them through the network) who are potentially exposed to other sources. In this sense, individuals in our network update their beliefs as a convex combination of the Bayesian posterior belief conditioned on their private signals and the opinion of their neighbors, as per the update rule proposed by Jadbabaie, Molavi, Sandroni, and Tahbaz-Salehi (2012) (JMST (2012) henceforth).

There are three types of agents in this society: *Sophisticated agents*, *unsophisticated agents* and *Internet bots*. Their characterization is to some extent interrelated because it depends not only on signals observed, but also on their network connectivities. In terms of signals received, both sophisticated and unsophisticated agents receive informative private signals every period of time. Internet bots, on the other hand, rely only on biased information and produce a stream of fake news. In terms of connectivities, Internet bots do not relay in the information of others (they are sinks in a Markov chain sense), but have a positive mass of followers. More specifically, bots' followers are exclusively composed by unsophisticated agents. These agents are unable to identify the bot as a source of misinformation, implying that they cannot detect and disregard *fake news*, which are incorporated when updating beliefs (therefore the adjective *unsophisticated*). The opinions generated from the exchange of information forms an inhomogeneous Markov process which may never lead to consensus among sophisticated agents since they are exposed

to unsophisticated agents. In such environment, it can be shown that society's beliefs fail to converge. Moreover, under some conditions, the belief profile can fluctuate in an ergodic fashion leading to misinformation and polarization cycles.

The structure of the graph representing the social media network and the degree of influence of unsophisticated agents shape the dynamics of opinion and the degree of misinformation and polarization in the long-run. More specifically, long-run misinformation and polarization depends on three factors: behavioral assumptions (e.g. the updating rule), communication technology (e.g. the speed at which network connections are active and creation of fake news), and the network topology (e.g. the degree of clustering, the share of unsophisticated agents on the population, how central they are, and the ability of bots to flood the network with fake news). Because a theoretical characterization of the relationship between the topology of the network and the degrees of misinformation and polarization is not trivial, we create a large set of random graphs with different behavioral assumptions, communication technologies and topologies. We then quantify how fake news, the degrees of centrality, and influence affect misinformation (e.g. how far agents beliefs are from the true state of the world) and long-run polarization, defined as in Esteban and Ray (1994).

We find that misinformation and polarization have an inverted u-shape relationship. This is to be expected: when individuals are able to effectively aggregate information and learn the true state of the world, polarization vanishes. At the other extreme, there are situations where there is no polarization because most individuals in the network converge to the wrong value of θ. This involves maximal misinformation with no polarization. Finally, there are cases in which individuals are on average correct but distributed symmetrically around the true state of the world, with large mass at the extremes of the belief distribution. Here, there are intermediate levels of misinformation and extreme polarization. Even though this implies somewhat better aggregation of information, it may lead to inefficient gridlock due to inaction. We find that when unsophisticated agents have a large number of followers in social media, misinformation rises but polarization is hardly affected. On the other hand, the clustering coefficient (i.e. a network statistics that says the extent through which friends of friends are also direct friends) is important for polarization (it actually reduces it) but irrelevant for misinformation. When unsophisticated agents are relatively more influential (because they manage to affect the opinions of influential followers), information is more efficiently aggregated. However, to the extent that agents do not fully learn the true state of the world, there is a significant amount of networks in which opinions become extreme. These are networks in which a bot with views at one extreme targets relatively

more influential unsophisticated agents than the bot with opposing views. This makes the bot more more efficient at spreading fake-news, since the speed at which each given piece of fake-news travels through the network rises, pulling opinions towards an extreme. We show that, for specific network topologies, significant levels of misinformation and polarization are possible in network in which as little as 10% of agents believe fake news. This is relevant, because it shows that the network externality effects are quantitatively important. In other words, only a relevant small number of unsophisticated agents is able to generate significant misinformation and polarization in our simulated networks.

Related Literature Our paper is related to a growing number of articles studying social learning with bounded rational agents and the spread of misinformation in networks.

The strand of literature focusing on social learning with bounded rational agents assumes that individuals use simple heuristic rules to update beliefs, like taking repeated averages of observed opinions. Examples are DeGroot (1974), Ellison and Fundenberg (1993, 1995), Bala and Goyal (1998,2001), De Marzo, Vayanos and Zwiebel (2003) and Golub and Jackson (2010). In most of these environments, under standard assumptions about the connectivity of the network and the bounded prominence of groups in growing societies, the dynamics of the system reaches an equilibrium and consensus emerges. In this sense, long-run polarization or misinformation would only arise in such models if those assumptions are relaxed. Common to most of these models is the fact that there is no new flow of information entering into the network. Agents are typically assumed to be bounded rational (naive) and do not observe private signals from external sources (and hence do not use standard Bayesian update rules). JMST (2012) extends these environments to allow for a constant arrival of new information over time in an environment in which agents also learn from their neighbors in a naive way. This feature allows agents to efficiently aggregate information even when some standard assumptions that ensure consensus are relaxed. Our paper uses an update rule based on JMST (2012) for gents, but introduces internet bots which break the connectivity of the network by basing their information exclusively on biased sources and disregard the information provided by others. The latter is a feature that we borrow from the literature on misinformation. More particularly, from the work by Acemoglu, Ozdaglar and ParandehGheibi (2010) (AOP henceforth) and Acemoglu, Como, Fagnani, and Ozdaglar (2013) (ACFO henceforth)

AOP (2010) focuses on understanding the conditions under which agents fail to reach consensus or reach wrong consensus. In their model, agents exchange opinion in a naive way conditional

on being pair-wise matched. Crucial to the emergence of misinformation in is the presence of *forceful* agents whose roles are to exert disproportional influence over regular agents and force them to conform with their opinions. ACFO (2013) consider the same naive learning model with random meetings dictated by a Poisson process, but allow for the existence of *stubborn* agents instead. These agents never update their opinions (they are sinks in a Markov chain sense) but influence other agents. Therefore, the information exchange dynamics never reaches a steady state and opinions fluctuate in a stochastic fashion. Both papers abstract from Bayesian learning. In our paper, we consider simultaneously the possibility that regular agents learn from unbiased sources while being exposed to fake news spread by Internet bots. Our learning rule follows JMST (2012) in the sense that agents learn from private signals in a a fully Bayesian fashion but also incorporate friends' opinions naively. The final belief is basically a convex combination of the Bayesian posterior and friends' posteriors. Moreover, we add the feature that agents meet randomly in the spirit of of AOP (2010) and ACFO (2013). Therefore, the main extensions with respect to JMST (2012) are i) the presence of Internet bots (sinks) seeded with biased information that spread fake news, which becomes the main source of misinformation in the system and ii) the fact that we allow for random meetings (inhomogeneous Markov chain). On the other hand, the main extension relative to ACFO (2013) is that we introduce Bayesian learning features. Our Internet bots can be understood as stubborn agents endowed with the capacity to countervail the flow of informative private signals that reaches regular agents every period of time. We call this feature *flooding capacity* and it basically consists in allowing these bots to spread a larger stream of fake news (signals) as other agents in the network.[2] Hence, our paper contributes to the social learning and spread of misinformation literatures by studying misinformation in an environment with informative signals.

Our main contribution relative to the existing literature, however, is that we simulate a large set of complex social networks and quantify the relative importance of behavioral assumptions, technological characteristics, and network topology on long-run polarization and misinformation. To the best of our knowledge, this is the first paper to quantify the relative importance of network characteristics on long-run misinformation and polarization.

Finally, there is a growing empirical literature analyzing the effects of social media in opinion formation and voting behavior (Halberstam and Knight, 2016). Because individual opinions are unobservable from real network data, these papers typically use indirect measures of ideology to

[2]Our model considers a Bernoulli rather than a Poisson process and restrict attention to a particular class of beliefs (Beta distributions) though.

back-out characteristics of the network structure (such as homophily) potentially biasing their impact. By creating a large number artificial networks, we can directly measure how homophily and other network characteristics affect opinion. Finally, our paper complements the literature on the role of biased media such as Campante and Hojman (2013), Gentzkow and Shapiro (2006, 2010, and 2011), and Flaxman et al. (2013) and the effects of social media on political polarization, such as Boxell et al (2017), Barbera (2016), and Weber at al (2013).

Basic notation: The notation and terminology introduced here is mostly employed in the appendix of this work but also serves as an important guide for the next sections. All vectors are viewed as column vectors, unless stated to the contrarily. Given a vector $v \in \mathbb{R}^n$, we denote by v_i its i-th entry. When $v_i \geq 0$ for all entries, we write $v \geq 0$. To avoid potential burden of notation, the summation $\sum v$ without index represents the sum of all entries of vector v. Moreover, we define $v^\top$ as the transpose of the vector v and for that, the inner-product of two vectors $x, y \in \mathbb{R}^n$ is denoted by $x^\top y$. Similarly for the product of conforming vectors and matrices. We denote by $\mathbf{1}$ the vector with all entries equal to 1. Finally, a vector v is said to be a stochastic vector when $v \geq 0$ and $\sum_i v_i = 1$. In terms of matrices, a square matrix M is said to be row stochastic when each row of M is a stochastic vector. A matrix M is said to be a square matrix of size n when the number of rows and columns is n. The identity matrix of size n is denoted by $\mathbb{I}_n$. For any matrix M, we write M_{ij} or $[M]_{ij}$ to denote the matrix entry in the i-th row and j-th column. The symbols M_{i*} and $[M]_{i*}$ are used to denote the i-th row of matrix M, while M_{*j} and $[M]_{*j}$ denote the j-th column of the matrix. Finally, the transpose of a matrix M is denoted by $M^\top$ and represents a new matrix whose rows are the columns of the original matrix M, i.e. $\left[M^\top\right]_{ij} = [M]_{ji}$.

2 Baseline Model

Agents, social bots and information structure The economy is composed by a finite number of agents $i \in N = \{1, 2, \ldots, n\}$ who interact in a social network over time for a large number T of periods (which need not be finite), date at which a one-dimensional policy needs to be determined. Individuals have imperfect information about $\theta \in \Theta = [0, 1]$, the optimal value of the policy. This parameter can be interpreted as the degree of government intervention in private markets (e.g. environmental control, enforcement of property rights, restrictions on the use of public land, gun control, etc.), as optimal fiscal or monetary policy (e.g. the inflation rate, tax rates on capital or labor income, tariffs, etc.), or as the best response to an unexpected shock (e.g. the

size of a bailout during a financial crisis, the response to a national security threat, the amount of aid given to a region that suffered a natural disaster, etc.). In period 0 individuals have a prior about θ and update their beliefs from private signals obtained up to period T, where the policy needs to be implemented. [3]

Agents obtain information from: (i) an unbiased source, (ii) other agents connected to them in a social network, and (iii) a *bot* spreading fake news. There are two types of bots, $L-bot$ and $R-bot$ with opposing agendas. Their objective is to manipulate opinions by sending extremely biased signals (e.g. close to 0 or 1). We assume that a majority of the population is *sophisticated*, meaning that they can identify bots and disregard fake news in their update process. There is small proportion μ_u of individuals, on the other hand, that can be influenced by fake news. We refer to them as *unsophisticated*. A key assumption is that individuals cannot back out the sources of information of other agents. As a result, sophisticated agents may be influenced by fake news indirectly through their social media friends.

To the extent that policy is chosen democratically (via direct voting or through representatives), the implemented policy may differ from the optimal one when bots are present. If a large number of voters have homogeneous beliefs but are *misinformed* (that is, have beliefs far away from θ), implemented policies will be inefficient. If, in addition, voters are *polarized*, then suboptimal delays in the response to shocks may arise. These result from gridlock or stalemate among policymakers representing individuals with opposing views. The welfare losses arising from informational frictions can be captured by a social welfare function $S(MI, P)$, which is decreasing in the aggregate degree of misinformation MI and the societal level of polarization in beliefs, denoted by P. Both MI and P depend critically on the distribution of voters' opinions. We will first describe how opinions evolve over time and then define how statistics obtained from this distribution can be used to compute misinformation and polarization, and hence quantify welfare losses associated to them.

Each agent starts with a prior belief $\theta_{i,0}$ assumed to follow a Beta distribution,

$$\theta_{i,0} \sim \mathcal{B}e\Big(\alpha_{i,0}, \beta_{i,0}\Big).$$

This distribution or *world-view* is characterized by initial parameters $\alpha_{i,0} > 0$ and $\beta_{i,0} > 0$. Note that individuals agree upon the parameter space Θ and the functional form of the probability distribution, but have different world-views as they disagree on $\alpha_{i,0}$ and $\beta_{i,0}$. Given prior beliefs,

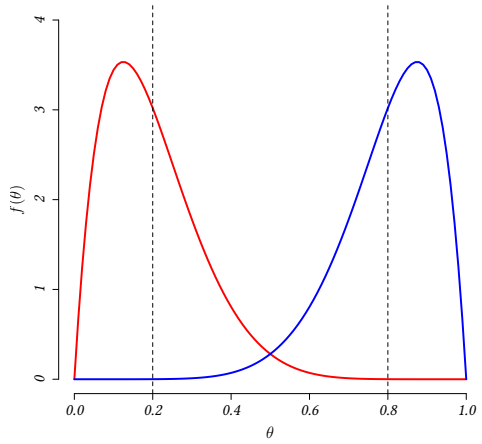
[3] In most of the analysis we will focus on the limiting case $T \to \infty$ to allow for convergence.

we define the initial *opinion* of agent i $y_{i,0}$ about the true state of the world as her best guess of θ given the available information,[4]

$$y_{i,0} = \mathbb{E}[\theta|\Sigma_0] = \frac{\alpha_{i,0}}{\alpha_{i,0} + \beta_{i,0}}$$

where $\Sigma_0 = \{\alpha_{i,0}, \beta_{i,0}\}$ denotes the information set available at time 0.

Example 1. *In the Figure below, we depict the world-views of two individuals (distributions) and their associated opinions (vertical lines). The world-view that is skewed to the right is represented by the distribution $\mathcal{B}e(\alpha = 2, \beta = 8)$. The one skewed to the left is represented by the distribution $\mathcal{B}e(\alpha = 8, \beta = 2)$. The opinions are, respectively, 0.2 and 0.8.*



We formalize the information obtained from unbiased sources as a draw $s_{i,t}$ from a Bernoulli distribution centered around the true state of the world θ,

$$s_{i,t} \sim \mathcal{B}ernoulli(\theta).$$

Through this channel, a majority of the population may learn θ in the limit. However, agents update their world-views and opinions based not only on $s_{i,t}$, but also through the influence of individuals connected to them in a social network, which may introduce misinformation. Social media thus generates an externality on the information aggregation process. To the extent that the social media externality is important, the true state of the world may not be uncovered by enough individuals and inefficient policies may be enacted or gridlock may arise. The network structure, and in particular the location of unsophisticated agents in it, will be important to de-

[4]Note that $\mathbb{E}[\theta|\Sigma_0]$ is the Bayesian estimator of θ that minimizes the mean squared error given a Beta distribution.

termine the quality of information and the degree of polarization in society. We formalize the social network structure next.

Social Network The connectivity among agents in the network at each point in time t is described by a directed graph $G_t = (N, g_t)$, where g_t is a real-valued $n \times n$ adjacency matrix. Each regular element $[g_t]_{ij}$ in the directed-graph represents the connection between agents i and j at time t. More precisely, $[g_t]_{ij} = 1$ if i is paying attention to j (i.e. receiving information from) at time t, and 0 otherwise. Since the graph is directed, it is possible that some agents pay attention to others who are not necessarily paying attention to them, i.e. $[g_t]_{ij} \neq [g_t]_{ji}$. The out-neighborhood of any agent i at any time t represents the set of agents that i is receiving information from, and is denoted by $N_{i,t}^{out} = \{j \mid [g_t]_{ij} = 1\}$. Similarly, the in-neighborhood of any agent i at any time t, denoted by $N_{i,t}^{in} = \{j \mid [g_t]_{ji} = 1\}$, represents the set of agents that are receiving information from i (e.g. i's audience or followers). We define a directed path in G_t from agent i to agent j as a sequence of agents starting with i and ending with j such that each agent is a neighbour of the next agent in the sequence. We say that a social network is *strongly connected* if there exists a directed path from each agent to any other agent.

