Badri, Saeed; Heidergott, Bernd; Lindner, Ines

Working Paper

Naïve Learning in Social Networks with Fake News: Bots as a Singularity

Tinbergen Institute Discussion Paper, No. TI 2022-097/II

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Badri, Saeed; Heidergott, Bernd; Lindner, Ines (2022) : Naïve Learning in Social Networks with Fake News: Bots as a Singularity, Tinbergen Institute Discussion Paper, No. TI 2022-097/II, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/273810

TI 2022-097/II
Tinbergen Institute Discussion Paper

Naïve Learning in Social Networks with Fake News: Bots as a Singularity

Saeed Badri[1,3]
Bernd Heidergott[2,3]
Ines Lindner[1,3]

1 Department of Economics, Vrije Universiteit Amsterdam

2 Department of Operations Analytics, Vrije Universiteit Amsterdam

3 Tinbergen Institute

Naïve Learning in Social Networks with Fake News: Bots as a Singularity [1]

Saeed Badri[a*], Bernd Heidergott[b*], Ines Lindner[a*]

December 18, 2022

[a] Department of Economics, Vrije Universiteit Amsterdam,
[b] Department of Operations Analytics, Vrije Universiteit Amsterdam,
[*] Tinbergen Institute Amsterdam, The Netherlands

Abstract. We study the impact of bots on social learning in a social network setting. Regular agents receive independent noisy signals about the true value of a variable and then communicate in a network. They naïvely update beliefs by repeatedly taking weighted averages of neighbors' opinions. Bots are agents in the network that spread fake news by disseminating biased information. Our main contributions are threefold. (1) We show that the consensus of the network is a mapping of the interaction rate between the agents and bots and is discontinuous at zero mass of bots. This implies that even a comparatively "infinitesimal" small number of bots still has a sizeable impact on the consensus and hence represents an obstruction to the "wisdom of crowds". (2) We prove that the consensus gap induced by the marginal presence of bots depends neither on the agent network or bot layout nor on the assumed connection structure between agents and bots. (3) We show that before the ultimate (and bot-infected) consensus is reached, the network passes through a quasi-stationary phase which has the potential to mitigate the harmful impact of bots.

Keywords: Fake news, Misinformation, Social networks, Social Media, Wisdom of Crowds
JEL: D83; D85; Z13.

[1] We would like to thank Jurjen Kamphorst and the participants of the NETWORKS Match Maker Seminar series for helpful comments on an early presentation.

1 Introduction

The last decade's events have led to a growing scientific interest in understanding how people form opinions and share information on social networks. In particular, the phenomenon of "fake news" and the growing emergence of "bots" has led to public concern after several high-impact political events, such as Brexit and the 2016 US presidential election. During the covid-19 pandemic, fake news posed a massive global health threat and caused extreme socioeconomic damage.[2] Fake news is diffused to a significant part by (social media) bots. These are automated accounts, steered by computer algorithms, that simulate human behavior on social networks and interact with regular users. Subrahmanian et al. (2016) report that bots represent approximately 8.5% of Twitter users, as disclosed by Twitter. The study by Varol et al. (2017) on social bots shows that out of all English-speaking active users on Twitter, 9% to 15% exhibit bot-like behaviors.
While bots can have benign or neutral intentions, such as chatbots or bots automatically disseminating news for news agencies[3], there is an increasing concern about malicious bots which are instrumentalized to deceive and manipulate public opinion. These attacks constitute organized crimes that pose potential threats to public opinion, democracy, public health, the stock market, and other disciplines.
The massive damaging effect of bots forms an interesting contrast to the phenomenon of the "wisdom of the crowds". The argument made by popular general-audience books such as Surowiecki (2004) states that large crowds are sometimes surprisingly good at aggregating individual partial information and detecting the truth. He offers the example that if many people are guessing independently, then the average of their guesses is often an excellent estimate of whatever they are guessing about (perhaps the number of jelly beans in a jar or the weight of a bull in a fair). Given that social media networks connect millions of people and establish not only large but huge crowds, the question is why these networks are so prone to non-wisdom and why bots have a damaging effect on such a large scale.
Surowiecki notes that the key to his argument about the "wisdom of the crowds" is that individuals each have private information (their signals), and they guess independently without knowing what the others have guessed. In the social process of opinion formation, in contrast, individuals are connected through a social network and keep forming new beliefs by repeatedly talking to or reading the beliefs of connected agents. Here, agents are not independent anymore, as they influence each other through the social network. This raises the question of which networks are efficient in social information aggregation. In particular, which social network architectures lead to "wisdom" such that the belief updating process leads to a consensus equal or close to the truth? This is one of the main topics of the large body of works on social learning in networks. Golub and Jackson (2010) discuss a simple learning setup, based on the seminal DeGroot (1974) network model. Here, individuals use simple heuristics to update beliefs, such as repeatedly taking weighted averages of neighbors' opinions. Under some standard assumptions about the network connectivity, the dynamic of the system reaches a consensus. Golub and Jackson (2010) call the network wise if agents succeed to aggregate independent noisy signals about the true value of a variable such that the society reaches consensus equal to the truth. Their main result states that large societies are wise if and only if the most influential agent in the network vanishes as the society grows. Does this mean that these networks are less affected by bots and hence more resilient to misinformation? More generally, can we come up with a theory about how the structure of social (media) networks and the network positions of the bots affect the degree of misinformation?

[2] See e.g. Grinberg et al. (2019) and de Moura et al. (2021).
[3] For categorization of bots see e.g. Stieglitz et al. (2017).

In this paper, we offer an analytical toolbox to tackle these questions. Our main contribution relative to the existing literature is that our theory is not based on numerical exercise. Instead, we offer analytical closed-form solutions to quantify the impact of bots in dynamic models of opinion formation in which individuals are connected through a social network. Therefore, we provide a tractable model that explains how the network topology determines the resilience to deliberate manipulation of the social learning process. These attacks can stem from various sources besides bots. Throughout the paper we use the term "bots" as a synonym for actual bots infesting social media platforms, stubborn agents spreading fake news, or ideological groups aiming to manipulate opinion formation in a social network.
For our analysis, we focus on two seminal benchmark models of social learning with bounded rational agents. However, our results can be applied to a large class of social learning models, as the benchmark models consist of formal components that are very common in this strand of literature.
We start with analyzing the impact of bots in the seminal DeGroot (1974) model of naïve learning as a simple and natural starting point. The social structure of a society is described by a weighted and possibly directed network. Bots form part of this network, such that there are two types of nodes: agents and bots. At each point in time, agents communicate with "neighbors" (humans and bots) in the social network and update their beliefs. The updating process is simple. An agent's new belief is the (weighted) average of his or her neighbors' beliefs from the previous period. The learning process starts with each real agent's initial belief, given by the true state of nature plus some idiosyncratic zero-mean noise. The opinion dynamics converges under some general standard conditions on network connectivity.
We first focus on the interaction rate between the set of agents and the set of bots. If this interaction rate is positive, the network will form a consensus influenced by bots. On the other hand, by setting this rate to zero, the agent network reaches a consensus without any bot influence.

Our main results are:

(1) The consensus of the network is given by a mapping of the interaction rate between the agents and bots. Most importantly, it is *discontinuous* at zero.

(2) The size of this wisdom loss, the consensus distance to the truth, depends neither on the agent's network layout nor on the assumed connection structure between agents and bots.

(3) We show that the layout of the agent network does affect the speed of convergence to consensus.

