Social Media Networks, Fake News, and Polarization

Marina Azzimonti
Department of Economics
Stony Brook University
marina.azzimonti@gmail.com

Marcos Fernandes
Department of Economics
Stony Brook University
marcos.fernandes@stonybrook.edu

PRELIMINARY AND INCOMPLETE
This version: July 2017

Abstract

We study how the structure of social media networks affects the degree of polarization in society. We analyze a dynamic model of opinion formation in which individuals have imperfect information about the true state of the world and suffer from bounded rationality. Key to the analysis is the presence of 'internet bots' that communicate fake news, modeled as extremely biased opinions. 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 simulate a large set of random networks with different characteristics and quantify how the number of bots, their degree of centrality, and ability to spread fake news affect polarization in the long-run.

Keywords: Learning, Polarization, Social Networks, Social Media, Fake News

JEL Classification: C63, D83, D85

1 Introduction

The United States has experienced an unprecedented surge in political polarization over the last two decades. 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).

What could be causing this increase in polarization? Traditional theories in economics and political science point to the recent rise in income inequality, the influence of PACs through cam-

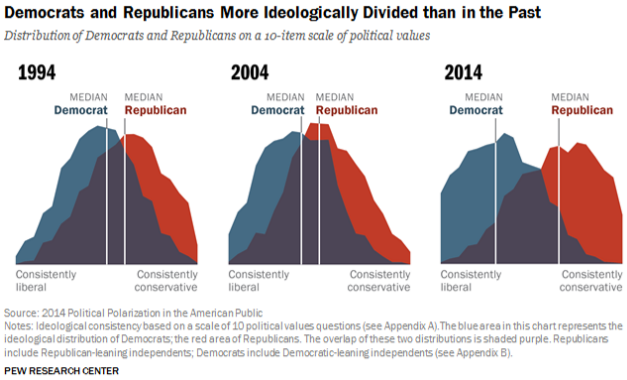


Figure 1: Political Polarization in the American Public (2014, Pew Research Center)

paign financing, party sorting among voters, re-districting (Gerrymanding), and changes in the media environment as potential determinants (see Barber and McCarthy, 2015 for an excellent discussion). More recently, attention has focused on the internet as an alternative candidate explanation. Cass Sunstein (2002) argues that the internet creates 'echo chambers' where individuals find their own biases and opinions endlessly reinforced, and writes that 'people restrict themselves to their own points of view—liberals watching and reading mostly or only liberals; moderates, moderates; conservatives, conservatives; Neo-Nazis, Neo-Nazis' (p. 5-6). This reduces the 'unplanned, unanticipated encounters central to democracy itself' and significantly increases polarization (p. 9).

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. With the dispersion of news through social media, and more generally the internet, and given 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 political polarization might be exacerbated by social media communication.

Another phenomena—of particular interest to the study of polarization—that became prevalent particularly around 2016 presidential election was the massive spread of *fake news* (also referred to as disinformation campaigns, cyber propaganda, cognitive hacking, and information

[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.

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, the arrival of the internet (particularly through social media) has made the spreading of ideas faster and more scalable, making it 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.

We study a dynamic model of opinion formation in the spirit of Jadbabaie, Molavi, Sandroni, and Tahbaz-Salehi (2012, JMST henceforth) in which individuals who are connected in a social network have imperfect information about the true state of the world. 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 (e.g. restrictions on immigration, imposition of tariffs, mandatory vaccination, etc.), the degree of government intervention (through the provision of public goods such as healthcare or education), etc. Individuals can obtain information (e.g. signals) about the true state of the world from unbiased sources (scientific studies, unbiased news media, reports from non-partisan research centers such as the CBO, etc.), but are unable process all the available information. They can also obtain information from their social neighbours (e.g. individuals connected to them through the network) who are potentially exposed to other sources.

Due to limited observability about the structure of the network and the probability distribution of signals observed by others, individuals would need to update opinions on the state of the world as well as on the topology of the network. This makes Bayesian updating complex and impractical. We assume instead that individuals suffer from bounded rationality, and update their opinions partly based on information obtained from their social network in an inhomogeneous stochastic gossip model of communication based on JMST(2012) and Acemoglu, Como, Fagnati, and Ozdaglar (2013, ACFO henceforth).

There are two types of individuals in this economy: *regular agents* and *internet bots*. Regular agents receive signals from unbiased sources and are also influenced by the opinion of their social neighbours. Internet bots, on the other hand, ignore the opinion of others and have the ability to produce *fake news*. The opinions generated from the exchange of information forms a Markov process which never leads to consensus among regular agents. In such environment, it can be shown that society's beliefs fail to converge almost surely. Moreover, under some conditions, the belief profile can fluctuate in an ergodic fashion leading to polarization cycles.

The structure of the graph representing the social media network and the degree of influence of bots in it shape the dynamics of opinion and the degree of polarization in the long-run. More specifically, long-run polarization depends on three factors: behavioral assumptions (e.g. the updating rule), communication technology (e.g. the speed at which information flows), and the network topology (e.g. the share of bots on the population, their centrality and ability to spread fake news, clustering among agents, etc.). Because a theoretical characterization of the relationship between the topology of the network and the degree of polarization is unfeasible, we simulate a large set of random networks with different characteristics. We then quantify how the degrees of centrality, connectedness, and influence affect long-run polarization, defined as in Esteban and Ray (1994) and Esteban (2007).

By connecting individuals through networks, social media provided a platform for individuals to share information in real time. On the flip side, it made it more difficult for them to assess the credibility of the content received allowing fake news to spread and contaminate the network. From our simulations, we find that to the extent that agents rely more heavily on the opinion of others (and less on incorporating signals from unbiased sources), polarization rises. The speed of communication, measured by the percentage of friends a given agent pays attention to at each point in time, *reduces* polarization, as agents are able to aggregate information even in the presence of bots. This is consistent with findings by Barbera (2015) who documents that the expansion of Twitter resulted in lower political polarization. As expected, a larger number of bots increases polarization, and so does an increase in their ability to stream fake news. In terms of the effects of network topology, we find that a higher degree of centrality of bots (measured by in-degree or PageRank) exacerbates polarization. However, as the ability of a given bot to reach most of the network (measured by high in-closeness centrality) rises, polarization declines. This happens because the internet bot is able to manipulate opinions more efficiently by pulling society's beliefs to one side of the political spectrum. Finally, we find that networks with a high degree of clustering and reciprocity tend to exhibit lower polarization in the long run, as greater

connectivity among agents facilitates consensus.

Related Literature Our paper is related to a growing number of articles studying information transmission in networks under both, bounded and fully rational agents.

The strand of the literature assuming that agents are fully rational typically considers a dynamic game where individuals interact sequentially and exchange opinions only once. Examples are Banerjee (1992), Smith and Sorensen (2000), Banerjee and Fundenberg (2004), and Acemoglu et. al (2011). Because the theoretical characterization of equilibria is complex, these papers restrict attention to very stylized networks. Moreover, they typically study environments in which society eventually reaches consensus, implying that polarization arises only in the short-run.

The strand of literature focusing on bounded rational agents (also referred to as 'De-Grootian') assumes that individuals follow simple heuristic rules to update beliefs. Examples are Ellison and Fundenberg (1993, 1995), Bala and Goyal (1998,2001), De Marzo et al (2003), Golub and Jackson (2010), and ACFO (2013). In these environments, long-run polarization arises in equilibrium because individuals receive information only once—at the outset of the initial period. There is no sense in which new information (form unbiased sources) may arrive and modify regular agents' opinions. JMST (2012) show that when this assumption is relaxed, that is, when individuals receive a constant flow of information, polarization eventually disappears. This occurs even though individuals are not fully Bayesian, but requires the network to be strongly connected (i.e. no internet bots are present).

In this paper, we consider simultaneously the possibility of learning from unbiased sources and being exposed to fake news spread by internet bots. As a result, our environment encompasses ACFO (2013) and JMST (2012) as special cases. We first show that their results can be replicated by an appropriate choice of parameters. That is, we can show that by shutting down the degree of influence of internet bots, regular agents eventually learn the truth. But if bots are influential, social media communication is not effective in aggregating information, and polarization persists in the long run. Our main contribution relative to the existing literature 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.

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 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).

2 Baseline Model

Agents and Information Structure The economy is composed by a finite number of agents $i \in N = \{1, 2, \ldots, n\}$ interacting through a social network. Individuals have imperfect information about the true state of the world θ belonging to a parameter space $\Theta = [0, 1]$. This parameter can be interpreted as the relative quality of two candidates, L and R, competing for office. A value of $\theta = 0$ would imply that candidate L is better suited for office, whereas $\theta = 1$ would imply that R is more qualified.

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

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

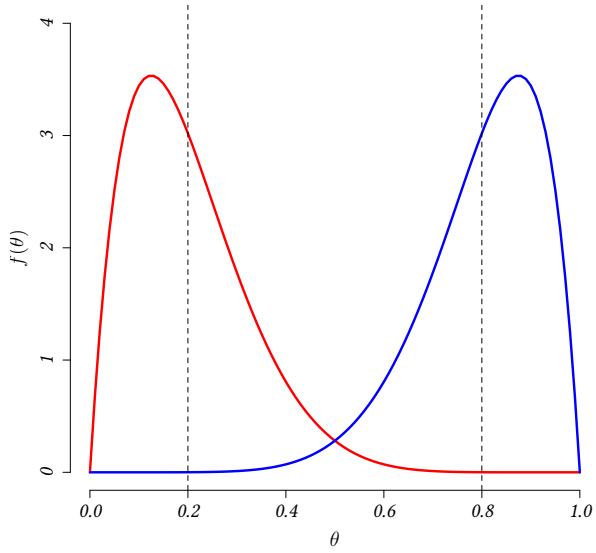
This distribution or *world-view* is characterized by initial parameters $\alpha_{i,0} > 0$ and $\beta_{i,0} > 0$. Note that individuals agree upon the parameter space Θ and the functional form of the probability distribution, but have different world-views as they disagree on $\alpha_{i,0}$ and $\beta_{i,0}$. Given prior beliefs, 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,[2]

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

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.*

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



At each point in time $t \geq 1$ regular agent i obtains information from unbiased sources that are jointly informative about the true state of the world. We formalize the information obtained from unbiased sources as a draw $s_{i,t}$ from a Bernoulli distribution with parameter θ. The signal is unbiased, as it is centered around the true state of the world.

