European Journal of Political Economy 76 (2023) 102256

Contents lists available at ScienceDirect

European Journal of Political Economy

journal homepage: www.elsevier.com/locate/ejpe

Social media networks, fake news, and polarization

Marina Azzimonti [a,*], Marcos Fernandes [b]

[a] Department of Economics, Stony Brook University, United States of America
[b] Department of Economics, University of São Paulo, Brazil

ARTICLE INFO

JEL classification:
C63
D83
D85

Keywords:
Learning
Misinformation
Polarization
Social networks
Social media
Fake news
Twitter
Synthetic network

ABSTRACT

We study how the structure of social media networks and the presence of fake news affects the degree of misinformation and polarization in a society. For that, we analyze a dynamic model of opinion exchange in which individuals have imperfect information about the true state of the world and exhibit bounded rationality. Key to the analysis is the presence of internet bots: agents in the network that spread fake news (e.g., a constant flow of biased information). We characterize how agents' opinions evolve over time and evaluate the determinants of long-run misinformation and polarization in the network. To that end, we construct a synthetic network calibrated to Twitter and simulate the information exchange process over a long horizon to quantify the bots' ability to spread fake news. A key insight is that significant misinformation and polarization arise in networks in which only 15% of agents believe fake news to be true, indicating that network externality effects are quantitatively important. Higher bot centrality typically increases polarization and lowers misinformation. When one bot is more influential than the other (asymmetric centrality), polarization is reduced but misinformation grows, as opinions become closer the more influential bot's preferred point. Finally, we show that threshold rules tend to reduce polarization and misinformation. This is because, as long as agents also have access to unbiased sources of information, threshold rules actually limit the influence of bots.

1. Introduction

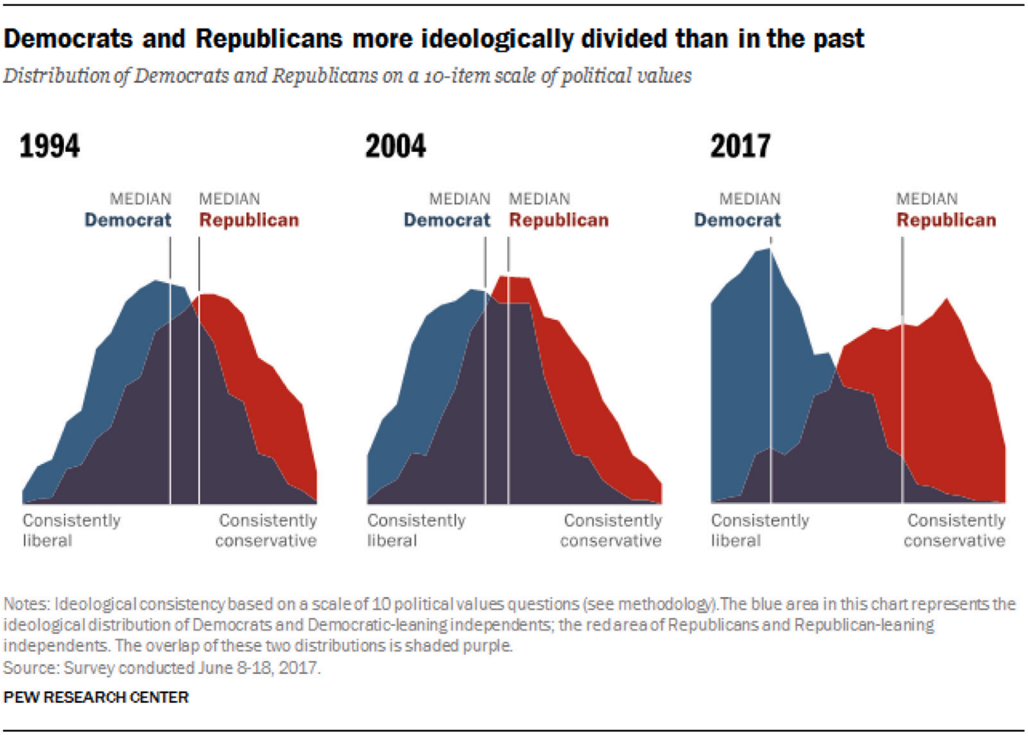
In the last decade, the United States has become more polarized than ever. A recent survey conducted by The Pew Research Center indicates that Republicans and Democrats are further apart ideologically than at any point since 1994 (see Fig. 1 Fig. 2).

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

There is a large literature trying to explain how slanted news and media bias may affect voters' opinions by generating misinformation and exacerbating polarization (see DellaVigna and Kaplan (2007) or (Martin and Yurukoglu, 2017)). While this

[*] Corresponding author.
E-mail addresses: marina.azzimonti@gmail.com (M. Azzimonti), mrf.ross@gmail.com (M. Fernandes).

https://doi.org/10.1016/j.ejpoleco.2022.102256
Received 31 October 2020; Received in revised form 24 May 2022; Accepted 8 June 2022
Available online 14 July 2022
0176-2680/Published by Elsevier B.V.



Democrats and Republicans more ideologically divided than in the past

Distribution of Democrats and Republicans on a 10-item scale of political values

Notes: Ideological consistency based on a scale of 10 political values questions (see methodology). The blue area in this chart represents the ideological distribution of Democrats and Democratic-leaning independents; the red area of Republicans and Republican-leaning independents. The overlap of these two distributions is shaded purple.
Source: Survey conducted June 8-18, 2017.

PEW RESEARCH CENTER

Fig. 1. Political polarization in the American Public (2017, Pew Research Center).

literature has been mostly focused on traditional media, such as newspapers, radio, and cable TV – broadly covered under the umbrella of 'broadcasting' – recent interest has shifted towards social media. There are several reasons for this shift. First, because individuals are increasingly obtaining information from social media networks. According to a 2021 study by the Pew Research Center, 53% of adults get their news from social media.[1] In addition, 59% of Twitter users regularly get their news from Twitter (Facebook users have a similar pattern).

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

This is relevant when coupled with another phenomena that became prevalent particularly around 2016 presidential election and has still not been resolved: the massive spread of *fake news* (also referred to as disinformation campaigns, cyber propaganda, cognitive hacking, and information warfare) through the internet. As defined by Gu et al. (2017), 'Fake news is the promotion and propagation of news articles via social media. These articles are promoted in such a way that they appear to be spread by other users, as opposed to being paid-for advertising. The news stories distributed are designed to influence or manipulate users' opinions on a certain topic towards certain objectives.' While the concept of propaganda is not new, social media has made the spreading of ideas faster and more scalable, making it potentially easier for propaganda material to reach a wider set of people. Relative to more traditional ways of spreading propaganda, fake news are extremely difficult to detect posing a challenge for social media users, moderators, and governmental agencies trying control their dissemination. A December 2016 Pew Research Center study found that 'about two-in-three U.S. adults (64%) say fabricated news stories cause a great deal of confusion about the basic facts of current issues and events.' Moreover, 23% admit to having shared a made-up news story (knowingly or not) on social media. Understanding how fake news spread and affect opinions in a networked environment is at the core of our work.

[1] According to Pew, 'Americans ages 18 to 29 stand out in that the most common digital way they get news is social media, with 42% saying they get news this way often versus 28% saying the same of either news websites or search engines.

In this context, we study a dynamic model of opinion formation in which individuals who are connected through a social network have imperfect information about the true state of the world, denoted by θ. For instance, the true state of the world can be interpreted as the relative quality of two candidates competing for office, the degree of vaccine efficacy, the optimality of a specific government policy or regulation, the need for environmental policies, the degree of government intervention in concentrated markets, etc. Why is this relevant? Because common beliefs about the value of our variable of interest may be a decisive factor in the implementation of certain policies under uncertainty. Consider, for example, the decision of whether to implement a mask mandate during the COVID-19 outbreak, with θ representing how effective masks are in preventing the spread of the disease. To the extent government representatives respond to their constituencies, implemented policies may differ from the optimal ones when bots are present. If a large number of voters have homogeneous beliefs but are *misinformed* (that is, have beliefs far away from the true θ), implemented policies will be inefficient. In our example, this would happen if the true θ is high, but a majority of voters believed their efficacy was low and no mandate would be in place. Another source of inefficiency arises when there is *polarization*. This happens when there are two sizeable groups with opposing beliefs and a status-quo that needed to be changed in a timely manner. In this case, there need not be a majority of people who are misinformed, but a mass large enough to stall the decision making process. Sub-optimal delays (or even inaction) in response to shocks would arise. Going back to our example, if masks were effective in protecting against COVID-19 but people were polarized about their efficacy, the government would not change the status-quo (no masks) when the optimal choice would have been to impose a mandate. This could potentially prolong the pandemic. Hence, polarization can also be detrimental for welfare.

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

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

The structure of the graph representing the social media network and the degree of influence of bot followers shape the dynamics of opinion and the degree of misinformation and polarization in the long-run. More specifically, long-run misinformation and polarization are determined by the network topology (e.g. the relative exposure to bots, how central they are, and the ability of bots to flood the network with fake news). Because a theoretical characterization of the relationship between the topology of the network and the degrees of misinformation and polarization is not trivial, we construct a synthetic large network (with around 4,000 nodes) and calibrate it to a real life social media network: Twitter. We then run multiple Monte Carlo simulations in which the location of bots and their followers is assigned randomly, and the process of communication exchange is simulated over long periods of time. While we fix the number of bot followers, we allow bots to have asymmetric influence (e.g. a different in-Degree). All other variables are kept constant across simulations. Our goal is to infer how the absolute and relative centrality of bots (and their followers) affect long-run polarization and misinformation.

For our calibrated synthetic network, we find that significant levels of misinformation and polarization are possible even though only 15% of agents believe fake news to be true and there are only 2 bots, who progressively become extreme. Even though most agents can detect fake news and exclude them from their information set, their views are indirectly affected through the opinions of other friends in the network. To the extent that bots are able to target a small amount of 'influencers,' biased signals will travel through the network affecting a large number of agents and hence generating misinformation and polarization. This is relevant, because it shows that *network externality effects* are quantitatively important. A summary of our quantitative results follow.

First, we find that misinformation and polarization have an inverted u-shape relationship when bots are symmetrically extreme (e.g. their preferred state is equidistant from the true state): on the one hand, when individuals are able to effectively aggregate information and learn the true state of the world, polarization vanishes. On the other hand, there are situations where there is no polarization because most individuals in the network converge to the wrong value of θ, i.e. they end up with the same (wrong) opinion and for that they do not polarize. This involves maximal misinformation with no polarization. In addition, there are cases in which individuals are on average correct but distributed symmetrically around the true state of the world, with large mass at the extremes of the belief distribution. Here, there are intermediate levels of misinformation and extreme polarization. Even though this implies somewhat better information aggregation, it may lead to inefficient gridlock due to inaction.

We distinguish between average centrality and relative centrality. The former captures how disruptive bots are, on average, to the aggregation of information. The larger its value, the higher the polarization and the lower the misinformation observed. The latter happens because when bots are equidistant from the truth, they offset each other and the unbiased signals become more important. Relative centrality, on the other hand, captures how much more influential one bot is relative to the other (keeping the total number of followers unchanged). As this rises, the more influential bot manages to pull opinions towards its extreme views. This significantly increases misinformation and reduces polarization in our benchmark case. This happens even in simulations in which polarization is significant.

Second, we find that the strength of fake news relative to informative news, as measured by the flooding capacity parameter of bots, increases both misinformation and polarization in a society. However, the technology seems to have decreasing returns as the effects of increasing the flooding parameter vanish at some point. We also show that increasing the number of signals each bot can send, or flooding, (keeping number of bots constant) is not equivalent to increasing the number of bots (keeping flooding constant). This is because flooding affects how fast the bot becomes extreme, whereas the latter affects the weight of bot's signals relative to other signals.

Finally, we experiment with a *threshold rule* (bounded confidence model) by which agents only pay attention to sufficiently like-minded agents. Interestingly, we find that this tends to reduce both polarization and misinformation. The result may seem counter-intuitive at first, as one could expect that if agents only communicate with like-minded friends, the differences between societal views would widen. However, these threshold rules make bots less relevant early on, which allows unbiased signals to move opinions towards the true state. While this may depend on initial conditions, and studying this is beyond the scope of this paper, we find this to be a promising avenue for research.

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

The strand of literature focusing on social learning with bounded rational agents assumes that individuals use simple heuristic rules to update beliefs, like taking repeated averages of observed opinions. Examples are (DeGroot, 1974), Ellison and Fudenberg (1993, 1995), Bala and Goyal (1998), Goyal (2005), DeMarzo et al. (2003) and Golub and Jackson (2010). In most of these environments, under standard assumptions about the connectivity of the network and the bounded prominence of groups in growing societies, the dynamics of the system reaches an equilibrium and consensus emerges. In this sense, long-run polarization or misinformation would only arise in such models if those assumptions are relaxed. Common to most of these models is the fact that there is no new flow of information entering into the network. Agents are typically assumed to be bounded rational (naive) and do not observe private signals from external sources. JMST (2012) extends these environments to allow for a constant arrival of new information over time in an environment in which agents also learn from their neighbors in a naive way. This feature allows agents to efficiently aggregate information even when some standard assumptions that ensure consensus are relaxed. Our paper uses the update rule proposed by JMST (2012), and introduces bots that produce a stream of fake news to countervail the effect of informative signals. The latter is a variation of the concept of stubborn (or forceful) agents in the literature on misinformation. See work by Acemoglu et al. (2010) (AOP henceforth), Acemoğlu et al. (2013) (ACFO henceforth), or (Como and Fagnani, 2016).