Our findings are of societal impact. Result (1) shows that even a comparatively "infinitesimal" small number of bots still has a sizeable impact on the consensus over the agents. Result (2) shows that the "wisdom of crowds" phenomenon is lost immediately in the presence of bots. Indeed, even for an arbitrarily large agent social network, having a structure known to be "wise", the relatively tiniest number of bots will have a sizeable impact on the consensus. Given that there are typically bots (or stubborn or manipulative agents) present in social networks, results (1) and (2) show that naïve learning fails for social networks. Fortunately, result (3) will show that for a small interaction rate, the speed of convergence of the agents' beliefs *does* depend on the structure of the agent network. Indeed, we identify agent network topologies that are resilient to bots in the following sense. Before the ultimate (and bot-infected) consensus is reached, belief dynamics passes through a quasi-stationary phase, with agents reaching consensus as if all bots were absent. We refer to these agent networks as short-term resilient. If, in addition, the agent network without bots is wise, it implies that the short-term consensus approximates the truth. We refer to these (agent) networks as short-term wise. Result (3) is hence tentatively optimistic,

showing that a resilient and wise agent network will maintain wisdom, at least in the short run. Our second benchmark model is an extension of the first benchmark DeGroot model. Here, regular agents, as opposed to bots, receive informative private signals every period of time. The idea is that individuals can obtain information about the true state of the world from unbiased sources external to the network, such as scientific studies, unbiased news media, etc. Agents update their beliefs by a convex combination of this private signal and the weighted average of neighbors' beliefs (the DeGroot model component). We discuss this setting for a very common and intuitive weighting specification of this convex combination, to which we refer as the *universal learning rate*. This rate assigns a weight to the current signal that ensures equal weight to all signals observed so far, and hence decreases with time. Our main finding for our second benchmark model is that consensus is reached (almost surely) and equals the truth in the absence of bots. This implies that external informative signals ensure that the society is wise, independent of the underlying agent networks. However, the main results (1)-(3) hold as in the first benchmark model. That is, despite the fact that agents keep receiving informative external signals at every point in time, the presence of bots has the same effect as in the first benchmark model. Already an "infinitesimal" small number of bots has a sizeable impact on consensus and obstructs "wisdom of crowds" immediately. However, similar to the first benchmark model, we characterize agent network topologies that can postpone the influence of the bots for a very long time by the emergence of quasi-stationary phase.

The paper is organized as follows. In Section 2 we introduce the two seminal benchmark models of social learning of our study. The model of the integrated network of agents and bots is introduced and analyzed in Section 3. The (instantaneous) wisdom loss as a measure of the impact of an "infinitesimal" small amount of bots is introduced in Section 4 An analysis of the speed of convergences towards its quasi-stationary consensus and how its dependence on the network architecture is presented Section 5. Finally, Section 6 shows how our findings can be applied to topics besides bots, such as information bubbles and media biases. We conclude with directions of further research.

Related Literature

This paper contributes to the growing field of social learning with bounded rational agents and misinformation in networks. This field is different to Bayesian learning models, in which individuals process observed information, such as beliefs of neighbors, in a sophisticated manner. While Bayesian updating has firm normative foundations, theories based on this learning soon become infeasible even for small numbers of agents. Bayesian updating assumes that agents adjust correctly their weighting of neighbors' belief for repetitions and dependencies in information they hear multiple times. This is way too complex to serve as a realistic behavioral rule for agents' learning, the individual belief updating, respectively. Due to the complexity, such full Bayesian learning is typically explored for very small networks, such as Gale and Kariv (2003) who study three-link networks.

The key assumption of bounded rationality, in contrast, is based on a much more naïve, but still natural, form of updating. In particular, it assumes that agents use simple heuristics such as updating their belief by taking a weighted average of what they hear from neighbors. The weights that agents place on other's opinions are assumed to be constant and used at every single time step[4]. Examples of social learning models in networks with bounded rational agents are DeGroot (1974), Ellison and Fudenberg (1993),Ellison and Fudenberg (1995), Bala and Goyal

[4]The bounded rationality argument is discussed at length in DeMarzo et al. (2003)

(1998), DeMarzo et al. (2003), and Golub and Jackson (2010). The results in Jadbabaie et al. (2012) extend these models to allow a constant arrival of private external signals, similar to our second benchmark model.
Perhaps closest to our work is Azzimonti and Fernandes (2022). The authors provide a simulation study to analyze the network impact of (left and right wing) bots on misinformation and polarization (segregation of the society). Similar to Jadbabaie et al. (2012) and our second benchmark model, they extend the DeGroot network approach of bounded rational social learning to allow a constant arrival of private unbiased signals. A shortcoming of their work, however, is the lack of a theoretical and hence analytical characterization of the relationship between network topology and degree of misinformation.

2 Benchmark Learning Models in Networks

Consider a set of agents in the society that represented by, $\mathcal{N} = \{1, \ldots, n\}$, $n \geq 1$, interacting as a social network. The interaction patterns are captured by a $n \times n$ row stochastic Markov matrix, where $P_{ij} > 0$ indicates that i pays attention to j. This refers to the weight or trust that agent i places on the current opinion or belief of agent j when i is forming her new belief for the next period.

There is a true state of nature, $\mu \in \mathbb{R}$, which we treat as fixed. Let the individual's "belief" of agent $i \in \mathcal{N}$ at time $t \geq 0$ be given by $b_i^{(t)}$, where belief represents the agents estimation of the truth. Following DeMarzo et al. (2003) and Golub and Jackson (2010), we assume that at $t = 0$, every agent i receives a noisy signal

$$b_i^{(0)} = \mu + e_i, \tag{1}$$

where e_i is a white noise, i.e., e_i has zero mean and homoscedastic finite variance.

Our first benchmark model is the seminal DeGroot model of opinion dynamics DeGroot (1974). Here, each agent starts with initial belief (1) and updates by repeatedly taking weighted averages of her neighbors' beliefs. It is a closed updating system in the sense that the agents don't receive any additional external signals during the learning dynamics.

Definition 1 (DeGroot Updating). *The DeGroot updating scheme is given by*

$$b^{(t)} = Pb^{(t-1)} = P^t b^{(0)}, \quad t \geq 1, \tag{2}$$

with the initial belief $b^{(0)}$ *as the* initial signal.

Our second benchmark model is an extension of the above model such that agents keep receiving individual unbiased signals at the beginning of each time period. To distinguish beliefs from signals, we denote the signal agent i receives at time t by $f_i^{(t)}$. With a small abuse of notation, we assume that $b^{(0)} = f^{(0)}$. Simply, we assume that the signals $f_i^{(t)}$ follow the same distribution as the initial beliefs in all time periods. Similar to what we assumed in (1), the signals are given by

$$f_i^{(t)} = \mu + e_i^{(t)}, \tag{3}$$

where $e_i^{(t)}$ is a white noise as discussed above, and $e_i^{(t)}$ are mutually independent for $1 \leq i \leq n$ and $t \geq 0$.[5] Agents update their belief by a convex combination of this private signal and the weighted average of neighbor's belief.

[5]This setup is more general than the simulation study of Azzimonti and Fernandes (2022), where the authors discuss the special case of external signals to be drawn from a Bernoulli distribution centered around true state of the nature.

Definition 2 (**Belief Updating with External Signals**). *The belief updating scheme with* external signals $\{f^{(t)}\}_{t=1}^{\infty}$ *is given by*

$$b^{(t)} = (1-\alpha^{(t)})Pb^{(t-1)} + \alpha^{(t)} f^{(t)}, \quad t \geq 1, \tag{4}$$

for some weight sequence $\{\alpha^{(t)}\}_{t=1}^{\infty}$, *with* $\alpha^{(t)} \in [0,1]$, *for* $t \geq 1$.