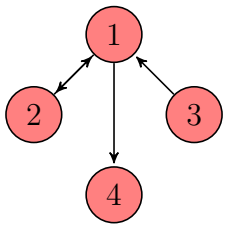
In the spirit of AOP (2010) and ACFO (2012), we allow the connectivity of the network to change stochastically over time. This structure captures rational inattention, incapacity of processing all information, or impossibility to pay attention to all individuals in the agent's social clique. More specifically, for all $t \geq 1$, we associate a *clock* to every directed link of the form (i,j) in the initial adjacency matrix g_0 to determine whether the link is activated or not at time t. The ticking of all clocks at any time is dictated by i.i.d. samples from a Bernoulli Distribution with fixed and common parameter $\rho \in (0, 1]$, meaning that if the (i,j)-clock ticks at time t (realization 1 in the Bernoulli draw), then agent i receives information from agent j. Hence, the parameter ρ measures the *speed of communication* in the network. The Bernoulli draws are represented by the $n \times n$ matrix c_t, with regular element $[c_t]_{ij} \in \{0, 1\}$. Thus, the adjacency matrix of the network evolves stochastically across time according to the equation

$$g_t = g_0 \circ c_t, \tag{1}$$

where the initial structure of the network, represented by the initial adjacency matrix g_0, remains unchanged.[5]

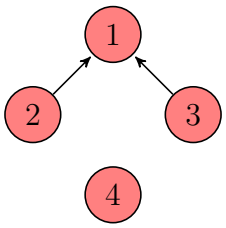
[5]The notation $\circ$ denotes the Hadamard Product, or equivalently, the element-wise multiplication of the matrices.

Example 2 (Bernoulli Clock)**.** *Panel* (2a) *represents the original network and its adjacency matrix, whereas Panel* (2b) *depicts a realization such that agent 1 does not pay attention to agents 2 and 4 in period* 1*. Agents 2 and 3, on the other hand, pay attention to agent 1 in both periods.*



$$g_0 = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(a) Original Network at $t = 0$



$$g_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(b) Potential Network at $t = 1$

Figure 2: Bernoulli Clock and Network Dynamics

Evolution of Beliefs Before the beginning of each period, both sophisticated and unsophisticated agents receive information from individuals in their out-neighbourhood, a set determined by the realization of the clock in period t and the initial network. All agents share their opinions and precisions, summarized by the shape parameters $\alpha_{i,t}$ and $\beta_{i,t}$. This representation aims at capturing communication exchanges through social media feeds. At the beginning of every period t, a signal profile is realized and an unbiased signal is privately observed by every regular agent, whereas bots observe a biased signal (see details below). In this way, part of the information obtained by unsophisticated agents from biased sources will be transmitted through the network in the following period, as it is incorporated during the belief updating process and shared with other agents in the network that naively internalize them.

We now explain the update rule of regular agents (sophisticated and unsophisticated) and bots. The full characterization of the update rules can be found in Appendix A.

Internet bots (Fake News source)

We assume that there are two types of Internet bots, a left wing bot (or L-bot) and right wing bot (or R-bot), both with extreme views. Internet bot i produces a stream of fake news $\kappa s^p_{i,t}$, for $p \in \{L, R\}$, where $s^L_{i,t} = 0$ for type L-bot and $s^R_{i,t} = 1$ for type R-bot for every t. The parameter

$\kappa \in \mathbb{N}^+$ measures the ability of bots to spread more than one fake-news article per period, which can be interpreted as their *flooding capacity* (i.e. how fast they can produce fake news compared to the regular flow of informative signals received by agents). Bots transmit the whole stream of information to agents paying attention to them. Hence, a value of $\kappa > 1$ gives them more de-facto weight in the updating rule of unsophisticated agents, emphasizing their degree of influence on the network. We can model the bot update as

$$\alpha^p_{i,t+1} = \alpha^p_{i,t} + \kappa s^p_{i,t} \tag{2}$$

$$\beta^p_{i,t+1} = \beta^p_{i,t} + \kappa - \kappa s^p_{i,t}. \tag{3}$$

Regular agents: sophisticated and unsophisticated

Sophisticated and unsophisticated agents share the same update rule. The only thing that distinguishes these agents is the composition of their neighborhood: while sophisticated agents only pay attention to regular agents, unsophisticated agents devote some share of their attention to bots and for that will be exposed to fake news. After observing the signal from unbiased sources, agents compute their Bayesian posteriors conditional on the observed signals. We assume that parameters $\alpha_{i,t+1}$ and $\beta_{i,t+1}$ are convex combinations between their Bayesian posterior parameters and the weighted average of the neighbors' parameters. In mathematical terms we have that

$$\alpha_{i,t+1} = (1 - \omega_{i,t})[\alpha_{i,t} + s_{i,t+1}] + \omega_{i,t} \sum_j [\hat{g}_t]_{ij} \, \alpha_{j,t} \tag{4}$$

$$\beta_{i,t+1} = (1 - \omega_{i,t})[\beta_{i,t} + 1 - s_{i,t+1}] + \omega_{i,t} \sum_j [\hat{g}]_{ij} \, \beta_{j,t}, \tag{5}$$

where $\omega_{i,t} = \omega$ when $\sum_j [g_t]_{ij} > 0$, and $\omega_{i,t} = 0$ otherwise.

Note that this rule assumes that agents exchange information (i.e. $\alpha_{j,t}$ and $\beta_{j,t}$) before processing new signals $s_{i,t+1}$.

A regular agent's full attention span is split between processing information from unbiased sources, $(1-\omega_{i,t})$, and that provided by their friends in the network, $\omega_{i,t}$ (e.g. reading a Facebook or Twitter feed). If no friends are found in the neighborhood of agent i, $\sum_j [\hat{g}_t]_{ij} = 0$, then the agent attaches weight 1 to the unbiased signal, behaving like a standard Bayesian agent. Conversely, if at least one friend is found, this agent uses a common weight $\omega \in (0, 1)$. The term $[\hat{g}_t]_{ij} = \frac{[g_t]_{ij}}{|N^{out}_{i,t}|}$ represents the weight given to the information received from her out-neighbor

j. As $\omega_{i,t}$ approaches 1, the agent only incorporates information from social media, making her update process closer to a DeGrootian in which individuals are purely conformists. In general, ω can be interpreted as the *degree of influence* of social media friends.

Finally, note we are assuming that the posterior distribution determining world-views of agents will also be a Beta distribution with parameters $\alpha_{i,t+1}$ and $\beta_{i,t+1}$. Hence, an agent's opinion regarding the true state of the world at t can be computed as

$$y_{i,t} = \frac{\alpha_{i,t}}{\alpha_{i,t} + \beta_{i,t}}.$$

Our heuristic rules resembles the one in JMST (2012), but there are three important distinctions. First, their adjacency matrix is fixed over time (homogeneous Markov chain), whereas ours is stochastic (inhomogeneous Markov chain), an element we borrowed mainly from ACFO (2013). Second, we restrict attention to a specific conjugated family (Beta-Bernoulli) and assume that individuals exchange shape parameters $\alpha_{i,t}$ and $\beta_{i,t}$ that characterize this distribution. So the heuristic rule involves updating two real valued parameters, whereas JMST (2012)'s heuristic rule involves a convex combination of the whole distribution function. Given their rule, the posterior distribution may not belong to the same family as the prior distribution, as the convex combination of two Beta distributions is not a Beta distribution. That is not the case in our environment, as the posterior will also belong to the Beta distribution family. Finally, we are considering the influence of fake news spread by bots and this feature is the main source of misinformation. Therefore, to the extent that bots reach unsophisticated agents who are influential, their presence will affect the existence and persistence of misinformation and polarization over time. This is due to the fact that they will consistently communicate fake news (biased signals) to some unsophisticated agents pushing them to extremes of the belief spectrum.[6]

3 Misinformation, Polarization, and Network Structure

An agent is misinformed when her beliefs are not close enough to the true state of the world θ. We can define the degree of 'misinformation' in society as the average distance between opinions

[6]We believe, even though we have not proved it, that the choice of modeling bots as agents in the network instead of simply biased signals reaching a subset of agents comes without any costs to our findings. Moreover, the decision of modeling bots as agents is similar to the idea of fanatics or stubborn agents in the spread of misinformation literature. Thus, the potential benefit of modeling in this way is the possibility of making direct comparisons to the current results in the literature. Finally, as pointed out by Gu, Kropotov, and Yarochkin (2016), fake news articles sometimes are promoted in such a way that they appear to be spread by other users. In this sense, modeling bots as agents seems to be a fair natural starting point. We get back to the resulting technical challenges later.

and the true state of the world θ.

Definition 1 (**Misinformation**). *The degree of misinformation is given by*

$$MI_t = \frac{1}{n}\sum_{i=1}^{n}\Big(y_{i,t} - \theta\Big)^2. \tag{6}$$

Given an arbitrarily small value $\epsilon > 0$, the proportion of misinformed agents in the population is given by

$$\#MI_t = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{|y_{i,t}-\theta|\geq\epsilon\}}. \tag{7}$$

The degree of misinformation grows when a large number of agents are far from the true state of the world. It is important to note that this is not measuring polarization. Moreover, it is not even capturing the variance of opinions, as the average opinion in a given society may be different from θ. For example, consider a network in which all agents believe that $y_{i,t} = 1$ in period t. Then the variance of opinions is null (there is no polarization either), yet the degree of misinformation MI_t will be large (at its maximum theoretical value). This variable measures the 'intensive margin' of misinformation (e.g. how far society is from the truth). The 'extensive margin' is represented by $\#MI_t$ which considers the percentage of individuals who are misinformed. In eq. (7), $\mathbb{1}_{\{\cdot\}}$ represents an indicator function that returns 1 whenever the condition within braces is met and 0 otherwise. These two definitions are related, as $\#MI_t = 0$ implies no misinformation.

We define a 'wise society' as one where there is no misinformation in the limit. Equivalently, when the maximum distance between the limiting opinion of agents and θ is arbitrarily small, as stated below.

Definition 2 (Wise society). *We say that a society is wise, for a fixed $\epsilon > 0$, if*

$$\lim_{t\to\infty} P\left(\max_{i\leq n}|y_{i,t} - \theta| \geq \epsilon\right) = 0.$$

We base our notion of polarization on the seminal work by Esteban and Ray (1994), adapted to the context of this environment. At each point in time, we partition the $[0, 1]$ interval into $K \leq n$ segments. Each segment represents significantly-sized groups of individuals with similar opinions. We let the share of agents in each group $k \in \{1, ..., K\}$ be denoted by $\pi_{k,t}$, with $\sum_k \pi_{k,t} = 1$.

Esteban and Ray (1994)'s polarization measure aggregates both 'identification' and 'alienation' across agents in the network. Identification between agents captures a sense of ideological

alignment: an individual feels a greater sense of identification if a large number of agents in society shares his or her opinion about the true state of the world. In this sense, identification of a citizen at any point in time is an increasing function of the share of individuals with a similar opinion. The concept of identification captures the fact that *intra*-group opinion homogeneity accentuates polarization. On the other hand, an individual feels alienated from other citizens if their opinions diverge. The concept of alienation captures the fact that *inter*-group opinion heterogeneity $|\tilde{y}_{k,t} - \tilde{y}_{l,t}|$ amplifies polarization. Mathematically, we have the following representation.

Definition 3 (Polarization). *Polarization P_t aggregates the degrees of 'identification' and 'alienation' across groups at each point in time.*

$$P_t = \sum_{k=1}^{K} \sum_{l=1}^{K} \pi_{k,t}^{1+\varsigma} \, \pi_{l,t} \, |\tilde{y}_{k,t} - \tilde{y}_{l,t}| \tag{8}$$

where $\varsigma \in [0, 1.6]$ and $\tilde{y}_{k,t}$ is the average opinion of agents in group k and $\pi_{k,t}$ is the share of agents in group k at time t.

Clearly, a society with no polarization may be very misinformed, as described above. On the other hand, we may observe a society in which there is a high degree of polarization but where opinions are centered around θ, so their degree of misinformation may be relatively small. In the latter case, individuals may be deadlocked on a policy choice despite relatively small differences in opinion. In terms of welfare, both variables capture different dimensions of inefficiency. Because of that, we are interested in characterizing both, misinformation and polarization in the limit,

$$\overline{MI} = \operatorname*{plim}_{t \to T} MI_t \quad \text{and} \quad \bar{P} = \operatorname*{plim}_{t \to T} P_t.$$

We can think of long-run misinformation and polarization as functions of: (i) the updating process (clock speed ρ and influence of friends ω), (ii) the initial network structure g_0, and (iii) the degree of influence of bots (flooding parameter κ, share μ_u and location of unsophisticated agents on the network). More formally,

$$\overline{MI} = \mathbf{MI}(\rho, \omega, \mu_u, \kappa, g_0) \quad \text{and} \quad \bar{P} = \mathbf{P}(\rho, \omega, \mu_u, \kappa, g_0).$$

We aim at characterizing the properties of the functions $\mathbf{P}$ and $\mathbf{MI}$. We will first show theoretical results for the limiting case $T \to \infty$ and then those obtained via computer simulations.

Non-influential Bots The following two results show conditions under which misinformation and polarization vanish in the limit. The first one is analogous to Sandroni et al (2012), whereas the second one extends it to a network with dynamic link formation as in Acemoglu et al (2010).

Proposition 1. *If the network G_0 is strongly connected, the directed links are activated every period (e.g., $\rho = 1$) and bots exert no influence, then the society is wise (i.e., all agents eventually learn the true θ). As a consequence, both polarization and misinformation converge in probability to zero.*

When the network is strongly connected all opinions and signals eventually travel through the network allowing agents to perfectly aggregate information. Since bots exert no influence (either because there are no bots or because all agents are sophisticated), individuals share their private signals who are jointly informative and eventually reach consensus (e.g. there is no polarization) uncovering the true state of the world, θ.

The result in Proposition (1) is in line with the findings in JMST (2012) despite the difference in heuristic rules being used. Proposition (2) shows that the assumption of a fixed listening matrix can be relaxed. In other words, even when G_t is not constant, the society is wise and polarization vanishes in the long run in strongly connected networks.

Proposition 2. *If the network G_0 is strongly connected, bots exert no influence, then even when the edges are not activated every period (i.e. $\rho \in (0, 1)$) society is wise. As a consequence, both polarization and misinformation converge in probability to zero.*

Influential Bots Influential bots cause misinformation by spreading fake news. This does not, however, necessarily imply that the society will exhibit polarization. The following example depicts two networks with three agents each: a sophisticated one (node 3) and two unsophisticated ones (nodes 1 and 2) who are influenced by only one bot—L-bot in panel (3a) and R-bot in the panel (3b)—.

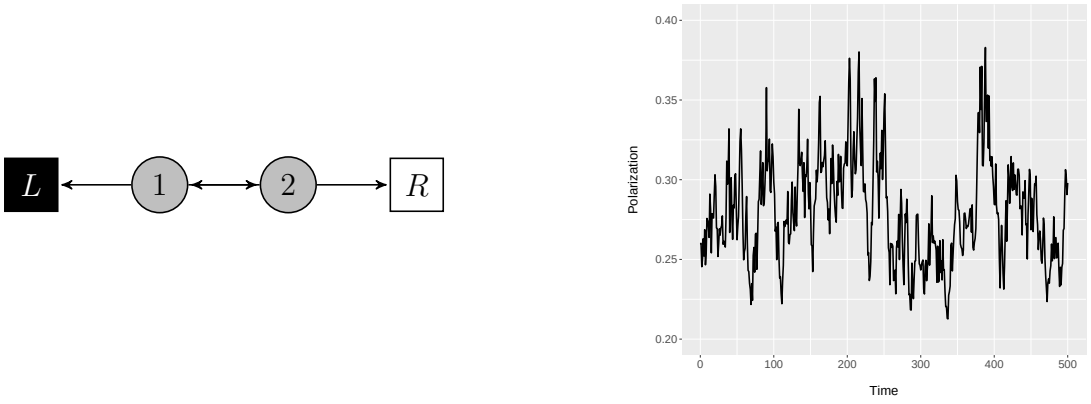
(a) Society influenced by L-bot (b) Society influenced by R-bot

Figure 3: Two societies with internet bots

Polarization in both societies converges to zero in the long-run. However, neither society is wise. This illustrates that the influence of bots may generate misinformation in the long run, preventing agents from uncovering θ, but does not necessarily create polarization. This insight is formalized in Proposition (3).

Proposition 3. *If a society is wise, then it experiences no social polarization in the long run. The converse is not true.*

More generally, when the relative influence of one type of bot is significantly larger than the other, it is possible for a society to reach consensus (i.e. experience no polarization of opinions) to a value of θ that is incorrect. This can happen when there is a sufficient amount of unsophisticated agents or when these unsophisticated agents, even if few, reach a large part of the network (i.e. when they are themselves influential). It is also necessary that one of the bots can reach a larger number of unsophisticated agents than the other; the example presented in Figure (3) is extreme in that one bot is influential whereas the other one is not. But, as we will see in Section (5), a society may converge to the wrong θ under less extreme assumptions. In order for a society to be polarized, individuals need to be sufficiently exposed to bots with opposing views.