Social Network We assume that regular agents update their world-views and opinions based not only on private signals $s_{i,t}$, but also through the influence of individuals connected to them in a social network.

The connectivity among agents in the network at each point in time is described by a directed graph $\mathbf{G}^t = (N, \mathbf{g}^t)$, where $\mathbf{g}^t$ is a real-valued $n \times n$ adjacency matrix. Each regular element g_{ij}^t in the directed-graph represents the connection between agents i and j at time t. More precisely, $g_{ij}^t = 1$ if i is paying attention to (e.g. receiving information from) j, and 0 otherwise. Since the graph is directed, it is possible that some agents pay attention to (e.g. receive information from) others who are not necessarily paying attention to (e.g. obtaining information from) them, i.e. $g_{ij}^t \neq g_{ji}^t$. The out-neighborhood of any agent i at any time t represents the set of agents that i is receiving information from (e.g. i's references), and is denoted by $N_i^{out}(\mathbf{g}^t) = \{j | g_{ij}^t = 1\}$. Similarly, the in-neighborhood of any agent i at any time t, denoted by $N_i^{in}(g^t)$, represents the set of agents that are receiving information from i (e.g. i's audience or followers), $N_i^{in}(\mathbf{g}^t) = \{j | g_{ji}^t = 1\}$. We define a directed path in $\mathbf{G}^t$ 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.

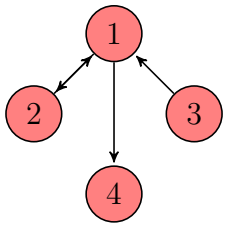
In the spirit of Acemoglu, Ozdaglar, and Parandeh Ghebi (2010) and ACFO (2012), we allow the connectivity of this graph g_{ij}^t to change over time stochastically. This structure captures rational inattention, incapacity of processing all information, or impossibility to pay attention to all individuals 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 then 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. The Bernoulli draws are represented by the $n \times n$ matrix $\mathbf{c}^t$, with regular element $c_{ij}^t \in \{0,1\}$. Thus, the adjacency matrix of the network evolves stochastically across time according to the following equation[3]:

$$\mathbf{g}^t = \mathbf{g}^0 \circ \mathbf{c}^t, \tag{1}$$

where the initial structure of the network, represented by the initial adjacency matrix $\mathbf{g}^0$, remains unchanged.

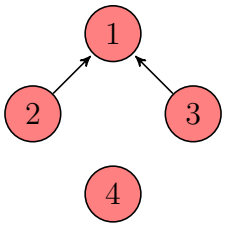
Example 2 (Bernoulli Clock)**.** *In this example we intend to illustrate the network dynamics. The figure in Panel 2a represents the original network and its adjacency matrix, whereas the figure in Panel 2b depicts a realization such that agent 1 does not pay attention to agents 2 and 4 in period* 1*.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

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



$$\mathbf{g}^0 = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(a) Original Network at $t = 0$



$$\mathbf{g}^1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(b) Potential Network at $t = 1$

Figure 2: Bernoulli Clock and Network Dynamics

Evolution of Beliefs Before the beginning of each period, agent i meets individuals in her out-neighbourhood $N_i^{out}(\mathbf{g}^t)$, a set determined by the realization of clock. These neighbors share their world-views, summarized by $\alpha_{j,t}$ and $\beta_{j,t}$ for all $j \in N_i^{out}(\mathbf{g}^t)$. At the beginning of period t, a signal profile is realized and the signal $s_{i,t}$ is privately observed by agent i.

Regular agents

After observing this signal from unbiased sources, regular agent i computes her Bayesian posterior conditional on $s_{i,t}$. We assume that parameters $\alpha_{i,t+1}$ and $\beta_{i,t+1}$ are convex combinations between her Bayesian posterior and the weighted average of the information obtained from her neighbors.

$$\alpha_{i,t+1} = b_{i,t}[\alpha_{i,t} + s_{i,t}] + (1 - b_{i,t}) \sum_{j \in N_i^{out}(\mathbf{g}^t)} \hat{g}_{ij}^t \alpha_{j,t} \tag{2}$$

$$\beta_{i,t+1} = b_{i,t}[\beta_{i,t} + 1 - s_{i,t}] + (1 - b_{i,t}) \sum_{j \in N_i^{out}(\mathbf{g}^t)} \hat{g}_{ij}^t \beta_{j,t}, \tag{3}$$

where

$$b_{i,t} = \mathbb{1}_{\{\sum_j \hat{g}_{ij}^t = 0\}} 1 + \left(1 - \mathbb{1}_{\{\sum_j \hat{g}_{ij}^t = 0\}}\right) b \tag{4}$$

denotes the reliance weight given to unbiased sources and $1 - b_{i,t}$ captures the influence of friends through social media. The parameter $b_{i,t} \in [0, 1]$ captures the attention span: a regular agent's full attention span is split between processing information from unbiased sources and that provided

by their friends in the network (e.g. reading a Facebook or Twitter feed). Equation (4) specifies that if no friends are found in the neighborhood of agent i, then this agent attaches weight 1 to the signal received. Conversely, if at least one friend is found, this agent uses a common weight $b \in [0, 1]$. The term $\hat{g}^t_{i,j} = \frac{g^t_{i,j}}{|N^{out}_i(\mathbf{g}^t)|}$ represents the weight given to the information received from her out-neighbor j. When $b_{i,t} = 1$ for some t agent i fully relies on her private signal behaving like a standard Bayesian agent. As $b_{i,t}$ approaches zero, she is more influenceable by social media, as more weight is given to her friends' opinions.

Finally, note that this updating rule implies that the posterior distribution determining world-views of agent i 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 rule resembles the one in JMST (2012), but there are two important distinctions. First, their adjacency matrix is fixed over time, whereas ours is stochastic (an element we borrowed from ACFO, 2013). Second, we restrict attention to a specific family of distributions (e.g. Beta) and assume that individuals exchange parameters that characterize this distribution (e.g. $\alpha_{i,t}$ and $\beta_{i,t}$). So the heuristic rule involves updating these parameters, whereas JMST (2012)'s heuristic rule involves a convex combination of the whole distribution. Given their rule, the posterior distribution may not belong to the same family as the prior distribution. That is not the case in our environment, as the posterior will also belong to the Beta distribution family.

Internet bots

We assume that there are two types of bots, L-bot and R-bot, with extreme views. Internet bot i disregards information from other agents in the network and has the ability to produce 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 at time t.

The parameter $\kappa \in \mathbb{N}^+$ measures the ability of bots to produce 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). When bumping into a regular agent in the network, bots transmit the whole stream of information to the agent. Hence, a value of $\kappa > 1$ gives them more de-facto weight in the updating rule of

agents, emphasizing their degree of influence on the network.[4]

To the extent that bots might be influential, their presence in the network will be key for both, the existence and persistence of polarization over time. This is due to the fact that they will consistently communicate fake news (biased signals) to other agents in the network and are ultimately the underlying force that pushes agents to the extreme of the political spectrum. Note that bots share similar characteristics with the 'stubborn' agents in ACFO (2013).

3 Polarization and Network Structure

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 amplifies polarization. Mathematically, we have the following representation.

Definition 1 (Polarization). *Polarization P_t aggregates the degrees of 'identification' and 'alienation' across groups at each point in time.*

$$P_t = \sum_{k=1}^{K} \sum_{l=1}^{K} \pi_{k,t}^{1+a} \, \pi_{l,t} \, |\tilde{y}_{k,t} - \tilde{y}_{l,t}| \tag{5}$$

[4]Under this rule, we can model the bot update as

$$\alpha_{i,t+1}^{p} = \alpha_{i,t}^{p} + \kappa s_{i,t}^{p}$$
$$\beta_{i,t+1}^{p} = \beta_{i,t}^{p} + \kappa - \kappa s_{i,t}^{p}.$$

where $a \in [0, 1.6]$ 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.

We are interested in understanding how the existence of bots and the structure of the network affect the evolution of polarization.

Polarization without Internet Bots The following two results show conditions under which polarization vanishes 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 $\boldsymbol{G}^0 = (N, \boldsymbol{g}^0)$ is strongly connected and if the directed links are activated every period (e.g., $\boldsymbol{g}^t = \boldsymbol{g}^0$), all agents eventually learn the true θ*

$$\max_i |\operatorname*{plim}_{t\to\infty} y_{i,t} - \theta| < \epsilon$$

As a consequence, polarization converges to zero,

$$\operatorname*{plim}_{t\to\infty} P_t = 0.$$

Proof. See Appendix A. □

When the network is strongly connected all opinions and signals eventually travel through the network allowing agents to perfectly aggregate information. Note that strong connectedness precludes the existence of bots, as these agents do not internalize other people's opinions. The proposition shows that the society reaches consensus (e.g. there is no polarization) and uncovers the true state of the world, θ. We refer to this as a 'wise' society, as defined below.