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

Our main contribution relative to the existing literature, however, is our numerical exercise. We construct a large synthetic network, calibrated to Twitter, and simulate communication exchange over a large period of time. Importantly, we allow the

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

location of bots and their followers to change across simulations, which allows us to estimate how network centrality (i.e, their degree of influence), potentially asymmetric, affects outcomes. To the best of our knowledge, this is the first paper to quantify the relative importance of network characteristics on long-run misinformation and polarization.

Finally, there is a growing empirical literature analyzing the effects of social media in opinion formation and voting behavior (Halberstam and Knight, 2016). Because individual opinions are unobservable from real network data, these papers typically use indirect measures of ideology to 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, 2011), and Flaxman et al. (2013) and the effects of social media on political polarization, such as (Boxell et al., 2017), Barberá (2014), and Webster and Ksiazek (2012).

2. Baseline model

Agents, social bots and information structure. The economy is composed by a finite number of agents $i \in N = \{1, 2, \dots, n\}$ who interact in a social network over time. Individuals have imperfect information about a variable of interest $\theta \in \Theta = [0, 1]$. This parameter can be interpreted as the optimal degree of government intervention in private markets (e.g. environmental control, enforcement of property rights, restrictions on the use of public land, gun control, etc.), as optimal fiscal or monetary policy (e.g. the inflation rate, tax rates on capital or labor income, tariffs, etc.), or as the best response to an unexpected shock (e.g. the size of a bailout during a financial crisis, the response to a national security threat, the amount of aid given to a region that suffered a natural disaster, etc.). Agents start period 0 with a prior about θ and update their beliefs with information obtained thereafter.

Individuals obtain information from: (i) an unbiased source (signals), (ii) other agents connected to them in a social network, and (iii) a biased source, which we interpret as a *bot* spreading fake news. There are two types of bots, L-bot and R-bot with opposing agendas. Their objective is to manipulate opinions by sending biased signals (e.g. close to 0 or 1). We assume that a majority of the population can identify bots and disregard fake news in their update process. We will refer to them as *regular agents*. There is small proportion μ of individuals, on the other hand, that are influenceable because they cannot distinguish fake news from real news (or, alternatively, that cannot distinguish bots from regular agents). We refer to them as *bot followers*. A key assumption is that nobody can back out the sources of information of other agents. As a result, regular agents may be influenced by fake news indirectly through their social media contacts. We first describe how opinions evolve over time and then define how statistics obtained from this distribution can be used to compute misinformation and polarization, and hence quantify welfare losses associated to them.

Each agent starts with a prior belief (μ) over θ, assumed to follow a Beta distribution, i.e.

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

This distribution (or world-view) is characterized by initial parameters $\alpha_{i,0} > 0$ and $\beta_{i,0} > 0$ and has the following functional form

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

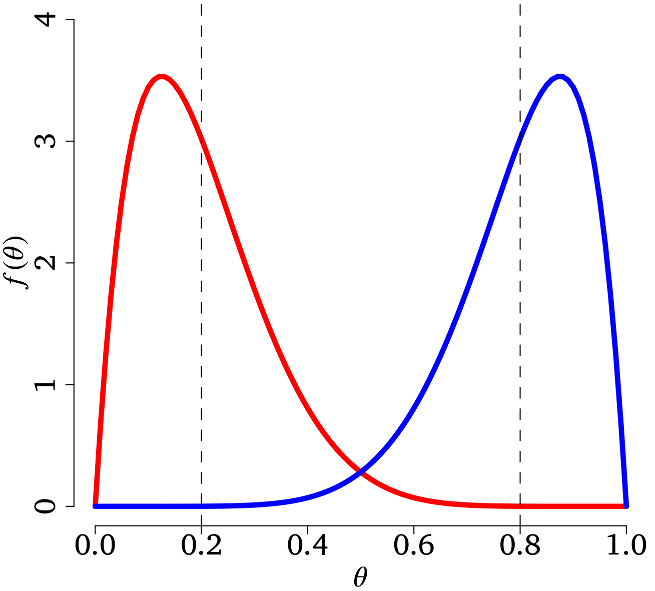
where $\Gamma(\cdot)$ is a Gamma function and the ratio of Gamma functions in the expression above is a normalization constant that ensures that the total probability integrates to 1.

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 about the true state of the world as her best guess of θ given the available information,[3]

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

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.

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



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

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

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

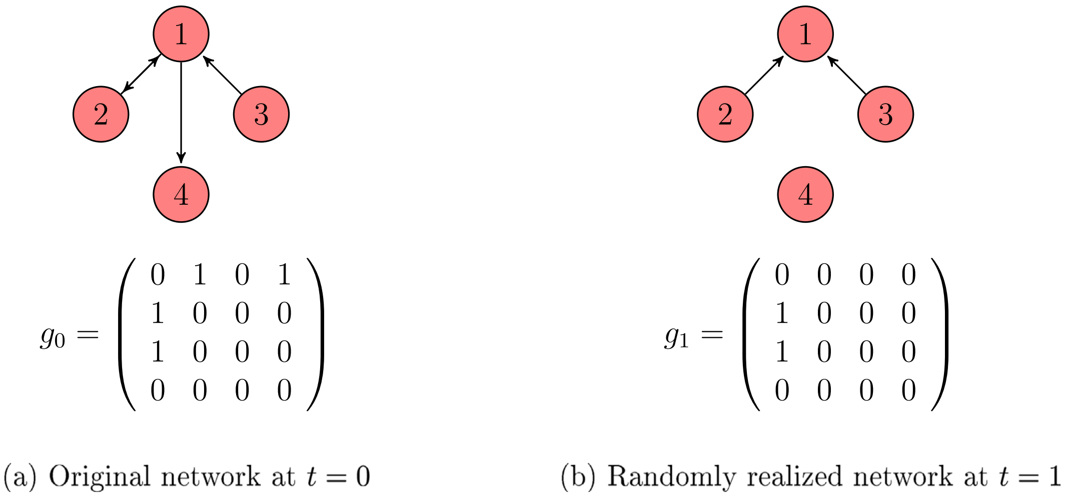
Social network and random meetings process. The connectivity among agents in the network at each point in time t is described by a directed graph $G_t = (N, g_t)$, where g_t is a real-valued $n \times n$ adjacency matrix. Each regular element $[g_t]_{ij}$ in the directed-graph represents the connection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

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

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

where the initial structure of the network, represented by the initial adjacency matrix g_0, remains unchanged.[4]

[4] Here $\circ$ denotes the Hadamard Product, the element-wise multiplication of matrices.



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

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

(a) Original network at $t = 0$

(b) Randomly realized network at $t = 1$

Fig. 2. Bernoulli clock and network dynamics.

Example 2 (*Bernoulli Clock*).

Panel 2(a) represents the original network and its adjacency matrix, whereas Panel 2(b) 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.

Evolution of beliefs. Before the beginning of each period, both regular agents and bot followers receive information from individuals in their out-neighborhood, a set determined by the realization of the clock in period t and the initial network. All agents share their opinions and precision, summarized by the shape parameters $\alpha_{i,t}$ and $\beta_{i,t}$. This representation aims at capturing communication exchanges through social media feeds. At the beginning of every period t, a signal profile is realized and an unbiased signal is privately observed by every agent. Bots, instead, disregard this unbiased signal. Hence, part of the information obtained by bot followers will be transmitted through the network in the following period, as it is incorporated during the belief updating process and shared with other agents in the network that naively internalize them.

We now explain the update rules of agents and bots. The full characterization of the update rules can be found in Appendix B.

Agents:

Regular agents and bot followers share the same update rule. The feature that distinguishes them is the composition of their neighborhood: while regular agents only pay attention to other regular agents, bot followers devote some share of their attention to the information transmitted by bots, and hence are exposed to fake news. After observing the signal from unbiased sources, agents compute their Bayesian posteriors conditional on the observed signals. We assume that parameters $\alpha_{i,t+1}$ and $\beta_{i,t+1}$ are convex combinations between their Bayesian posterior parameters and the weighted average of their neighbors' parameters. In mathematical terms we have that

$$\alpha_{i,t+1} = (1-\omega_{i,t})[\alpha_{i,t} + s_{i,t+1}] + \omega_{i,t} \sum_j \left[\hat{g}_t\right]_{ij} \alpha_{j,t} \tag{2}$$

$$\beta_{i,t+1} = (1-\omega_{i,t})[\beta_{i,t} + 1 - s_{i,t+1}] + \omega_{i,t} \sum_j \left[\hat{g}_t\right]_{ij} \beta_{j,t}, \tag{3}$$

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

Note that this rule assumes that agents exchange information (i.e. $\alpha_{j,t}$ and $\beta_{j,t}$) before processing new signals $s_{i,t+1}$. Agents' full attention span is split between processing information from unbiased sources, $(1-\omega_{i,t})$, and that provided by their friends in the network, $\omega_{i,t}$ (e.g. reading a Twitter feed). If no friends are found in the neighborhood of agent i, $\sum_j \left[\hat{g}_t\right]_{ij} = 0$, then the agent attaches weight 1 to the unbiased signal, behaving exactly as a standard Bayesian agent. Conversely, if at least one friend is found, this agent uses a common weight $\omega \in (0,1)$. The term $\left[\hat{g}_t\right]_{ij} = \frac{[g_t]_{ij}}{|N_{i,t}^{out}|}$ represents the weight given to the information received from her out-neighbor j. As $\omega_{i,t}$ approaches 1, the agent only incorporates information from social media, making her update process closer to a DeGrootian in which individuals are purely conformists. In general, ω can be interpreted as the degree of influence of social media friends.

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

$$y_{i,t} = \frac{\alpha_{i,t}}{\alpha_{i,t} + \beta_{i,t}}. \tag{4}$$

Bots

We assume that there are two bots, a left wing bot (or L-bot) and a right wing bot (or R-bot), both with biased views. They ignore unbiased signals and those provided by other individuals in the network. The beliefs of the R-bot evolve according to $\alpha^R_{t+1} = \alpha^R_t + \kappa$ and $\beta^R_{t+1} = \beta^R_t$, whereas those of the L-bot evolve according $\alpha^L_{t+1} = \alpha^L_t$ and $\beta^L_{t+1} = \beta^L_t + \kappa$. The parameter $\kappa > 0$ measures the ability of bots to spread fake-news at a different rate than other agents, which can be interpreted as their *flooding capacity* (i.e. how fast/slow they can produce fake news compared to the regular flow of informative signals received by agents). Bots transmit the whole stream of information to agents paying attention to them. Hence, a value of $\kappa > 1$ gives them more de-facto weight in the updating rule of their followers, emphasizing their degree of influence on the network. Given initial beliefs α^i_0 and β^i_0 for bot i, their update rule can simply be written as

$$\text{L-bot:} \quad \alpha^L_t = \alpha^L_0 \qquad \text{and} \qquad \beta^L_t = \beta^L_0 + \kappa t \tag{5}$$

$$\text{R-bot:} \quad \alpha^R_t = \alpha^R_0 + \kappa t \qquad \text{and} \qquad \beta^R_t = \beta^R_0. \tag{6}$$

From Eq. (4), these imply that the L-bot's opinion converges to $\lim_{t\to\infty} y^L_t = 0$ whereas $\lim_{t\to\infty} y^R_t = 1$. The flooding parameter κ dictates the speed of convergence to the extremes. This assumption aims to capture, in reduced form, the fact that bots try to disguise themselves as regular agents. Because of this, they do not typically start out with completely extreme views, but instead converge to them over time.

Our heuristic rules differ from those analyzed in Como and Fagnani (2016), who assume that individuals exchange opinions $y_{i,t}$ directly. They resemble, instead, the ones in JMST (2012). But there are three important distinctions. First, their adjacency matrix is fixed over time (homogeneous Markov chain), whereas ours is stochastic (in-homogeneous Markov chain), an element we borrowed mainly from ACFO (2013). Second, we restrict attention to a specific conjugated family (Beta-Bernoulli) and assume that individuals exchange shape parameters $\alpha_{i,t}$ and $\beta_{i,t}$ that characterize this distribution. So the heuristic rule involves updating two real valued parameters, whereas JMST (2012)'s heuristic rule involves a convex combination of the whole distribution function. Given their rule, the posterior distribution may not belong to the same family as the prior distribution, as the convex combination of two Beta distributions is not a Beta distribution. That is not the case in our environment, as the posterior will also belong to the Beta distribution family. Finally, we are considering the influence of fake news spread by bots and this feature is the main source of misinformation. Therefore, to the extent that bots are followed by agents who are influential (e.g. those with a large number of followers), their presence will affect the existence and persistence of misinformation and polarization over time. This is due to the fact that they will consistently communicate fake news (biased signals) to their followers pushing them to extremes of the belief spectrum.[5]

3. Misinformation and polarization: Limiting results

An agent i is misinformed when her opinion y_i is sufficiently far from the true state of the world θ. We can define the 'degree of misinformation' in society – at time t – as the average square distance between agents' opinions and the true state of the world.