(a) Society with both L-bot and R-bot (b) Cycles

Figure 4: Two societies with internet bots

Consider the social network of two agents depicted in Figure (4a), in which both bot-types are present: agent 1 is influenced by the L-bot whereas agent 2 is influenced by the R-bot. Even though unsophisticated agents 1 and 2 receive unbiased signals and communicate with each other, this society exhibits polarization in the long run. This happens because bots subject to different biases are influential. The degree of misinformation may be lower than in the previous example

(as opinions end up being averaged out and potentially closer to the true state of the world), but to the extent policy is chosen by majority voting may still lead to inaction and hence inefficiencies.

A noticeable characteristic of the evolution of P_t over time, depicted in Panel (4b) of Figure (4), is that rather than settling at a constant positive value, it fluctuates in the interval [0.2,0.4]. The example illustrates that polarization cycles are possible in this environment.

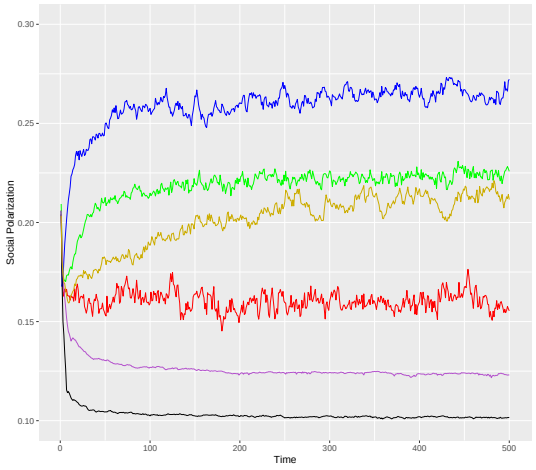


Figure 5: Different polarization levels

Finally, we want to point out that whether misinformation and polarization increase, decrease, or fluctuate over time depend importantly on the topology of the network, the number and degree of influence of bots, the frequency of meetings between individuals (e.g. the clock) and the degree of rationality of agents. Figure (5) depicts the behavior of P_t over time for a series of larger random networks (e.g. there are 100 nodes, an arbitrary number of bots, and different rationality levels). The next section is devoted to uncovering what drives these different dynamics.

4 Numerical Simulation

One of the biggest challenges when using network analysis is to ascertain analytical closed forms and tractability for our objects of interest, namely $\mathbf{MI}$ and $\mathbf{P}$. The combinatorial nature of social networks that exhibit a high degree of heterogeneity makes them very complex objects, imposing a natural challenge for theoretical analysis. In our work, limiting properties can be characterized only when we assume strong connectivity and absence of internet bots' influence. As we drop these assumptions, we observe that different networks might experience different limiting misinformation and polarization levels, even if departing from the very same belief distribution.

To understand such differences, we resort to numerical methods where a large number M of random networks is generated and limiting properties of the distribution of beliefs, namely long run average misinformation and polarization, are computed through simulations. In Section (4.1) we describe the algorithm used to generate the initial network g_0 from a combination of the Barabasi-Albert and the Erdos-Renyi random graph models. In Section (4.2) we describe the simulation exercise in which key parameters determining the initial network, the location of unsophisticated agents in it, the communication technology, the influence of bots, and update process are drawn to generate a synthetic sample of social media networks. In each network $j \in \{1, ..., M\}$, we simulate the evolution of beliefs and compute the limiting misinformation and polarization measures, $\overline{MI}_j$ and $\overline{P}_j$. In Section (5) we characterize how $\mathbf{MI}$ and $\mathbf{P}$ depend on key statistics of the social network topology and parameters determining the learning process by estimating the conditional expectation of $\overline{MI}_j$ and $\overline{P}_j$ conditional on the parameters.

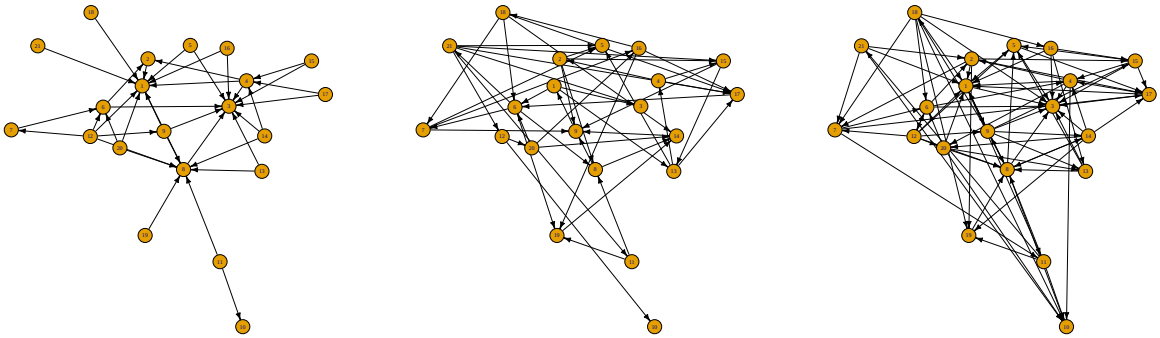
4.1 Network Algorithm

Here we describe the algorithm used to create and populate each initial network $\mathbf{G}_j^0$ in our sample, and define a series of statistics which characterize its topology.

Generating g_0: Social networks have two important characteristics. First, there is reciprocity in the sense that the exchange of information among individuals is bi-directional. Second, some agents are more influential than others (that is, they have a larger in-neighborhood). Barabasi and Albert's (1999) random graph model has the preferential attachment property, generating a few nodes in the network which are very popular relative to others. This is illustrated in Figure (6a) for a network with $n = 21$ nodes. Unfortunately, it exhibits no reciprocity. Their algorithm better captures characteristics of broadcasting, where newspapers, tv, or radio stations (e.g. nodes 1, 3, and 8) send a signal received by a large—and passive—audience. The Erdos-Renyi's random graph model, on the other hand, allows for reciprocity but presumes that all agents have similar degrees of influence, as can be seen in Figure (6b).

Because we want our exercise to incorporate both characteristics, the initial network is constructed as the union of these two random graph models. More specifically, we first create a random graph with n nodes using the Barabasi-Albert algorithm. We then create another one (also with n nodes) following Erdos-Renyi's algorithm. Finally, we combine them into a "BAUER network," which is simply the union of these two graphs. The resulting network is illustrated in

Figure (6c). We can see that information flows in both directions (e.g. there is reciprocity) and that some agents are more influential than others (e.g. there is preferential attachment).



(a) Barabasi-Albert (b) Erdos-Renyi (c) BAUER

Figure 6: Random graph models

There are five key parameters affecting the topology of g_0 in the BAUER model, defined by the set $\Omega = \{m, \alpha, a, p, n\}$. Three of them are associated to the construction of the Barabasi-Albert network, namely, the number of meetings of incoming nodes m, the power of preferential attachment α, and the attractiveness of nodes with no adjacent edges a. The last two define $P[i] = k[i]^{\alpha} + a$, the probability that an old node i is chosen to be linked to a new node at each iteration of the network formation algorithm of Barabasi and Albert. The parameter p represents the probability of drawing an edge between two arbitrary nodes in the Erdos-Renyi algorithm. Finally, n denotes the number of nodes which determines the size of the network. When creating our dataset, we fix n and vary the remaining four parameters.

Populating g_0: We populate each network with two types of agents, sophisticated and unsophisticated, and define which unsophisticated agent is influenced by the L-bot and the R-bot. We do this in two steps. First, using a uniform distribution we randomly select a number $u = \mu_u n$ of agents. This defines the location of unsophisticated agents in g_0. Note that from an ex-ante perspective, every node in the network has the same probability of being populated by an unsophisticated agent. Second, we assign a probability 0.5 to each unsophisticated agent to receive signals from L-bot (exclusively). The remaining unsophisticated agents are assumed to receive signals from the R-bot. This ensures that, on average, bots' messages reach the same number of unsophisticated agents.[7] This does not imply, however, that bots will have the same degree of

[7]We experimented allowing some unsophisticated agents to follow both bots at the same time but it proved to be too complicated compared to the little gain of resulting insights.

influence, as the audience of unsophisticated agents may differ depending on their location in the network. The remaining individuals, $n - u$, are assigned the update rule in eq. ((5)).

4.2 Generating the dataset

We fix the number of agents (or nodes) to $n = 37$ and the true state of the world at $\theta = 0.5$. We also fix the initial distribution of beliefs so that the same mass of the total population lies in the middle point of each one of 7 groups. This rule basically distributes our agents evenly over the political spectrum $[0, 1]$ such that each of the 7 groups contains exactly $\frac{1}{7}$ of the total mass of agents. Moreover, we set the same variance for each agent world-view to be $\sigma^2 = 0.03$. With both opinion and variance, we are able to compute the initial parameter vector (α_0 , β_0). [8] Given these parameters, we draw and populate a large number $M = 8,248$ of initial random networks g_0 following the BAUER model described in Section (4.1). To generate heterogeneity, we consider values for m (number of meetings) in the set $\{1, .., 5\}$, the preferential attachment α in $[0.5, 1.5]$, the attractiveness of nodes with no adjacent edges $a \in \{1, ...4\}$, and the probability $p \in [0.01, 0.1]$.

Characteristics of g_0: Since the process of generating g_0 and populating it with unsophisticated agents involves randomness, a given set of parameters Ω may lead to significantly different network graphs g_0. These graphs can be characterized by a series of standard network statistics, such as diameter, average clustering, and reciprocity. Average values across the M networks are reported in Table (1), which contrasts them to those observed on real-life social media networks such as Twitter, Facebook and Google +.

Table 1: Network Topology

	Simulation	Twitter	Facebook	Google +
Diameter	6.5	7	6	8
Reciprocity	0.04	0.03		
Avg Clustering	0.36	0.57	0.49	0.6

Diameter captures the shortest distance between the two most distant nodes in the network. The average diameter in our network is 6.5, very much in line with the values observed in real-life

[8] In this case, we only need to use the relationships $\mu = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ to fully determine α and β. Algebraic manipulation yields $\alpha = -\frac{\mu(\sigma^2+\mu^2-\mu)}{\sigma^2}$ and $\beta = \frac{(\sigma^2+\mu^2-\mu)(\mu-1)}{\sigma^2}$.

networks. *Reciprocity* is the proportion of all possible pairs (i, j) which are reciprocal (e.g. have edges between them in both directions), provided there is at least one edge between i and j,

$$R(g_0) = \frac{\sum_i \sum_j \left(g_0 \circ g_0^\top\right)_{ij}}{\sum_i \sum_j g_{ij}}$$

Our synthetic sample mean of 0.04 is similar to that observed in Twitter (reciprocity cannot be computed in Facebook and Google + because it is impossible to back out who is following whom). *Average clustering* captures the tendency to form circles in which one's friends are friends with each other. We use an extension to directed graphs of the clustering coefficient proposed by Fagiolo (2007), defined as the average, over all nodes i, of the nodes-specific clustering coefficients

$$cl(g_0) = \frac{1}{n} \sum_i \frac{\left(g_0 + g_0^\top\right)^3_{ii}}{2\left(D_i^{\text{tot}}\left(D_i^{\text{tot}} - 1\right) - 2\left(g_0^2\right)_{ii}\right)},$$

where D_i^{tot} is the total degree, i.e. in-degree plus out-degree, of agent i. Average clustering across networks is 0.36, somewhat smaller than values observed in real-life networks as seen in Table (1). Finally, we can also compute the initial *homophily* of opinions of agents, measured by an assortativity coefficient, as in Newman (2003), which takes positive values (maximum 1) if nodes with similar opinion tend to connect to each other, and negative (minimum -1) otherwise. In the sample, initial homophily ranges from -0.4 and 0.5, with an average value of -0.03. Notice though, that the degree of homophily in the long-run is endogenously determined. In an environment with no bots, for example, all agents converge to the same opinion (in which case limiting homophily is 1).

Influence of unsophisticated agents: The amount and location of unsophisticated agents in g_0 are important determinants of limiting $\overline{MI}$ and $\overline{P}$, as it is through them that fake news spread in the network. If the bot manages to manipulate an unsophisticated agent who is very influential (e.g. central), it may be able to effectively affect the opinion of others. We vary the share of unsophisticated agents by drawing μ_u from a uniform distribution in the interval $[0.1, 0.4]$. The average share on the sample is 0.25 as shown in Table (2) with a maximum percentage of unsophisticated agents of 40% and a minimum of 10%.

Table 2: Centrality of unsophisticated agents

	Mean	Std Dev.	Min	Max
Share of unsophisticated	0.250	0.100	0.1	0.4
in-Degree	0.12	0.06	0	0.54
out-Degree	0.15	0.05	0.03	0.3
PageRank	0.03	0.01	0.01	0.11
in-Closeness	0.18	0.06	0.01	0.29
out-Closeness	0.22	0.14	0.01	0.56
Betweenness	0.05	0.02	0	0.24

There are several statistics in the literature that can proxy for the degree of influence or centrality of an unsophisticated agent. *Degree* is the simplest centrality measure, which consists on counting the number of neighbors an agent has. The in-degree is defined as the number of incoming links to a given unsophisticated agent,

$$D_i^{\text{in}} = \frac{1}{n-1} \sum_j \left[g_0\right]_{ji}.$$

The out-degree is defined analogously. Table (2) reports the average in-Degree and out-Degree of all unsophisticated agents in the network (regardless of which type of bot influences them). The average in-degree is 0.12. There is a large dispersion across networks, with cases in which unsophisticated agents are followed by about 54% of agents in the network. Out-degree of these agents is on average 0.15. A larger value of this measure increases the influence of friends in the network for each given unsophisticated agent, reducing the influence of bots.

While this measure of influence is intuitive, it is not necessarily the only way in which a bot can be efficient at manipulating opinion, and hence affecting misinformation and polarization. There are networks in which unsophisticated agents have very few followers (and hence a low in-degree) but each of their followers is very influential. An alternative measure of centrality that incorporates these indirect effects is Google's *PageRank* centrality.[9] PageRank tries to account not only by quantity (e.g. a node with more incoming links is more influential) but also by quality (a node with a link from a node which is known to be very influential is also influential).

[9]This measure is a variant of eigenvector centrality, also commonly used in network analysis.

Mathematically, the PageRank centrality PR_i of a node i is represented by

$$PR_i = \alpha \sum_j \frac{[g_0]_{ji}}{D_j^{\text{out}}} PR_j + \frac{1-\nu}{n},$$

where D_j^{out} is the out-degree of node j if such degree is positive and ν is the damping factor.[10] Note that the PageRank of agent i depends on the PageRank of its followers in the recursion above. Summary statistics for the average PageRank of unsophisticated agents are shown in Table (2).

An alternative measure of centrality is given by *closeness* centrality. This measure keeps track of how close a given agent is to each other node in the network. High proximity of the unsophisticated agent to all other agents in the network makes the bot more efficient in spreading fake news, as they reach their targeted audience more quickly. To compute closeness, we first measure the mean distance between the unsophisticated agent and every other agent in the network. Define d_{ji} as the length of the shortest path from agent j to unsophisticated i in the network G_0.[11] *In-closeness* centrality is defined as the inverse of the mean distance d_{ji} across agents to reach unsophisticated agent i,

$$C_i^{\text{in}} = \frac{n}{\sum_j d_{ji}}.$$

Out-closeness is similarly defined. Finally, *betweenness* centrality measures the frequency at which a node acts as a bridge along the shortest path between two other nodes. Statistics for these measures can be found in the last two rows of Table (2).

Table 3: Correlation across centrality measures

	in-Degree	out-Degree	in-Closeness	out-Closeness	PageRank	Betweenness
in-Degree	1					
out-Degree	0.54	1				
in-Closeness	0.63	0.65	1			
out-Closeness	0.43	0.66	0.78	1		
PageRank	0.64	−0.005	0.33	0.06	1	
Betweenness	0.18	0.004	0.33	0.09	0.41	1

[10] The damping factor tries to mitigate two natural limitations of this centrality measure. First, an agent can get "stuck" at the nodes that have no outgoing links (bots) and, second, nodes with no incoming links are never visited. The value of $\nu = 0.85$ is standard in the literature and it is the one we will use in the simulations.

[11] In many networks sometimes one agent may find more than one path to reach the unsophisticated agent). In such case, the shortest path is the minimum distance among all possible distances.