Definition 2 (Wise Society). *We say that a society is wise if*

$$\max_i |\operatorname*{plim}_{t\to\infty} y_{i,t} - \theta| < \epsilon.$$

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 $\mathbf{g}^t$ is not constant, polarization vanishes in strongly connected networks.

Proposition 2. *If the network $\boldsymbol{G}^0 = (N, \boldsymbol{g}^0)$ is strongly connected, even when the edges are not activated every period, polarization still converges to zero,* $\operatorname{plim}_{t\to\infty} P_t^a = 0$.

Proof. See Appendix B. □

Polarization with Internet Bots The presence of bots breaks the strong connectivity in the network, but this does not necessarily imply that the society will exhibit polarization. The following example depicts two networks, with three regular agents (2, 3, and 4) and one bot—L-bot in panel (a) and R-bot in the panel (b)—.

(a) Society with L-bot (b) Society with R-bot

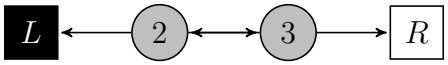
Figure 3: Two societies with internet bots

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 mis-information in the long run, preventing agents from uncovering θ, but does not necessarily create polarization. This insight is formalized in Proposition 3

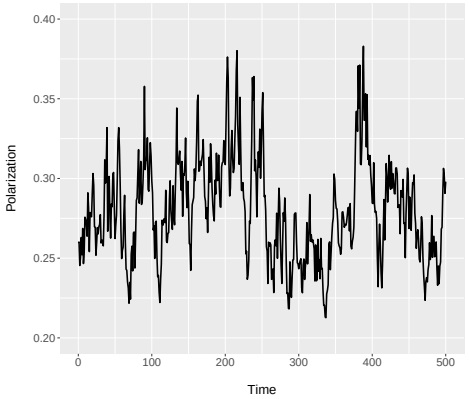
Proposition 3. *A wise society experiences null social polarization. However, not all societies that experience null social polarization are wise.*

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. □

In other words, it is possible for a society to reach consensus (i.e. experience no polarization of opinions) to a value of θ that is incorrect. In order for a society to be polarized, individuals need to be 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 depicted in Figure 4a, in which both L-bots and R-bots are present. Even though agents 2 and 3 receive unbiased signals and communicate with each other (e.g. update their beliefs according to eqs. 2 and 3), this society exhibits polarization in the long run. This happens because bots subject to different biases (e.g. L and R) are influential.

Another noticeable characteristic of the evolution of P_t over time 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.

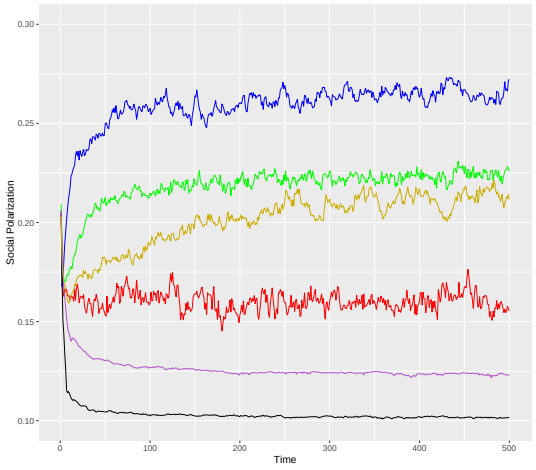


Figure 5: Different polarization levels

Finally, we want to point out that whether polarization increases, decreases, or fluctuates over time depends 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. 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. As we drop these assumptions, we observe that different networks might experience different limiting polarization levels, even if departing from the very same initial level of polarization.

To understand such differences, we resort to computer simulations where a large number of random networks is mainly generated according to a classical random graph model based on Barabasi and Albert (1999). Besides emulating real-world networks characteristics, this model allow us to create a variety of initial networks with different characteristics (e.g. different degrees of influence, etc) and learning standards (i.e., exposition to signals from unbiased sources and/or social media). The simulation exercise helps us to better understand the relative importance of the network topology and other social characteristics in driving polarization by producing enough variability in a controlled environment.

4.1 Network Topology: Augmented Barabasi-Albert Random Graph

Barabasi and Albert (1999) were mainly motivated by the emergence of the World Wide Web and the evolution of popularity of some web pages. They noted that popular web pages would show a tendency to get more popular over time. The popularity of web pages in this context refers to the number of other web pages pointing a direct link to them. This characteristic means that new entrant nodes (web pages) tend to link themselves to already existent nodes that are very well connected (popular web pages), indicating that the probability with which a new node connects to the existing nodes is not uniform. Contrarily, there is a higher probability that it will be linked to a node that already has a large number of connections. An implication of this characteristic is that a few nodes in the network are very well connected while most of the other nodes are not

as well connected and "hubs" are formed.

In this context, Barabasi and Albert (1999) developed an algorithm to generate random networks with such characteristics using a process called *preferential attachment*. In this process, starting with a small number n_0 of nodes, at every time step a new node with $m(\leq n_0)$ edges is added to the network. Thus, the new node links to m different nodes already present in the system. To incorporate preferential attachment, they assume that the probability Π that a new node will be connected to node i depends on the connectivity k_i (in-degree, or the number of nodes pointing to them) of that node, so that $\Pi = \frac{k_i}{\sum_j k_j}$. After t periods, this protocol leads to a random network with $t + n_0$ nodes and mt edges. Figure 6 illustrates a random network generated following this procedure. In it, there is a small subset of agents in the center of the network with a relatively large audience. In our context, each node represents an agent. Individuals with a larger audience are more influential.

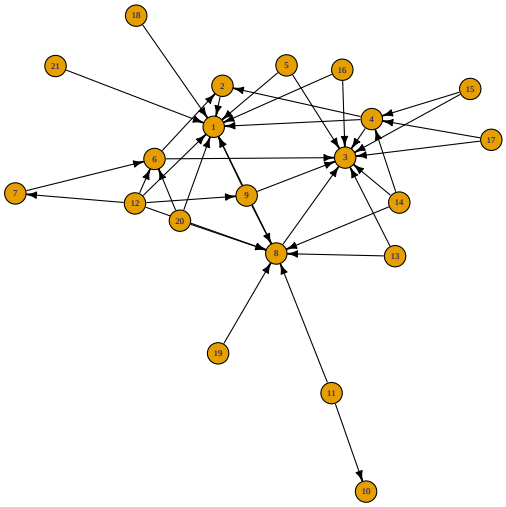


Figure 6: Barabasi-Albert
$n = 28$, Power= 1.5,
Out Dist = $\mathcal{G}(3, 1)$

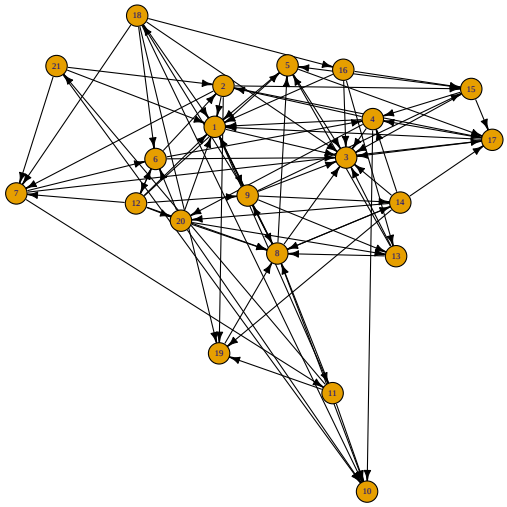


Figure 7: Barabasi-Albert w/ reciprocity
$n = 28$, Power= 1.5
Out Dist = $\mathcal{G}(3, 1)$

While this model allows us to introduce influential agents, it rules out (i) reciprocity, i.e. the chance that nodes are both paying attention to each other; and (ii) heterogeneity in the degree centrality of agents, both key characteristics of social media interactions. To capture this, we augment the Barabasi-Albert algorithm to introduce different degrees of reciprocity through the heterogeneity of connections. To produce networks with such characteristics, we implement two extra routines, respectively: (i) similarly to the Erdos-Renyi random graph model, we fix a set of nodes $\bar{n} \subseteq n$ in each simulation and assume that each link between two agents is formed with a

given probability. The link formation is independent across links; and (ii) in each step, instead of allowing only a fixed number $m(\leq n_0)$ of links to be formed, the number of edges of an entrant node at any time t is given by the realization of a draw from a gamma distribution defined over the potential number of edges to add in each time step. If this rule is not implemented, we would implicitly assume that every regular agent would be paying attention to exactly $m(\leq n_0)$ agents.

Figure 7 illustrates the network generated from the augmented model. In it, 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). This procedure usually produces networks in which regular agents, i.e. all agents but bots, are not strongly connected.[5] We disregard from our analysis the few instances in which strong connectivity is observed among these agents.

In addition, we select one agent, uniformly at random among agents $i \in N$ with $|N_i^{in}(\mathbf{g}^0)| \geq 1$ and $|N_i^{out}(\mathbf{g}^0)| \geq 1$, to be a bot. After such node is transformed into a bot, we break its outlinks. We repeat this routine sequentially until the pre-determined number of bots is reached (see next subsection). By proceeding this way, we guarantee that the bot is not disconnected of the resulting network and that it can be located in any node that displays at least a minimal amount of followers.