Definition 1. The degree of misinformation is given by

$$MI_t = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,t} - \theta\right)^2 . \tag{7}$$

While MI_t grows with the number of agents whose beliefs are far from θ, it does not capture disagreement. For example, consider a network in which $y_{i,t} = 1$ in period t for all i. In such case, MI_t reaches its maximum theoretical value, but with all agents agreeing on the wrong value of θ (e.g. there is zero disagreement but maximal misinformation). To capture disagreement among individuals, we use the definition of polarization constructed by Esteban and Ray (1994).

Definition 2 (*Polarization*). Polarization is defined as

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

for some constant $C > 0$, where K is a pre-determined number of groups in society, $\tilde{y}_{k,t}$ is the average opinion of agents in each group $k \in \{1,\dots,K\}$, $\pi_{k,t}$ is the share of agents in group k at time t and ς is a positive parameter that reflects group identification.

Basically, this measure computes polarization over a discretized version of the domain of θ (e.g. the interval $[0,1]$), defining the share of agents in each group $k \in \{1,\dots,K\}$ by $\pi_{k,t}$, with $\sum_k \pi_{k,t} = 1$. The higher degree of heterogeneity of opinions across groups, $|\tilde{y}_{k,t} - \tilde{y}_{l,t}|$ (*alienation*), the greater the level of polarization according to this measure. Polarization also increases with *intra*-group opinion homogeneity, which is given by the mass of individuals $\pi_{k,t}$ that share similar opinion. Esteban and Ray (1994) restrict the

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

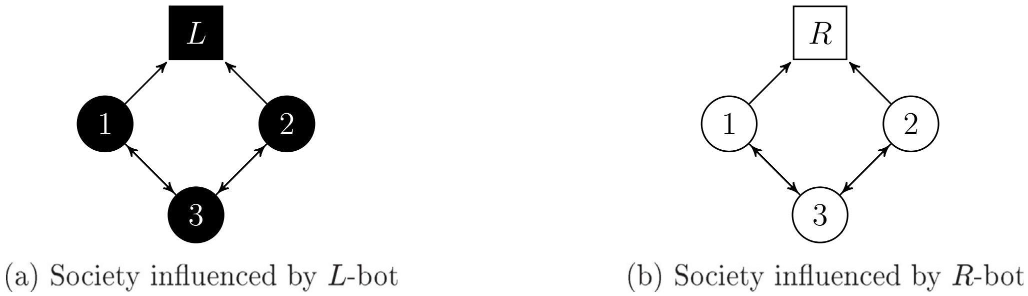

L
1
2
3
(a) Society influenced by L-bot
R
1
2
3
(b) Society influenced by R-bot

Fig. 3. Two societies with bots.

group identification parameter ς to the interval $(0, 1.6]$. The higher ς, the lower polarization due to the greater sense of identification among people within the same group. Note that when $\varsigma = 0$, the polarization measure is similar to a standard Gini coefficient of opinions.

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

We can think of long-run misinformation and polarization as functions of: (i) the initial network structure g_0, (ii) the location of bots and their followers in it, and (iii) other parameters (such as clock speed ρ, influence of friends ω, share of bot followers μ, and flooding parameter κ),

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

We aim at characterizing the properties of the functions **P** and **MI** in the limit. We first show some (limited) theoretical results, to illustrate that this model shares the properties of other well-known models in the literature, and then present results obtained via computer simulations.

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

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

Proof. See Appendix E.1. □

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

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

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

Proof. See Appendix E.2. □

Influential bots. Influential bots cause misinformation by spreading fake news. This does not imply necessarily that the society will polarize. The following example depicts two networks with three agents each: a regular one (node 3) and two bot followers (nodes 1 and 2) influenced by only one bot—L-bot in panel 3 and R-bot in the panel 3(b)—.

Polarization in both societies converges to zero in the long-run. However, neither society is wise if $\theta \neq 1$ or $\theta \neq 0$. This illustrates that the influence of bots may generate misinformation in the long run, preventing agents from uncovering θ, but does not necessarily create polarization. In general, if a society is wise, then it experiences no social polarization in the long run. The converse, on the other hand, is not true.

More generally, when the relative influence of one type of bot is significantly larger than the other, it is possible for a society to reach consensus (i.e. experience no polarization of opinions) to a value of θ that is incorrect. This can happen when there is a

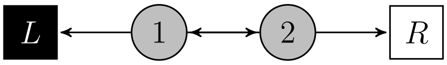


Fig. 4. Society with both L-bot and R-bot.

Table 1
Ego-networks: data vs fit.

Network statistics	Ego-Twitter SNAP (Data)	Synthetic Ego-net Best Fit (Model)
Average in-Degree	21.74	21.83
Average clustering	0.60	0.57
Diameter (directed)	15	12
Avg. path length	4.91	3.55
Avg. path length to Diameter (ratio)	0.33	0.30
Reciprocity	0.32	0.29
Clusters	1	1
Degree distribution (log normal)		
Mean	1.94	1.98
St. Dev	1.51	1.49

large number of bot followers or when bot followers, even if few, reach a large part of the network (i.e. when they are themselves influential). In order for this case to arise, it is also necessary that one of the bots has a relatively larger number of followers (or in-Degree) than the other bot; the example presented in Fig. 3 is extreme in that one bot is influential whereas the other one is not. A society may converge to the wrong θ under less extreme assumptions. In order for a society to be polarized, individuals need to be sufficiently exposed to bots with opposing views.

Consider the social network of two agents depicted in Fig. 4, in which both bot-types are present: agent 1 is influenced by the L-bot whereas agent 2 is influenced by the R-bot. Even though agents 1 and 2 receive unbiased signals and communicate with each other, this society exhibits polarization in the long run. This happens because bots subject to different biases are influential. The degree of misinformation may be lower than in the previous example (as opinions end up being averaged out and potentially closer to the true state of the world), but to the extent policy is chosen by majority voting may still lead to inaction and hence inefficiencies.

Finally, we want to point out that whether misinformation and polarization are relevant even in the long run depends importantly on the topology of the network, the number, and degree of influence of bots and their followers. The next section is devoted to uncovering what drives these different dynamics.

4. Monte Carlo Simulations

One of the biggest challenges when using network analysis is to ascertain analytical closed forms and tractability for long-run polarization and misinformation. 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. To understand the relative importance of the network structure and belief formation process on limiting polarization and misinformation, we resort to numerical methods, where a synthetic large-scale ego-network capturing communication via Twitter is generated. Limiting properties of the distribution of beliefs, namely long run average misinformation and polarization, are computed through Monte Carlo simulations for different location of bot followers in the network, while keeping all other parameters unchanged at our benchmark values. This exercise allows us to quantify how the centrality of bot followers affects misinformation and polarization.

4.1. Calibrating the network structure: g_0 choice

To construct g_0, we use the ego-networks from Twitter identified by Leskovec and Mcauley (2012).[6] Their full dataset consists of 81,306 nodes and 1,768,149 direct edges capturing social circles in Twitter. Each synthetic ego-network can be easily be constructed with the R-package *igraph*. Selected statistics capturing the network topology are displayed in the first column of Table 1.

We build a smaller calibrated synthetic network with 3,991 nodes and 87,134 edges to capture the dynamics of complex, real life networks in a model, while at the same time keeping the computation time of centrality measures manageable. The calibration algorithm consists on first drawing a large number of ego-networks and then selecting the one that minimizes the overall distance between selected network topology characteristics from the model to those in the data. More specifically, we match: (i) the average clustering coefficient, (ii) average in-Degree, and (iii) reciprocity ratio of links.

The empirical in-Degree distribution (in logs) is plotted in the left panel of Fig. 5.

[6] These are available from SNAP at http://snap.stanford.edu/data/ego-Twitter.html.

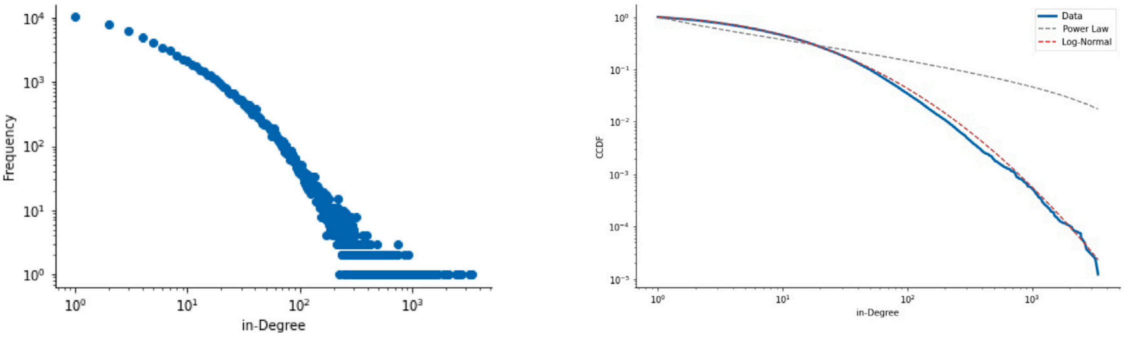


Fig. 5. Twitter in-Degree distribution.

Table 2
Benchmark parameters.

Parameters	Symbol	Value
Influence of friends	ω	0.5
Speed of communication	ρ	0.5
Flooding capacity	κ	25
Number of bots	b	2
Share of bot followers	μ	15%
Number of nodes	n	3,991
Number of simulations (different networks)	M	1,233

Because this is a high-dimensional object, we first approximate it with a Log-Normal distribution and then use the estimated mean and variance, equal to 1.94 and 1.51, respectively, as targets in our calibration. The CCDFs of the fitted log-normal (dashed red line) and the data (solid blue line) are plotted in the right panel of Fig. 5. We also include the fitted Power function distribution (dashed gray line), to illustrate that the log-normal fits the Twitter data better. This eyeball test is corroborated by computing the likelihood ratio between the two candidate distributions, and rejecting the hypothesis that the power function has a superior fit (see Alstott et al. (2014)).

The network statistics obtained from our best fit are displayed in the second column of Table 1. The diameter of our calibrated network (12) is naturally below its data counterpart, since the network is smaller. The average path length to diameter ratio and the reciprocity parameter in our calibrated network are both relatively close to the data. The next step consists on stating which agents in our synthetic network are bot followers.

We populate each network $j \in \{1, \ldots, M\}$ with two bots and two types of individuals, regular agents and bot followers. The only difference across networks is the location of these agents and the identity of the bot they follow. The share of bot followers is set to $\mu = 0.15$ (approximately 600 nodes), consistent with the percentage of agents reporting that they are 'only a little confident' in their ability to detect fake news from a Pew Research Center research poll in 2019. We define which agent follows the L-bot and which one follows the R-bot in two steps. First, using a uniform distribution, we randomly select 600 nodes out of our total of 3,991 nodes and assign them a "bot follower" label. This defines their location in g_0. From an ex-ante perspective, every node in the network has the same probability of being populated by a bot follower. Second, out of this set, we randomly pick $d_L \leq 600$ nodes to receive signals from the L-bot (exclusively). The remaining agents $600 - d_L$ follow the R-bot. Through this procedure, the in-Degree of the two bots can be different across simulations (e.g. they could have asymmetric influence over opinions), even when the total number of bot followers remains the same. Finally, the remaining individuals are assumed to be regular agents, and hence are not connected to (e.g. do not follow) any bots. For the benchmark case, we construct $M = 1,233$ networks.

4.2. Communication process: benchmark case

The true state of the world is assumed to be $\theta = 0.5$. In the long-run, this implies that both bots have preferred points which are equidistant from the true state (e.g. they are symmetrically biased). There is no obvious way to discipline ω and ρ using data, so we assume that agents place half of the weight in the unbiased signal, $\omega = 0.5$, and that the clock parameter is $\rho = 0.5$. While these choices are somewhat arbitrary, the comparative statics relative to these two parameters are straightforward. Finally, we set the bot's flooding capacity at $\kappa = 25$. We summarize the benchmark parameters in Table 2.

Assigning initial beliefs $\alpha_{i,0}$ and $\beta_{i,0}$. We fix the initial distribution of beliefs so that the same mass of the total population lies in the middle point of each one of $K = 7$ groups.[7] This rule basically distributes our agents evenly over the belief spectrum $[0, 1]$ such that

[7] We also experimented with $K = 3$ groups and $K = 5$ groups, and the resulting average levels of long-run polarization remain unchanged. See Appendix H.2.