It is worth noticing that even though all of these are alternative measures of centrality, they capture slightly different concepts. An unsophisticated agent is central according to in-degree when it has a large number of followers, whereas she is central according betweenness if she is in the information path of many agents. The correlation between these two variables is just 0.18, as seen in Table (3) which reports the correlation coefficient across centrality measures in our sample. Moreover, the correlation of betweeness and almost all other centrality measures is relatively low. Note that even though the relationship between in-degree and PageRank is 64% and the correlation between in-degree and in-closeness is 63%, PageRank and in-closeness have a relatively low correlation of 0.33. This illustrates the challenges in finding a minimum set of measures to characterize influence in a random graph model.

Belief heterogeneity: The variability in the behavioral dimension is given by changes in the parameter ω, capturing the degree to which agents rely more or less heavily on the opinion of others. We draw ω from a Uniform distribution on the interval $(0.1, 1]$ with band-with 0.05. The average value og ω is 0.54, with a standard deviation of 0.29 in our sample. These are reported in Table (4).

Table 4: Belief heterogeneity and bot influence

	Mean	Std Dev.	Min	Max
Influence of friends ω	0.55	0.29	0.10	1
Speed of communication ρ	0.54	0.26	0.10	1
Flooding capacity κ	92	109	1	300

The two parameters capturing communication technology are ρ, which controls the speed at which links are activated and κ, which determines the ability of bots to flood the network with fake news. We draw ρ from a uniform distribution in $(0, 1]$. Its average value is 0.54. A standard deviation of 0.29 ensures that there is significant variability in the artificial dataset. Note that we are excluding cases in which nodes are never activated, $\rho = 0$, as the network would exhibit no dynamics.

We consider alternative values for the flooding parameter, $\kappa \in \{1, 10, 50, 100, 300\}$. The average number of signals sent by each bot in our sample is 92, with a minimum of one (which is the number of signals sent by unsophisticated agents per encounter) and a maximum of 300, indi-

cating that bots can place fake news 300 times faster than an unbiased source of news per period. Note that increasing κ is analogous to increasing the number of bots in the model (assuming the number and location of unsophisticated agents does not change).

4.3 Simulation

We simulate communication in each social media network for a large number of periods ($T = 2000$) and use the resulting opinions to compute misinformation and polarization. For each network j, we draw a signal $s_{i,t}^{j}$ for individual $i \in N$ at time $t \leq T$ from a Bernoulli distribution with parameter $\theta = 0.5$. We also draw the $n \times n$ matrix $\mathbf{c}^t$ at each period t from a Bernoulli distribution with parameter ρ_j, which determines the evolution of the network structure according to eq. (1). Together, the signals and the clock determine the evolution of world-views according to eqs. (4) and (5). With these, for each network j, we compute a time series for opinions $y_{i,t,j}$ for individual i at period t. Figure (7) displays the distribution of average opinions $\overline{y}_j$ across networks. We define $\overline{y}_j$ as

$$\overline{y}_j = \frac{1}{500}\frac{1}{n}\sum_{t\geq 1500}\sum_{i=1}^{n} y_{i,t,j},$$

We choose this threshold because simulations converge after about 500 periods to an ergodic set (most statistics and results are unchanged when using the last 200 periods instead).

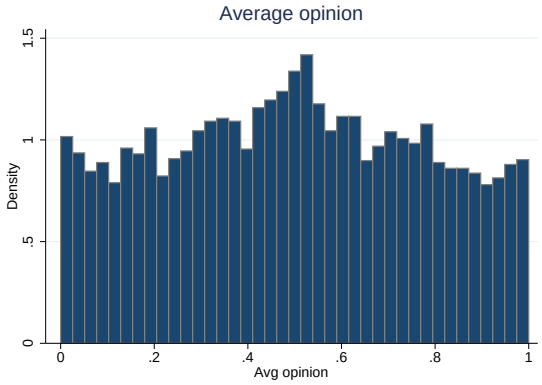


Figure 7: Average opinions

There is a large number of networks in which individuals are on average correct (e.g. close to $\theta = 0.5$), but there exists dispersion around this value (the standard deviation of $\overline{y}_j$ across networks is reported in Table (5)). Moreover, there is a non-trivial amount of networks in which agents' opinions become extreme implying that bots are successful at manipulating options to-

wards their own.

Our variables of interest are the long-run degree of misinformation $\overline{MI}_j$, the share of misinformed agents $\#\overline{MI}_j$, and limiting polarization $\bar{P}_j$ in each network.

The degree of misinformation is simply,

$$\overline{MI}_j \equiv \frac{1}{n}\sum_i \left(y_{i,j,T} - \theta\right)^2 .$$

where $y_{i,j,T}$ is the opinion of agent i at the last period $T = 2000$. The average degree of long-run misinformation across all networks, reported in Table (5), is 0.09, with a standard deviation of 0.07. The distribution of misinformation across networks is depicted in Figure (8). It is skewed to the left, indicating that there is a significant amount of networks with low degrees of misinformation. However, there is some mass around 0.25, which constitutes the theoretical upper bound for this measure. There are 147 networks in which $\overline{MI}_j \sim 0.25$, so individuals are fully misinformed.

Table 5: Simulation results

	Mean	Std Dev.	Min	Max
Average opinion $\overline{y}_j$	0.50	0.28	0	1
Misinformation $\overline{MI}_j$	0.09	0.07	0.0006	0.25
% Misinformed $\#\overline{MI}_j$	0.97	0.045	0.26	1
Polarization $\overline{P}_j$	0.11	0.09	0	0.64

Computing the proportion of misinformed individuals is challenging in this environment. The reason being that there are very few informed individuals at any given point in time. Moreover, small shocks (e.g. signals from either biased or unbiased sources) can easily move opinions ϵ away from the true θ. Instead of simply computing the average percentage of individuals in the last period (as we did with misinformation), we compute the smallest proportion of misinformed individuals over an interval $[1500, 2000]$. Mathematically,

$$\#\overline{MI}_j \equiv \max_t \{\#MI_{j,t}\},$$

where $\#M_{j,t}$ is calculated from eq. ((6)) assuming $\epsilon = 0.00025$ (we do not assume $\epsilon = 0$ to allow for computational rounding error). In other words, we restrict attention to an interval in the long-run and look for the lowest number of individuals who are misinformed in that period (or

equivalently, the largest percentage of wise agents). Even under this very generous definition, the proportion of misinformed agents is on average 0.97 across simulations, implying that most individuals in the network have opinions which are ϵ away from $\theta = 0.5$. In 30% of our networks all individuals are misinformed, whereas in about 40% of the cases only one or two agents learn the true θ. The lowest value for $\#\overline{MI}_j$ in our sample is 0.26 (i.e. only 26% of agents are misinformed). The result is not driven by high values of the flood parameter or a high proportion of unsophisticated agents in the network, as $\#\overline{MI} \sim 0.96$ when $\kappa = 1$ and $\mu_u = 0.1$. These values are likely to result from the fact that we are considering small networks (there are only 37 nodes in them), in which at least 4 agents are unsophisticated, so misinformation is bound to be large by construction. In addition, we are ignoring cases in which agents are fairly Bayesian, as the smallest value of ω is 0.1 in the simulations.

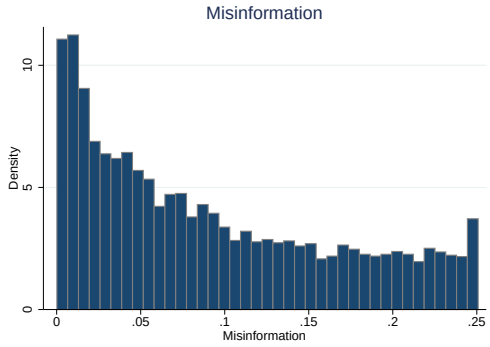


Figure 8: Misinformation

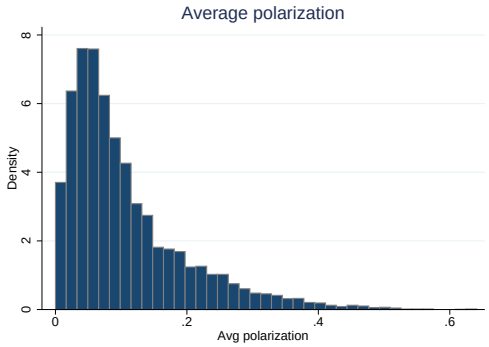


Figure 9: Polarization

For each network j, and given $\varsigma = 0.5$, we compute

$$\overline{P}_j \equiv \frac{1}{500\hat{p}} \sum_{t \geq 1500} P_{j,t},$$

which is normalized by $\hat{p}$ to belong to the interval $[0, 1]$. [12] About 98% of our sample exhibits positive polarization in the long run. Figure (9) depicts the distribution of polarization resulting from our simulation exercise. There is a significant degree of variability in our sample, even though the polarization levels are relatively small, with most $\bar{P}_j$ observations lying below 0.5 (recall that maximum polarization has been normalized to 1, yet the maximum polarization level observed in our sample is just 0.64). The average value of $\bar{P}_j$ across networks is 0.11, with a standard

[12] With $\varsigma = 0.5$ the maximum possible level of polarization (theoretically) is $\hat{p} = 0.707$. We divide all values of polarization by this number to normalize the upper bound to 1. This is without loss of generality and aims at easing interpretation.

deviation of 0.09. Interestingly, we also observe some mass near 0, indicating that agents reach quasi-consensus. Unfortunately, most of these cases involves consensus around extreme values of θ rather than efficient aggregation of information to the true θ. Figure (10) shows a scatterplot of misinformation and polarization in networks with a small number of unsophisticated agents $\mu_u = 0.1$. We see that there is an inverted U-shape relationship between these variables. Polarization can be low either because individuals learned the true (e.g low misinformation) or because they converged to the wrong value of θ. There are, however, situations where misinformation is relatively low but polarization is extremely high.

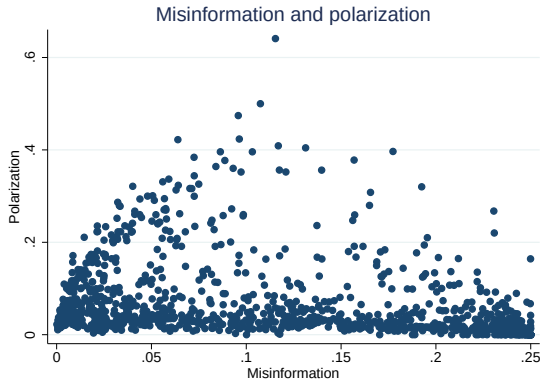


Figure 10: Relationship between misinformation and polarization (with 4 unsophisticated agents)

5 Regression Analysis

We are interested in estimating the effect of network characteristics on limiting misinformation and polarization. To assess the quantitative importance of each explanatory variable, we estimate the coefficients of an OLS model,

$$Y_j = \mathbf{X}_j\beta + \epsilon_j. \tag{9}$$

where the $M \times 1$ vector Y_j denotes either long-run misinformation $\overline{MI}_j$ or polarization P_j obtained from simulation $j \in \{1 \dots M\}$, $\mathbf{X}_j$ denotes the matrix of network characteristics per simulation j, and ϵ_j is the error term.

5.1 Benchmark case

The set of explanatory variables includes BAUER parameters in Ω, parameters of the heuristic update rule and the technology of communication (see Table (4)), as well as those characterizing

the network topology (see Table (1)) and the average centrality of unsophisticated agents (see Table (2)). In addition, we also consider the *relative* influence of bots. An R-bot may create more misinformation if it is more effective at manipulating influential individuals than the L-bot. We compute 'relative centrality' as the absolute value of the difference in the centrality measure of individuals who are influenced by bots of different types. For example, in the case of PageRank, we compute

$$\text{Relative PageRank} = |PageRank(L) - PageRank(R)|.$$

Other relative measures of centrality are computed analogously.

The results are summarized in Table (6), where we omit the estimated coefficients on reciprocity and homophily as they are statistically insignificant across regressions models. In the first two columns, variables have been normalized by their sample standard deviation in order to simplify the interpretation of coefficients and ease comparison across covariates. Hence, each estimated coefficient represents by how many standard deviations (st devs) misinformation or polarization change when the respective independent variable increases by one standard deviation.

The positive coefficient on ω implies that as agents place more weight on the opinions of their social media friends (and less on the unbiased signal), both misinformation (column 1) and polarization (column 2) rise. This is expected as higher values of this variable increases (indirectly) the influence of bots. The effect on polarization is larger, as it rises by 0.14 st devs, whereas misinformation only rises by 0.08 standard deviations.

The overall effect of a higher clock parameter ρ is a priori ambiguous: on the one hand, it is more likely that a sophisticated agent will incorporate fake news from those paying attention to the extreme views of bots as the speed of communication rises; on the other hand, a faster flow of information allows agents to incorporate unbiased private signals obtained by friends at a faster rate. Under the current specification, we find that misinformation declines with ρ, suggesting that the effect of internalizing a larger number of opinions outweighs the effect of higher fake news exposure. Clearly, this result would change in larger networks or those with a smaller share of unsophisticated agents in them. The effect of the clock on polarization is much stronger, as $\bar{P}$ declines by 0.15 st devs when ρ rises by one st deviation.

In terms of the network topology, we find that larger networks (measured by their diameter) are associated with higher misinformation and polarization, whereas high clustering is important

Table 6: Regression results: Benchmark case

	Misinformation (1)	Polarization (2)	Average opinion (3)
Communication technology			
Influence of friends ω	0.08***	0.14***	0.002
	(0.009)	(0.007)	(0.007)
Speed of communication ρ	-0.05***	-0.15***	-0.004
	(0.008)	(0.007)	(0.007)
Network Topology			
Diameter	0.05***	0.006	-0.003
	(0.01)	(0.01)	(0.01)
Clustering	0.01	-0.16***	-0.006
	(0.02)	(0.02)	(0.02)
Ω	yes	yes	yes
Bot influence			
Share of unsophisticated μ	0	+	0
Flooding κ	+	+	0
in-Degree	0.13***	-0.003	
	(0.023)	(0.018)	
PageRank	-0.44***	0.26***	
	(0.02)	(0.02)	
in-Closeness	-0.10***	-0.70***	
	(0.02)	(0.02)	
Betweeness	0.05***	0.02*	
	(0.016)	(0.013)	
out-Degree	-0.08***	-0.03	
	(0.024)	(0.019)	
Out-Closeness	-0.002	0.06***	
	(0.019)	(0.015)	
Relative PageRank	0.62***	-0.21***	31***
	(0.017)	(0.014)	(0.61)
Relative in-Degree	-0.05***	0.09***	0.22*
	(0.016)	(0.013)	(0.12)
Relative Betweeness	-0.10***	0.02*	-6.4***
	(0.01)	(0.01)	(0.24)
Relative in-Closeness	0.14***	-0.24***	3.8***
	(0.013)	(0.01)	(0.19)
Observations	8,248	8,248	8,248
R-squared	0.39	0.60	0.57

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

for polarization (it actually reduces it) but irrelevant for misinformation. The negative coefficient on clustering suggests that the implied higher connectivity reached with higher clustering countervails the bias reinforcement associated with echo-chambers (Sunstein 2002, 2009). The effects of initial homophily on misinformation and polarization vanish over time, as their coefficient is statistically insignificant (result omitted from the table).

The coefficients associated with the parameters Ω are not reported to make the exposition

of the regression results clearer, but are available upon request. It is worth mentioning that once centrality measures and network topology statistics are included, most of these parameters become statistically insignificant.[13]

The most interesting results are those related to measures of centrality. When unsophisticated agents have a large number of followers in social media, misinformation rises: the coefficient of in-Degree, reported on Table (6), is 0.13 and statistically different from zero. The effect on polarization is statistically insignificant. Keeping the number of followers constant, we can consider how unsophisticated agents who are followed by influential agents affect our variables of interest through the coefficient of PageRank (see definition on Section (4.2)). Interestingly, a higher PageRank is associated with less misinformation but more polarization. A one st dev increase in PageRank is associated with a reduction of 0.44 st devs in misinformation and an increase of 0.26 st devs in polarization, both quantitatively more important than in-Degree. This suggests that when unsophisticated agents are relatively more influential (because they manage to affect the opinions of influential followers), information is more efficiently aggregated. However, to the extent that agents do not fully learn the true θ, there is a significant amount of networks in which opinions become extreme. This is illustrated in the left panel of Figure (11), which depicts the distribution of average opinions across networks for PageRank below average. We see that this is basically a three point distribution (e.g., $\overline{y}_j = 0$, $\overline{y}_j = 0.5$, and $\overline{y}_j = 1$) with small mass in intermediate values. The right panel of that figure shows $\overline{y}_j$ for PageRank above average. The distribution is much more uniform in that case. There is a significantly smaller mass at θ (e.g. more misinformation) but also at the extremes (indicating less polarization).