An intuitive choice for $\alpha^{(t)}$ is

$$\alpha^{(t)} = \frac{1}{t+1} \tag{5}$$

to which we will refer as the *universal learning rate*, in short *universal learning*. This term is also known from the literature on stochastic approximation, and we note that Delyon (2000) mentions that Bru has traced back the first known description of this learning rate to 1890; see Bru (1996). The universal learning rate (5) gives weight α^t to the period t signal and due to the rescaling of the accumulated beliefs by $(1-\alpha^{(t)})$ this implies assigning equal weight to all t signals observed so far.

A fundamental question is under what conditions the heuristic rules of our two benchmark models lead the society to have a common belief to which we refer as *consensus*, in particular under which conditions the consensus is correct.

Definition 3. *A belief updating scheme reaches a consensus if almost surely*

$$\lim_{t\to\infty} b^{(t)} = b$$

where $b = \beta\mathbf{1}$, *i.e., all entries of* b *are equal. Moreover, a belief updating scheme is "wise" if* $\beta = \mu$, *where* μ *is the true state of nature.*

We start with recalling some well-known results from Markov Chain theory, see Seneta (1981) for details.

Theorem 1. *For any finite Markov chain* P *a finite matrix* Π *exists such that*

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} P^k = \Pi.$$

If P *is, in addition, aperiodic and irreducible*[6], *then*

$$\lim_{t\to\infty} P^t = \lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} P^k = \Pi.$$

Moreover, $\Pi_{i,*} = \Pi_{j,*}$ *for all* i, j, *equal* π *which represents the unique stationary distribution of* P.

The following result characterizes consensus and wisdom for our two benchmark models.

Theorem 2. *Let be* P *be irreducible and aperiodic.*

(i) The associated DeGroot updating scheme (2) converges to a consensus, given by

$$\lim_{t\to\infty} b^{(t)} = \Pi b^{(0)} \text{ and } (\Pi b^{(0)})_i = \pi b^{(0)}, \text{ for all } i \in \mathcal{N}, \tag{6}$$

where $b^{(0)}$ *is the realization of the initial signal* $f^{(0)}$. *The expectation of the consensus is therefore given by*

$$\pi\mathbb{E}[f^{(0)}] = \mu. \tag{7}$$

[6] An equivalent conditions is for P to be a *primitive* matrix.

(ii) The associated universal learning scheme (4) and (5) reaches a consensus and is wise.

Proof. The proof of (i) follows immediately from Theorem 1. As for (ii), note that (4) implies

$$b^{(t)} = \sum_{k=0}^{t}(1-\alpha^{(t)})...(1-\alpha^{(k+1)})\alpha^{(k)}P^{t-k}f^{(k)} = \sum_{k=0}^{t}\prod_{i=k+1}^{t}(1-\alpha^{(i)})\alpha^{(k)}P^{t-k}f^{(k)}. \tag{8}$$

With the universal rate (5), the products in the sum of (8) simplify considerably as follows

$$\prod_{i=k+1}^{t}(1-\alpha^{(i)})\alpha^{(k)} = \frac{t}{t+1}\frac{t-1}{t}....\frac{k+2}{k+3}\frac{k+1}{k+2} = \frac{k+1}{t+1}\frac{1}{k+1} = \frac{1}{t+1},$$

such that for (8) follows

$$b^{(t)} = \sum_{k=0}^{t}\frac{1}{t+1}P^{t-k}f^{(k)} = \frac{1}{t+1}\sum_{k=0}^{t}P^{t-k}f^{(k)}. \tag{9}$$

From (9) follows for any large and fixed $0 \ll T$ and $T < t$,

$$\begin{aligned} b^{(t)} &= \frac{1}{t+1}\sum_{k=0}^{t}P^{t-k}f^{(t)} = \frac{1}{t+1}\left(\sum_{k=T+1}^{t}P^{t-k}f^{(k)} + \sum_{k=0}^{T}P^{t-k}f^{(k)}\right) \\ &= \underbrace{\frac{1}{t+1}P^{t-T}\sum_{k=0}^{T}P^{T-k}f^{(k)}}_{\text{term 1}} + \underbrace{\frac{t-T}{t+1}\left(\frac{1}{t-T}\sum_{k=T+1}^{t}P^{t-k}f^{(k)}\right)}_{\text{term 2}}. \end{aligned} \tag{10}$$

We will complete the proof by discussing term 1 and 2 of (10) for $t \to \infty$. Note that the sum in term 1 is fixed for any fixed $T > 0$. Therefore, as t increases the factor $\frac{1}{t+1}$ tends to 0 such that

$$\lim_{t\to\infty}\frac{1}{t+1}P^{t-T}\sum_{k=0}^{T}P^{T-k}f^{(k)} = 0.$$

On the other hand, for term 2, note that the factor $\frac{t-T}{t+1}$ tends to 1 as t increases. Furthermore, Theorem 1 ensures that for $0 \ll T \ll t$ the term P^{t-T} is approximately equal to Π and converges to Π as $t \to \infty$. This provides

$$\lim_{t\to\infty}\text{term 2} = \lim_{t\to\infty}\left(\frac{1}{t-T}\sum_{k=T+1}^{t}\Pi f^{(k)}\right) = \Pi\lim_{t\to\infty}\left(\frac{1}{t-T}\sum_{k=T+1}^{t}f^{(k)}\right). \tag{11}$$

Since the signals $f^{(k)}$ are independent and identically distributed with $\mathbb{E}[f^{(k)}] = \mathbb{E}[f^{(0)}]$ for all $0 \leq k$, the Central Limit Theorem ensures that the term in brackets in (11) converges to $\mathbb{E}[f^{(0)}] = \mu$ almost surely.

□

Theorem 2 (i) implies that consensus is a weighted average of the individual initial beliefs. In particular, the relative weight π_i is the weight that all agents assign to the initial belief of agent i and is therefore a measure of *influence* of agent i on consensus. Recall that the initial

beliefs are unbiased such that consensus is correct (equals the truth) in expectation as stated in (7). The question, however, is for which networks the consensus beliefs are correct, not only in expectation, but for sure. Golub and Jackson (2010) analyze this question in the context of the DeGroot model (2). Their main result states that for networks increasing in size n to infinity, the DeGroot updating scheme is wise almost surely if the influence π_i of each individual agent i tends to 0 as n grows illustrated in the following example.

Example 1 (Golub and Jackson (2010))**.** *Consider a network P with uniform stationary distribution such that the influence of any agent is given by $\pi_n = 1/n$, for all $i = 1, ..., n$. Consensus in the DeGroot model (6) is given by*

$$\pi_1 b_1^{(0)} + ... + \pi_n b_n^{(0)} = \frac{1}{n} \sum_{i=1}^{n} b_i^{(0)} \tag{12}$$

which is the average of all individual initial beliefs. Now consider the effect on (12) for increasing size n of the network. Since all initial beliefs are iid, we conclude from the Central Limit Theorem that (12) converges to μ almost surely.
On a more general note, let $(P(n))_{n=1}^{\infty}$ denote a sequence of networks growing in size with respective stationary distributions $\pi_{P(n)}$. If $\pi_{P(n)}$ approaches uniformity as n tends to ∞, then the associated consensus $\pi_{P(n)} b^{(0)}$ approaches truth as n tends to ∞, see Golub and Jackson (2010).