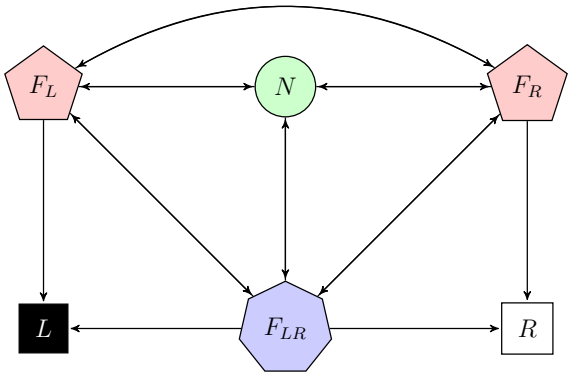


Figure 8: Possible

Finally, our algorithm produces agents which can be related to bots in four different ways, for any size n of the network. These are: Follower of L-bot (F_L), Follower of R-bot (F_R), Follower of both R-bot and R-bot simultaneously (F_{LR}) and Neutral agent N (e.g. someone who is not following either bot but is connected to other regular agents). Figure 8 depicts these agents in a hypothetical complete network characterized by exactly one agent of each type.

[5]This characteristic is basically driven by a small probability of link formation between two nodes.

4.2 Generating the dataset

The set of networks constituting our artificial dataset is created as follows. We fix the number of agents (or nodes) $n = 35$ and the average degree of polarization P^0 across networks. We restrict attention to networks with a symmetric number of L-bots and R-bots, considering $1, 2, 3$, or 4 bots of each type per network. Given these parameters, we draw a large number $M = 5435$ of initial random networks $\mathbf{G}^0$ following our augmentation of the Barabasi-Albert model. We then assign initial conditions and other characteristics that vary across networks. More specifically, we vary the speed of communication through the clock parameter ρ and the flooding parameter κ, the weight given to unbiased sources b, the location of bots in the network, the degree of initial homophily, clustering, and reciprocity as described in more detail below. This produces the basic structure for social communication and determines the initial dispersion of beliefs about θ. Finally, we simulate social media communication for a large number of periods ($T = 2000$) given the network structure, and use the resulting opinions to compute the evolution of polarization.

4.2.1 Network Heterogeneity

For each network m, we fix the initial distribution of opinions 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, as shown in Figure 9.

Figure 9: Initial distribution of opinions in all simulations

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). [6] Among the agents populating our network, a predetermined number of agents is chosen, uniformly at random among those with at least one individual in their in-neighborhood at $t = 0$, to be internet bots in each simulation. For symmetry, we simulate networks with the same number of bots of each type.

[6] 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 $\beta = \frac{(\sigma^2+\mu^2-\mu)(\mu-1)}{\sigma^2}$.

We also allow the degree of rationality b and the parameter ρ of the Bernoulli distribution determining the ticking of the clock (e.g. the persistence of connections) to vary across networks. We draw b_m and ρ_m from discrete Uniform distributions with $b_m \in [0, 1]$ and $\rho_m \in (0, 1]$ (with bandwith 0.05) for each network $m \leq M$. Note that we are excluding cases in which nodes are never activated, $\rho = 0$ as the network would exhibit no dynamics in such case. We vary the flooding parameter κ across simulations from the set $\kappa \in \{1, ..., 8, 10, 15, 20\}$.

It is informative to analyze the variability in network characteristics arising from our random network generation exercise. Network characteristics can be split in three categories: *behavioral, technological*, and *topological*.

Table 1: Network characteristics

	Mean	Std Dev.	Min	Max
Behavioral				
Weight unbiased source b	0.475	0.319	0	1
Technological				
Clock ρ	0.522	0.275	0.100	1
Flooding parameter κ	7.42	5.77	1	20
Topological				
in-Degree L-bot	0.170	0.142	0.024	0.917
in-Degree R-bot	0.170	0.140	0.024	0.861
Page Rank L-bot	0.027	0.021	0.005	0.212
Page Rank R-bot	0.027	0.021	0.005	0.237
in-Closeness L-bot	0.155	0.089	0.024	0.486
in-Closeness R-bot	0.155	0.089	0.024	0.473
% Following both bots	0.149	0.148	0	0.829
Reciprocity	0.070	0.043	0	0.257
Clustering	0.276	0.049	0.090	0.405
Homophily	−0.023	0.057	−0.263	0.231

The variability in the *behavioral* dimension is given by changes in the parameter b, capturing the degree to which agents rely more or less heavily on the opinion of others. Recall that a higher value of b gives more weight to the Bayesian posterior from unbiased signals. Table 1 shows that the parameter ranges between 0 (emulating De-Grootian agents) and 1 (emulating purely Bayesian agents), with an average value of b is 0.475 and a standard deviation of 0.319.

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. A higher value of ρ could be interpreted as higher speed of information flow, i.e. the easiness to access all friends opinions. We restrict this parameter to be strictly positive and no greater than 1. Given that it has been drawn from a uniform distribution, the average ρ in our sample is 0.52. A standard deviation of 0.275 ensures that there is significant variability in the artificial dataset. The average number of signals sent by each bot in our sample is 7.4, with a minimum of one (which is the number of signals sent by regular agents per encounter) and a maximum of 20, indicating that bots can place fake news 20 times faster than an unbiased source of news per period. The main difference between increasing κ vis-a-vis increasing the number of bots, is that the former keeps the location of bots in the network constant whereas the latter doesn't.

In terms of the network *topology*, of particular interest to us is the location of bots in the network, as this affects their degree of influence. The more central a bot is, the easier it is for it to spread fake news and manipulate regular agents' opinions. There are several ways in which centrality can be measured according to the literature. *Degree* is the simplest centrality measure, which consists on counting the number of neighbors an agent has. We focus on in-degree, defined as the number of incoming links to a given bot (their out-degree is null by construction). This measure is normalized by the size of the network (minus 1),

$$D_i^{\text{in}} = \frac{1}{n-1} \sum_j g_{ji}.$$

The average in-degree in our sample is 0.14, indicating that bots of a given type are followed, on average, by 14% of regular agents. There is a large dispersion across networks, with cases in which bots are being followed by around 90% of agents in the network.

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 polarization. There are networks in which a bot has very few followers (and hence a low in-degree) but each of their followers is very influential. If the bot manages to manipulate a regular agent who is itself very central, it may be able to affect the opinion of others to a large extent. An alternative measure of centrality that incorporates these indirect effects is Google's *PageRank* centrality.[7] 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).

[7]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

$$PR_i = \alpha \sum_j \frac{g_{ji}}{D_j^{\text{out}}} PR_j + \frac{1-\alpha}{n},$$

where D_j^{out} is the out-degree of node j if such degree is positive and α is the damping factor, set to $\alpha = 0.85$.[8] Note that the PageRank of bot i depends on the PageRank of its followers in the recursion above. Summary statistics for the average PageRank across bots of each type are shown in Table 1. As this is a more sophisticated version of degree centrality, it's correlation with in-degree is high (about 0.8 in our sample).

An alternative measure of centrality is given by *closeness* centrality. This measure keeps track of how close a given bot is to each other node in the network. High proximity to all 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 bot and every other agent in the network. Define d_{ji} as the length of the shortest path from regular agent j to bot i.[9] *In-closeness* centrality is defined as the inverse of the mean distance d_{ji} across regular agents to reach bot i,

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

Table 1 shows summary statistics for the in-closeness of each bot. It is worth noticing that even though in-degree and in-closeness are related measures of centrality, they capture slightly different concepts. A bot is central according to in-degree, because it has a large number of followers, whereas a bot is central according to in-closeness because it is easier for it to reach a large number of agents in the network. The correlation between these two variables is just 0.44. Moreover, there is disagreement between these two metrics when ranking the most influential bot in some of our networks.

Finally, different than the measures of centrality computed for the set of agents, we present three network statistics we judge relevant for our analysis. For these network statistics, we use the subgraph of $\mathbf{G}^0$ induced by the set of regular agents $\mathcal{R} = \{i \in N : |N_i^{out}(\mathbf{g}^0)| > 0\}$, i.e. all nodes, but bots. The reason we compute these statistics for this subgraph is because these measures would be under-reported if we had kept bots in the network. This is because bots do

[8] 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 0.85 is standard in the literature.

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

not connect to any agent in the network. Moreover, any comparison across simulations could be harmed by the average connectivity of each bot, implying that we could lose some degree of comparability of simulations, particularly when changing the number of bots across networks.

The first measure is *reciprocity*, which defines the proportion of mutual connections in a network. A nodes pair (i, j) is said to be reciprocal if there are edges between them in both directions. The reciprocity of a directed graph is the proportion of all possible pairs (i, j) which are reciprocal, provided there is at least one edge between i and j. In mathematical terms

$$R(m) = \frac{\sum_i \sum_j (\mathbf{g} \circ \mathbf{g}')_{ij}}{\sum_i \sum_j g_{ij}}$$

Even though reciprocity is bounded between 0 and 1, it is possible to see that our data generation process only allow reciprocity to reach the maximum value of 0.257. This is mainly because we have not strongly connected networks and this measure could never reach 1 in our data. On average, this metric tells us that around 7% of the connections are mutual, with standard deviation of 4.3%.

Another important aspect of a social networks is how tightly clustered they are. Many empirical networks display an inherent tendency to form circles in which one's friends are friends with each other. In order to assess clustering in our directed networks, we use an extension to directed graphs of the clustering coefficient proposed by Fagiolo (2007). This quantity is defined as the average, over all nodes i, of the nodes-specific clustering coefficients and is defined as follows

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

where D_i^{tot} is the total degree, i.e. in-degree plus out-degree, of agent i. In our data, we observe an average clustering of 0.276, with standard deviation of 0.049.