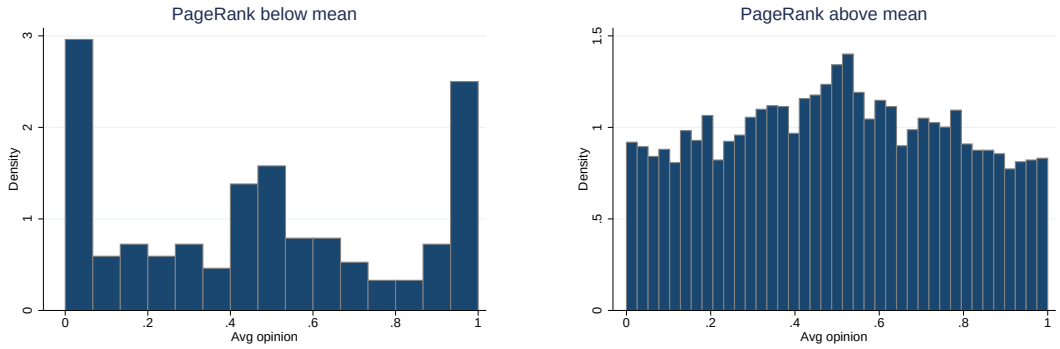


Figure 11: Average opinion for different PageRank

[13]Only a and α, determining the preferential attachment (and hence influence of any given node), are significant for polarization. This would suggest that our set of centrality measures may not be sufficient to capture influence in a network.

It is useful to analyze this coefficient together with that of Relative PageRank, which is positive for misinformation and negative for polarization. As unsophisticated agents following the R-bot become relatively more influential (e.g. $PageRank(R)$ rises) they tend to pull opinions towards 1. This is clearly seen in first panel in Figure (12) which depicts limiting values of average opinion for networks in which $PageRank(R) > PageRank(L)$. The distribution is clearly skewed to the right (analogously, when $PageRank(R) < PageRank(L)$ the distribution is skewed to the left). In column 3 of Table (6) we present the results of an estimated regression equation similar to eq. ((9)), in which the dependent variable is average opinions, $\bar{y}_j$, and we consider relative centrality as the difference between unsophisticated agents following R vs those following L (nor in absolute value neither normalized by its standard deviation, so that a higher centrality of R would be associated with $\bar{y}_j$ closer to 1). Interestingly, the only coefficients which are statistically significant are those related to relative centrality.

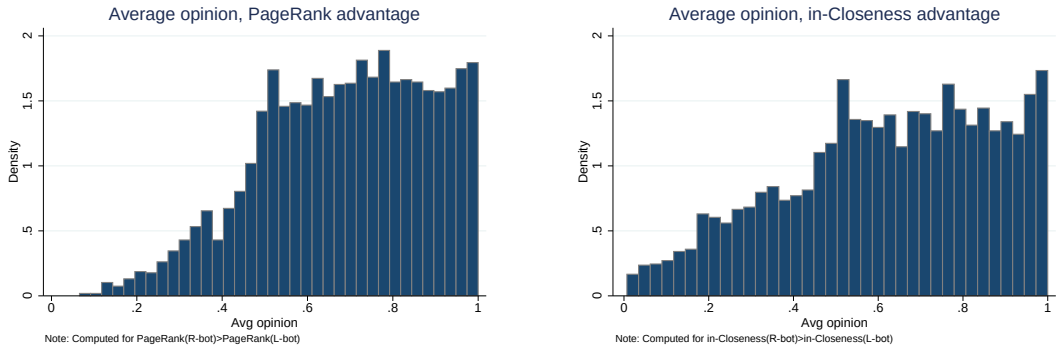


Figure 12: Average opinion for alternative measures of relative centrality

This phenomenon is also observed when considering relative in-Closeness instead, as evident from analyzing the right panel of Figure (12). As an R-bot gets relatively closer to the rest of the network through their influence on unsophisticated agents (with high in-Closeness scores), it becomes more efficient at spreading fake-news, since the speed at which each given piece of fake-news travels through the network rises, pulling opinions towards 1. This increases misinformation and decreases polarization. The effects of relative in-Closeness are much smaller than those of relative PageRank, as they increase misinformation by 0.64 and 0.14 st devs, respectively. Interestingly, when in-Closeness increases *on average* both misinformation and polarization decline. The effect on polarization is opposite to that of PageRank (a coefficient of -0.7 st devs in the former vs 0.26 st devs in the latter), indicating that higher in-closeness allows for better aggregation of information without having to incur the cost of greater disagreement. This happens

because higher in-Closeness is associated with smaller distance between each node in the network and that of the unsophisticated agent. A larger average value is then proxying for a more connected network in which information travels faster. As we have seen from the analysis of ρ, as speed of communication rises, it is easier for agents to learn the true θ (which implies both lower misinformation and polarization). The effects of Betweenness, out-Degree, and out-Closeness are quantitatively smaller and in some cases statistically insignificant.

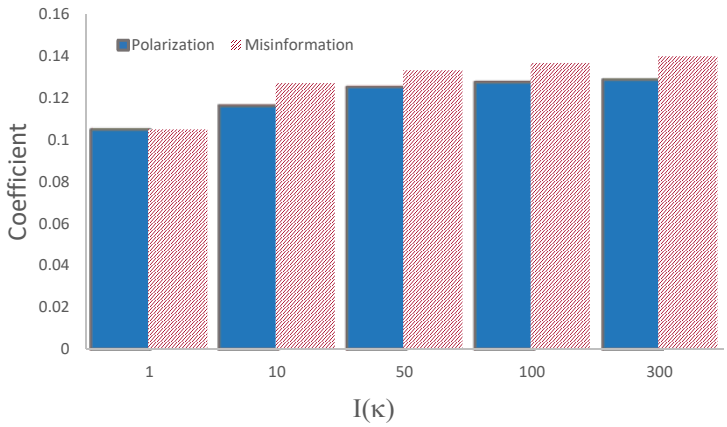


Figure 13: Estimated coefficient on the indicator $I(\kappa)$.

Through our simulations, we consider networks in which bots can send multiple fake-news articles each period (e.g. by allowing $\kappa > 1$). We control for this greater ability to spread fake-news by introducing a set of dummy variables $I(\kappa)$, one for each κ in the regression equation. To ease readability, we plot the resulting coefficients in Figure ((13)). The solid bars represent the effects of κ on misinformation, whereas the dashed pattern their effect on polarization (these variables are not normalized by the standard deviations). As evident from the graph, all coefficients are positive (and significant with p-values lower than 1%.), indicating that the greater ability to spread fake news by each bot, keeping everything else constant, results in less information aggregation and more variability of opinions in the long run. In addition, note that the larger effects on misinformation are observed for relatively small values of κ, with the effects remaining more or less stable for $\kappa \geq 50$. This suggests decreasing marginal returns to the introduction of fake-news on misinformation levels. In contrast to the flooding parameter, increases in the proportion of unsophisticated agents μ_u do not significantly affect misinformation in our estimation. They do, on the other hand, increase polarization (results omitted).

5.2 No flooding: $\kappa = 1$

We restrict the flooding technology of bots such that they can only send one signal each period. Because this significantly reduces their ability to spread fake news, average misinformation declines by 17% (from 0.09 to 0.075) and average polarization is 22% lower (from 0.11 to about 0.08). We re-estimated eq. ((9)) by restricting attention to networks with $\kappa = 1$, which lowers the number of observations from 8,248 to 1,651. The resulting coefficients for misinformation are presented in Specification (2) and for polarization in Specification (5) of Table (7). Columns (1) and (4) replicate estimated coefficient under the benchmark (e.g. unrestricted) case for reference. All variables are normalized by their standard deviations, as in the previous section. Note, however, that the normalizing st devs in the benchmark and restricted cases need not be the same, as the volatility of some variables (particularly the dependent variables) may change when bots are constrained.

We have omitted coefficients on Betweeness, Out-Closeness, and Out-Degree as they are statistically insignificant for the restricted sample. The influence of friends, ω is now much more important in increasing polarization. This is intuitive: if a bot can only spread one fake news per period, the only way in which this would influence opinions is if individuals pay more attention to it and relatively less attention to unbiased signals. A similar pattern is observed for polarization. The effects of ρ are also larger, on both MI and P, when bots' technology of spreading fake news is slower. Information aggregation becomes more efficient when the speed of communication rises, as bots are less likely to clutter signals received by unsophisticated agents.

Relative PageRank, average in-Closeness and relative in-Closeness are all significantly less important in reducing polarization (their coefficients are more than halved) when compared to the benchmark case. This is probably happening because bots are less effective into pulling opinions towards their preferred value when the volume of fake news they are able to broadcast declines.

5.3 Small share of uninformed: $\mu_u < 0.1$

Figure (14) depicts the distribution of misinformation (top) polarization (bottom) across networks for two cases: a small share of unsophisticated agents $\mu_u < 0.1$ (left) and a high share of unsophisticated agents $\mu_u > 0.1$. As expected, there is a larger number of networks in which agents aggregate information better when μ_u is relatively low. This can be seen by the fact that the mass near zero MI and P is higher in the left panels. Despite of this, misinformation is 22% higher and polarization 35% smaller than in the benchmark case.

Table 7: Regression results: $\kappa = 1$ and $\mu_u < 0.1$ cases

	Misinformation			Polarization		
	Benchmark	$\kappa = 1$	$\mu_u = 0.1$	Benchmark	$\kappa = 1$	$\mu_u = 0.1$
	(1)	(2)	(3)	(4)	(5)	(6)
Communication technology						
Influence of friends ω	0.08***	0.14***	0.04	0.14***	0.23***	0.04
	-0.01	-0.02	-0.02	-0.01	-0.01	-0.02
Speed of communication ρ	-0.05***	-0.06**	-0.03	-0.15***	-0.21***	-0.08***
	-0.01	-0.02	-0.02	-0.01	-0.01	-0.02
Network Topology						
Diameter	0.05***	0.07*	0.05	0.01	0.09***	0.05
	-0.01	-0.03	-0.03	-0.01	-0.02	-0.03
Clustering	0	-0.04	-0.06	-0.16***	-0.18***	-0.13*
	-0.02	-0.05	-0.05	-0.02	-0.04	-0.05
Ω	yes	yes	yes	yes	yes	yes
Bot influence						
in-Degree	0.13***	0.09	0.14*	0	-0.03	0.06
	-0.02	-0.05	-0.06	-0.02	-0.04	-0.05
PageRank	-0.44***	-0.32***	-0.51***	0.26***	0.32***	0.35***
	-0.02	-0.05	-0.06	-0.02	-0.03	-0.05
in-Closeness	-0.10***	0.22***	0.15**	-0.70***	-0.21***	-0.83***
	-0.02	-0.05	-0.05	-0.02	-0.03	-0.05
Relative PageRank	0.61***	0.59***	0.58***	-0.21***	-0.11***	-0.24***
	-0.02	-0.04	-0.05	-0.01	-0.03	-0.04
Relative in-Degree	-0.04**	-0.08*	-0.06	0.09***	0	0.08*
	-0.02	-0.04	-0.04	-0.01	-0.03	-0.04
Relative Betweeness	-0.10***	-0.03	-0.11**	0.02	0.01	-0.02
	-0.01	-0.03	-0.04	-0.01	-0.02	-0.03
Relative in-Closeness	0.15***	0.20***	0.40***	-0.23***	-0.09***	-0.44***
	-0.01	-0.03	-0.04	-0.01	-0.02	-0.03
Observations	8248	1651	1147	8248	1651	1147
R-squared	0.392	0.338	0.444	0.596	0.677	0.528

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This is potentially driven by the non-trivial number of networks in which misinformation becomes extreme with opinions converging to either 0 or 1. This is surprising given the smaller share of uninformed agents. A potential explanation is that when a large proportion of agents are targeted by bots in a symmetric way (in a relatively small network such as ours), agents average out opposing views from their friends muting the impact of extreme signals. Alternatively, an unsophisticated agent is very likely to receive both types of extreme signals (directly or indirectly), and hence not be susceptive to modifying their own views to a large extent. When the share of unsophisticated agents is small, each one of them is more likely to receive only one type of extreme signal: the one sent by the bot targeting them. Because of this, their views become more extreme (as there is no counterbalancing information received), particularly if the bot can

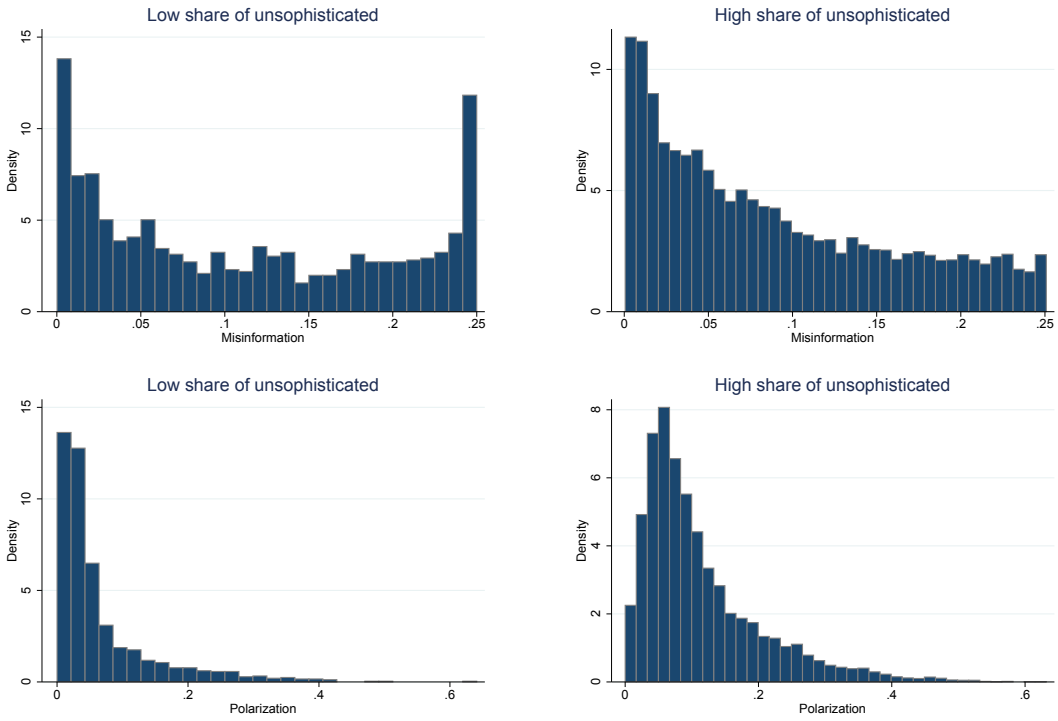


Figure 14: Misinformation and polarization for $\mu_u \leq 0.1$ (left) and $\mu_u \geq 0.1$ (right)

flood them. To the extent that one bot manages to manipulate a relatively more influential set of agents, it will be significantly more likely to succeed in modifying average opinions. Hence, a bot may be more effective by targeting a small set of influential agents, rather than by targeting the whole population.

In Specifications (3) and (6) we re-estimate our regression restricting analysis to networks with a small share of uninformed agents $\mu_u < 0.1$. Interestingly, both ρ and ω do not affect misinformation in this case, whereas polarization is only marginally reduced when the speed of communication rises (a considerable difference relative to the benchmark case). As suggested by the previous analysis, a one standard deviation increase in relative Page-Rank, relative in-Closeness, and even average in-Closeness create significantly more misinformation than in the benchmark case.

5.4 Welfare analysis

Misinformation and polarization may introduce inefficiencies in the decision-making process through different channels. A highly misinformed society will agree on selecting the wrong policy. A polarized society—which is on average correct—may not react fast enough to shocks

due to inefficient gridlock. A micro-founded political-economy model would dictate how much individuals would be willing to trade-off polarization for misinformation when opinions do not converge to the true θ. Because elaborating such model is beyond the scope of our paper, we now take a reduced-form approach and consider instead the following social welfare function

$$SW(MI_j, P_j) = -\left[\lambda MI_j + (1-\lambda)P_j\right],$$

which is decreasing in MI and P, with λ capturing the relative importance of misinformation in society j. Given this ad-hoc function, we can analyze how our explanatory variables affect societal welfare for alternative values of λ. The results are presented in Table (8), with variables normalized by their standard deviations (similarly to the procedure followed in Table (6)).