After having defined and characterized "wisdom of the crowd" for our two benchmark models (DeGroot learning and universal learning), the next section discusses how the emergence of bots forms an obstacle to wisdom.

3 Network Approach of Social Learning with Bots

Bots are agents in the network that try to disguise themselves as regular agents. For this reason, they do not typically start out with extreme views, but instead converge to them over time. From a network perspective, they can be understood as stubborn agents with a high self link trying to countervail the aggregation of unbiased individual information of regular agents. For our analytical framework, however, we do not assume a specific architecture of how the bots are embedded in the network. Instead, we distinguish between two kinds of agents - regular agents and bots - and merely assume a weak coupling between these two types.

Consider the set of agents $\mathcal{N} = \{1, \ldots, n\}$. Without loss of generality we refer to the first $n_A \leq n$ nodes as the regular agents and the last n_B as bots such that $n = n_A + n_B$. With respect to a matrix representation of the network, we let matrix A of size $n_A \times n_A$ represent the weights that regular agents assign to each other. We assume that A is strongly connected and aperiodic. Analogously, B represents the network among the population bots.[7] Let the *joint network* of agents and bots be given by the $n \times n$ Markov chain

$$P(\epsilon, \rho) = \begin{bmatrix} (1-\epsilon)A & \epsilon\, C \\ \rho\, D & (1-\rho)B \end{bmatrix}, \tag{13}$$

where $1 > \epsilon, \rho \geq 0$ are the parameters representing the coupling between regular agents and bots and vice versa; and C and D present the weights that nodes assign to nodes of the other type

[7] One could argue, for example, that bots do not listen to each other which suggests setting B equal to the identity matrix.

such that C_{ij} measures the weight that regular agent i assigns to bot $j - n_A$, and D_{ij} measures the weight that bot $i - n_A$ assigns to agent j.

Note that for $\rho = \epsilon = 0$, the chain (13) is reducible as there is no link between the agents and bots. Moreover, small values of $\epsilon, \rho > 0$ imply that the spectral radius of $(1-\epsilon)A$ and $(1-\rho)B$ are close to one. Markov chains, so that the diagonal blocks have eigenvalues close to one, are called *weakly decomposable* or *nearly reducible*. Weak decomposability is a major source for slow convergence of matrix powers, see Meyer (1989).

Example 2. *Consider the simple example of two regular agents and one bot. Let the interaction matrix among the agents be given by*

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

with stationary distribution $\pi_A = (1/2, 1/2)$ *such that the two regular agents have equal influence* $1/2$ *on the consensus. Now assume that both agents are coupled to one bot and put* $\rho = \epsilon^2$*. Then,*

$$P(\epsilon, \epsilon^2) = \begin{pmatrix} \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & \epsilon \\ \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & \epsilon \\ \frac{\epsilon^2}{2} & \frac{\epsilon^2}{2} & 1-\epsilon^2 \end{pmatrix}, \tag{14}$$

with stationary distribution

$$\pi(\epsilon, \epsilon^2) = \left(\frac{\epsilon}{2(1+\epsilon)}, \frac{\epsilon}{2(1+\epsilon)}, \frac{1}{1+\epsilon} \right). \tag{15}$$

For the regular agents 1 *and* 2*, the effect of introducing bots is hence an immediate influence loss from* $1/2$ *to* $\epsilon/(2(1+\epsilon))$ *for any (arbitrarily small)* $\epsilon > 0$*.*

Example 2 shows that the effect of coupling to bots has a sizeable impact on the influence of the regular agents on consensus. In the following, we will characterize the impact of influence for more general settings. Our first result shows that the stationary distribution of (13) can be characterized by a macro chain which models how the set of regular agents is coupled to the set of bots. For ease of notation, we write $i \in A$, for $1 \leq i \leq n_A$, and $i \in B$, for $n_A + 1 \leq i \leq n$. We define a macro random walk Y_t with state-space $\{a, b\}$, where

$$Y_t \quad = \quad \begin{cases} a & \text{if and only if} \quad X_t \in A, \\ b & \text{if and only if} \quad X_t \in B. \end{cases} \tag{16}$$

Here, Y_t is the Markov chain that only changes its state if X_t moves from A to B and vice versa. From (13), the associated matrix representation of the Markov chain (16) is

$$P_Y(\epsilon, \rho) = \begin{bmatrix} (1-\epsilon) & \epsilon \\ \rho & (1-\rho) \end{bmatrix}, \tag{17}$$

with stationary distribution

$$(\nu_{\epsilon,\rho}(a), \nu_{\epsilon,\rho}(b)) = \left(\frac{\rho}{\epsilon+\rho}, \frac{\epsilon}{\epsilon+\rho} \right). \tag{18}$$

For ease of reference, we summarize the following assumptions of our main theorem:

(V1) $P(\epsilon, \rho)$ in (13) is aperiodic and irreducible for $\epsilon, \rho \in (0, 1)$.

(V2)' A is a aperiodic and irreducible stochastic matrix with unique stationary distribution π_A on $1 \leq i \leq n_A$; and B is a stochastic matrix such that B^n converges element-wise to some limiting matrix Π_B.

Note that conditions **V(1)** and **V(2)'** are satisfied for A being an aperiodic and irreducible stochastic matrix, and C, D being non-squared matrices and B being identity matrix of appropriate size. While the results developed in the following will hold in the general setting of **V(2)'**, we will work for ease of presentation with the somewhat more restrictive condition

(V2) A, B are aperiodic and irreducible stochastic matrices with unique stationary distributions π_A, π_B.

Under assumption **(V1)** and **(V2)** we can trace the impact of the stationary distribution for the agents and bots on the overall stationary distribution.

Theorem 3. *Assume (V1) and (V2) hold. For all $i \in \mathcal{N}$ and $\epsilon, \rho > 0$ it holds*

$$\pi_i(\epsilon, \rho) = \nu_{\epsilon,\rho}(a)\pi_A(i)\mathbf{1}(i \in A) + \nu_{\epsilon,\rho}(b)\pi_B(i)\mathbf{1}(i \in B), \tag{19}$$

where $\nu_{\epsilon,\rho}$ denotes the stationary distribution of (16).

Proof. Let $\{X_t\}_{t=1}^{\infty}$ be a $P(\epsilon, \rho)$ Markov chain, that is,$P_{ij} = \mathbb{E}(X_{t+1} = j | X_t = i)$. Provided that $P(\epsilon, \rho)$ is aperiodic and irreducible, the unique stationary distribution of X_t, denoted by $\pi(\epsilon, \rho)$, is related to X_t through

$$\pi_i(\epsilon, \rho) = \lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} \mathbf{1}(X_t = i) \tag{20}$$

with probability one.

At any point in time $T \geq 0$, let $Z_A(T)$ count the number of time epochs X_t was in A, and let $Z_B(T)$ count the number of time epochs X_t was in B; in formula

$$Z_A(T) = \sum_{t=1}^{T} \mathbf{1}(X_t \in A) \quad \text{and} \quad Z_B(T) = \sum_{t=1}^{T} \mathbf{1}(X_t \in B),$$

and it holds that $T = Z_A(T) + Z_B(T)$.