Finally, many social networks exhibit a characteristic named homophily. This concept refers to the fact that people are more prone to maintain relationships with people who are similar to themselves. In our context, the level of initial homophily of opinions of agents is 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[10].

[10] In political science and economic networks literatures, homophily is a characteristic that drives link formation. In our case, initial homophily is simply a statistic of assortativity computed over opinions after the initial random network is fully characterized and populated with different agents and beliefs. 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.

Homophily is an important aspect of social networks since it might be related to the degree of political polarization. Our simulated societies display an average of homophily of -0.023, with standard deviation of 0.057.

4.2.2 Simulation

We fix the true state of the world, $\theta = 0.5$. For each network m, we draw a signal $s_{i,t}^m$ for individual $i \in N$ at time $t \in T$ from a Bernoulli distribution with parameter $\theta = 0.5$ for regular agents, or parameters $\theta^L = 0$ for L-bots and $\theta^R = 1$ for R-bots. We also draw the $n \times n$ matrix $\mathbf{c}^t$ at each period t from a Bernoulli distribution with parameter ρ_m, 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. (2) and (3). Polarization $P_{m,t}$ is computed according to eq. (5) at each point in time, assuming a parameter $a = 0.5$. Our variable of interest is the level of polarization in the long-run, $\bar{P}_m$, which is normalized to belong to the interval $[0, 1]$.[11] This is defined as the average value of $P_{m,t}$ for t larger than a threshold $\bar{t}$, which in our simulations is set to be $\bar{t} = 500$. We chose 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). Mathematically, for each network m we compute

$$\bar{P}_m \equiv \sum_{t>\bar{t}} \frac{1}{T - \bar{t} - 1} P_{m,t}.$$

It is important to recall that this measure of polarization is computed from the limiting opinion of regular agents (that is, we do not consider the opinion of bots). This is done in order to allow a fair comparison of polarization levels across networks with different number of bots; if bots' opinions were included, we would mechanically rise polarization as we rise the number of bots.

The resulting sample consists of 5435 networks with associated observations for polarization $\bar{P}_m$ (where m indicates a particular network). About 91% of the sample exhibits positive polarization levels $\bar{P}_m > 0$, while the remaining 9% is composed of networks in which agents have weight parameters b close to 1 (so their opinions eventually converge to the true state of the world implying that $\bar{P}_m = 0$). Figure 10 depicts the distribution of polarization (conditioned on being positive) in our sample. There is a significant degree of variability in our sample, even though the polarization levels are relatively small, with most $\bar{P}_m$ observations lying below 0.5

[11] With $a = 0.5$ the maximum possible level of polarization is around 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.

(recall that maximum polarization has been normalized to 1 while the maximum polarization level in our sample is 0.63). The average value of $\bar{P}_m$ across networks is 0.113, with a standard deviation of 0.09. Interestingly, we also observe some mass near 0, indicating that agents reach quasi-consensus (e.g. disagreement is relatively small among regular agents) about θ in 3% of the networks.

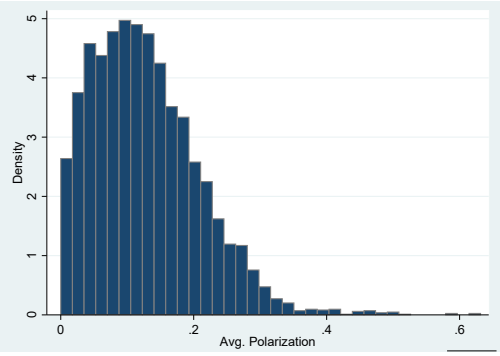


Figure 10: Average polarization

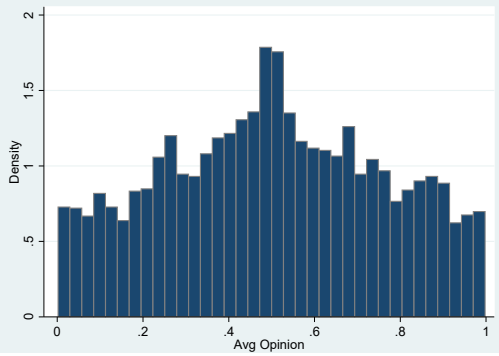


Figure 11: Average opinions

The distribution of opinions across networks that exhibit positive polarization (also computed over the last 500 periods) is shown in Figure 11. Two remarks are in place. First, we see that the distribution centers around 0.5. This implies that there is a non-negligible set of networks in which individuals' opinions about the true state of the world are –on average—correct, even though there is some disagreement among regular agents. Second, there is a non-trivial amount of networks in which agents' opinions become extreme on average. This indicates that there are cases in which bots are successful at manipulating options towards their own, as even though there is some disagreement, average beliefs are centered around 0 or 1.

5 Regression Analysis

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

$$\bar{P}_m = \mathbf{X}_m \beta + \epsilon_m.$$

where the $m \times 1$ vector $\bar{P}_m$ denotes long-run polarization obtained from simulation $m \in \{1 \dots M\}$, $\mathbf{X}_m$ denotes the matrix of network characteristics per simulation m, and ϵ_m is the error term.

The set of explanatory variables in our benchmark specification is composed of the behavioral, technological, and topological characteristics of our networks listed in Table 1. Estimated coefficients are reported in Table 2.

Table 2: Regression results

	Dependent Variable: Average Polarization $\bar{P}_m$			
	in-Degree, $\kappa = 1$	in-Degree	PageRank	+ PR Aud
	(1)	(2)	(3)	(4)
Weight unbiased source b	-0.078***	-0.065***	-0.065***	-0.065***
	(0.007)	(0.003)	(0.003)	(0.003)
Clock ρ	-0.081***	-0.080***	-0.082***	-0.082***
	(0.007)	(0.003)	(0.003)	(0.003)
Degree/PageRank bot	0.096***	0.091***	0.707***	0.680***
	(0.022)	(0.009)	(0.067)	(0.068)
% Following both bots F_{LR}	0.083***	0.048***	0.167**	0.267***
	(0.026)	(0.0099)	(0.076)	(0.082)
in-Closeness bot	-0.045	-0.187***	-0.165***	-0.178***
	(0.045)	(0.025)	(0.025)	(0.024)
Initial Homophily	0.039	-0.011	-0.016	-0.015
	(0.045)	(0.016)	(0.016)	(0.016)
Reciprocity	-0.194***	-0.258***	-0.203***	-0.216***
	(0.0455)	(0.0223)	(0.0228)	(0.0229)
Cluster of Direction	-0.395***	-0.352***	-0.249***	-0.246***
	(0.042)	(0.017)	(0.018)	(0.018)
Proportion of type i bots				
$2/35$	0.033***	0.009***	0.012***	0.009***
	(0.006)	(0.003)	(0.003)	(0.003)
$3/35$	0.061***	0.019***	0.027***	0.022***
	(0.007)	(0.004)	(0.004)	(0.004)
$4/35$	0.074***	0.030***	0.040***	0.034***
	(0.008)	(0.005)	(0.005)	(0.005)
PageRank Audience L-bot				0.283***
				(0.073)
PageRank Audience R-bot				0.297***
				(0.066)
κ	$= 1$	≥ 1	≥ 1	≥ 1
Observations	752	4,816	4,816	4,816
R-squared	0.63	0.55	0.53	0.53

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

In the first column, we consider the case in which each bot can send at most one message by restricting the sample to networks in which $\kappa = 1$. The negative coefficient on b implies that as regular agents place more weight on the unbiased signal (and less on the social media friends' opinions), polarization falls. In other words, agents are able to partially "mute" the network channel through which fake news permeate, facilitating information aggregation and reducing polarization as b raises. The overall effect of a higher clock parameter ρ is a priori ambiguous: on the one hand, it is more likely that a regular agent will (indirectly) 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 makes it more likely to form consensus among

regular agents. Under the current specification, we find that the elasticity of polarization to the parameter ρ (evaluated at their average values $\bar{\rho}$ and $\bar{P}_m$) is $\frac{\partial \bar{P}_m}{\partial \bar{\rho}} \frac{\bar{\rho}}{\bar{P}_m} = -0.37$. This result suggests that the effect of internalizing a larger number of opinions outweighs the effect of higher fake news exposure.

As the degree of influence of bots— proxied by their in-degree— rises, polarization is exacerbated. This follows from the the positive and significant coefficient on degree centrality, reported only for the R-bot (the L-bot is basically identical due to symmetry). The elasticity of polarization to the number of followers, computed at their mean values is $\frac{\partial \bar{P}_m}{\partial \bar{D}^{\text{in}}} \frac{\bar{D}^{\text{in}}}{\bar{P}_m} = 0.17$. It is also interesting to note that as the percentage of individuals following both bots, A_{LR} rises, polarization increases as well. This happens because only a subset of all the notes in a given agents' in-neighborhood are activated at each point in time. In a relatively small network, this means that individuals will be exposed at opposing extreme views over time causing fluctuations in their beliefs.[12] The coefficient of in-closeness is insignificant in this specification, a result that is not robust to networks in which bots can send more than one message. Reciprocity decreases polarization, as expected from that fact that it facilitates consensus between any two agents exchanging information. 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 polarization vanish over time, as seen by the fact that the coefficient is statistically insignificant. As expected, as we increase the percentage of bots in the population, polarization among regular agents raises.