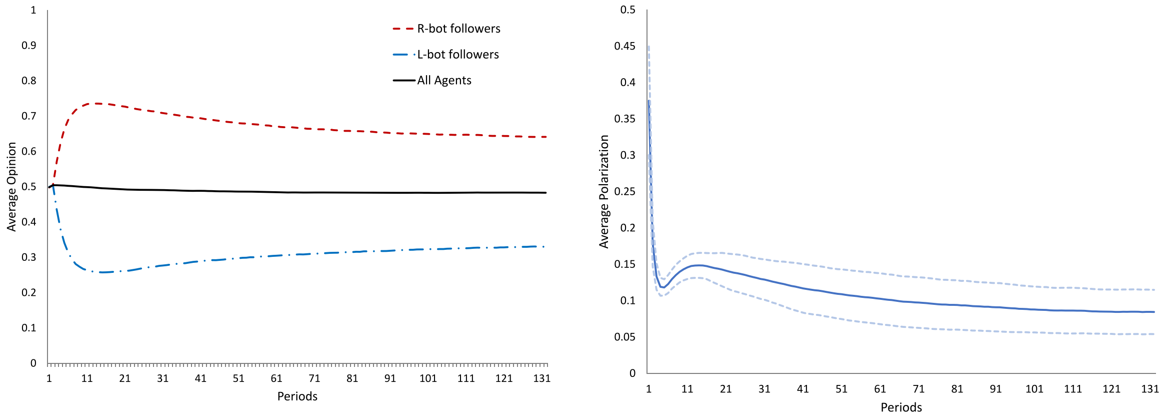


Fig. 6. Evolution of opinions and polarization.

each of the 7 groups contains exactly $\frac{1}{7}$ of the total mass of agents in the initial period. We set the same variance for each agent world-view to be 0.03. With both opinion and variance, we are able to compute the initial parameter vector for each agent i.[8]

Simulating opinion exchange over time. For each network $j \in \{1, \dots, M\}$, we allow agents to interact (following Eqs. (2) and (3)) for a large number of periods ($T = 1,000$) and use the resulting opinions to compute misinformation and polarization in the long run. For each network $j \in \{1, \dots, M\}$, we draw a signal $s^j_{i,t}$ for individual $i \in N$ at time t from a Bernoulli distribution with parameter $\theta = 0.5$. We also draw the $n \times n$ matrix $\mathbf{c}^t$ at each period t from a Bernoulli distribution with parameter ρ, 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). For each network specification $j \in \{1, \dots, M\}$, we compute a time series for opinions $y_{i,t,j}$ for individual i at period t and use this to compute average opinions $y_{j,t}$, polarization $P_{j,t}$, and misinformation $M_{j,t}$.[9]

The left panel of Fig. 6 displays the evolution over time of average opinion of all agents (black solid line), of L-bot followers (dashed blue line) and R-bot followers (dashed red line). These plots are averaging out results across network configurations j at each point in time t. We can see that agents start, on average, at $\theta = 0.5$ (this is by construction). Immediately after, there is a wide dispersion between the beliefs of R-bot and L-bot followers that grows over time up until period 71, where the difference starts shrinking. This happens because agents are still receiving the unbiased signal, which tends to moderate them. Despite of that, differences in opinions diverge over time, which generates positive polarization. The right panel of the figure displays the evolution of polarization over time (averaged out across network configurations), with dashed lines indicating ± 1 standard-deviation. Polarization starts at a large value because we assigned initial opinions uniformly over the $[0, 1]$ spectrum. It goes down subsequently as agents share information and incorporate their unbiased signals, with some agents getting closer to the beliefs of the bot they follow. This results in about three distinct groups of agents with positive mass: the L-bot followers, the R-bot followers, and regular agents not directly connected to these. It is interesting to note that while polarization declines, it settles at a constant value. Moreover, it remains basically unchanged after a couple of hundred interactions.

***Long-run results* :.** Because average opinions and polarization tend to stabilize after a large number of periods in each network j, we can approximate their limiting values by calculating an average over the last 200 periods (recall that we simulate opinion exchange for a total of 1,000 periods),

$$\overline{y}_j = \frac{1}{200} \sum_{t>800} y_{j,t} \qquad \text{and} \quad \overline{P}_j \equiv \frac{1}{200\hat{p}} \sum_{t>800} P_{j,t},$$

where polarization is normalized by $\hat{p}$ to belong to the interval $[0, 1]$. The degree of misinformation is computed similarly.[10] Table 3 reports relevant statistics for our three variable of interest.[11]

Fig. 7 displays the distribution of long-run average opinions $\overline{y}_j$ and polarization $\overline{P}_j$ across networks. While there are cases in which individuals are on average correct (e.g. close to $\theta = 0.5$), there exists significant dispersion around this value with a

[8] In our model, the mean opinion in group K at date zero is $\mu_k = \frac{\alpha}{\alpha+\beta}$ and its variance $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Using these two conditions, we find that $\alpha = -\frac{\mu_k(\sigma^2+\mu_k^2-\mu_k)}{\sigma^2}$ and $\beta = \frac{(\sigma^2+\mu_k^2-\mu_k)(\mu_k-1)}{\sigma^2}$. Since $\sigma^2 = 0.03$, the draws of μ_k determine initial beliefs for agents in each group.

[9] Note that the intervals used to compute polarization are pre-determined. We first set $K = 7$ and split the interval $[0, 1]$ in 7 groups, so the first sub-interval is $[0, 1/7]$, the second one is $(1/7, 2/7]$ and so on. In the first period of the simulation, each group has as a mass $\pi_{i,0} = 1/7$, with $i \in \{1, \dots, 7\}$. This is because we start initial opinions from a uniform distribution. Over time, $\pi_{i,t}$ (the proportion of the population that belongs to each group) changes endogenously. Hence if we end up in a situation where the whole population converges to 1, we would have $\pi_i = 0$ for $i < 7$ and $\pi_7 = 1$.

[10] With $\varsigma = 0.5$ the maximum possible level of polarization (theoretically) is $\hat{p} = 2\left(\frac{1}{2}\right)^{2.5} = 0.35$. 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.

[11] These are computed across simulations. For example, mean average opinions is $\sum_{j=1}^{1,233} y_j$.

Table 3
Long-run opinions, misinformation, and polarization.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Average opinions $\overline{y}_j$	0.51	0.28	0.03	0.26	0.75	0.97
Misinformation $\overline{MI}_j$	0.08	0.06	0.01	0.03	0.14	0.23
Average polarization $\overline{P}_j$	0.08	0.02	0.03	0.06	0.09	0.16

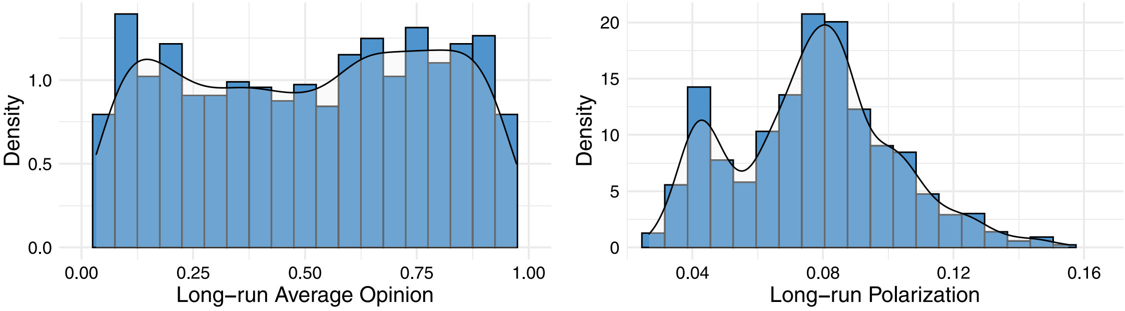


Fig. 7. Long-run average opinions and polarization.

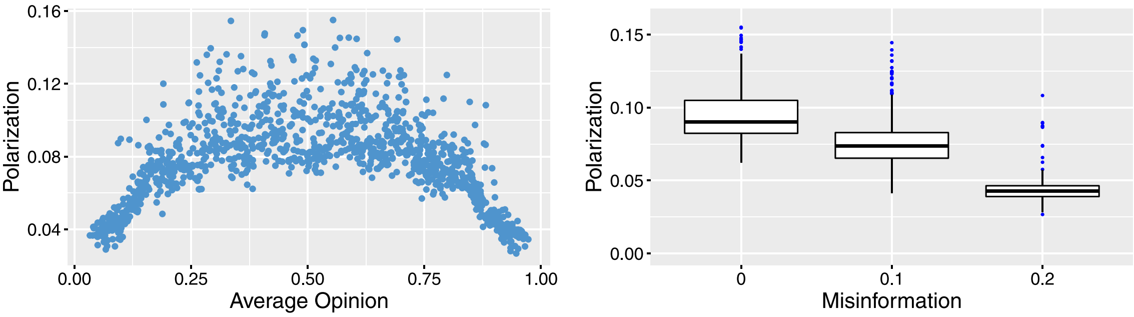


Fig. 8. Relationship between $\bar{P}_j$ and $\bar{y}_j$ (left) and $\overline{MI}_j$ (right).

distribution that looks almost uniform. Additionally, there is a non-trivial amount of networks in which agents' opinions become extreme, implying that bots are successful at manipulating options towards their own. As a result, the degree of misinformation could be quite large (recall that the theoretical maximum value for misinformation is 0.25).

The right panel of Fig. 7 depicts the distribution of polarization resulting from our simulation exercise. There is a significant degree of variability in our sample, even though the polarization levels are relatively small (recall that maximum polarization has been normalized to 1, yet the maximum polarization level observed in our sample is just 0.16). The average value of $\bar{P}_j$ across networks is 0.08, with a standard deviation of 0.02. Interestingly, we also observe some mass near 0, indicating that agents reach quasi-consensus. Unfortunately, most of these cases involves consensus around extreme values of θ rather than efficient aggregation of information to the true θ.[12]

Fig. 8 (left panel) shows a scatter-plot of long-run polarization and average opinions for our benchmark parameterization. There is an inverted U-shape relationship between these variables, indicating that polarization is generally low when individuals converge to the wrong value of θ (e.g. when long-run opinions are close to 0 or 1) and it is larger when average opinions are close to 0.5. Because of this, polarization and misinformation are negatively related, as seen in the right panel of the figure. The bars inside each box correspond to the median values of polarization, the top of the box indicates the 75th percentile, the bottom the 25th percentile, and the dots capture outliers.

5. Fake news and bot centrality

Bots need to be influential enough (or reach influential enough followers) in order to prevent society from learning. The number and location of bots (and their followers) in g_0 are thus important determinants of limiting $\overline{MI}$ and $\overline{P}$, as it is through them that fake news spread in the network. In our first set of experiments, we keep the share of bot followers constant.

[12] While we work with Esteban and Ray's measure of polarization, using the variance of opinions in the long-run is also a feasible alternative. The two measures of disagreement are highly correlated, with a correlation coefficient of 0.88. See Figure 20. in Appendix G.

Table 4
Centrality of bots and followers.

	Mean	Std Dev.	Min	Max
Average Centrality				
Avg in-Degree	22	5	12	96
Avg out-Degree	23	2	13	64
Avg Betweenness	0.009	0.006	0.003	0.033
Avg Page Rank	0.07	0.007	0.055	0.105
Avg Page Rank (bot)	0.007	0.002	0.003	0.013
Relative Centrality				
Rel in-Degree (bot)	296	172	1	596
Rel in-Degree	5	8	0	153
Rel out-Degree	3	3	0	80
Rel Page Rank	0.07	0.007	0.055	0.105
Rel Page Rank (bot)	0.007	0.004	0	0.022
Rel Betweenness	0.01	0.01	0	0.05

The variability in our long-run outcomes arises from two main sources. The first one is how much impact bots can have, on average. This depends on how influential the bot followers are. If bots can reach very central followers, they will tilt opinions towards the extremes, generating significant disagreement on the population, and hence high polarization. Their *average* centrality matters. The second one arises because the L-bot and the R-bot are competing for the attention of a potential (fixed) pool of followers. A bot will be able to tilt the population's opinions towards its preferred value if it can successfully: (i) reach a larger audience than the other bot (e.g. have a higher in-Degree) and/or (ii) reach more influential followers. This force could, potentially, reduce polarization but would increase misinformation. The *relative* centrality, then, should also impact long-run outcomes.

***Average centrality*:.** We first calculate the centrality measure of interest: in-Degree, out-Degree, Betweenness or Page Rank of all the agents following a bot (formal definitions of these centrality variables can be found in Appendix F). We then average them out across bots. For example, for Avg in-Degree, we first compute the total in-Degree (normalized) of all the R-bot followers as

$$\text{inDegree(R)} = \sum_{i \in N^{in}_{i,R}} \sum_{j} \frac{[g_0]_{j,i}}{n-1}.$$

The inner-sum computes the number of agents j who pay attention to i (e.g., the in-Degree of agent i, normalized by the potential number of followers). The outer-sum adds up the in-Degree of all i agents who are R-bot followers. Hence, it is affected by two forces: how popular the R-bot followers are (measured by their in-Degree) and how many followers the R-bot has. The in-Degree(L) is analogously computed. The Average in-Degree is just

$$\text{Avg in-Degree}(bot) = \frac{\text{inDegree(L)} + \text{inDegree(R)}}{2},$$

Note that we are taking the average across bots (not across followers). We can interpret this statistic as the overall indirect influence of bots in the network: higher values indicate that bots are influencing a large number of regular agents via bot followers. The results are reported at the top of Table 4. The first two rows indicate that bot followers are representative, with average in-Degree and out-Degree measures close to those in the population. However, we see significant difference across simulations. There are networks in which bots capture followers who have up to 96 followers themselves.

While average in-Degree is an intuitive measure of centrality, it is not necessarily the only way in which a bot can be efficient at manipulating opinions, and hence affecting misinformation and polarization. There are networks in which bot followers have very few followers (and hence a low in-Degree) but each of their followers is very influential. An alternative measure of centrality that incorporates these indirect effects is Google's Page Rank centrality.[13] Page Rank tries to account not only for 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). Individuals with a high Page Rank score are key agents of the network because other relevant network agents interact often with them. In the table, we report Page Rank centrality statistics for both, the bots and their followers.