Denote by

$$i(A, T) = \{t : 1 \leq t \leq T \quad \text{and} \quad X_k \in A\}$$

the set of indices for which X_t is in A, and by

$$i(B, T) = \{t : 1 \leq t \leq T \quad \text{and} \quad X_t \in B\};$$

where $|i(A, T)| = Z_A(T)$ and $|i(B, T)| = Z_B(T)$. For any $\epsilon > 0$, X_t and Y_t are ergodic. From (20) follows for the stationary distribution of X_t

$$\begin{aligned}
\pi_i(\epsilon, \rho) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \Big(\mathbf{1}(X_t = i)\mathbf{1}(X_t \in A) + \mathbf{1}(X_t = i)\mathbf{1}(X_t \in B) \Big) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \Big(\mathbf{1}(X_t = i)\mathbf{1}(Y_t = a) + \mathbf{1}(X_t = i)\mathbf{1}(Y_t = b) \Big) \\
&= \lim_{T\to\infty} \frac{1}{T} \left(\sum_{t\in i(A,T)} \mathbf{1}(X_t = i)\mathbf{1}(Y_t = a) + \sum_{t\in i(B,T)} \mathbf{1}(X_t = i)\mathbf{1}(Y_t = b) \right)
\end{aligned}$$

noting that $\mathbf{1}(Y_t = a) = 1$ for all $t \in i(A,T)$ and $\mathbf{1}(Y_t = b) = 1$ for all $t \in i(B,T)$, we continue

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{T} \left(\sum_{t\in i(A,T)} \mathbf{1}(X_t = i) + \mathbf{1} \sum_{t\in i(B,T)} \mathbf{1}(X_t = i) \right) \\
&= \lim_{T\to\infty} \frac{Z_A(T)}{T} \frac{1}{Z_A(T)} \sum_{t\in i(A,T)} \mathbf{1}(X_t = i) + \lim_{T\to\infty} \frac{Z_B(T)}{T} \frac{1}{Z_B(T)} \sum_{t\in i(B,T)} \mathbf{1}(X_t = i).
\end{aligned}
$$

Note that for $i \in A$

$$\pi_A(i) = \lim_{T\to\infty} \frac{1}{Z_A(T)} \sum_{t\in i(A,T)} \mathbf{1}(X_t = i),$$

and, using $T = Z_A(T) + Z_B(T)$,

$$\nu(a) = \lim_{T\to\infty} \frac{Z_A(T)}{T} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}(Y_t = a).$$

In the same way, for $i \in B$,

$$\pi_B(i) = \lim_{T\to\infty} \frac{1}{Z_B(T)} \sum_{t\in i(B,T)} \mathbf{1}(X_t = i)$$

and

$$\nu(b) = \lim_{T\to\infty} \frac{Z_B(T)}{T} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}(Y_t = b).$$

Inserting this in the above limit, we arrive at

$$\pi_i(\epsilon,\rho) = \nu_{\epsilon,\rho}(a)\pi_A(i)\mathbf{1}(i \in A) + \nu_{\epsilon,\rho}(b)\pi_B(i)\mathbf{1}(i \in B),$$

for all i. □

Example 3 (Example 2 revisited). *The setup (14) provides*

$$\pi_A = \pi_{(\epsilon,\rho;A)} = \left(\frac{1}{2}, \frac{1}{2}\right), \quad \text{and} \quad \pi_B = \pi_{(\epsilon,\rho;B)} = 1.$$

From (18) follows $\nu_{\epsilon,\epsilon^2}(a) = \left(\frac{\epsilon}{1+\epsilon}, \frac{1}{1+\epsilon}\right)$. *Inserting into (19) of Theorem 3 provides*

$$\pi_i(\epsilon,\epsilon^2) = \nu_{\epsilon,\epsilon^2}(a)\pi_A(i) = \frac{\epsilon}{1+\epsilon} \cdot \frac{1}{2}, \quad i = 1,2,$$

and

$$\pi_3(\epsilon,\epsilon^2) = \nu_{\epsilon,\epsilon^2}(a)\pi_B(1) = \frac{1}{1+\epsilon} \cdot 1$$

which confirms our earlier finding (15).

Our model allows for diversifying the rate ϵ with which the bots influence the agents and the rate ρ with which the bots adjust to the current beliefs of the agents in order to come across as human. A case of particular interest is $\epsilon = \rho$, so that (18) yields

$$\lim_{\epsilon\to 0} \nu_{\epsilon,\epsilon} = \left(\frac{1}{2}, \frac{1}{2}\right). \tag{21}$$

Note that in this case, $P_Y(\epsilon,\epsilon)$ has a unique stationary distribution on $\{a,b\}$, which is given by $\nu_{\epsilon,\epsilon} = (1/2, 1/2)$ and is *independent of* ϵ.

We introduce the following assumption on the limit of $\nu_{\epsilon,\rho}$

(V3) Assume that

$$\lim_{(\epsilon,\rho)\downarrow(0,0)} \nu_{\epsilon,\rho} = (\nu(a), \nu(b)),$$

with $v(a) + v(b) = 1$.

We summarize our findings in the following result.

Theorem 4. *Under (V1) - (V3), the limit of $\pi(\epsilon,\rho)$ is given by*

$$\pi^+(i) := \lim_{(\epsilon,\rho)\downarrow(0,0)} \pi_i(\epsilon,\rho) = \nu(a)\pi_A(i)\mathbf{1}(i \in A) + \nu(b)\pi_B(i)\mathbf{1}(i \in B). \tag{22}$$

Proof. The proof follows from inserting the limit for $\nu_{\epsilon,\rho}$ in the representation of $\pi(\epsilon,\rho)$ by Theorem 3. □

The following section discusses how our results allows to introduce a measure of the damaging impact of bots.

4 Measuring the Impact of Bots

In this section, we will apply our results to develop a measure of the impact of bots on the collective learning process. The idea is to compare the result of the learning process of regular agents without bots, $\epsilon = \rho = 0$, with the setting with bots, $\epsilon, \rho > 0$, for $(\epsilon,\rho) \to (0,0)$. We refer to this expected distance as the *instantaneous truth gap*. From a conceptual point of view, it is plausible to assume that ρ goes to zero faster than ϵ. The reason is that bots are programmed to be perceived as regular humans while bots only adjust to beliefs of regular users for strategic reasons. We will come back to this setting as a special case of our analytical toolbox.

Recall that the initial beliefs of the regular agents are unbiased and independently drawn with mean μ (representing the truth) given by (1). In contrast, we set the initial belief of bots equal to $w \in [0,1], w \neq \mu$. The same setting applies for the second benchmark model (4) for all external signals. We summarize this setting for ease of reference:

(V4) $\{f^{(t)} : t \geq 0\}$ is iid, so that $f_i^{(t)}$ has mean μ and finite variance for all t and $1 \leq i \leq n_A$; and $f_i^t = w$ for all t and $n_A + 1 \leq i \leq n_A + n_B$.

This assumption aims to capture that bots try to disguise themselves as regular agents. Updating according to (4) means that they follow partly the current prevalent beliefs, however, updating with a biased signals allows a subtle draw of the collective belief dynamics towards the target w.

Definition 4. *Let $b^{(t)}(\epsilon,\rho)$ denote the t-th belief vector under $P(\epsilon,\rho)$. Then, for i, with $1 \leq i \leq n_A$,*

$$\xi(i) = \lim_{t\to\infty} b_i^{(t)}(0,0) - \lim_{(\epsilon,\rho)\downarrow(0,0)} \lim_{t\to\infty} b_i^{(t)}(\epsilon,\rho) \tag{23}$$

is the instantaneous truth gap for agent i, provided the limit exits.

Note that $\xi(i)$ is well defined under condition (V1) and (V2). The instantaneous truth gap expresses the impact of an arbitrarily small coupling of bots to agents on the limiting belief of the regular agents. The next theorem shows that under our standard conditions the instantaneous truth gap can be expressed in closed form.