In Specification (2), we also consider networks in which bots can send multiple fake-news articles each period (e.g. by allowing $\kappa > 1$). The number observations raises significantly (from 752 to 4816) as we consider values of $\kappa \in \{1, ..., 8, 10, 15, 20\}$. 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 (with the exception of $\kappa = 1$, which is the reference value). To ease readability, we plot the resulting coefficients in Figure (12).

[12] A discussion of the determinants of polarization cycles is deferred to future work.

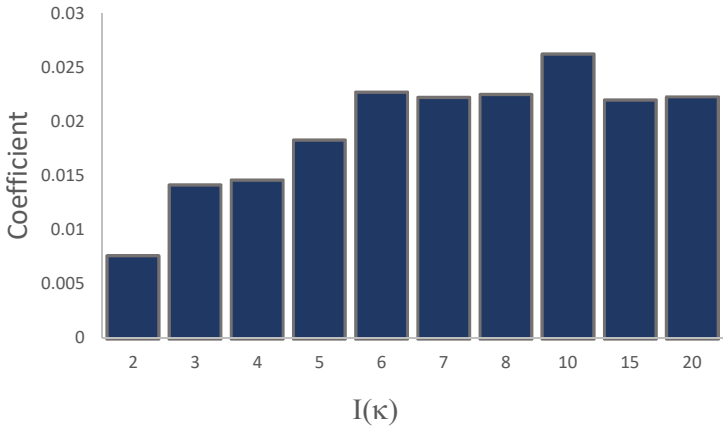


Figure 12: Estimated coefficient on the indicator $I(\kappa)$ (using Specification 2).

As evident from the graph, all coefficients are positive, indicating a greater ability to spread fake news by each bot, keeping everything else constant, results in greater polarization levels. It is worth noting that all coefficients are significant with p-values lower than 1%. In addition, note that the larger effects on polarization are observed for relatively small values of κ, with the effects remaining more or less stable for $\kappa > 6$. This suggests decreasing marginal returns to the introduction of fake-news on polarization levels (although the returns on tilting opinions could well be increasing).

The effects of bayes, clock, homophily, in-degree, and clustering are basically unchanged once the full set of κs is included. One important difference relative to the first specification arises on the significance of the in-closeness coefficient. Interestingly, higher centrality as measured by in-closeness dampens polarization, as suggested by the negative sign of the estimated coefficient. We interpret this as suggesting that the speed at which each given piece of fake-news travels through the network allows the bot to effectively manipulate opinions pulling individuals towards their preferred point. Note that this is very different from the effect of in-degree: a larger number of followers increases polarization whereas greater proximity to most agents in the network decreases it. A second difference between this specification and the previous one is that the effect of the number of bots is smaller. This is intuitive, as each bot is now able to send several messages, so the marginal effect of a given bot is now lower.

5.1 Robustness

In this sub-section, we would like to study the robustness of the results presented so far.

PageRank: We first investigate whether the effects of centrality are robust to our measure

of degree centrality. In Specification (3) of Table 2, we replace in-Degree by PageRank as an explanatory variable and instead of considering the $\%$ of followers to both, we use the page rank of agents connected to both bots in $t = 0$. The basic message remains unchanged: the more influential the bot is, the greater the polarization it introduces when spreading fake news. The size of the coefficient is significantly larger, but the elasticity is about the same (0.13 under this specification versus 0.17 in the previous one). The magnitude of the estimated coefficients of all other explanatory variables are basically unchanged, indicating that the results are robust to alternative degree centrality measures. The goodness of fit decreases slightly from 0.55 in Specification (1) to 0.53 in Specification (2).

In Specification (4) we include, in addition to all the regressors from Specification (3), the page rank of individuals in the in-degree of each bot. The coefficient is positive and significant, suggesting that a bot with low page-rank can still affect polarization if the average page-rank of regular agents connected to it is large.

Bot Asymmetry: The previous specification studied the effects of network characteristics on polarization considering the centrality of each bot separately. In this sub-section, we want to consider the effects of their *relative* centrality. To that end, we construct two auxiliary variables. The first one is the Relative PageRank, defined as the absolute difference in the page rank of L-bot and R-bots

$$RelPageRank = |PageRank(L) - PageRank(R)|.$$

The second one is Relative in-Closeness, defined in a similar way, as the absolute difference in the in-closeness measure of the two types of bot. Using a specification similar to (3), in which individual PageRank and in-Closeness are substituted by relative ones, we find that the coefficient on relative page rank is statistically insignificant. That is, polarization is unresponsive to increases in the relative number of followers of a given bot. The coefficient of relative in-closeness is negative. This suggests that as a bot gets relatively closer to the rest of the network, he becomes more efficient at spreading fake-news decreasing polarization levels. Results are omitted due to space constraints but are available upon request.

True State: Here, we study whether assuming $\theta = 0.5$ has consequences for the size or sign of estimated coefficients. To that end, we simulate an additional 4317 networks using the same procedure as before, but assuming that the true state of the world is $\theta = 0.7$ instead. The first column of Table 3 replicates the results from Specification (3) above (e.g. when $theta = 0.5$) whereas the second column displays the result for the sample where $\theta = 0.7$.

Table 3: Regression results

Dependent Variable: Average Polarization P_m		
	$\theta = 0.5$	$\theta = 0.7$
Weight unbiased source b	-0.065***	-0.072***
	(0.003)	(0.004)
Clock ρ	-0.082***	-0.086***
	(0.003)	(0.004)
PageRank L-bot	0.703***	0.763***
	(0.0627)	(0.0691)
PageRank R-bot	0.680***	0.427***
	(0.068)	(0.062)
PageRank F_L	0.283***	0.270***
	(0.0726)	(0.0740)
PageRank F_R	0.297***	0.513***
	(0.065)	(0.069)
Page Rank F_{LR}	0.267***	0.586***
	(0.082)	(0.113)
in-Closeness L-bot	-0.206***	-0.154***
	(0.0256)	(0.0217)
in-Closeness R-bot	-0.178***	-0.147***
	(0.024)	(0.023)
Reciprocity	-0.216***	-0.281***
	(0.023)	(0.024)
Cluster of Direction	-0.246***	-0.190***
	(0.018)	(0.019)
Proportion of type i bots		
$2/35$	0.009***	0.014***
	(0.003)	(0.004)
$3/35$	0.022***	0.037***
	(0.004)	(0.005)
$4/35$	0.034***	0.049***
	(0.005)	(0.005)
κ	≥ 1	≥ 1
Observations	4,816	4,317
R-squared	0.53	0.54

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

The only significant difference is on the magnitude of coefficient on PageRank of the R-bot, which is now smaller (this is robust to excluding the page rank R-bot's audience for the regression equation). Overall, the results presented in the benchmark case are robust to having one bot's messages being closer to the true state of the world (i.e. news being less fake).

Network Size: Finally, we wanted to analyze the effects of network size on our results. We run an additional 1501 simulations using the same procedure but considering twice the number of nodes, $n = 70$. We reduced the set of κ to $\{1, 5, 10, 15, 20\}$ and the number of fanatics of each type to 1, 2, and 3. Summary statistics for the dataset with large networks can be found in Table 5 at Appendix C. Average polarization declines from $\bar{P}_m = 0.12$ when $n = 35$ (and the number of fanatics and κ are restricted to the sets above) to $\bar{P}_m = 0.09$ when $n = 70$.

We estimate specifications (2) and (3) from Table 2 for small ($n = 35$) and large ($n = 70$)

networks. The resulting coefficients are displayed in Table 4. Note that the sample size for small networks is smaller than in our benchmark case due to the fact that we restricted attention to networks with comparable κ and number of fanatics to the ones in large networks.

Table 4: Regression results

Dependent Variable: Average Polarization $\bar{P}_m$				
	Small $n = 35$	Large $n = 70$	Small $n = 35$	Large $n = 70$
Weight unbiased source b	-0.064***	-0.053***	-0.065***	-0.053***
	(0.005)	(0.005)	(0.005)	(0.005)
clock	-0.077***	-0.078***	-0.078***	-0.080***
	(0.005)	(0.006)	(0.005)	(0.006)
Degree/PageRank L-bot	0.086***	0.055**	0.725***	0.869***
	(0.012)	(0.024)	(0.083)	(0.203)
Degree/PageRank R-bot	0.108***	0.050**	0.648***	0.574***
	(0.013)	(0.024)	(0.098)	(0.175)
% Following both bots F_{LR}	0.052***	0.162***	0.182*	0.349**
	(0.015)	(0.026)	(0.107)	(0.171)
in-Closeness L-bot	-0.165***	-0.234***	-0.188***	-0.213**
	(0.035)	(0.090)	(0.036)	(0.091)
in-Closeness R-bot	-0.220***	-0.445***	-0.174***	-0.362***
	(0.034)	(0.101)	(0.035)	(0.102)
Initial Homophily	-0.0038	0.059	-0.014	0.070*
	(0.024)	(0.036)	(0.025)	(0.037)
Reciprocity	-0.203***	-0.352***	-0.142***	-0.314***
	(0.032)	(0.070)	(0.033)	(0.072)
Cluster of Direction	-0.301***	-0.320***	-0.199***	-0.269***
	(0.025)	(0.032)	(0.026)	(0.034)
Proportion of type i bots				
$2/35$	0.004	-0.011	0.008*	0.007
	(0.004)	(0.007)	(0.004)	(0.007)
$4/35$	0.022***	0.002	0.033***	0.047***
	(0.006)	(0.012)	(0.006)	(0.012)
κ	≥ 1	≥ 1	≥ 1	≥ 1
Observations	2,337	1,501	2,337	1,501
R-squared	0.53	0.55	0.51	0.50

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

The effects of the weighting parameter b and the clock ρ on polarization are the same regardless of network size. Interestingly, the effects of in-degree decrease sharply whereas the effects of in-closeness increase significantly when we move from $n = 35$ to $n = 70$. This implies that having a larger set of followers is less important in driving polarization when networks are large. Being closer to most followers by one unit, on the other hand, proves more important in manipulating opinions when $n = 70$. Another characteristic that gains importance in large networks is the degree of reciprocity, which also becomes more relevant in reducing polarization when networks are large. These are not the result of differences in network characteristics, as the elasticities of reciprocity and in-closeness are more negative when $n = 70$ than when $n = 35$.[13]

[13]The elasticity of reciprocity is -0.12 in small networks and -0.18 in large networks. The elasticity of in-

6 Conclusions

We simulated a large number of social media networks by varying their characteristics in order to understand what the most important drivers of polarization are. A premise in all of them is the presence of bots with opposite extreme views who purposely spread fake news in order to manipulate the opinion of other agents. To the extent that regular agents can be partially influenced by these signals—directly by 'following' the bot, or indirectly by following friends who are themselves influenced by fake news—, this generates polarization in the long run. In other words, fake news prevent information aggregation and consensus in the population.