To illustrate how this measure can affect outcomes, we plotted the distribution of polarization across simulations in Fig. 9 for different levels of bot's Average Page Rank.[14] The left panel—which includes cases with Avg Page Rank(bot) in the bottom 20th percentile—exhibits significantly less polarization than the right panel —which conditions the sample to the top 20th percentile of Avg Page Rank(bot)—. Hence, networks with more influential bots tend to be more polarized.

Our final measure is average Betweenness. Betweenness centrality computes how often an agent (or node) lies on the shortest path between any two agents in the network. Individuals with high Betweenness centrality have the potential to influence individuals near them and quickly spread fake news.

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

[14] The measure simply averages out the Page Rank of the L-bot and the R-bot. We use the average because we want to measure overall influence of bots, rather than relative influence.

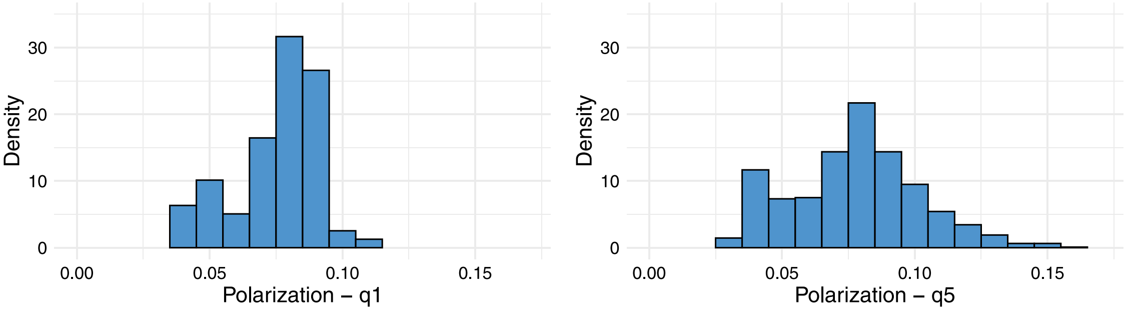


Fig. 9. Average Page Rank (bot) and long-run polarization.

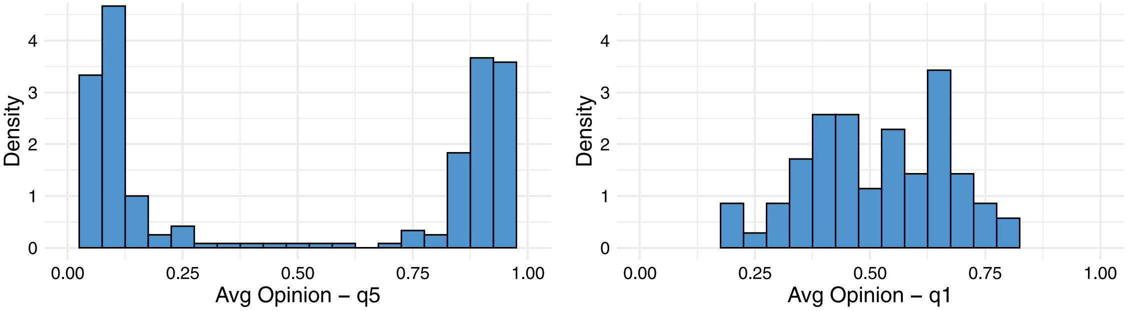


Fig. 10. Relative in-Degree (bot) and long-run opinions.

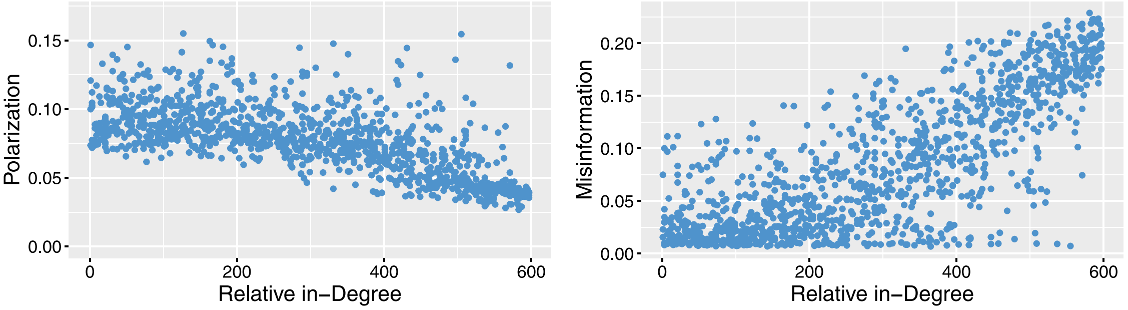


Fig. 11. Polarization and misinformation Relative in-Degree(bot).

Relative centrality :. While we keep the total number of followers constant, we allow the number of followers each bot has to change across simulations. This changes how central a bot is relative to the other. We compute Relative in-Degree as

$$\text{Relative in-Degree}(bot) = |\text{inDegree(L) - inDegree(R)}|.$$

When the variable is zero, the two bots are equally influential, whereas maximum centrality is reached when all bot followers pay attention to just one of the bots (e.g. the measure is 600). One average, the relative in-Degree in our sample is 296 agents – with a standard deviation of 172 followers –, suggesting large variability in the configuration of our different networks.

To better understand how this affects communication in Twitter, we plotted the distribution of long-run average opinion for two different levels of relative in-Degree in Fig. 10. When one of the bots has a significant larger number of followers (e.g. the upper 20th percentile of relative in-Degree), society converges to extremely biased opinions about θ (see the left panel). When the relative in-Degree is in the lowest 20th percentile, average opinions are centered around the true $\theta = 0.5$ (see the right panel). As a result of this, polarization is negatively correlated to the relative in-Degree of bots (see Fig. 11), whereas misinformation increases with it. The relationship between misinformation and in-Degree advantage is less stark, indicating that this measure of centrality is not enough to explain the variability in our simulations.

Table 4 also reports the relative in-Degree of bot followers (i.e, the difference between the in-Degree of a follower of bot L and a follower of bot R, in absolute value). While the average advantage is just 5 agents, it can reach up to a maximum of 153 (for reference, recall that agents have on average 21 followers each). The relative out-degree is 3; a larger value of this measure increases the influence of friends in the network for each given bot follower, reducing the influence of bots.

It is worth noticing that even though all of these are alternative measures of centrality, they capture slightly different concepts. An agent is central according to in-Degree when it has a large number of followers, whereas she is central according Betweenness if she is in the information path of many agents. The correlation between these two variables is just 0.2. Moreover, the correlation of these centrality measures is relatively low, with the exception of Relative in-Degree(bot) and Page Rank (bot follower), with a correlation coefficient of 0.93. This indicates that when a bot has a significantly larger number of followers it is capable of attracting the attention of relatively influential ones. See Appendix G.1 for an illustration of these correlations.

5.1. Regression analysis: assessing centrality measures

We are interested in estimating which of these relative centrality measures has the largest effect on long-run misinformation and polarization. To assess the quantitative importance of each explanatory variable, we estimate the coefficients of an OLS model,

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

where the $M \times 1$ vector Y_j denotes either long-run misinformation $\overline{MI}_j$ or polarization $\bar{P}_j$ obtained from simulation $j \in \{1 \dots M\}$, $\mathbf{X}_j$ denotes the matrix of network characteristics per simulation j, and ϵ_j is the error term. The set of explanatory variables includes: (i) the average and relative centrality measures from Table 4, (ii) initial opinions homophily, (iii) initial polarization, and (iv) average clustering, as those are the ones that vary as we randomly select bots' audience.

The results are displayed in Table 5. All variables have been normalized by their sample standard deviation (computed for our benchmark case) in order to simplify the interpretation of coefficients and ease comparison across covariates. Hence, each estimated coefficient represents by how many standard deviations misinformation or polarization change when the respective independent variable increases by one standard deviation. The columns under 'All' include the full sample, whereas those under $\overline{P}_{High}$ restrict the sample to networks in which polarization is high (in the top 25th percentile of $\overline{P}_j$). Analogously, columns under $\overline{MI}_{High}$ restrict the sample to networks with the top 25th percentile of $\overline{MI}_j$.

Full sample *:.* As bots become more central, subgroups of agents converge to extreme views about θ and polarization increases. This is evidenced by the finding that a one-standard deviation increase in Average Page Rank (bot) increases polarization by 0.32 s.d. Relative centrality, measured by the bot's Page Rank, has the greatest impact by lowering polarization (by 0.75 s.d) and increasing misinformation (by 0.87 s.d.). When a bot is relatively more influential than the other, it is able to nudge most opinions towards its extreme view. This lowers disagreement at the expense of higher misinformation.

In simulations where bot followers are popular (i.e high Avg in-Degree), the problem worsens because they influence a large number of people as well, increasing polarization by 0.15 s.d. Increases in Avg out-Degree have the expected effect but are quantitatively small. At the same time, networks in which both bots and their followers are influential experience lower misinformation levels, as evidenced by the regression coefficients of -0.22 (Avg Page Rank of bot) and -0.12 (Avg in-Degree of follower). This result is consistent with Fig. 8, where we saw that polarization and misinformation are negatively related. As we show in later (in Section 6.4), this is a particular feature of the benchmark case. When bots are equally biased (as assumed when $\theta = 0.5$) and equally influential, they fail at nudging agents' opinions away from the middle of the $[0, 1]$ spectrum. Thus, the bots' balanced influence and informative signals jointly reduce misinformation.

Increases in the relative in-Degree advantage of followers has similar effects by reducing polarization and increasing misinformation by about 0.10 s.d. each. Initial polarization has no significant effects on long-run polarization, suggesting that initial conditions do not matter. The effects of opinions' Homophily are relatively small and those of Average clustering statistically insignificant.

We want to point out that the largest effects on long-run outcomes arise from the variability in the relative number of bot followers. That is, in cases where bots are *asymmetric*. We have experimented with cases in which bots are symmetric (with in-Degree of 300 each) and found that the coefficients on the centrality of bot followers are smaller and the explanatory power for polarization decreases significantly.

High polarization subset *:.* In the table we also replicated the regression for a subset of our sample in which high levels of polarization were reached in the long run (second and fourth columns, under $\overline{P}_{High}$). The most important predictor for polarization is the average in-Degree of bot followers, which increases it by 0.3 sd (which is twice what we obtained using the full sample), followed by the average Page Rank of the bot. Recall that high levels of polarization are typically attained when bots are influential, but in a symmetric way. This is why the coefficients on relative centrality measures have a lower impact, and the goodness of fit shrinks. Interestingly, Avg Betweenness affects polarization in this case: a one s.t. reduces disagreement by 0.16 s.d. Conditional on high polarization levels, a one-s.d. in the relative Page Rank advantage (bot) increases misinformation by 0.54 s.d. (the value for the full sample was 0.87). Other measures of centrality are quantitatively small or insignificant, which probably happens because simulations in which polarization is high tend to exhibit low misinformation levels.

High misinformation subset *:.* In the third and sixth columns (those under $\overline{MI}_{High}$), we instead condition our sample to include cases in which the level of misinformation is high. In these cases, the explanatory power of our centrality measures for both, polarization and misinformation, increases. Moreover, the effects of bot centrality are exacerbated when compared to the full sample. In other words, when misinformation is already high, marginal increases in the centrality have very significant effects in reducing polarization and increasing misinformation even further, evidence of increasing returns in the spread of fake news technology.

Table 5
Regression results: benchmark case.

	Polarization			Misinformation		
	Full sample	$\overline{P}_{High}$	$\overline{MI}_{High}$	Full sample	$\overline{P}_{High}$	$\overline{MI}_{High}$
Bot Centrality						
Avg Page Rank (bot)	0.32***	0.20***	1.59***	−0.22***	−0.09***	−0.83***
	(0.028)	(0.041)	(0.130)	(0.014)	(0.020)	(0.072)
Rel in-Degree (bot)	−0.24***	−0.19*	0.16	0.13***	0.00	−0.06
	(0.065)	(0.112)	(0.125)	(0.034)	(0.056)	(0.069)
Rel Page Rank (bot)	−0.75***	−0.27***	−2.27***	0.87***	0.54***	1.42***
	(0.034)	(0.061)	(0.165)	(0.018)	(0.030)	(0.092)
Bot Follower Centrality						
Avg in-Degree	0.15***	0.30**	−0.02	−0.12***	0.04	−0.04**
	(0.040)	(0.140)	(0.026)	(0.021)	(0.070)	(0.014)
Avg out-Degree	−0.04*	0.02	0.01	0.03**	−0.07*	0.00
	(0.026)	(0.078)	(0.016)	(0.013)	(0.039)	(0.009)
Avg Betweenness	−0.02	−0.16**	0.22***	0.05*	−0.00	0.06*
	(0.053)	(0.072)	(0.060)	(0.027)	(0.036)	(0.033)
Avg Page Rank	−0.04	−0.05	−0.11**	−0.00	0.02	0.00
	(0.038)	(0.076)	(0.053)	(0.020)	(0.038)	(0.030)
Rel in-Degree	−0.11***	−0.10	0.01	0.10***	0.06	0.02*
	(0.034)	(0.091)	(0.021)	(0.017)	(0.045)	(0.012)
Rel out-Degree	−0.01	0.09	−0.01	0.02*	−0.00	0.01
	(0.023)	(0.070)	(0.013)	(0.012)	(0.035)	(0.008)
Rel Betweenness	0.02	0.08	−0.21***	−0.03	−0.02	−0.04
	(0.051)	(0.072)	(0.059)	(0.026)	(0.036)	(0.033)
Rel Page Rank	0.08	0.19	0.36**	0.09**	0.07	0.00
	(0.077)	(0.123)	(0.148)	(0.040)	(0.061)	(0.082)
Network Structure						
Opinions Homophily	−0.04**	−0.01	−0.02	0.01	0.01	−0.01
	(0.017)	(0.032)	(0.016)	(0.009)	(0.016)	(0.009)
Initial Polarization	−0.02	−0.04	−0.00	0.01	−0.01	0.01
	(0.017)	(0.028)	(0.015)	(0.009)	(0.014)	(0.008)
Avg Clustering	−0.02	−0.01	−0.00	−0.02	−0.02	−0.02*
	(0.026)	(0.049)	(0.021)	(0.013)	(0.024)	(0.012)
Observations	1,233	308	308	1,233	308	308
R-squared	0.640	0.192	0.695	0.904	0.609	0.846

Standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

6. Parameter analysis

In this section, we study how changes in selected model parameters affect polarization and misinformation in the long run.