Theorem 5. *Suppose that (V1) to (V4) hold. Then, the instantaneous truth gap $\xi(i)$ is independent of i and equals*

- *for DeGroot updating*

$$\sum_{i=1}^{n_A} \pi_A(i) f_i^{(0)} - \pi^+ f^{(0)} \tag{24}$$

  and

- *for the belief updating scheme with external signals (4) and learning rate $\alpha^{(t)}$ as in (5)*

$$(1 - v(a))(\mu - w). \tag{25}$$

Moreover, (24) equals (25) in expectation.

Proof. For the DeGroot model we have by Theorem 2

$$\lim_{t\to\infty} b_i^{(t)}(0,0) = \sum_{i=1}^{n_A} \pi_A f_i^{(0)}$$

and by Theorem 4 that

$$\lim_{(\epsilon,\rho)\downarrow(0,0)} \lim_{t\to\infty} b_i^{(t)}(\epsilon,\rho) = \sum_{i\in\mathcal{N}} \pi^+(i) f_i^{(0)}.$$

Writing the sum on the above right hand side as $\pi^+ f^0$, proves (i).

We now turn to the proof of (ii). Following the same line of argument, we obtain for belief updating with external signals by Theorem 2

$$\lim_{t\to\infty} b_i^{(t)}(0,0) = \sum_{i=1}^{n_A} \pi_A \mathbb{E}[f_i^{(0)}] = \sum_{i=1}^{n_A} \pi_A \mu = \mu \quad a.s., \tag{26}$$

and by Theorem 4

$$\lim_{(\epsilon,\rho)\downarrow(0,0)} \lim_{t\to\infty} b_i^{(t)}(\epsilon,\rho) = \sum_{i\in\mathcal{N}} \pi^+(i) \mathbb{E}[f_i^{(0)}].$$

Using that the mean of $f_i^{(0)}$ is either μ or w and using the explicit expression for π^+ in Theorem 4 gives

$$\begin{aligned} \sum_{i\in\mathcal{N}} \pi^+(i) \mathbb{E}[f_i^{(0)}] &= \nu(a) \sum_{i=1}^{n_A} \pi_A(i)\mu + (1-\nu(a)) \sum_{i=1}^{n_B} \pi_B(i) w \\ &= \mu\nu(a) + w(1-\nu(a)). \end{aligned}$$

Thus, combing the above with (26) gives

$$\xi(i) = \mu - \mu\nu(a) - w(1-\nu(a)) = (\mu - w)(1-\nu(a)). \tag{27}$$

Finally, applying (26) and (27) to the expectation of (24) concludes the proof. □

The result put forward in Theorem 5 shows that the limiting belief as a mapping of (ϵ,ρ) is discontinuous in $(\epsilon,\rho) = 0$. This result is related to the theory of singular perturbation of Markov processes, and we refer for more details Yin and Zhang (2012), Avrachenkov et al. (2002).

We now discuss the special setting of ρ going exponentially faster to zero than ϵ.

Example 4 (Example 1 revisited). *Consider a general version of Example 1 with n_A agents and one bot. In particular, set $\rho = \epsilon^2$. For the interaction matrix follows*

$$P(\epsilon, \epsilon^2) = \begin{pmatrix} \frac{1-\epsilon}{n_A} & \dots & \frac{1-\epsilon}{n_A} & \epsilon \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1-\epsilon}{n_A} & \dots & \frac{1-\epsilon}{n_A} & \epsilon \\ \frac{\epsilon^2}{n_A} & \dots & \frac{\epsilon^2}{n_A} & 1-\epsilon^2 \end{pmatrix}. \tag{28}$$

From (18) follows

$$(\nu_{\epsilon,\epsilon^2}(a), \nu_{\epsilon,\epsilon^2}(b)) = \left(\frac{\epsilon}{1+\epsilon}, \frac{1}{1+\epsilon}\right)$$

and hence for (22)

$$\pi^+(i) := \lim_{\epsilon \downarrow 0} \pi_i(\epsilon, \epsilon^2) = \mathbf{0}(i \in A) + \mathbf{1}(i \in B) \tag{29}$$

We conclude for the truth gap (24)

$$(1/n_A) \sum_{i=1}^{n_A} f_i^{(0)} - w. \tag{30}$$

The above example shows that there is no way of restoring wisdom in case of ρ going exponentially faster to zero than ϵ, even if the number of bots n_B is arbitrarily small. The reason is that the macro coupling (17) generates a non-fading memory of the bots in the limit case. In light of Theorem 5, the network is insensitive with respect to the bots if and only if

$$\lim_{(\epsilon,\rho)\to(0,0)} \frac{\rho}{\epsilon + \rho} = \nu_{\epsilon,\rho}(a) = 1.$$

The above relates to the case where, for example, $\rho = f(\epsilon)$ for f being some power function $f(x) = x^\alpha$, for $\alpha < 1$. It is worth noting that a power rate $\alpha < 1$ models a network where the bots assign significantly more weight to the agents than the agents to the bots, which is a unrealistic setting (and in a way contradicts the very definition of a bot). Hence, we can conclude that in practice networks are sensitive to the influence of bots.

Lemma 1. *Under (V1) to (V4) and homoscedasticity, it holds for the DeGroot updating scheme that*

$$\mathrm{Var}(\xi(i)) = (1 - \nu(a))\sigma^2,$$

where σ^2 denotes the variance of $f_i^{(0)}$ for $1 \le i \le n_A$.

Proof. By Theorem 5 and (V4)

$$\begin{aligned} \mathrm{Var}(\xi(i)) &= \mathrm{Var}\left(\sum_{i=1}^{n_A} \pi_A(i) f_i^{(0)} - \pi^+ f^{(0)}\right) \\ &= \mathrm{Var}\left(\sum_{i=1}^{n_A} (1 - \nu(a)\pi_A(i) f_i^{(0)} + \sum_{j=n_A+1}^{n_A+n_B} (1 - \nu(a))\pi_B(i) f_j^{(0)}\right) \\ &= ((1 - \nu(a))\sigma^2. \end{aligned}$$

□

The main takeaway from Theorem 5 is that the instantaneous truth gap is independent of the network topology of the sub-matrices A,...D, and the size of the agent and bot network. Instead, the instantaneous truth gap is entirely determined by the macro coupling (17).

In the next section we will bring some relativism to this negative result, as we show that the *speed of convergence* does depend on the network (i) topology of the sub-matrices A,...D, and (ii) the size of the agent and bot network. Moreover, we will show that for some structure of the agent networks, a quasi-stationary belief behaviour can be reached that approximates the truth well.

5 Characterization of Convergence and Resilience

Our analysis so far has focused on how the network position of bots impacts the long-run consensus. In practice, however, we often observe disagreement, in contrast to consensus, even within connected communities. We will provide an explanation of this phenomenon by analyzing how the convergence behavior depends on network structure in presence of bots. In particular, we show how the weak coupling between parts of the network fosters the emergence of quasi-stationary states of the belief updating dynamics. These are states in which parts of the network seem to have reached a local consensus which is stable for a very long time, while the overall convergence behavior to the final consensus sets in much later. If, in addition, this local consensus is close to the truth and stays in this state for a very long time, it would mitigate the damaging impact of bots - at least in the short run.