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. In addition, they could 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.

closeness is -0.22 for small networks and -0.31 for large networks.

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.

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.

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.

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.

A Proof of Proposition 1

Lemma 1. *The matrix* $W_t = B_t + (\mathbb{I}_n - B_t)\,\hat{g}_t$ *is row-stochastic in any period* t*, where* $B_t = diag(b_{1,t}, b_{2,t}, \ldots, b_{n,t})$.

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 the vector $W_t\mathbf{1}$, for every t, has all entries equal to

$$b_{i,t} + (1 - b_{i,t})\hat{g}^t_{i,*}\mathbf{1} = \begin{cases} b_{i,t} = \mathbb{1}_{\{\hat{g}^t_{i,*}\mathbf{1}=0\}}1 + \left(1 - \mathbb{1}_{\{\hat{g}^t_{i,*}\mathbf{1}=0\}}\right) b = 1 & \text{, if } \hat{g}^t_{i,*}\mathbf{1} = 0 \\ 1 & \text{, if } \hat{g}^t_{i,*}\mathbf{1} = 1 \end{cases}$$

, where $\hat{g}^t_{i,*}$ is the i-th row of matrix $\hat{g}^t$. □

Lemma 2. *The iteration of the row-stochastic matrix* W *is convergent and therefore there exists a threshold* $\bar{\tau} \in \mathbb{N}$ *such that* $|W^{\tau+1}_{ij} - W^{\tau}_{ij}| < \epsilon$ *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 $\boldsymbol{D} = (d_1, d_2, \ldots, d_n)'$ the eigenvector of size n associated to the unity eigenvalue $\lambda_1 = 1$[14]. Without loss of generality, we assume the following normalization $\mathbf{1'}\boldsymbol{D} = 1$. 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 genreally, 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

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

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

[14] 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.

With these two auxiliary lemmas, we start by considering the parameter update process described in the Section (XX). Since the network's edges are activated every single period, $\hat{g}_t = \hat{g}$ and $B_t = B_{n\times n} = B = \text{diag}(b, b, \ldots, b)$, where $b \in [0, 1]$, since $\sum_j g_{ij}^t \neq 0$ for any i and t. Thus, the update process for the parameter-vector α of size n 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

$$\alpha_\tau = W^\tau\alpha_0 + \sum_{t=0}^{\tau-1} W^t Bs_{\tau-t} \tag{6}$$

Similarly for the parameter β, we have

$$\beta_\tau = W^\tau\beta_0 + \sum_{t=0}^{\tau-1} W^t B(\mathbf{1} - s_{\tau-t}). \tag{7}$$

where **1** is the vector of ones of size n. From Equations (6) and (7), the sum of this two parameter-vectors is given by the following expression

$$\begin{aligned}\alpha_\tau + \beta_\tau &= W^\tau(\alpha_0+\beta_0) + \sum_{t=0}^{\tau-1} W^t B\mathbf{1} \\ &= W^\tau(\alpha_0+\beta_0) + \sum_{t=0}^{\tau-1} W^t \mathbf{b} \\ &= W^\tau(\alpha_0+\beta_0) + \tau\mathbf{b}. \end{aligned} \tag{8}$$

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 (6), we write

$$\begin{aligned}\alpha_{i,\tau} &= W^\tau_{i*}\alpha_0 + \sum_{t=0}^{\tau-1} W^t_{i*} b s_{\tau-t} \\ &= W^\tau_{i*}\alpha_0 + \tau b \frac{1}{\tau}\sum_{t=0}^{\tau-1} W^t_{i*} s_{\tau-t} \\ &= W^\tau_{i*}\alpha_0 + \tau b\tilde{\theta}_i(\tau), \end{aligned} \tag{9}$$

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 (9) and (8), we write $y_{i,\tau}$ as

$$\begin{aligned} y_{i,\tau} &= \frac{W^\tau_{i*}\alpha_0 + \tau b\tilde{\theta}_i(\tau)}{W^\tau_{i*}(\alpha_0+\beta_0) + \tau b} \\ &= \frac{\tau}{\tau}\left(\frac{\frac{1}{\tau}W^\tau_{i*}\alpha_0 + b\tilde{\theta}_i(\tau)}{\frac{1}{\tau}W^\tau_{i*}(\alpha_0+\beta_0) + b}\right), \end{aligned} \tag{10}$$

From Equation (10), we have that the limiting opinion (in probability) of any agent i, at any point in time τ, is described as

$$\begin{aligned}\operatorname*{plim}_{\tau\to\infty} y_{i,\tau} &= \operatorname*{plim}_{\tau\to\infty} \tilde{\theta}_i(\tau) \\ &= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=0}^{\tau-1} W^t_{i*} s_{\tau-t} \\ &= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=0}^{\bar{\tau}} W^t_{i*} s_{\tau-t} + \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau}\sum_{t=\bar{\tau}+1}^{\tau} W^t_{i*} s_{\tau-t}. \end{aligned} \tag{11}$$

From Lemma 2, we can split the series in Equation (12) 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

$$\begin{aligned}
\operatorname*{plim}_{\tau\to\infty} y_{i,\tau} &= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau} \sum_{t=\bar{\tau}+1}^{\tau} W_{i*}^{t} s_{\tau-t} \\
&= \operatorname*{plim}_{\tau\to\infty} \frac{1}{\tau} \sum_{t=\bar{\tau}+1}^{\tau} \boldsymbol{W}_{i*} s_{\tau-t} \quad \text{(by Lemma 2)} \\
&= \operatorname*{plim}_{\tau\to\infty} \boldsymbol{W}_{i*} \frac{1}{\tau} \sum_{t=\bar{\tau}+1}^{\tau} s_{\tau-t} \\
&\overset{\text{asy}}{=} \operatorname*{plim}_{\tau\to\infty} \boldsymbol{W}_{i*} \frac{1}{\tau-\bar{\tau}} \sum_{t=\bar{\tau}+1}^{\tau} s_{\tau-t} \\
&\overset{\text{asy}}{=} \boldsymbol{W}_{i*}\boldsymbol{\theta}^* = \theta^*, \qquad \text{(i.i.d. Bernoulli signals)} \qquad (12)
\end{aligned}$$

where $\boldsymbol{W} = \boldsymbol{D}\boldsymbol{1}'$. From equation (12), 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 $\operatorname{plim}_{t\to\infty} P_t = |\theta^* - \theta^*| = 0$. **(Q.E.D.)**

B Proof of Proposition 2

Consider again the update process described in the Section (XX)

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

Notice that B_t is not fixed over time now. We re-write the stochastic matrix (see lemma 1) inside the squared bracket as

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

When $t = 0$,

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

When $t = 1$,

$$\begin{aligned}\alpha_2 &= W_1\alpha_1 + B_1 s_2 \\ &= W_1\left(W_0\alpha_0 + B_0 s_1\right) + B_1 s_2 \\ &= W_1W_0\alpha_0 + W_1B_0 s_1 + B_1 s_2.\end{aligned}$$

When $t = 2$,

$$\begin{aligned}\alpha_3 &= W_2\alpha_2 + B_2 s_3 \\ &= W_2(W_1W_0\alpha_0 + W_1B_0 s_1 + B_1 s_2) + B_2 s_3 \\ &= W_2W_1W_0\alpha_0 + W_2W_1B_0 s_1 + W_2B_1 s_2 + B_2 s_3.\end{aligned}$$

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

Following Chaterjee and Seneta (1977), Seneta (2006) and Tahbaz-Salehi and Jadbabaie (2008), we let $\{W_k\}$, for $k \geq 0$, be a fixed sequence of stochastic matrices (see lemma 1), and let $U_{r,k}$ be the stochastic matrix defined by the following *backward product*

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

where $W_k = \{w_{ij}(k)\}$, $U_{r,k} = \{u_{ij}^{(r,k)}\}$[15].

[15]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.