6.1. Alternative bots' flooding capacity

In our model, the parameter κ controls the speed at which bots' beliefs become extreme (e.g. L-bot's opinion moves towards 0 and R-bot's opinion moves towards one). A low value of the flooding parameter implies high persistence in the bot's initial opinion, which need not be extreme, and hence a lower degree of influence on the dynamics of communication. To study how κ affects communication, we computed long-run average opinions and polarization for $\kappa \in \{0.5, 1, 5, 25\}$. Their long-run distributions across network simulations are plotted in Fig. 12.

As κ declines from our benchmark value of 25, average opinions become more concentrated around the true value of θ (left panel) and polarization shrinks (right panel). Intuitively, a low κ reduces the influence of the bot and allows more efficient aggregation of information and consensus to the true θ. In Appendix H.1, we show results for flooding parameters above 25 and note that the changes in long-run average opinions and polarization are not as sensitive to increases in κ, as values are higher than the benchmark. Qualitatively, more flooding above 25 is associated with a wider distribution for opinion and higher polarization levels, but the distributions look very similar to those in our benchmark case.

6.2. Alternative number of bots

In Fig. 13 we display the long-run outcomes as the number of bots increases, with the flooding parameter constant at our benchmark value of $\kappa = 25$ and $\mu = 0.15$ (so the number of its followers does not change). When the number of bots increases from

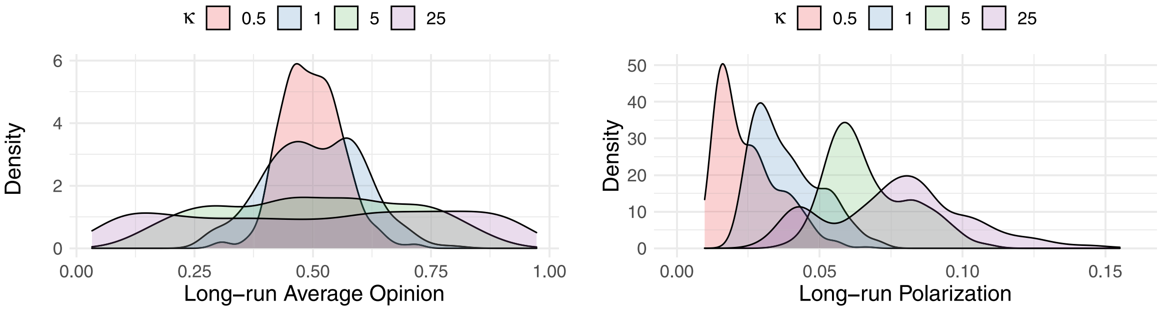


Fig. 12. The effects of bot's flooding capacity κ.

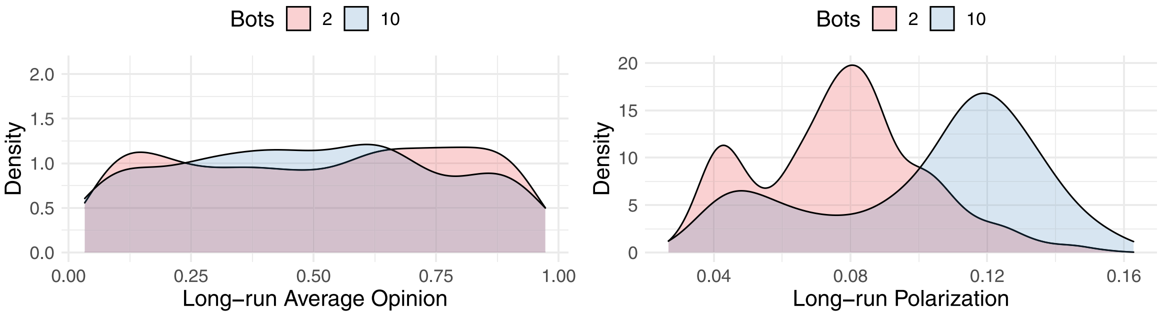


Fig. 13. Long-run outcomes and number of bots.

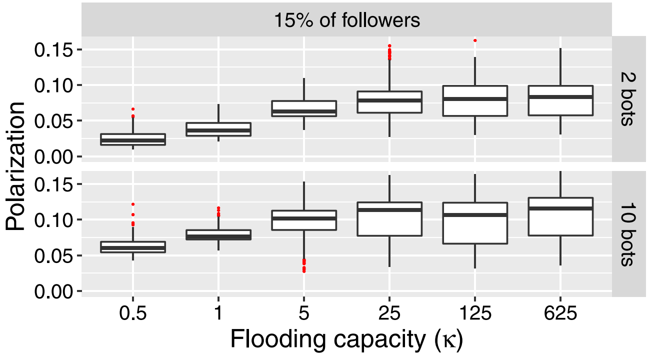


Fig. 14. Number of bots vs flooding and polarization.

2 to 10, opinions become slightly more concentrated (left panel) but polarization increases significantly, as seen by the fact that whole distribution shifts to the right (right panel).

Notice that this comparative static is different from the previous one: adding more bots to influence the same number of agents is analogous to increasing the weight of bots in the updating function of agents. Before, we were affecting the speed at which bots became extreme. Now, we are increasing the attention span they receive from their followers instead.

Number of bots vs flooding **:.** That increasing the number of bots is not the same as increasing the flooding parameter can be seen more clearly in Fig. 14, which shows long-run polarization for different combinations of κ and number of bots. The clearest contrast is in the case in which there are 2 bots with $\kappa = 25$ (as in the benchmark), vs the case with 10 bots and $\kappa = 5$, so that each type of bot has roughly 25 signals per period. In the first case, each bot is more extreme, and hence able to pull the network towards the edge values. As a result, polarization goes down. In the latter, agents pay more attention to bots than to other friends, so larger groups with opposite views arise. However, bots have more symmetric power and manage to create a more polarized society (e.g. it is more difficult for one individual bot to become more influential than the other). Polarization is on average 20% larger in the case in which there are 10 bots with 5 signals each.

The figure also illustrates that more bots result in more polarization for all levels of flooding and that more flooding results in more polarization, although polarization levels out for $\kappa \geq 25$.

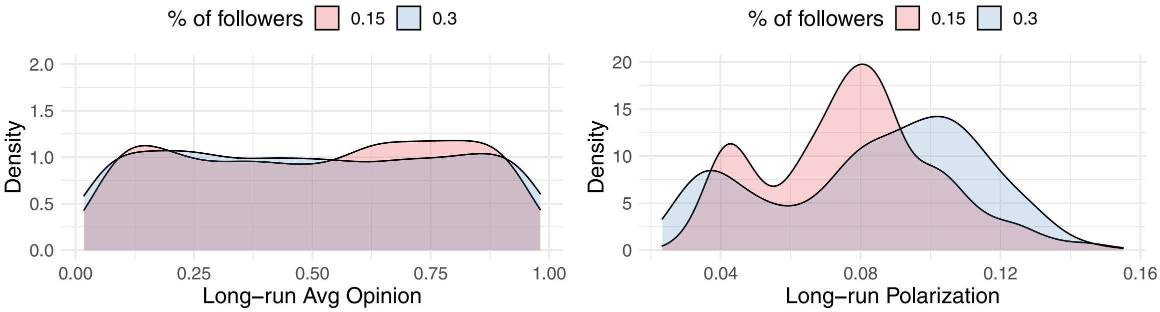


Fig. 15. Long-run outcomes and share of bot followers.

6.3. Alternative bots' audience size

In this experiment, we keep the number of bots and the flooding as in the benchmark (e.g. 2 bots and $\kappa = 25$), but double the share of bot followers (e.g. from $\mu = 0.15$ to $\mu = 0.3$).

Fig. 15 depicts the distributions of long-run average opinions (left) and polarization (right). As bots become more proficient in creating undetectable fake news, they generate more polarization on average. That the dispersion of polarization increases indicates that they are also more likely to end up in scenarios where there is either a large proportion of individuals with extreme views (large polarization) or most of the network pulled to one extreme view (low polarization).

The distribution of opinions becomes slightly flatter, but it is not significantly affected by the increase in μ. This happens because bots' opinions become extreme relatively quickly when $\kappa = 25$. Under $\kappa = 0.5$, for example, the difference between the distribution of opinions for alternative values of μ is more pronounced (see Figure 22 in Appendix G).

6.4. Regression analysis: assessing centrality measures

We repeat the main regression exercise from Table 5, but pooling results from a wider configuration of networks and parameters. In particular, we consider alternative values of flooding capacity $\kappa \in \{0.5, 1, 5, 25, 125, 625\}$, number of bots $b \in \{2, 10\}$, and share of bot followers $\mu \in \{0.15, 0.3\}$. To control for these, we include a set of dummy variables, in addition to the explanatory variables included in our previous estimation. Our total sample is now $M = 10{,}980$ observations. The estimated coefficients for selected explanatory variables are presented in Table 6, also normalized by their sample standard deviation (the full table is shown in Appendix H.3 Table 8).

The first set of regressors are the average and relative centrality of bots and their followers. The coefficients for both polarization and misinformation have the same sign but are smaller in magnitude if compared to the benchmark case. For example, a one s.d. increase in the relative Page Rank of the bot reduces polarization by 0.53 s.d. (vs 0.75 in the benchmark case) and increases misinformation by just 0.44 s.d. (vs 0.87 in the benchmark). Here we can see that bots' relative Page Rank and followers' relative Page Rank are the main drivers of misinformation and are equally relevant to nudge public opinion away from the true state. Unlike in the benchmark case, Avg Page Rank (bot) is now neutral to misinformation.

Average clustering, which was statistically insignificant in the benchmark case, now becomes a relevant explanatory variable (because we change the number of bot followers in some specifications, there is now enough variability in clustering). Higher clustering means that bot followers are communicating with each other. Therefore, there is a reinforcement aspect which reduces polarization, while increasing misinformation.

We also include a set of dummy variables to account for potential differential effects in parameters changed. The first one relates to the share of bot followers, and it is equal to 1 when $\mu = 0.3$. The second one captures the number of bots, and it is equal to 1 when there are 10 bots. Finally, we add a set of dummy variables for different values of κ. The results also show that, interestingly, a larger share of followers (dummy for $\mu = 0.30$) and more bots (dummy for $b = 10$) reduce misinformation relative to the benchmark case. This counter-intuitive result has to do with the fact that equally influential bots induce information aggregation when $\theta = 0.5$. Since bots are, on average, equally balanced, they fail at pulling people's opinions to the extreme points of the $[0, 1]$ spectrum, leaving room for informative signals to nudge agents opinions towards the true state.

Polarization and misinformation increase monotonically with respect to the bots' flooding capacity κ (see Figs. 17 and 18 below), but display decreasing marginal returns. Moreover, the benchmark flooding capacity $\kappa = 25$ seems to be at a saturation point.

6.5. Alternative state of the world

Fixing the true state at the level 0.5 can be seen as a very particular special case because, as in Como and Fagnani (2016), equally influential bots may exercise homogeneous influence, i.e. may nudge public opinion to a common value which is a convex combination of bots' extreme opinions. In this case, this common value would be 0.5, i.e. the true state. Thus, this particular situation could suggest that the bots' presence is conducive to efficient information aggregation. But one may ask what happens when the

Table 6
Regression results for $\theta = 0.5$: varying μ, κ, b.