Regardless of the values of λ considered, a higher degree of influence of bots resulting from higher weight on opinions of social media friends (ω), the ability of bots to flood the network (κ), or the fact that they manage to influence agents with a large number of followers (in-Degree) decreases societal welfare. By comparing the size of different coefficients in Table (8), we can observe that the largest effects on welfare are caused by PageRank scores. To the extent that bots are symmetric, higher average PageRank results in higher welfare as it allows more information aggregation. However, when one bot is relatively more influential (through their targeting of agents with high PageRank scores), the negative effects on welfare are extremely large. This has policy implications: a society that is successful in eliminating a source of fake news promoting one extreme of the political spectrum may end up worse off due to the unintended consequences of making the other extreme relatively more powerful. This, in the end, would generate greater misinformation and lower welfare, despite effectively reducing polarization.

5.5 Robustness

There were a few statistics measuring network topology and centrality of unsophisticated agents who were either statistically insignificant in the analysis above, or have very small effects on polarization and misinformation. These are Diameter, out-Degree, out-Closeness, Initial homophily, and Reciprocity. In Specifications (2) and (5) of Table (9) we re-estimate the model excluding these variables. It is evident that the results are basically unchanged: the goodness of fit is identical to the second decimal and the size of the coefficients are basically unchanged relative to the Benchmark case (replicated in columns (1) and (4)).

It may also be of interest to consider how the results change when we do not include the pa-

Table 8: Regression results: Welfare

	$\lambda = 1$	$\lambda = 0.8$	$\lambda = 0.5$
Communication technology			
Influence of friends ω	-0.08***	-0.12***	-0.17***
	(0.009)	(0.008)	(0.007)
Speed of communication ρ	0.05***	0.10***	0.16***
	(0.009)	(0.008)	(0.007)
Network Topology			
Diameter	-0.05***	-0.05***	-0.04***
	(0.01)	(0.01)	(0.01)
Clustering	-0.01	0.04*	0.12***
	(0.02)	(0.02)	(0.02)
Ω	yes	yes	yes
Bot influence			
Share of unsophisticated μ	0	-	-
Flooding κ	-	-	-
in-Degree	-0.13***	-0.13***	-0.08***
	(0.02)	(0.02)	(0.02)
PageRank	0.44***	0.35***	0.08***
	(0.02)	(0.02)	(0.02)
in-Closeness	0.10***	0.30***	0.63***
	(0.02)	(0.02)	(0.02)
Betweeness	-0.05***	-0.05***	-0.05***
	(0.02)	(0.02)	(0.01)
out-Degree	0.08***	0.08***	0.07***
	(0.02)	(0.02)	(0.02)
Out-Closeness	0.002	-0.02	-0.05***
	(0.02)	(0.02)	(0.02)
Relative PageRank	-0.62***	-0.54***	-0.24***
	(0.02)	(0.02)	(0.01)
Relative in-Degree	0.04***	0.02	-0.05***
	(0.02)	(0.02)	(0.01)
Relative Betweeness	0.10***	0.10***	0.05***
	(0.01)	(0.01)	(0.01)
Relative in-Closeness	-0.14***	-0.07***	0.10***
	(0.01)	(0.01)	(0.01)
Observations	8,248	8,248	8,248
R-squared	0.392	0.436	0.572

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

rameters in Ω as regressors. This is useful to know because when considering real-life networks, it is typically impossible to back out the parameters in Ω (which determine how the initial network is created). A statistician may only be able to compute statistics from observable variables, such links determining centrality and clustering. As in the previous case, estimated coefficients and the R-squared are unchanged when Ω is excluded from the regression model. This suggests that the set of network statistics in columns (3) and (6) are sufficient to describe how misinformation

Table 9: Robustness Exercises

	Misinformation			Polarization		
	Benchmark	Less regressors	No Ω	Benchmark	Less regressors	No Ω
	(1)	(2)	(3)	(4)	(5)	(6)
Communication technology						
Influence of friends ω	0.08***	0.08***	0.08***	0.14***	0.14***	0.14***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Speed of communication ρ	-0.05***	-0.05***	-0.05***	-0.15***	-0.15***	-0.15***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Network Topology						
Diameter	0.05***			0.01		
	(0.01)			(0.01)		
Clustering	0.01	-0.00	0.01	-0.16***	-0.17***	-0.16***
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)
Ω	yes	yes	no	yes	yes	no
Bot influence						
in-Degree	0.13***	0.14***	0.13***	-0.00	-0.00	-0.04*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
PageRank	-0.44***	-0.44***	-0.43***	0.26***	0.25***	0.28***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
in-Closeness	-0.10***	-0.12***	-0.18***	-0.70***	-0.68***	-0.69***
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)
Betweeness	0.05**	0.06***	0.07***	0.02	0.02	0.03*
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Out-Degree	-0.08**			-0.03		
	(0.02)			(0.02)		
Out-Closeness	-0.00			0.06***		
	(0.02)			(0.02)		
Relative PageRank	0.62***	0.62***	0.62***	-0.21***	-0.20***	-0.21***
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Relative in-Degree	-0.04**	-0.05**	-0.04**	0.09***	0.10***	0.11***
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Relative Betweeness	-0.10***	-0.10***	-0.10***	0.02	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Relative in-Closeness	0.14***	0.16***	0.15***	-0.24***	-0.24***	-0.25***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	8248	8248	8248	8248	8248	8248
R-squared	0.392	0.390	0.389	0.597	0.596	0.593

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

and polarization react to changes in the environment.

6 Conclusions

We created a large sample of synthetic social media networks by varying their characteristics through simulations in order to understand what the most important drivers of misinformation and polarization are. A premise in all of them is the ability of bots (with opposite extreme views)

who purposely spread fake news in order to manipulate the opinion of a small share of agents in the network. To the extent that agents can be partially influenced by these signals—directly by not filtering out fake news, or indirectly by following friends who are themselves influenced by bots—, this can generate misinformation and polarization in the long run. In other words, fake news prevent information aggregation and consensus in the population. We find that when bots at one extreme are relatively more efficient at manipulating news (by targeting a small number of influential agents), the may be able to generate full misinformation in the long run, where beliefs are at one end of the political spectrum. There would be no polarization in that case, but at the expense of agents converging to the wrong value of θ, the parameter of interest. There are other situations where agents are on average correct, but have nonetheless very heterogeneous opinions. These cases would still be sub-optimal, as they may result in inefficient gridlock and inaction.

An important assumption is that the links in the network evolve stochastically. It would be interesting to extend the model to consider a case in which links are endogenously determined. This could achieved by allowing agents to place a higher weight on individuals who share similar priors and choose to 'unfollow' (e.g. break links) agents who have views which are relatively far from their own.

Having identified the main determinants of polarization, it would be interesting to parameterize a real-life social media network (e.g. calibrate it) in order to back out the amount of fake news necessary to produce the observed increase in polarization between two periods of time. It would also be possible to carry forward a key-player analysis on the location of internet bots to better understand what is the most efficient way to reduce polarization.

Finally, we do observe polarization cycles in some of our networks. Analyzing their determinants could be a fruitful avenue for future research.

References

Acemoglu, D., V. Chernozhukov, and M. Yildiz (2008): "Fragility of Asymptotic Agreement Under Bayesian Learning," *SSRN eLibrary*.

Acemoglu, D., G. Como, F. Fagnani, and A. Ozdaglar (2013): "Opinion Fluctuations and Disagreement in Social Networks," *Mathematics of Operations Research*, 38, 1–27.

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

Acemoglu, D. and A. Ozdaglar (2011): "Opinion Dynamics and Learning in Social Networks," *Dynamic Games and Applications*, 1, 3–49.

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

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," Tech. rep., National Bureau of Economic Research.

Andreoni, J. and T. Mylovanov (2012): "Diverging opinions," *American Economic Journal: Microeconomics*, 4, 209–232.

Aumann, R. J. . (1976): "Agreeing to Disagree," *The Annals of Statistics*, 4, 1236–1239.

Azzimonti, M. (2015): "Partisan Conflict and Private Investment," *NBER Working Paper*, 21273.

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

Baldassarri, D. and P. Bearman (2007): "Dynamics of political polarization," *American sociological review*.

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

Banerjee, A. V. (1992): "A simple model of herd behavior," *The Quarterly Journal of Economics*, 797–817.

Barabási, A.-L. and R. Albert (1999): "Emergence of scaling in random networks," *science*, 286, 509–512.

Barber, M. and N. McCarty (2015): "Causes and consequences of polarization," *Solutions to Political Polarization in America*, 15.

Barberá, P. (2014): "How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the US," *Working Paper, New York University*, 46.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017): "Is the internet causing political polarization? Evidence from demographics," Tech. rep., National Bureau of Economic Research.

Chandrasekhar, A. G., H. Larreguy, and J. P. Xandri (2012): "Testing Models of Social Learning on Networks," *Working paper*, 1–54.

Chatterjee, S. and E. Seneta (1977): "Towards consensus: some convergence theorems on repeated averaging," *Journal of Applied Probability*, 89–97.

Conover, M., J. Ratkiewicz, and M. Francisco (2011): "Political Polarization on Twitter," *ICWSM*.

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

DeMarzo, P. M., D. Vayanos, and J. Zwiebel (2003): "Persuasion Bias, Social Influence, and Unidimensional Opinions," *The Quarterly journal of economics*, 118, 909–968.

Dixit, A. K. and J. W. Weibull (2007): "Political polarization," *Proceedings of the National Academy of Sciences of the United States of America*, 104, 7351–7356.

Duclos, J.-Y., J. Esteban, and D. Ray (2004): "Polarization: Concepts, Measurement, Estimation," *Econometrica*, 72, 1737–1772.

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

——— (1995): "Word-of-mouth communication and social learning," *The Quarterly Journal of Economics*, 93–125.

Epstein, L. G., J. Noor, and A. Sandroni (2010): "Non-Bayesian Learning," *The B.E. Journal of Theoretical Economics*, 10.

Erdös, P. and A. Rényi (1959): "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, 6, 290–297.

Esteban, J., C. Gradín, and D. Ray (2007): "An extension of a measure of polarization, with an application to the income distribution of five OECD countries," *The Journal of Economic Inequality*, 5, 1–19.

Esteban, J. and D. Ray (1994): "On the Measurement of Polarization," *Econometrica*, 62, 819–851.

——— (2010): "Comparing Polarization Measures," *Journal Of Peace Research*, 0–29.

Fiorina, M. and S. Abrams (2008): "Political Polarization in the American Publlic," *The Annual Review of Political Science*, 49–59.

Gentzkow, M. and J. M. Shapiro (2006): "Media bias and reputation," *Journal of political Economy*, 114, 280–316.

——— (2010): "What drives media slant? Evidence from US daily newspapers," *Econometrica*, 78, 35–71.

——— (2011): "Ideological segregation online and offline," *The Quarterly Journal of Economics*, 126, 1799–1839.

Golub, B. and M. Jackson (2010): "Naive Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, 2, 112–149.

Goyal, S. (2005): "Learning in networks," *Group formation in economics: networks, clubs and coalitions*, 122–70.

Groseclose, T. and J. Milyo (2005): "A Measure of Media Bias," *The Quarterly Journal of Economics*, CXX.

Gruzd, A. and J. Roy (2014): "Investigating political polarization on Twitter: A Canadian perspective," *Policy & Internet*.

Guerra, P. H. C., W. M. Jr, C. Cardie, and R. Kleinberg (2013): "A Measure of Polarization on Social Media Networks Based on Community Boundaries," *Association for the Advancement of Artificial Intelligence*, 1–10.

Jackson, M. (2010): *Social and Economic Networks*, vol. 21, Princeton University Press.

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

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

Kelly, J., D. Fisher, and M. Smith (2005): "Debate, division, and diversity: Political discourse networks in USENET newsgroups," *Online Deliberation Conference*.

Lee, J. K., J. Choi, C. Kim, and Y. Kim (2014): "Social Media, Network Heterogeneity, and Opinion Polarization," *Journal of Communication*, 64, 702–722.

Messing, S. and S. J. Westwood (2012): "Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online," *Communication Research*, 41, 1042–1063.

Meyer, C. D., ed. (2000): *Matrix Analysis and Applied Linear Algebra*, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Mohammad, S. M., X. Zhu, S. Kiritchenko, and J. Martin (2015): "Sentiment, emotion, purpose, and style in electoral tweets," *Information Processing & Management*, 51, 480–499.

Mossel, E., A. Sly, and O. Tamuz (2012): "On Agreement and Learning," *Arxiv preprint arXiv:1207.5895*, 1–20.

Roux, N. and J. Sobel (2012): "Group Polarization in a Model of Information Aggregation," .

Seneta, E. (1979): "Coefficients of ergodicity: structure and applications," *Advances in applied probability*, 576–590.

——— (2006): *Non-negative matrices and Markov chains*, Springer Science & Business Media.

Sethi, R. and M. Yildiz (2013): "Perspectives, Opinions, and Information Flows," *SSRN Electronic Journal*.

Shapiro, J. M. and N. M. Taddy (2015): "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech," .

Smith, L. and P. Sørensen (2000): "Pathological outcomes of observational learning," *Econometrica*, 68, 371–398.

Sobkowicz, P., M. Kaschesky, and G. Bouchard (2012): "Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web," *Government Information Quarterly*, 29, 470–479.

Sunstein, C. R. (2002): *Republic.com*, Princeton University Press.

——— (2009): *Republic.com 2.0*, Princeton University Press.

Tahbaz-Salehi, A. and A. Jadbabaie (2008): "A Necessary and Sufficient Condition for Consensus Over Random Networks," *IEEE Transactions on Automatic Control*, 53, 791–795.

Watts, D. J. and P. S. Dodds (2007): "Influentials, Networks and Public Opinion Formation," *Journal of Consumer Research*, 34, 441–458.

Webster, J. G. and T. B. Ksiazek (2012): "The Dynamics of Audience Fragmentation: Public Attention in an Age of Digital Media," *Journal of Communication*, 62, 39–56.

Yardi, S. and D. Boyd (2010): "Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter," *Bulletin of Science, Technology & Society*, 30, 316–327.

Appendix A: Beta-Bernoulli model and the update rule

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

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

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

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

The idiosyncratic likelihood induced by the vector of length K of observed signals $s_{i,t+1}$ is

$$\ell(s_{i,t+1}|\theta) = \theta^{\sum s_{i,t+1}}(1-\theta)^{K-\sum s_{i,t+1}},$$

and therefore the standard Bayesian posterior is computed as

$$\mu_{i,t+1}(\theta|s_{i,t+1}) = \frac{\ell(s_{i,t+1}|\theta)\,\mu_{i,t}(\theta)}{\displaystyle\int_{\Theta}\ell(s_{i,t+1}|\theta)\,\mu_{i,t}(\theta)\,d\theta}.$$

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

$$\begin{aligned}\mu_{i,t+1}(\theta|s_{i,t+1}) &\propto \ell(s_{i,t+1}|\theta)\,\mu_{i,t}(\theta) \\ &\propto \theta^{\alpha_{i,t}+\sum s_{i,t+1}-1}(1-\theta)^{\beta_{i,t}+K-\sum s_{i,t+1}-1}.\end{aligned}$$

Therefore, the posterior distribution is

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

where

$$\alpha_{i,t+1} = \alpha_{i,t} + \sum s_{i,t+1} \tag{10}$$

$$\beta_{i,t+1} = \beta_{i,t} + K - \sum s_{i,t+1}. \tag{11}$$

Equations (10) and (11) are used to update the shape parameters of both regular agents (sophisticated and unsophisticated, by setting $K = 1$) and bots (left and right, by setting $K = \kappa$) as per Equations (4), (5), (2) and (3) in subsection *Evolution of beliefs* in section 2.