Then, the update process of both parameters can be represented in the following form for any period τ

$$\alpha_\tau = U_{0,\tau-1}\alpha_0 + \left(\sum_{r=1}^{\tau-1} U_{r,\tau-1-r}B_{r-1}s_r\right) + B_{\tau-1}s_\tau \tag{14}$$

$$\beta_\tau = U_{0,\tau-1}\beta_0 + \left(\sum_{r=1}^{\tau-1} U_{r,\tau-1-r}B_{r-1}\left(\mathbf{1}-s_r\right)\right) + B_{\tau-1}\left(\mathbf{1}-s_\tau\right) \tag{15}$$

From equation (14), we write its entries as

$$\begin{aligned}
\alpha_{i,\tau} &= \sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}s_{j,r}\right) + b_{i,\tau-1}s_{i,\tau} \\
&= \sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \tau\frac{1}{\tau}\left[\left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}s_{j,r}\right) + b_{i,\tau-1}s_{i,\tau}\right] \\
&= \sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0} + \tau\tilde{\theta}_{i,1}(\tau)
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
\alpha_{i,\tau}+\beta_{i,\tau} &= \sum_j u_{ij}^{(0,\tau-1)}\left(\alpha_{j,0}+\beta_{j,0}\right) + \left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}\right) + b_{i,\tau-1} \\
&= \sum_j u_{ij}^{(0,\tau-1)}\left(\alpha_{j,0}+\beta_{j,0}\right) + \tau\frac{1}{\tau}\left[\left(\sum_j\sum_{r=1}^{\tau-1} u_{ij}^{(r,(\tau-1-r))}b_{j,r-1}\right) + b_{i,\tau-1}\right] \\
&= \sum_j u_{ij}^{(0,\tau-1)}\left(\alpha_{j,0}+\beta_{j,0}\right) + \tau\tilde{\theta}_{i,2}(\tau)
\end{aligned} \tag{17}$$

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:

$$\begin{aligned}\operatorname*{plim}_{\tau\to\infty} y_{i,\tau} &= \operatorname*{plim}_{\tau\to\infty}\left(\frac{\sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0}+\tau\tilde{\theta}_{i,1}(\tau)}{\sum_j u_{ij}^{(0,\tau-1)}(\alpha_{j,0}+\beta_{j,0})+\tau\tilde{\theta}_{i,2}(\tau)}\right)\\ &= \operatorname*{plim}_{\tau\to\infty}\frac{\tau}{\tau}\left(\frac{\frac{\sum_j u_{ij}^{(0,\tau-1)}\alpha_{j,0}}{\tau}+\tilde{\theta}_{i,1}(\tau)}{\frac{\sum_j u_{ij}^{(0,\tau-1)}(\alpha_{j,0}+\beta_{j,0})}{\tau}+\tilde{\theta}_{i,2}(\tau)}\right)\\ &= \operatorname*{plim}_{\tau\to\infty}\frac{\tilde{\theta}_{i,1}(\tau)}{\tilde{\theta}_{i,2}(\tau)} \qquad (18)\end{aligned}$$

Our main concern in order to prove that equation (18) converges 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 $\{W_k\}$ is said to be *weakly ergodic*, as $k \to \infty$, if for all $i, j, s = 1, 2, \ldots, n$ and $r \geq 0$

$$|u_{i,s}^{(r,k)} - u_{j,s}^{(r,k)}| \to 0$$

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

$$\lim_{k\to\infty} U_{r,k} = \mathbf{1}D_r'$$

Where $\mathbf{1}$ is a size n vector of ones and D_r is a probability vector in which $D_r \geq 0$ and $D_r'\mathbf{1} = \mathbf{1}$.

Both weak and strong ergodicity 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.

The three following lemmas are auxiliary helps to conclude the proof. Lemma 1 do xxx, Lemma 2 do yyy, whereas Lemma 3 do zzz.

Lemma 3. *For the backward product* (13)*, weak and strong ergodicity are equivalent.*

Proof. Following Seneta (1977)'s Theorem 1, we only need to prove that weak ergodicity implies strong ergodicity. Fix $r \geq 0$ and $\epsilon > 0$. Then, by *weak* ergodicity, we have

$$-\epsilon \leq u_{i,s}^{(r,k)} - u_{j,s}^{(r,k)} \leq \epsilon \iff u_{i,s}^{(r,k)} - \epsilon \leq u_{j,s}^{(r,k)} \leq u_{i,s}^{(r,k)} + \epsilon$$

for $k \geq \bar{\tau}$ for all $i, h, s = 1, \ldots, n$. Since $U_{r,k+1} = W_{r+k+1}U_{r,k}$,

$$\sum_{j=1}^{n} w_{hj}(r+k+1)(u_{i,s}^{(r,k)} - \epsilon) \leq \sum_{j=1}^{n} w_{hj}(r+k+1)u_{j,s}^{(r,k)} \leq \sum_{j=1}^{n} w_{hj}(r+k+1)(u_{i,s}^{(r,k)} + \epsilon).$$

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

$$u_{i,s}^{(r,k)} - \epsilon \leq u_{h,s}^{(r,k)} \leq u_{i,s}^{(r,k)} + \epsilon.$$

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

$$|u_{h,s}^{(r,k+q)} - u_{i,s}^{(r,k)}| \leq \epsilon.$$

By setting $i = h$, it is clear that $u_{i,s}^{k,r}$ is a Cauchy sequence that aproaches a limit as $k \to \infty$. □

Definition 3. *The scalar function $\mu(\cdot)$ continuous on the set of $n \times n$ stochastic matrices W and satisfying $0 \leq \mu(W) \leq 1$ is called a coefficient of ergodicity. It is said to be proper if $\mu(W) = 1 \Leftrightarrow W = \mathbf{1}\boldsymbol{v}'$, where $\boldsymbol{v}'$ is any probability vector (i.e. whenever W is a row-stochastic matrix with unity rank).*

In particular, we will focus on the *proper* coefficient of ergodicity $\mu(W) = 1 - a(W)$, where

$$a(W) = \frac{1}{2} \max_{i,j} \sum_{s=1}^{n} |w_{is} - w_{js}|.$$

Therefore, weak ergodicity is then equivalent to $\mu(U_{r,k}) \to 1$ as $k \to \infty$ and $r \geq 0$.

Lemma 4. *Suppose that $1 - a(\cdot)$ and $\mu(\cdot)$ are both proper coefficients of ergodicity. Then $\{W_k\}$, $k \geq 0$, is ergodic if and only if there exists a strictly increasing subsequence $\{i_j\}$, $j = 1, 2, \ldots$ of the positive integers such that*

$$\sum_{j=1}^{\infty} \mu\left(U_{i_j, i_{j+1} - i_j}\right) = \infty$$

Proof. Soon □

Lemma 5. *The weak ergodicity of the sequence $\{W_k\}$, $k \geq 0$ is a trivial event when $\mathbf{g}^t$ follows equation* (1).

Proof. Soon □

With the results of these three lemmas, we can proceed with

$$
\begin{aligned}
\operatorname*{plim}_{\tau\to\infty} \frac{\tilde{\theta}_{i,1}(\tau)}{\tilde{\theta}_{i,2}(\tau)} &= \operatorname*{plim}_{\tau\to\infty} \frac{\frac{1}{\tau}\sum_j \sum_{r=1}^{\tau-\bar{\tau}} u_{ij}^{(r,(\tau-1-r))} b_{j,r-1} s_{j,r}}{\frac{1}{\tau}\sum_j \sum_{r=1}^{\tau-\bar{\tau}} u_{ij}^{(r,(\tau-1-r))} b_{j,r-1}} \qquad \text{(by lemma 5)} \\
&= \operatorname*{plim}_{\tau\to\infty} \frac{\sum_j \bar{u}_{ij} \frac{1}{\tau}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1} s_{j,r}}{\sum_j \bar{u}_{ij} \frac{1}{\tau}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1}} \qquad \text{(by lemma 4)} \\
&\overset{\text{asy}}{\equiv} \operatorname*{plim}_{\tau\to\infty} \frac{\sum_j \bar{u}_{ij} \frac{1}{\tau-\bar{\tau}}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1} s_{j,r}}{\sum_j \bar{u}_{ij} \frac{1}{\tau-\bar{\tau}}\sum_{r=1}^{\tau-\bar{\tau}} b_{j,r-1}} \\
&= \frac{\sum_j \bar{u}_{ij} \mathbb{E}(b_j s_j)}{\sum_j \bar{u}_{ij} \mathbb{E}(b_j)} \qquad \text{(by weak law of large numbers)} \\
&= \frac{\sum_j \bar{u}_{ij} \mathbb{E}(b_j)\mathbb{E}(s_j)}{\sum_j \bar{u}_{ij} \mathbb{E}(b_j)} \qquad \text{(by independence of } b_j \text{ and } s_j\text{)} \\
&= \frac{\theta^* \sum_j \bar{u}_{ij} \mathbb{E}(b_j)}{\sum_j \bar{u}_{ij} \mathbb{E}(b_j)} = \theta^* \quad \text{(since } \mathbb{E}(s_j) = \theta^*, \forall j \text{)} \qquad (19)
\end{aligned}
$$

C Summary statistics for big networks

Table 5: Network characteristics when $n = 70$

	Mean	Std Dev.	Min	Max
Behavioral				
Weight unbiased source b	0.416	0.282	0	1
Technological				
Clock ρ	0.468	0.250	0.100	1
Flooding parameter κ	9.76	6.69	1	20
Topological				
in-Degree L-bot	0.094	0.082	0.013	0.568
in-Degree R-bot	0.098	0.086	0.013	0.691
Page Rank L-bot	0.013	0.011	0.002	0.106
Page Rank R-bot	0.014	0.012	0.002	0.104
in-Closeness L-bot	0.122	0.046	0.013	0.224
in-Closeness R-bot	0.122	0.047	0.013	0.231
% Following both bots	0.09	0.111	0	0.651
Reciprocity	0.048	0.023	0	0.120
Clustering	0.226	0.045	0.096	0.336
Homophily	-0.010	0.036	-0.144	0.191