	Polarization	Misinformation
Avg Page Rank (bot)	0.48***	0.04
	(0.024)	(0.025)
Rel in-Degree (bot)	−0.12***	−0.03
	(0.027)	(0.028)
Rel PageRank (bot)	−0.53***	0.44***
	(0.012)	(0.013)
Avg in-Degree	0.12***	−0.15***
	(0.010)	(0.010)
Avg Page Rank	−0.11**	0.01
	(0.045)	(0.047)
Rel in-Degree	−0.08***	0.12***
	(0.008)	(0.009)
Rel PageRank	−0.05	0.47***
	(0.030)	(0.031)
Avg Clustering	−0.34***	0.12***
	(0.030)	(0.031)
$\mu = 0.3$	0.36***	−0.44***
	(0.085)	(0.089)
$\kappa = 0.5$	−1.01***	−1.10***
	(0.015)	(0.016)
$\kappa = 1$	−0.63***	−0.93***
	(0.015)	(0.016)
$\kappa = 5$	−0.19***	−0.37***
	(0.015)	(0.015)
$\kappa = 125$	−0.01	0.10***
	(0.016)	(0.017)
$\kappa = 625$	0.02	0.11***
	(0.016)	(0.017)
Bots = 10	0.27***	−0.18***
	(0.039)	(0.041)
Observations	10,980	10,980
R-squared	0.764	0.741

Standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

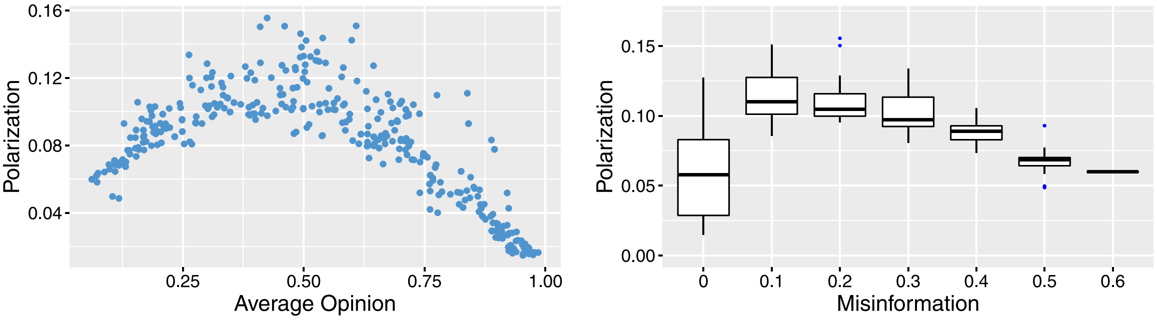


Fig. 16. Long-run outcomes for alternative state of the world ($\theta = 0.8$) .

true state is not exactly 0.5. To assess this case, we consider an alternative scenario where $\theta = 0.8$ (and keep all other parameters at the benchmark case). Despite being another special case, it illustrates how our main results are sensitive to the parameter choice, i.e. when the true state moves away from the center towards extreme points in the [0.1] spectrum.

Fig. 16 is the counterpart of Fig. 8, but assuming $\theta = 0.8$. As before, we can see an inverted U-shape relationship between long-run polarization and long-run average opinions (left). Differently from before, polarization levels are not at the maximum when opinions are close to the true value of θ. Moreover, polarization and misinformation have also an inverted U-shape relationship now (before, the two were positively correlated). There are cases in which polarization is low and misinformation is basically nil (these correspond to efficient information aggregation). In most of our simulations, increases in misinformation are related to *decreases* in polarization. These correspond to those in which the L-bot is relatively more influential, pulling beliefs towards 0. Another interesting observation is that now the range of possible misinformation levels widens significantly (before, it was between 0 and 0.2, now the upper bound increases to 0.6). This happens because, theoretically, we can also have more misinformation when one of the bots is disproportionately far away form the true θ (in this case the L-bot). One way to interpret this experiment is to think of the L-bot as being more extreme than the R-bot. In our benchmark case, they had biases that were symmetric from the true state of the world.

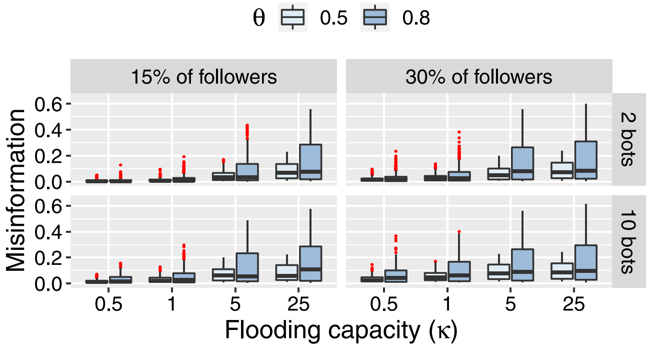


Fig. 17. Misinformation for alternative specifications for benchmark ($\theta = 0.5$) and alternative state ($\theta = 0.8$).

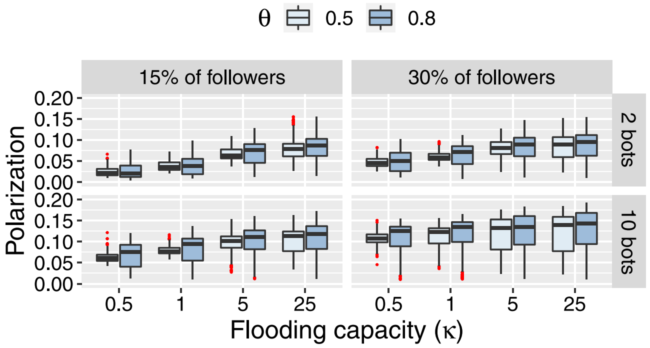


Fig. 18. Polarization for alternative specifications for $\theta = 0.5$ and $\theta = 0.8$.

6.5.1. *Parameter analysis for* $\theta = 0.8$

Long-run misinformation levels for different combinations of flooding κ, number of bots, and share of bot followers μ are displayed in Fig. 17. The bars correspond to the median values, and the boxes represent the variability in each specification produced by the fact that bots and bot followers may have different centrality (as before, for each set of parameters, we populate the network with bots and their followers randomly).

The most striking difference in experiments arises when flooding is large enough. While average misinformation levels are about the same, the variance of misinformation (size of the boxes) is significantly wider in the $\theta = 0.8$ case. When $\kappa \leq 1$, there is basically no difference in long-run misinformation. This happens because when bot-L is relatively more influential, it can pull most of the network towards 0, and hence further away from $\theta = 0.8$ (than in the $\theta = 0.5$ scenario). The larger κ, the more likely the outcome.

Polarization outcomes across network specifications are displayed in Fig. 18. Polarization grows with flooding, number of followers, and number of bots regardless of the value of θ. Hence, making one bot more extreme than the other does not change this previous finding. Long-run polarization is, on average, slightly larger with $\theta = 0.8$. The variability of polarization outcomes (size of the boxes) is larger with 2 bots, but not necessarily with 10 bots.

6.5.2. *Regression analysis for* $\theta = 0.8$

The regression results for this alternative value of θ are shown in Table 9 of Appendix H.3.2.. The results in the first column (polarization) are very much in line with those in Table 8: the parameters are similar in magnitude and significance, and the goodness of fit is around the same. This suggests that bot centrality affects polarization in about the same way for both values of θ considered (note that we are repeating the exact same simulations, but with a different value of θ). In terms of misinformation, our linear regression model does significantly worse (the R^2 is just 0.24 now, versus 0.74 in the previous case). This is probably due to the fact that our agents also update beliefs according to the unbiased signal, which pushes them towards 0.8. Because this is close to the preferred point of the R-bot, cases in which bots are not very central (so individuals converge to the true) and cases in which the R-bot is relatively more central (so agents get close to 1) are confounded in the sample. In other words, we have observations that deliver similar levels of misinformation (e.g. average opinions closer to 1) in which the L-bot is either very influential or not influential at all.

Another important point is that now a larger share of followers (dummy for $\mu = 0.30$) and more bots (dummy for $b = 10$) are no longer relevant to reduce misinformation, which reinforces the claim that $\theta = 0.5$ is a very special case where bots may help decreasing misinformation. The bot followers' popularity (as measured by the average in-Degree) still decreases misinformation. Higher average in-Degree means that bot followers are likely following each other and, therefore, they are balancing out their effect, which leaves room for informative signals to instruct the society. This interpretation is corroborated by the positive effect of relative in-Degree on misinformation, e.g. unbalanced bots tend to harm more the information aggregation process.

7. Threshold rules: a bounded confidence model

Our baseline model considered an opinion formation process where agents unreservedly place equal weight on friends they meet in every period of time. Despite its simplicity, such rule does not allow agents to place more attention to the opinion of like-minded friends. In this section we explore a variation of the update rule described on Eqs. (2) and (3) to assess how polarization and misinformation respond to such homophilic interaction.

We augment our model by considering the opinion exchange process proposed by Krause et al. (2000) (see also Hegselmann et al., 2002) and Deffuant et al. (2000) to allow for an agent to pay attention only to other agents whose opinions do not differ much from his or her own. The normalized adjacency matrix in Eqs. (2) and (3) is now replaced by the following:

$$\left[\hat{g}_t\right]_{ij} = \begin{cases} \frac{1}{|N_{i,t}^{out}(d)|} & \text{, if } |y_{i,t} - y_{j,t}| < d \text{ , where } |N_{i,t}^{out}(d)| = \left| \left\{ k : |y_{i,t} - y_{k,t}| < d \right\} \right| \\ 0 & \text{, otherwise.} \end{cases}$$

According to this weighting rule, an agent places equal weight on all opinions that are within some distance d of his or her own current opinion, and no weight to opinions that are further apart. We refer to it as a 'threshold rule.' Although agents place equal weight on his or her friends, this rule implies that agents end up placing more weight on the opinion of like-minded friends. The threshold rule substantially complicates the opinion formation process (and increases computation time), as the updating depends on the specifics of the opinions (signals received, friends met, and bots' influence) and the distance between an agents' beliefs and those of the people in his or her neighborhood. Moreover, the weights change endogenously over time.

In order to study how the threshold rule impacts both long-run polarization and misinformation, we again ran multiple Monte Carlo simulations for different distance thresholds d in the grid $\{0.10, 0.15, 0.30\}$ and compare them to the benchmark case. For that, we fixed all other parameters as in the benchmark case (see Table 2). Fig. 19 displays the distribution of polarization and misinformation across simulations of the four cases considered.[15]

The simulations above suggest that as the distance d shrinks, i.e. as agents become more attached to like-minded friends, both median polarization and median misinformation decrease monotonically. The intuition behind this result is as follows: (i) since agents are very restrictive with respect to whom to exchange information with, they tend to behave as fully Bayesian agents since they place weight 1 to the informative signal if no agent meets the (stringent) attention requirement imposed by the distance d; (ii) since most agents are proportionally more exposed to other regular agents rather than bots, there is a higher chance that signals will induce these agents to aggregate information efficiently instead of being influenced (and trapped) by bots; (iii) given the points (i) and (ii), agents opinions start clustering around θ and, despite how restrictive agents are with respect to the distance d chosen, they resume communication again with agents that are more influenced by signals and end up reinforcing the relative importance of signals.

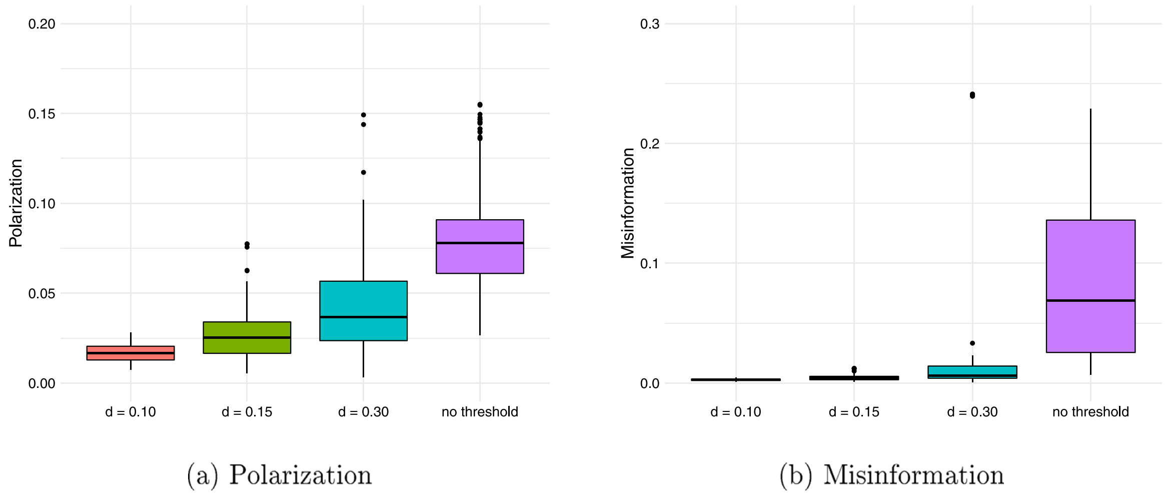
It is worth highlighting that these results contrast with the ones by Krause et al. (2000), Hegselmann et al. (2002) and Deffuant et al. (2000) where there is an emergence of opinion clusters (therefore, polarization) as agents become more attached to like-minded individuals. This is mainly because agents, in our model, receive a stream of informative signals over time. Thus, instead of clustering opinions, the threshold rule acts as a shield against bots influence, and opens room for agents to learn from signals. Moreover, besides shielding public opinion from bots' influence, the threshold rule also reinforces learning by allowing these individuals to resume communication after they collect many signals and become relatively wise.

A final remark regarding this exercise is that these results are definitely sensitive to the initial endowment of opinions. In our case, we start the process at $t = 0$ with agents' opinions uniformly distributed over the $[0, 1]$ spectrum. If one assigns extreme opinions to agents, the odds may change in favor of the bots. Such particular case may also depend on the κ (flooding capacity) and μ (share of followers) choices, i.e. as bots become more prominent, they may overcome the informativeness of signals and nudge public opinion to the extremes (either 0 or 1) with higher probability. For the case where $\theta = 0.5$, both polarization and misinformation would tend to the extremes. In our view, this is an interesting research strand.