Appendix B: Auxiliary lemmas and propositions regarding the properties of the sequence $\{W_t\}_{t=1}^{\infty}$ and the average weight matrix $\bar{W}$

Before proceeding, we implement here a slight change of notation: we let $1-\omega_{i,t} = b_{i,t}$. Then, as explained in section (2), the weight given by agent i to the bayesian update at any time t depends on whether agent i finds any other agent j in his neighborhood. In algebraic terms

$$b_{i,t} = \mathbb{1}_{\{\sum_j [\hat{g}_t]_{ij}=0\}} 1 + \left(1 - \mathbb{1}_{\{\sum_j [\hat{g}_t]_{ij}=0\}}\right) b. \tag{12}$$

Lemma 1. *The adjacency matrix g_0 is an irreducible matrix if and only if G_0 is strongly connected.*

Proof. By assumption, g_0 is strongly connected, for the completeness of the argument see **?** (Ch. 8, page 671). □

Lemma 2. *For all $t \geq 0$, the matrix W_t is row-stochastic.*

Proof. It is sufficient to show that $W_t\mathbf{1} = B_t\mathbf{1} + (\mathbb{I}_n - B_t)\,\hat{g}_t\mathbf{1} = \mathbf{1}$. For that we can show that for every period t the vector $W_t\mathbf{1}$ has all entries equal to

$$b_{i,t} + (1-b_{i,t})\sum_j [\hat{g}_t]_{ij} = \begin{cases} b_{i,t} = \mathbb{1}_{\{\sum_j [\hat{g}_t]_{ij}=0\}} 1 + \left(1 - \mathbb{1}_{\{\sum_j [\hat{g}_t]_{ij}=0\}}\right) b = 1 & \text{, if } \sum_j [\hat{g}_t]_{ij} = 0 \\ 1 & \text{, if } \sum_j [\hat{g}_t]_{ij} = 1, \end{cases}$$

as per the equation (12). □

Lemma 3. *The matrix $\bar{W}$ has diagonal entries $\left[\bar{W}\right]_{ii} = b + (1-b)(1-\rho)^{|N_{i,0}^{out}|}$ for all i and off-diagonal entries*

$$\left[\bar{W}\right]_{ij} = 0 \textit{ when } [g_0]_{ij} = 0 \textit{ and}$$

$$\left[\bar{W}\right]_{ij} = \left(1 - b - (1-b)(1-\rho)^{|N_{i,0}^{out}|}\right)[\hat{g}_0]_{ij} \textit{ when } [g_0]_{ij} \neq 0.$$

Proof. For any agent $i \in N$, the number of neighbors met at time t is a binomial random variable with parameters $|N_{i,0}^{out}|$ and ρ. Therefore, the probability that agent i finds no other agent in his neighborhood at time t (denoted as p_{it}^0) is

$$p_{it}^0 = \binom{|N_{i,0}^{out}|}{0} \rho^0 (1-\rho)^{|N_{i,0}^{out}|} = (1-\rho)^{|N_{i,0}^{out}|}.$$

Notice that the right hand side of the expression above does not depend on time t, thus, we can establish that $p_{it}^0 = p_i^0$.

Therefore, according to equation (12), we conclude that the elements in the main diagonal of the matrix $\bar{W}$ are

$$\begin{aligned}
\left[\bar{W}\right]_{ii} &= E(b_{i,t}) \\
&= 1p_i^0 + b(1-p_i^0) \\
&= (1-\rho)^{|N_{i,0}^{out}|} + b\left(1-(1-\rho)^{|N_{i,0}^{out}|}\right) \\
&= b + (1-b)(1-\rho)^{|N_{i,0}^{out}|}
\end{aligned}$$

In contrast, the elements off-diagonal can be written as

$$\begin{aligned}
\left[\bar{W}\right]_{ij} &= E\left((1-b_{i,t})[\hat{g}_t]_{ij}\right) \\
&= E\left(1-b_{i,t}\right) E\left([\hat{g}_t]_{ij}\right) \\
&= \left(1-E\left(b_{i,t}\right)\right) E\left(\frac{[g_t]_{ij}}{\sum_j [g_t]_{ij}}\right) \\
&= \left(1-\left(b+(1-b)(1-\rho)^{|N_{i,0}^{out}|}\right)\right) E\left(\frac{[g_0]_{ij}[c_t]_{ij}}{\sum_j \left([g_0]_{ij}[c_t]_{ij}\right)}\right) \\
&= \begin{cases} 0 & \text{, if } [g_0]_{ij} = 0 \\ \left(1-b-(1-b)(1-\rho)^{|N_{i,0}^{out}|}\right)\dfrac{[g_0]_{ij}\rho}{\rho\left(\sum_j [g_0]_{ij}\right)} & \text{, if } [g_0]_{ij} \neq 0 \end{cases} \\
&= \begin{cases} 0 & \text{, if } [g_0]_{ij} = 0 \\ (1-b)\left(1-(1-\rho)^{|N_{i,0}^{out}|}\right)\frac{1}{|N_{i,0}^{out}|} & \text{, if } [g_0]_{ij} \neq 0 \end{cases}
\end{aligned}$$

□

The next Lemma shows that the matrix $\bar{W}$ is primitive, i.e. there is a positive integer m such that $\bar{W}^m > 0$.

Lemma 4. *The average weight matrix $\bar{W}$ is irreducible and primitive.*

Proof. The irreducibility of $\bar{W}$ comes from the fact that g_0 is irreducible by the Assumption 1 (strongly connectedness and aperiodicity). By the Perron-Frobenius theorem, the largest eigenvalue of $\bar{W}$ in absolute terms is 1 and it has algebraic multiplicity of 1 (i.e. it is the only eigenvalue in the spectral circle of $\bar{W}$). By the Frobenius' test for primitivity (see Meyer (2000), ch. 8, page

678) it can be shown that any nonnegative irreducible matrix having only one unity eigenvalue on its spectral circle is said to be a primitive matrix. The converse is always true. □

Before introducing the next lemma, we will introduce two definitions. First, the *distance* between any two agents i and j is defined as the number of connections in the shortest path connecting them, i.e. the minimum number of "steps" that agent i should take to reach agent j. The *diameter* of the network is the largest distance between any two agents in the network, i.e. the maximum shortest path length in the network.

The next lemma provides a positive uniform lower bound on the entries of the matrix $\bar{W}^d$ as a function of the diameter of the network d induced by $\bar{W}$ and the minimum (non-zero) expected share of attention received by an agent i from any other agent $j \in N$ (i.e. including i himself), ω, defined as

$$\begin{aligned}\omega &= \overset{+}{\min_{i,j}} \left[\bar{W}\right]_{ij} \\ &= \min\left\{\min_{i\in N}\left(b+(1-b)(1-\rho)^{|N_{i,0}^{out}|}\right),\ \min_{i\in N}\left((1-b)\left(1-(1-\rho)^{\left|N_{i,0}^{out}\right|}\right)\frac{1}{\left|N_{i,0}^{out}\right|}\right)\right\}.\end{aligned}$$

Lemma 5. *Let d denote the diameter of the network induced by the social interaction matrix $\bar{W}$ and $\omega > 0$ be the scalar defined above. Then the entries of the matrix $\bar{W}^d$ are bounded below by the scalar ω^d.*

We will omit the proof of Lemma (5). For further details, the reader can refer to Theorems 1.3 and 1.4 in Seneta (2006, pgs. 18 and 21, respectively) and its related Lemmas.

Consider the update process described in the equation (4) of Section (2) in its matrix form

$$\begin{aligned}\alpha_{t+1} &= B_t(\alpha_t + s_{t+1}) + (\mathbb{I}_n - B_t)\hat{g}_t\alpha_t \\ &= \left[B_t + (\mathbb{I}_n - B_t)\hat{g}_t\right]\alpha_t + B_t s_{t+1}.\end{aligned}$$

Notice that B_t is not fixed over time as it depends on the realization of encounters in every period t. The stochastic matrix (see Lemma (2)) inside the squared bracket is denoted by W_t from now on and we re-write the previous update process as

$$\alpha_{t+1} = W_t\alpha_t + B_t s_{t+1}.$$

By forward iteration, we have that when $t = 0$,

$$\alpha_1 = W_0\alpha_0 + B_0 s_1.$$

When $t = 1$,

$$\begin{aligned}\alpha_2 &= W_1\alpha_1 + B_1 s_2 \\ &= W_1\left(W_0\alpha_0 + B_0 s_1\right) + B_1 s_2 \\ &= W_1 W_0\alpha_0 + W_1 B_0 s_1 + B_1 s_2.\end{aligned}$$

When $t = 2$,

$$\begin{aligned}\alpha_3 &= W_2\alpha_2 + B_2 s_3 \\ &= W_2(W_1 W_0\alpha_0 + W_1 B_0 s_1 + B_1 s_2) + B_2 s_3 \\ &= W_2 W_1 W_0\alpha_0 + W_2 W_1 B_0 s_1 + W_2 B_1 s_2 + B_2 s_3,\end{aligned}$$

so on and so forth and similarly for the shape parameter vector β.

Following Chaterjee and Seneta (1977), Seneta (2006) and Tahbaz-Salehi and Jadbabaie (2008), we let $\{W_t\}$, for $t \geq 0$, be a fixed sequence of stochastic matrices, and let $U_{r,k}$ be the stochastic matrix defined by the *backward product* of matrices

$$U_{r,k} = W_{r+k} \cdot W_{r+(k-1)} \ldots W_{r+2} W_{r+1} W_r.^{14} \tag{13}$$

With this definition in hand, we show some important properties of the expected backward product that will help us to prove convergence of opinions in probability to θ.

Proposition 4. *Let d be the diameter of the network induced by the matrix $\bar{W}$ and ω be the minimum expected share of attention received by some agent i from any other agent $j \in N$. Then, for all $r \geq 1$, i and j, and given d*

$$p_{ij} = P\left(\left[U_{r,d}\right]_{ij} \geq \frac{\omega^d}{2}\right) \geq \frac{\omega^d}{2} > 0.$$

Proof.

$$\begin{aligned}P\left(\left[U_{r,d}\right]_{ij} \geq \frac{\omega^d}{2}\right) &= P\left(1 - \left[U_{r,d}\right]_{ij} \leq 1 - \frac{\omega^d}{2}\right) \\ &= 1 - P\left(1 - \left[U_{r,d}\right]_{ij} \geq 1 - \frac{\omega^d}{2}\right)\end{aligned} \tag{14}$$

[14] Our backward product has last term equals to W_r, rather than W_{r+1}. This is because our *first* period is 0, rather than 1. This notation comes without costs or loss of generality.

By the Markov inequality, the probability in the right hand side of the equation (14) can be written as

$$P\left(1-\left[U_{r,d}\right]_{ij}\geq 1-\frac{\omega^d}{2}\right)\leq\frac{E\left(1-\left[U_{r,d}\right]_{ij}\right)}{1-\frac{\omega^d}{2}},$$

and therefore,

$$1-P\left(1-\left[U_{r,d}\right]_{ij}\geq 1-\frac{\omega^d}{2}\right)\geq 1-\frac{E\left(1-\left[U_{r,d}\right]_{ij}\right)}{1-\frac{\omega^d}{2}}. \tag{15}$$

Using equation (15), I rewrite equation (14) as

$$P\left(\left[U_{r,d}\right]_{ij}\geq\frac{\omega^d}{2}\right)\geq 1-\frac{E\left(1-\left[U_{r,d}\right]_{ij}\right)}{1-\frac{\omega^d}{2}}. \tag{16}$$

From the functional form of the backward product (see eq. (13)) and given that $\{W_t\}_{t=1}^{\infty}$ is a sequence of independent matrices over all t (see eq. (1)), the expectation above can be written as

$$E\left(U_{r,d}\right)=E\left(W_{r+d-1}W_{r+d-2}\cdots W_r\right)=\bar{W}^d,$$

thus, from Lemma (5), this implies that for all i and j

$$E\left(\left[U_{r,d}\right]_{ij}\right)\geq\omega^d.$$

Therefore, eq. (16) becomes

$$P\left(\left[U_{r,d}\right]_{ij}\geq\frac{\omega^d}{2}\right)\geq 1-\frac{1-\omega^d}{1-\frac{\omega^d}{2}}=\frac{\omega^d}{2-\omega^d}\geq\frac{\omega^d}{2},$$

proving that the (i,j) entry of the matrix represented by the backward product $U_{r,d}$ is positive with non-zero probability, i.e $p_{ij}>0$. □

Therefore, Proposition (4), together with the assumption that the sequence of matrices $\{W_t\}_{t=1}^{\infty}$ are i.i.d. and have positive diagonals (see Lemma (3)), ensures that the matrix represented by the backward product U_{1,n^2d+1} of length n^2d is positive with at least probability $\Pi_{i,j}p_{ij}>0$. The choice n^2d is a conservative one as in AOP (2010) and Tahbaz-Salehi and Jadbabaie (2008).

Lemma 6. *Consider $\rho=1$, i.e. $W_t=W$ for every t. The iteration of the row-stochastic matrix W is convergent and therefore there exists a threshold $\bar{\tau}\in\mathbb{N}$ such that $|W_{ij}^{\tau+1}-W_{ij}^{\tau}|<\epsilon$ for any $\tau\geq\bar{\tau}$ and $\epsilon>0$*

Proof. In order to see how W^τ behaves as τ grows large, it is convenient to rewrite W using its diagonal decomposition. In particular, let v be the squared matrix of left-hand eigenvectors of W and $D = (d_1, d_2, \ldots, d_n)^\top$ the eigenvector of size n associated to the unity eigenvalue $\lambda_1 = 1$.[15] Without loss of generality, we assume the following normalization $\mathbf{1}^\top D = 1$. Therefore, $W = v^{-1}\Lambda v$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the squared matrix with eigenvalues on its diagonal, ranked in terms of absolute values. More generally, for any time τ we write

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

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

$$[W^\tau]_{ij} = d_j + \sum_r \lambda_r^\tau v_{ir}^{-1} v_{rj},$$

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

[15] This is a feature shared by all stochastic matrices because having row sums equal to 1 means that $\|W\|_\infty = 1$ or, equivalently, $W\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the unity n-vector.

Appendix C: Backward product ergodicity

Our main concern in order to prove that agents' opinions converge in probability to θ is the behavior of $U_{r,k}$ when $k \to \infty$ for each $r \geq 0$. For that, we need to define two concepts of ergodicity. The sequence $\{W_t\}_{t=1}^{\infty}$ is said to be *weakly ergodic* if for all $i, j, s = 1, 2, \ldots, n$ and $r \geq 0$,

$$\left| [U_{r,k}]_{is} - [U_{r,k}]_{js} \right| \to 0$$

as $k \to \infty$.

On the other hand, we say that this very same sequence is *strongly ergodic* for all $r \geq 0$, and element-wise if

$$\lim_{k \to \infty} U_{r,k} = \mathbf{1} P_r^{\top},$$

where $\mathbf{1}$ is a vector of ones of size n and P_r is a probability vector in which $P_r \geq 0$ and $P_r^{\top} \mathbf{1} = 1$ for all $r \geq 0$.

Both weak and strong ergodicity (in the backward direction) describe a tendency to consensus. In the strong ergodicity case, all rows of the stochastic matrix $U_{r,k}$ are becoming the same as k grows large and reaching a stable limiting vector, whereas in the weak ergodicity case, every row is converging to the same vector, but each entry not necessarily converges to a limit. In our case, we can show that there is an equivalence between both concepts since each row of $U_{r,k+1}$ is a weighted average of the rows of $U_{r,k}$.

Lemma 7. *For the backward product* (13)*, weak and strong ergodicity are equivalent.*

Proof. Following Theorem 1 in Chatterjee and Seneta (1977), it suffices to show that weak ergodicity implies strong ergodicity. For that, we fix an arbitrary $r \geq 0$ and a small $\epsilon > 0$ and assume that there is a $k = \bar{k}$ such that W_k has the form $\mathbf{1} P^{\top}$, where P is a probability vector. Then by the definition of *weak* ergodicity we have that

$$-\epsilon \leq [U_{r,k}]_{is} - [U_{r,k}]_{js} \leq \epsilon \iff [U_{r,k}]_{is} - \epsilon \leq [U_{r,k}]_{js} \leq [U_{r,k}]_{is} + \epsilon$$

for $k \geq \bar{k}$ and for all $i, h, s = 1, \ldots, n$. Since each row of $U_{r,k+1}$ is a weighted average of the rows

of $U_{r,k}$, we have

$$\sum_{j=1}^{n} [W_{r+k+1}]_{hj} \left([U_{r,k}]_{is} - \epsilon \right) \leq \sum_{j=1}^{n} [W_{r+k+1}]_{hj} [U_{r,k}]_{js} \leq \sum_{j=1}^{n} [W_{r+k+1}]_{hj} \left([U_{r,k}]_{is} + \epsilon \right).$$

The inequality above shows that for any h and $k \geq \bar{k}$

$$[U_{r,k}]_{is} - \epsilon \leq [U_{r,k+1}]_{hs} \leq [U_{r,k}]_{is} + \epsilon.$$

Thus, by induction, for any $i, h, s = 1, 2, \ldots, n$, for any $k \geq \bar{k}$ and for any integer $q \geq 1$

$$\left| [U_{r,k+q}]_{js} - [U_{r,k}]_{is} \right| \leq \epsilon.$$

By setting $i = j$ in the expression above and taking $k \geq \bar{k}$, we see that $[U_{k,r}]_{i,s}$ converges to a limit as $k \to \infty$ for all s. □

Definition 4. *The scalar function $\tau(\cdot)$ continuous on the set of $n \times n$ stochastic matrices and satisfying $0 \leq \tau(\cdot) \leq 1$ is called coefficient of ergodicity. It is said to be proper if $\tau(W) = 0$ if and only if $W = \mathbf{1}v^\top$, where $v^\top$ is any probability vector (i.e. whenever W is a row-stochastic matrix with unit rank), and improper otherwise.*

Two examples of coefficients of ergodicity, in terms of W, drawn from Seneta (2006) p. 137 are

$$a(W) = \max_j \left(\max_{i,i'} \left| [W]_{ij} - [W]_{i'j} \right| \right)$$
$$c(W) = 1 - \max_j \left(\min_i [W]_{ij} \right),$$

where the first coefficient is proper and the second improper. Moreover, it can be shown that, i) for any stochastic matrix W, $a(W) \leq c(W)$ and ii) if $\tau(\cdot)$ is a proper coefficient of ergodicity, the inequality

$$\tau \left(W_m W_{m-1} \cdots W_2 W_1 \right) \leq \prod_{t=1}^{m} \tau \left(W_t \right) \tag{17}$$

is satisfied for any $m \geq 1$.[16] In this sense, for a proper coefficient of ergodicity $\tau(\cdot)$, weak ergodicity is equivalent to $\tau(U_{r,k}) \to 0$ as $k \to \infty$ and $r \geq 0$.