A key insight in the theory of Markov chains is that the second-largest eigenvalue $\lambda_2(P) < \lambda_1(P) = 1$ of a stochastic matrix P is related in magnitude to the convergence time of the iterated process (see e.g. Seneta (1981)). The networks setting (13) we study in this paper means $P(\epsilon, \rho)$ behaves for (ϵ, ρ) small as if agents of part A and B are only weakly interacting. In particular, we are interested in characterizing network configurations with $\lambda_2(P(\epsilon, \rho))$ close to one while $\lambda_2(A)$ is close to zero. The following theorem shows that this implies that there is a $T \gg 0$ such that the agent part behaves as if bots are absent. The agent's beliefs converge quickly to a quasi stationary state. This state is the consensus of the agent network A. According to Theorem 2, this means that agents beliefs settle at $\pi_A f^{(0)}$ in the DeGroot model and at μ for the second benchmark model with external signals.

Theorem 6. *Suppose (V1) to (V4) hold and $\lambda_2(A) \ll 1$. Provided that (ϵ, ρ) is sufficiently small, there exists a $T \gg 0$ such that $b_A^{(t)}$, the agent part of the belief vector, approaches for $t \leq T$ the consensus characterized by Theorem 2 with agent network A.*

Proof. For the network (13) it holds that $\lim_{(\epsilon,\rho)\downarrow(0,0)} \lambda_2(P(\epsilon, \rho)) = 1$ such that $\lambda_2(P(\epsilon, \rho)) \approx 1$ for sufficiently small (ϵ, ρ), Meyer (1989).
For the setting of the theorem, the Simon-Ando theory, see Simon and Ando (1961), shows that there is $0 \ll T$ so that the agent part behaves during the first T updates as if independent of the bots. By ergodicity of A together with $0 < \lambda_2(A) \ll 1$, it follows that $(A^T f^{(0)})(i)$ is close to $\pi_A f^{(0)}$ for all $i \in A$, which establishes the first part of the theorem. The proof of the second part follows from the same line of argument and is therefore omitted. □

We conclude that $\lambda_2(A) \ll 1$ eliminates the harmful impact of bots in the short run and can postpone the influence of the bots for a very long time. We will refer to such agent networks A as *short-term resilient*. In order to characterize such networks, we employ some useful estimation techniques for $\lambda_2(A)$.

For A irreducible and aperiodic, the *Coefficient of Ergodicity*

$$\tau(A) = 1 - \min_{i,j} \sum_{k=1}^{n} \min\{A_{i,k}, A_{j,k}\}, \tag{31}$$

provides an upper bound for $\lambda_2(A)$ (see Seneta (1979)). From (31) it can easily be seen that $\tau(A)$ is small when the nodes in the network have similar connection patterns such that the matrix has similar rows. As a simple example, consider the stylized case where all agents weight each other equally such that all elements of A are given by $1/n_A$. Here, $\tau(A) = 0 = \lambda_2(A)$. This society shows immediate (quasi-stationary) consensus within A. Since A is wise for large n_A, this local consensus will be close to the truth μ and will stay there for a long time until the impact of the bots sets in and drives the agents' consensus away from μ. The agent network A is hence *short-term wise*.
Note, however, that a small $\lambda_2(A) \approx 0$ does not ensure short term wisdom. A simple counterexample is a society where all agents weight just one agent, say agent 1. Here, as in the previous setting, we conclude from (31) $\tau(A) = 0 = \lambda_2(A)$ and hence immediate (quasi-stationary) consensus within A. This network, however, is not short-term wise as A is not wise in the first place. The network A is short-term resilient, as the consensus among agents is independent of bots for a very long time.

Another approach to estimate $\lambda_2(A)$ goes back to Cheeger (1970) and employs the graph interpretation of a Markov chain A. In particular, note that any A can be related to a graph (E, V) with node set $E = \{1, \dots n\}$ and edge set $V = \{(i,j) \in E^2 : A_{i,j} > 0\}$. The subsequent definition is taken from Cheeger (1970).

Definition 5. *Let (E, V) be the graph related to Markov chain A. For $W \subset V$ let $\partial W = \{(i,j) \in V, i \in W, j \in V \setminus W\}$, then the Cheeger constant is given by*

$$h(A) := \min \left\{ \frac{|\partial W|}{|W|} : W \subset V, 0 < |W| < \frac{1}{2}|V| \right\}, \tag{32}$$

where $|W|$ denotes the cardinality of W.

As shown in Cheeger, the Cheeger constant $h(A) \leq 1$ provides bounds for the second-largest eigenvalue of A as follows

$$1 - 2h(A) \leq \lambda_2(A) \leq 1 - \frac{(h(A))^2}{2\Delta}, \tag{33}$$

where Δ denotes the maximum degree of the graph of A. We call

$$h_l(A) = 1 - 2h(A)$$

the *lower Cheeger bound* and

$$h_u(A) = 1 - \frac{(h(A))^2}{2\Delta} \tag{34}$$

the *upper Cheeger bound*.

The Cheeger constant $h(A) \leq 1$ can be interpreted as a measure of "seperability" of a network A. Indeed, the Cheeger constant is small if the network can be split into two big components by cutting only a few links. Hence, a disassortative network with two evenly large clusters (read, two hubs) has a low Cheeger bound. This phenomenon is very frequent, such as in political blogs

supporting opponent candidates (see for example Sasahara et al. (2021)).

A simple heuristic for the quality of estimation provided by the Cheeger constant is the distance between the upper and lower bound of (33). This distance is the smallest for $h(A) = 0$ and keeps increasing for increasing $h(A) \leq 1$. Moreover, this distance increases with increasing Δ. In summary, the rule of thumb is that estimation (33) works best for networks showing patterns of separability and bounded maximal degree.

The following example illustrates the suitability of estimation techniques (31) and (33) depending on the network A.

Example 5. *Consider again the simple stylized case where all agents weight each other equally. We refer to this case as complete uniform with network A_{cu} with all equal elements $1/n_A$. As noted earlier, (31) provides $\lambda_2(A_{cu}) = 0$ due to the upper bound $\tau_{cu} = 0$. The Cheeger constant, however, is at its maximal value $h(A_{cu}) = 1$. Moreover, the maximal degree is at its highest possible value $\Delta = n_A$. As conjectured by the rule of thumb, (33) will not be sharp and indeed reads,*

$$-1 \leq \lambda_2(A_{cu}) \leq 1 - \frac{1}{2n_A},$$

which provides no information about the actual value of $\lambda_2(A_{cu})$ and hence no indication about the resistance to bots.
Now consider the network in Figure 1 to which we refer as the splitted network A_s. Moreover, the highest degree $\Delta = 3$ is small compared to $n_A = 19$ such that we expect the Cheeger inequality (33) to provide a sharper estimation than (31). Table (1) confirms this intuition for $n_A = 19$.