8. Conclusions

We created a synthetic social media network, calibrated to Twitter, and populated it with bots and bot followers, considering alternative configurations of their location in the network. We then simulated the opinion exchange process in order to identify the most important drivers of misinformation and polarization. A premise in all of them is the ability of bots (with opposite biased views) to purposely spread fake news in order to manipulate the opinion of a small share of agents in the network. To the extent that agents can be partially influenced by these signals – directly by not filtering out fake news, or indirectly by following friends who are themselves influenced by bots –, this can generate misinformation and polarization in the long run. In other words, fake news prevent information aggregation and consensus in the population.

[15] The case where $\theta = 0.8$ can be found in Appendix H.4 All results and conclusions are in line with the ones for the benchmark case above.



(a) Polarization (b) Misinformation

Fig. 19. Polarization and Misinformation under a homophilic weight-rule and $\theta = 0.5$.

We find that when one of the bots is relatively more efficient at manipulating news (by targeting a small number of influential agents), it may be able to generate full misinformation in the long run, where beliefs are at one end of the spectrum. There would be no polarization in that case, but at the expense of agents converging to the wrong value of θ, the parameter of interest. These findings have important policy implications: Policymakers could either try to identify all bots and eliminate them from the network, and this would restore the efficient accumulation of information. However, such policy is difficult to implement because the cost of another bot popping in is relatively small. Alternatively, policymakers could introduce a 'counter-bot' sending signals at the other end of the spectrum to counter-balance the effect of the first bot. In this environment, simply reporting the true state of the world may not be sufficient. The counter-bot should be itself somewhat extreme, or influential enough to tilt opinions towards the true state. This may, however, end up exacerbating polarization and potentially inducing gridlock and inaction. This suggests the existence of a trade-off when combating misinformation through the use of counter-bots: the potential to enhance polarization. Reducing the amount of bot followers (e.g. training people on the detection of fake news) is probably a more effective strategy. The current practice of simply eliminating bots from a network, rather than training the social media users, may have undesired consequences because it may increasing asymmetries (e.g. by eliminating bots at one extreme, we are increasing the relative centrality of the bot at the other extreme).

In most of our paper, the links in the network evolve stochastically. In the last section, we analyzed threshold rules where links are endogenously determined, as agents place a higher weight on individuals who share similar priors. We find that – in our benchmark case – polarization and misinformation tend to be lower under such rules. However, we have not done an extensive analysis of threshold rules under initially high polarization levels. We believe this is a promising direction for future research.

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

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejpoleco.2022.102256.

References

Acemoğlu, D., Como, G., Fagnani, F., Ozdaglar, A., 2013. Opinion fluctuations and disagreement in social networks. Math. Oper. Res. 38 (1), 1–27.
Acemoglu, D., Ozdaglar, A., ParandehGheibi, A., 2010. Spread of (mis) information in social networks. Games Econom. Behav. 70 (2), 194–227.
Alstott, J., Bullmore, E., Plenz, D., 2014. Powerlaw: a python package for analysis of heavy-tailed distributions. PLoS One 9 (1), e85777.
Bala, V., Goyal, S., 1998. Learning from neighbours. Rev. Econom. Stud. 65 (3), 595–621.

Barberá, P., 2014. How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. In: Job Market Paper, vol. 46. New York University, pp. 1–46.

Boxell, L., Gentzkow, M., Shapiro, J.M., 2017. Is the Internet Causing Political Polarization? Evidence from Demographics. National Bureau of Economic Research.

Campante, F.R., Hojman, D.A., 2013. Media and polarization: Evidence from the introduction of broadcast TV in the United States. J. Publ. Econom. 100, 79–92.

Como, G., Fagnani, F., 2016. From local averaging to emergent global behaviors: The fundamental role of network interconnections. Systems Control Lett. 95, 70–76.

Deffuant, G., Neau, D., Amblard, F., Weisbuch, G., 2000. Mixing beliefs among interacting agents. Adv. Complex Syst. 3 (01n04), 87–98.

DeGroot, M.H., 1974. Reaching a consensus. J. Amer. Statist. Assoc. 69 (345), 118–121.

DellaVigna, S., Kaplan, E., 2007. The fox news effect: Media bias and voting. Q. J. Econ. 122 (3), 1187–1234.

DeMarzo, P.M., Vayanos, D., Zwiebel, J., 2003. Persuasion bias, social influence, and unidimensional opinions. Q. J. Econ. 118 (3), 909–968.

Ellison, G., Fudenberg, D., 1993. Rules of thumb for social learning. J. Polit. Econ. 612–643.

Ellison, G., Fudenberg, D., 1995. Word-of-mouth communication and social learning. Q. J. Econ. 93–125.

Esteban, J., Ray, D., 1994. On the measurement of polarization. Econometrica 62, 819–851.

Flaxman, S., Goel, S., Rao, J.M., 2013. Ideological segregation and the effects of social media on news consumption. Available At SSRN, 2363701.

Gentzkow, M., Shapiro, J.M., 2006. Media bias and reputation. J. Polit. Econ. 114 (2), 280–316.

Gentzkow, M., Shapiro, J.M., 2010. What drives media slant? Evidence from US daily newspapers. Econometrica 78 (1), 35–71.

Gentzkow, M., Shapiro, J.M., 2011. Ideological segregation online and offline. Q. J. Econ. 126 (4), 1799–1839.

Golub, B., Jackson, M.O., 2010. Naive learning in social networks and the wisdom of crowds. Am. Econ. J. Microecon. 2 (1), 112–149.

Goyal, S., 2005. Learning in networks. In: Group Formation in Economics: Networks, Clubs and Coalitions. Cambridge: Cambridge University Press, pp. 122–170.

Gu, L., Kropotov, V., Yarochkin, F., 2017. The fake news machine. In: How Propagandists Abuse the Internet and Manipulate the Public. Pobrane, vol. 25.

Halberstam, Y., Knight, B., 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. J. Publ. Econom. 143, 73–88.

Hegselmann, R., Krause, U., et al., 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. J. Artif. Soc. Soc. Simul. 5 (3).

Jadbabaie, A., Molavi, P., Sandroni, A., Tahbaz-Salehi, A., 2012. Non-Bayesian social learning. Games Econom. Behav. 76 (1), 210–225.

Krause, U., et al., 2000. A discrete nonlinear and non-autonomous model of consensus formation. Commun. Diff. Equat. 2000, 227–236.

Leskovec, J., Mcauley, J., 2012. Learning to discover social circles in ego networks. Adv. Neural Inf. Process. Syst. 25.

Martin, G.J., Yurukoglu, A., 2017. Bias in cable news: Persuasion and polarization. Amer. Econ. Rev. 107 (9), 2565–2599.

Webster, J.G., Ksiazek, T.B., 2012. The dynamics of audience fragmentation: Public attention in an age of digital media. J. Commun. 62 (1), 39–56.

Further reading

Acemoglu, D., Chernozhukov, V., Yildiz, M., 2016. Fragility of asymptotic agreement under Bayesian learning. Theor. Econ. 11 (1), 187–225.

Acemoglu, D., Dahleh, M.A., Lobel, I., Ozdaglar, A., 2011. BayesIan learning in social networks. Rev. Econom. Stud. 78 (4), 1201–1236.

Acemoglu, D., Ozdaglar, A., 2011. Opinion dynamics and learning in social networks. Dynam. Games Appl. 1 (1), 3–49.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., Wacziarg, R., 2003. Fractionalization. J. Econ. Growth 8 (2), 155–194.

Andreoni, J., Mylovanov, T., 2012. Diverging opinions. Am. Econ. J. Microecon. 4 (1), 209–232.

Aumann, R.J., 1976. Agreeing to disagree. Ann. Statist. 4 (6), 1236–1239.

Azzimonti, M., 2018. Partisan conflict and private investment. J. Monetary Econ. 23.

Baldassarri, D., Bearman, P., 2007. Dynamics of political polarization. Am. Sociol. Rev. 72 (5), 784–811.

Banerjee, A.V., 1992. A simple model of herd behavior. Q. J. Econ. 797–817.

Banerjee, A., Fudenberg, D., 2004. Word-of-mouth learning. Games Econom. Behav. 46 (1), 1–22.

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

Barber, M., McCarty, N., 2015. Causes and consequences of polarization. In: Solutions To Political Polarization in America, vol. 15. Cambridge University Press.

Buechel, B., Hellmann, T., Klößner, S., 2015. Opinion dynamics and wisdom under conformity. J. Econom. Dynam. Control 52, 240–257.

Chandrasekhar, A.G., Larreguy, H., Xandri, J.P., 2020. Testing models of social learning on networks: Evidence from two experiments. Econometrica 88 (1), 1–32.

Chatterjee, S., Seneta, E., 1977. Towards consensus: Some convergence theorems on repeated averaging. J. Appl. Probab. 89–97.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A., 2011. Political polarization on twitter. In: Proceedings of the International Aaai Conference on Web and Social Media, vol. 5. pp. 89–96.

Dixit, A.K., Weibull, J.W., 2007. Political polarization. Proc. Natl. Acad. Sci. 104 (18), 7351–7356.

Duclos, J.-Y., Esteban, J., Ray, D., 2004. Polarization: concepts, measurement, estimation. Econometrica 72 (6), 1737–1772.

Epstein, L.G., Noor, J., Sandroni, A., 2010. Non-Bayesian learning. B. E. J. Theor. Econ. 10 (1).

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

Esteban, J., Gradín, C., Ray, D., 2007. An extension of a measure of polarization, with an application to the income distribution of five OECD countries. J. Econom. Inequal. 5 (1), 1–19.

Esteban, J., Ray, D., 2012. Comparing polarization measures. In: Oxford Handbook of Economics of Peace and Conflict. Oxford University Press Oxford, pp. 127–151.

Fernandes, M., 2019. Confirmation bias in social networks. Available At SSRN 3504342.

Fiorina, M.P., Abrams, S.J., et al., 2008. Political polarization in the American public. Ann. Rev. Polit. Sci. 11, 563.

Golub, B., Jackson, M.O., 2012. How homophily affects the speed of learning and best-response dynamics. Q. J. Econ. 127 (3), 1287–1338.

Grabisch, M., Rusinowska, A., 2020. A survey on nonstrategic models of opinion dynamics. Games 11 (4), 65.

Groseclose, T., Milyo, J., 2005. A measure of media bias. Q. J. Econ. 120 (4), 1191–1237.

Gruzd, A., Roy, J., 2014. Investigating political polarization on Twitter: A Canadian perspective. Policy Internet 6 (1), 28–45.

Guerra, P., Meira Jr., W., Cardie, C., Kleinberg, R., 2013. A measure of polarization on social media networks based on community boundaries. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 7. pp. 215–224.

Jackson, M.O., 2010. Social and Economic Networks. Princeton University Press.

Kelly, J., Fisher, D., Smith, M., 2005. Debate, division, and diversity: Political discourse networks in USENET newsgroups. In: Online Deliberation Conference, vol. 2005. Stanford University, 4–3.

Lee, J.K., Choi, J., Kim, C., Kim, Y., 2014. Social media, network heterogeneity, and opinion polarization. J. Commun. 64 (4), 702–722.

Messing, S., Westwood, S.J., 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. Commun. Res. 41 (8), 1042–1063.

Meyer, C.D., 2000. Matrix Analysis and Applied Linear Algebra, vol. 71. Siam.

Mohammad, S.M., Zhu, X., Kiritchenko, S., Martin, J., 2015. Sentiment, emotion, purpose, and style in electoral tweets. Inf. Process. Manage. 51 (4), 480–499.

Mossel, E., Sly, A., Tamuz, O., 2012. On agreement and learning.

Roux, N., Sobel, J., 2015. Group polarization in a model of information aggregation. Am. Econ. J. Microecon. 7 (4), 202–232.

Rusinowska, A., Berghammer, R., De Swart, H., Grabisch, M., 2011. Social networks: prestige, centrality, and influence. In: International Conference on Relational and Algebraic Methods in Computer Science. Springer, pp. 22–39.

Seneta, E., 1979. Coefficients of ergodicity: Structure and applications. Adv. Appl. Probab. 576–590.

Seneta, E., 2006. Non-Negative Matrices and Markov Chains. Springer Science & Business Media.

Sethi, R., Yildiz, M., 2013. Perspectives, opinions, and information flows. In: MIT, Department of Economics' Working Paper Series. Department of Economics, massachusetts Institute of Technology, Cambridge, MA.

Shapiro, J.M., Taddy, N.M., 2015. Measuring polarization in high-dimensional data: Method and application to congressional speech. In: NBER Working Paper. 22423.

Smith, L., Sørensen, P., 2000. Pathological outcomes of observational learning. Econometrica 68 (2), 371–398.

Sobkowicz, P., Kaschesky, M., Bouchard, G., 2012. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. Gov. Inf. Q. 29 (4), 470–479.

Sunstein, C.R., 2001. Republic.Com. Princeton University Press.

Sunstein, C.R., 2009. Republic.Com 2.0. Princeton University Press.

Tahbaz-Salehi, A., Jadbabaie, A., 2008. A necessary and sufficient condition for consensus over random networks. IEEE Trans. Automat. Control 53 (3), 791–795.

Watts, D.J., Dodds, P.S., 2007. Influentials, networks, and public opinion formation. J. Consum. Res. 34 (4), 441–458.

Yardi, S., Boyd, D., 2010. Dynamic debates: An analysis of group polarization over time on twitter. Bull. Sci. Technol. Soc. 30 (5), 316–327.