[16] More specifically, it can be shown that for any two proper coefficients of ergodicity $\tau_i(\cdot) \leq \tau_j(\cdot)$, the inequality holds with $\tau_i \left(W_m W_{m-1} \cdots W_2 W_1 \right) \leq \prod_{t=1}^{m} \tau_j \left(W_t \right)$.

Lemma 8. *The sequence $\{W_t\}_{t=1}^{\infty}$ is weakly ergodic if there exists a strictly increasing subsequence of the positive integers $\{i_x\}$, $x = 1, 2, \ldots$ such that*

$$\sum_{x=1}^{\infty} \left(1 - \tau(W_{i_{x+1}} \cdots W_{i_x+1})\right) \tag{18}$$

diverges.

Proof. Let $\theta_x = \tau(W_{i_{x+1}} \cdots W_{i_x+1})$. The standard inequality $z - 1 \geq \log z$ (or equivalently $1 - z \leq -\log z$) implies that $1 - \theta_x \leq -\log \theta_x$, for any x. Summing up across index x in both sides yields

$$\begin{aligned} \sum_{x=1}^{\infty} (1 - \theta_x) &\leq -\sum_{x=1}^{\infty} \log \theta_x \\ &\leq -\log \left(\prod_{x=1}^{\infty} \theta_x \right). \end{aligned} \tag{19}$$

Since equation (17) holds, the sum in the left hand side of equation (19) diverges, implying that $\log \left(\prod_{x=1}^{\infty} \theta_x\right) = -\infty$. For that, it must be the case that $\prod_{x=1}^{\infty} \theta_x \to 0$ as $x \to \infty$. Because $\tau(\cdot)$ is a proper coefficient of ergodicity, equation (17) ensures weak ergodicity of the sequence $\{W_t\}_{t=1}^{\infty}$

□

Appendix D: Proofs of propositions

Proof of proposition (1)

Proof. We start by considering the parameter update process described in eq. (5) of section (2). Since the network's edges are activated every single period (i.e. $\rho = 1$), $\hat{g}_t = \hat{g}$ and $B_t = B = \text{diag}(b, b, \ldots, b)$, where $b \in (0, 1)$. Moreover, since we are assuming strong connectivity, $\sum_j [g]_{ij} \neq 0$ for any i. Thus, the update process for the parameter vector α (of size n) in its matrix form is

$$\begin{aligned} \alpha_{t+1} &= B(\alpha_t + s_{t+1}) + (\mathbb{I}_n - B)\hat{g}\alpha_t \\ &= [B + (\mathbb{I}_n - B)\hat{g}]\,\alpha_t + Bs_{t+1}. \end{aligned}$$

We define the matrix inside the squared bracket as W for any t. We re-write the update process above as follows

$$\alpha_{t+1} = W\alpha_t + Bs_{t+1}$$

When $t = 0$,

$$\alpha_1 = W\alpha_0 + Bs_1$$

When $t = 1$,

$$\begin{aligned} \alpha_2 &= W\alpha_1 + Bs_2 \\ &= W\left(W\alpha_0 + Bs_1\right) + Bs_2 \\ &= W^2\alpha_0 + WBs_1 + Bs_2 \end{aligned}$$

When $t = 3$,

$$\begin{aligned} \alpha_3 &= W\alpha_2 + Bs_3 \\ &= W\left(W^2\alpha_0 + WBs_1 + Bs_2\right) + Bs_3 \\ &= W^3\alpha_0 + W^2Bs_1 + WBs_2 + Bs_3 \end{aligned}$$

So on and so forth, resulting in the following expression for any particular period τ

$$\alpha_\tau = W^\tau\alpha_0 + \sum_{t=0}^{\tau-1} W^t Bs_{\tau-t} \tag{20}$$

Similarly for the parameter β, we have

$$\beta_\tau = W^\tau \beta_0 + \sum_{t=0}^{\tau-1} W^t B(\mathbf{1} - s_{\tau-t}). \tag{21}$$

where $\mathbf{1}$ is the vector of ones of size n. From Equations (20) and (21), the sum of this two parameter-vectors is given by the following expression

$$\begin{aligned} \alpha_\tau + \beta_\tau &= W^\tau (\alpha_0 + \beta_0) + \sum_{t=0}^{\tau-1} W^t B\mathbf{1} \\ &= W^\tau (\alpha_0 + \beta_0) + \sum_{t=0}^{\tau-1} W^t \mathbf{b} \\ &= W^\tau (\alpha_0 + \beta_0) + \tau \mathbf{b}. \end{aligned} \tag{22}$$

Therefore, at any point in time τ, the opinion of any agent i is given by $y_{i,\tau} = \dfrac{\alpha_{i,\tau}}{\alpha_{i,\tau} + \beta_{i,\tau}}$. From equation (20), we write

$$\begin{aligned} \alpha_{i,\tau} &= W^\tau_{i*}\alpha_0 + \sum_{t=0}^{\tau-1} W^t_{i*} b s_{\tau-t} \\ &= W^\tau_{i*}\alpha_0 + \tau b \frac{1}{\tau} \sum_{t=0}^{\tau-1} W^t_{i*} s_{\tau-t} \\ &= W^\tau_{i*}\alpha_0 + \tau b \tilde{\theta}_i(\tau), \end{aligned} \tag{23}$$

where the symbol W^τ_{i*} is used to denote the i-th row of matrix W^τ and $W^0 = \mathbb{I}_n$. From equations (23) and (22), we write $y_{i,\tau}$ as

$$\begin{aligned} y_{i,\tau} &= \frac{W^\tau_{i*}\alpha_0 + \tau b \tilde{\theta}_i(\tau)}{W^\tau_{i*}(\alpha_0 + \beta_0) + \tau b} \\ &= \frac{\tau}{\tau} \left(\frac{\frac{1}{\tau} W^\tau_{i*}\alpha_0 + b\tilde{\theta}_i(\tau)}{\frac{1}{\tau} W^\tau_{i*}(\alpha_0 + \beta_0) + b} \right), \end{aligned} \tag{24}$$

From Equation (24), we have that the limiting opinion (in probability) of any agent i, at any

point in time τ, is described as

$$\begin{aligned}
\operatorname*{plim}_{\tau\to\infty} y_{i,\tau} &= \operatorname*{plim}_{\tau\to\infty} \tilde{\theta}_i(\tau) \\
&= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=0}^{\tau-1} W_{i*}^t s_{\tau-t} \\
&= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=0}^{\bar{\tau}} W_{i*}^t s_{\tau-t} + \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=\bar{\tau}+1}^{\tau} W_{i*}^t s_{\tau-t}. \qquad (25)
\end{aligned}$$

From Lemma (6), we can split the series in Equation (26) into two parts. The first term describes a series of $\bar{\tau}$ terms that represent the "most recent" signals coming in to the network. Notice that every weight-matrix W^t in the interval from $t = 0$ to $t = \bar{\tau}$ is different from one another, since the matrix W^t does not converge to a row-stochastic matrix with unity rank for low t. It is straight-forward to see that this term converges to zero as $\tau \to \infty$. The second term represents describes a series of $\tau - \bar{\tau}$ terms that represent the"older signals" that entered in the network and fully reached all agents. As $\tau \to \infty$, this term becomes a series with infinite terms. From the i.i.d. property of the Bernoulli signals, we can conclude that

$$\begin{aligned}
\operatorname*{plim}_{\tau\to\infty} y_{i,\tau} &= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=\bar{\tau}+1}^{\tau} W_{i*}^t s_{\tau-t} \\
&= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=\bar{\tau}+1}^{\tau} \boldsymbol{W}_{i*} s_{\tau-t} \\
&= \operatorname*{plim}_{\tau\to\infty} \boldsymbol{W}_{i*}\frac{1}{\tau}\sum_{t=\bar{\tau}+1}^{\tau} s_{\tau-t} \\
&\overset{\text{asy}}{=} \operatorname*{plim}_{\tau\to\infty} \boldsymbol{W}_{i*}\frac{1}{\tau-\bar{\tau}}\sum_{t=\bar{\tau}+1}^{\tau} s_{\tau-t} \\
&\overset{\text{asy}}{=} \boldsymbol{W}_{i*}\boldsymbol{\theta}^* = \theta^*, \qquad \text{(i.i.d. Bernoulli signals)} \qquad (26)
\end{aligned}$$

where $\boldsymbol{W} = D\mathbf{1}^\top$. From equation (26), we conclude that society is wise and because of that, $\operatorname{plim}_{t\to\infty} |\tilde{y}_{k,t} - \tilde{y}_{l,t}| = 0$, i.e. the K groups reach consensus, impliying $\operatorname{plim}_{t\to\infty} P_t = |\theta^* - \theta^*| = 0$. □

Proof of proposition (2)

Proof. The update process of both shape parameters can be represented in their matrix form for any period τ as

$$\alpha_\tau = U_{0,\tau-1}\alpha_0 + \left(\sum_{r=1}^{\tau-1} U_{r,\tau-1-r}B_{r-1}s_r\right) + B_{\tau-1}s_\tau, \tag{27}$$

$$\beta_\tau = U_{0,\tau-1}\beta_0 + \left(\sum_{r=1}^{\tau-1} U_{r,\tau-1-r}B_{r-1}\left(\mathbf{1}-s_r\right)\right) + B_{\tau-1}\left(\mathbf{1}-s_\tau\right). \tag{28}$$

To reduce the burden of notation, consider $\left[U_{r,k}\right]_{ij} = u_{ij}^{(r,k)}$ for any r and k. Therefore, from equation (27), we write its entries as

$$\begin{aligned}
\alpha_{i,\tau} &= \sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}s_{j,r}\right) + b_{i,\tau-1}s_{i,\tau} \\
&= \sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \tau\frac{1}{\tau}\left[\left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}s_{j,r}\right) + b_{i,\tau-1}s_{i,\tau}\right] \\
&= \sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \tau\tilde{\theta}_{i,1}(\tau).
\end{aligned} \tag{29}$$

Each entry of the parameter vector β is written in a similar way

$$\beta_{i,\tau} = \sum_j u_{ij}^{(0,\tau-1)}\beta_{j,0} + \left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}(1-s_{j,r})\right) + b_{i,\tau-1}(1-s_{i,\tau}).$$

The sum of both parameters $\alpha_{i,\tau}$ and $\beta_{i,\tau}$ yields

$$\begin{aligned}
\alpha_{i,\tau}+\beta_{i,\tau} &= \sum_j u_{ij}^{(0,\tau-1)}\left(\alpha_{j,0}+\beta_{j,0}\right) + \left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}\right) + b_{i,\tau-1} \\
&= \sum_j u_{ij}^{(0,\tau-1)}\left(\alpha_{j,0}+\beta_{j,0}\right) + \tau\frac{1}{\tau}\left[\left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}\right) + b_{i,\tau-1}\right] \\
&= \sum_j u_{ij}^{(0,\tau-1)}\left(\alpha_{j,0}+\beta_{j,0}\right) + \tau\tilde{\theta}_{i,2}(\tau).
\end{aligned} \tag{30}$$

In which $\sum_j u_{ij}^{(r,(\tau-1))} = 1$, for all $r \geq 0$ since $U_{r,k}$ is a stochastic matrix. Therefore, the opinion of each agent i in this society, at some particular time τ, is $y_{i,\tau} = \frac{\alpha_{i,\tau}}{\alpha_{i,\tau}+\beta_{i,\tau}}$, where each

entry of the parameter vectors can be written as follows:

$$y_{i,\tau} = \frac{\sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \tau\tilde{\theta}_{i,1}(\tau)}{\sum_j u_{ij}^{(0,\tau-1)}\left(\alpha_{j,0}+\beta_{j,0}\right) + \tau\tilde{\theta}_{i,2}(\tau)}$$

Asymptotically we have:

$$\begin{aligned}
\operatorname*{plim}_{\tau\to\infty} y_{i,\tau} &= \operatorname*{plim}_{\tau\to\infty}\left(\frac{\sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \tau\tilde{\theta}_{i,1}(\tau)}{\sum_j u_{ij}^{(0,\tau-1)}(\alpha_{j,0}+\beta_{j,0}) + \tau\tilde{\theta}_{i,2}(\tau)}\right) \\
&= \operatorname*{plim}_{\tau\to\infty}\frac{\tau}{\tau}\left(\frac{\frac{\sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0}}{\tau} + \tilde{\theta}_{i,1}(\tau)}{\frac{\sum_j u_{ij}^{(0,\tau-1)}(\alpha_{j,0}+\beta_{j,0})}{\tau} + \tilde{\theta}_{i,2}(\tau)}\right) \\
&= \operatorname*{plim}_{\tau\to\infty}\frac{\tilde{\theta}_{i,1}(\tau)}{\tilde{\theta}_{i,2}(\tau)}
\end{aligned} \tag{31}$$

With the results of Lemmas in Appendices B and C, weak law of large numbers and the assumption of independence between b_t and s_t we can can show that

$$\begin{aligned}
\operatorname*{plim}_{\tau\to\infty}\frac{\tilde{\theta}_{i,1}(\tau)}{\tilde{\theta}_{i,2}(\tau)} &= \operatorname*{plim}_{\tau\to\infty}\frac{\frac{1}{\tau}\sum_j\sum_{r=1}^{\tau-\bar{\tau}} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}s_{j,r}}{\frac{1}{\tau}\sum_j\sum_{r=1}^{\tau-\bar{\tau}} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}} \\
&= \operatorname*{plim}_{\tau\to\infty}\frac{\sum_j \bar{u}_{ij}\frac{1}{\tau}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1}s_{j,r}}{\sum_j \bar{u}_{ij}\frac{1}{\tau}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1}} \\
&\overset{\text{asy}}{\equiv} \operatorname*{plim}_{\tau\to\infty}\frac{\sum_j \bar{u}_{ij}\frac{1}{\tau-\bar{\tau}}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1}s_{j,r}}{\sum_j \bar{u}_{ij}\frac{1}{\tau-\bar{\tau}}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1}} \\
&= \frac{\sum_j \bar{u}_{ij}\mathbb{E}(b_j s_j)}{\sum_j \bar{u}_{ij}\mathbb{E}(b_j)} \\
&= \frac{\sum_j \bar{u}_{ij}\mathbb{E}(b_j)\mathbb{E}(s_j)}{\sum_j \bar{u}_{ij}\mathbb{E}(b_j)} \\
&= \frac{\theta^*\sum_j \bar{u}_{ij}\mathbb{E}(b_j)}{\sum_j \bar{u}_{ij}\mathbb{E}(b_j)} = \theta^*
\end{aligned} \tag{32}$$

□

Proof of proposition (3)

Proof. If perfect information aggregation is reached at any particular time $\bar{t}$, then we know that $y_{i,\bar{t}} = \theta$ for all $i \in G$, thus all alienation terms in the polarization function are zero because $|y_{i,\bar{t}} - y_{j,\bar{t}}| = |\theta - \theta| = 0$, for all i and j in N. Therefore, Polarization $P_{\bar{t}}$ is zero for any particular

choice of parameter a. Conversely, if polarization at time $\bar{t}$ is zero, then all alienation terms are necessarily zero, since the measure of groups is non-negative. This means that $|y_{i,\bar{t}} - y_{j,\bar{t}}| = 0$ implies $y_{i,\bar{t}} = y_{j,\bar{t}}$ and, therefore, any opinion consensus of the form $y_{i,\bar{t}} = y_{j,\bar{t}} = \tilde{\theta}$, such that $\tilde{\theta} \in \Theta = [0, 1]$ and $\tilde{\theta} \neq \theta$, meets this requirement. □