Figure 1: Splitted network with informative Cheeger bound

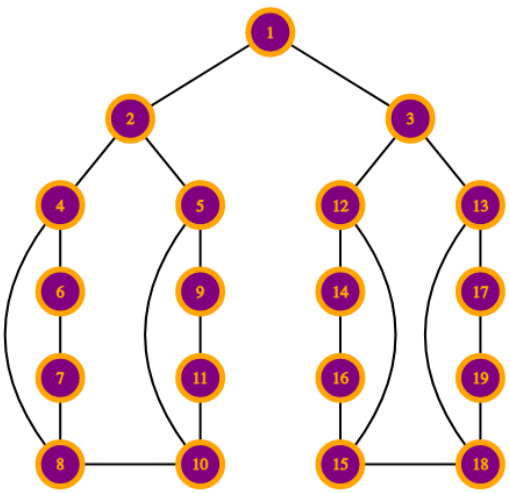

Table 1: Coefficient of Ergodicity and Cheeger Bound in Comparison

Network Structure	$\lambda_2(A)$	$\tau(A)$	$h_l(A)$	$h_u(A)$
Complete uniform $A = A_{cu}$	0	0	-1	0.98
Splitted network $A = A_s$	0.992	1	0.84	0.99

The numerical examples put forward in Table 1 illustrate the validity of our qualitative argumentation. For cohesive networks with similar connection patterns among agents, as in the extreme case A_{cu}, the coefficient of ergodicity tends to provide sharp estimations as upper bound while the Cheeger bounds are not informative. For networks with patterns of separability, in contrast, it is the other way around. Note that $\tau(A_s) = 1$ and hence provides no information as an upper bound for the splitted network. The high value of the lower Cheeger bound, however, is very informative as it precludes $\lambda_2(A_s) \ll 1$ and therefore forecloses this indicator of short-term resilience.

6 Applications: Echo Chambers, Filter Bubbles and False Balance

Our model speaks to the widespread concern about information bubbles. One proposed explanation for the occurrence of these bubbles is that many agents have homophilic preferences (see e.g. Hart et al. (2009); Kastenmüller et al. (2010); Boucher (2015); and Del Vicario et al. (2017)) and that internet and social media facilitate the self-selection of like-minded contacts and information sources, resulting in echo chambers or information cocoons (see e.g. Cinelli et al. (2021)). Another explanation for the occurrence of information bubbles is that search algorithms employed by search engines and social media tend to select information that is likely to agree with the user's ideas (e.g. Mobasher et al. (2000); Nikolov et al. (2015); and Levy (2021)) Both echo chambers and filter bubbles create situations where agents update their beliefs, mostly according to the information they receive from within the bubble. Our model explains why information bubbles are harmful even in the absence of bots. The main argument is that bubbles represent a partition of the set of agents into weakly coupled subgroups of various sizes. Our analysis, however, shows that every bubble has the same impact on the global consensus, regardless of its size. This is an obstruction to wisdom, as the wisdom of crowds phenomenon deploys for large sizes of agents. A partition of the nodes into smaller subgroups will decrease the likelihood of social learning arriving at a consensus close to the truth. Moreover, one echo chamber can be biased and less open to beliefs outside the bubble. Our study shows that this introspection tendency obstructs social learning, since introspective chambers neglect information outside the bubble. In addition, the influence of subgroups grows with introspection, which amplifies the harmful effect of introspective subgroups on social learning.

Finally, this paper provides a theoretical foundation of the media term "false balance". The latter emerges from the ideal of journalistic objectivity of reporting all (credible or reasonable) opposing positions. Similar to information bubbles, the harmful effect is that every position of an ideological community has the same impact on consensus, irrespective of its number of supporters and proportion of actual evidence.

7 Conclusion

We study the impact of bots on social learning in social networks. In particular, our model addresses the impact of bots (or stubborn agents) on consensus forming. We introduce the concept of instantaneous truth gap, which captures the impact of bots in case there is a comparatively "infinitesimal" small number of bots. We show that even the smallest number of bots has a sizeable impact on the consensus and hence represents an obstruction of the "wisdom of crowds'. However, the agent's network architecture impacts the speed of the learning process. If the convergence process of the agent network is sufficiently fast, the learning process of the agents can reach a quasi-stationary consensus independent of the bots. Identifying these resilient network structures gives rise to some optimism, as it shows means to preserve the wisdom of a crowd for a very long time.

Our framework allows various extensions for future work. A fruitful avenue could be to study a non-stationary model approach, where the truth value may slowly vary over time due to exogenous effects. Moreover, extending the model to consider a case in which links are endogenously determined would be interesting. Such a setting would allow agents to place a higher weight on individuals who share similar priors and choose to "unfollow" (e.g. break links) agents who have views that are relatively far from their own.

Finally, it is of high societal relevance to find counteractive measures to bots. Our model explains the underlying mechanics of misinformation and could be a fruitful toolbox for identifying effective policy measures.

References

Avrachenkov, K., Filar, J., and Haviv, M. (2002). Singular perturbations of Markov chains and decision processes. In *Handbook of Markov Decision Processes*, pages 113–150. Springer.

Azzimonti, M. and Fernandes, M. (2022). Social media networks, fake news, and polarization. *European Journal of Political Economy* (to appear).

Bala, V. and Goyal, S. (1998). Learning from neighbours. *The Review of Economic Studies*, 65(3):595–621.

Boucher, V. (2015). Structural homophily. *International Economic Review*, 56(1):235–264.

Bru, B. (1996). Problème de l'efficaciteté du tir a l'école d'artillerie de Metz. *Math. Inf. Sci. Hum.*, 34:29–42.

Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the Laplacian. *Problems in Analysis*, 625(195-199):110.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

de Moura, Y., Desidério, G., G.A., and et al (2021). *The Impact of Fake News on Social Media and its Influence on Health during the COVID-19 Pandemic: a Systematic Review*. Springer.

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

Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., and Quattrociocchi, W. (2017). Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*, 50:6–16.

Delyon, B. (2000). Stochastic approximation with decreasing gain: Convergence and asymptotic theory. *Unpublished lecture notes, Université de Rennes*, page 26.

DeMarzo, P., Vayanos, D., and Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *Quarterly Journal of Economics*, 118:909–968.

Ellison, G. and Fudenberg, D. (1993). Rules of thumb for social learning. *Journal of Political Economy*, 101(4):612–643.

Ellison, G. and Fudenberg, D. (1995). Word-of-mouth communication and social learning. *The Quarterly Journal of Economics*, 110(1):93–125.

Gale, D. and Kariv, S. (2003). Bayesian learning in social networks. *Games and Economic Behavior*, 75(2):329–346.

Golub, B. and Jackson, M. O. (2010). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. (2009). Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4):555.

Jadbabaie, A., Molavi, P., Sandroni, A., and Tahbaz-Salehi, A. (2012). Non-bayesian social learning. *Games and Economic Behavior*, 76(1):210–225.

Kastenmüller, A., Greitemeyer, T., Jonas, E., Fischer, P., and Frey, D. (2010). Selective exposure: The impact of collectivism and individualism. *British Journal of Social Psychology*, 49(4):745–763.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70.

Meyer, C. D. (1989). Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Review*, 31(2):240–272.

Mobasher, B., Cooley, R., and Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151.

Nikolov, D., Oliveira, D. F., Flammini, A., and Menczer, F. (2015). Measuring online social bubbles. *Peer Journal of Computer Science*, 1:e38.

Sasahara, K., Chen, W., and et al (2021). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4:381—-402.

Seneta, E. (1979). Coefficients of ergodicity: structure and applications. *Advances in Applied Probability*, 11(3):576–590.

Seneta, E. (1981). *Non-negative Matrices and Markov Chains*. Springer-Verlag, 2 edition.

Simon, H. A. and Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica: Journal of the Econometric Society*, pages 111–138.

Stieglitz, S., Brachten, F., Ross, B., and Jung, A.-K. (2017). Do social bots dream of electric sheep? a categorisation of social media bot accounts. *arXiv preprint arXiv:1710.04044*.

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., and Menczer, F. (2016). The DARPA Twitter bot challenge. *Computer*, 49(6):38–46.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Little, Brown.

Varol, O., Ferrara, E., Davis, C., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 280–289.

Yin, G. and Zhang, Q. (2012). *Continuous-Time Markov Chains and Applications: a Singular Perturbation Approach*, volume 37. Springer.