

Swiss Finance Institute

Research Paper Series

N°22-58

Fake News in Social Networks



Christoph Aymanns

QuantCo

Jakob Foerster

University of Oxford

Co-Pierre Georg

EDHEC Business School

Matthias Weber

University of St. Gallen and Swiss Finance Institute

Fake News in Social Networks^{*}

Christoph Aymanns[†] Jakob Foerster[‡] Co-Pierre Georg[§]
Matthias Weber[¶]

July 22, 2022

Abstract

We propose multi-agent reinforcement learning as a new method for modeling fake news in social networks. This method allows us to model human behavior in social networks both in unaccustomed populations and in populations that have adapted to the presence of fake news. In particular the latter is challenging for existing methods. We find that a fake-news attack is more effective if it targets highly connected people and people with weaker private information. Attacks are more effective when the disinformation is spread across several agents than when the disinformation is concentrated with more intensity on fewer agents. Furthermore, fake news spread less well in balanced networks than in clustered networks. We test a part of these findings in a human-subject experiment. The experimental evidence provides support for the predictions from the model. This suggests that our model is suitable to analyze the spread of fake news in social networks.

^{*}We thank Olimpia Carradori for excellent research assistance and Martina Dosser for help with programming the experiment. IRB approval was obtained from the University of St. Gallen. Financial support from GFF Project Funding from the University of St. Gallen is gratefully acknowledged.

[†]QuantCo. Email: christoph.aymanns@gmail.com.

[‡]University of Oxford. Email: jakob.foerster@eng.ox.ac.uk.

[§]EDHEC Business School. Email: co-pierre.georg@edhec.edu.

[¶]University of St. Gallen and Swiss Finance Institute. Email: matthias.weber@unisg.ch.

1 Introduction

Fake news are spread in social networks to manipulate users’ perception of facts. The 2016 US presidential election has highlighted the threat that fake news pose to open societies. Similarly, Brexit, Covid-19, and the war in Ukraine are further recent contexts in which fake news have played a crucial role. Yet, only little is known about the spread of fake news in social networks, mainly because decision making on social networks is highly complex and not well captured by existing models. In this paper, we use deep multi-agent reinforcement learning to solve a game of information aggregation on social networks. The advantage of this approach is that it allows for much richer learning dynamics, but it is not clear whether this is a more *realistic* model of human behaviour. Therefore, we conduct a controlled experiment, which shows that predictions of our model regarding the spread of fake news are consistent with human behavior in the laboratory.

We model the spread of fake news as a sequential game of information aggregation by N agents who interact via an exogenously given social network. Agents are presented with a claim that is either factually true or wrong and decide on a binary action. They are rewarded if their action matches the veracity of the claim. Agents receive private signals about the state of the world and can observe the previous actions of others. The game is subtle, because it is not necessarily optimal for agents to choose the action that best matches their posterior information since this might induce other agents to simply copy the action, in which case their actions reveal no further information. Because of this subtlety, this game is an ideal setting for independent Q-learning (IQL), which we use to computationally solve the game. Our computational approach is flexible, tractable, and yields a number of relevant insights.

First, our baseline is information aggregation without an attack. We show that the accuracy of information aggregation using IQL is higher than in the benchmark of DeGroot learning. Second, our numerical approach allows us to study various determinants of how effective an attack is, including the network position of the attacked citizen, the global structure of the network, and the distribution of bias across agents. Targeting highly connected agents in the network or those with weak private signals, makes biased signal attacks more effective. Clustered networks are more susceptible to attack than balanced networks and spread attacks are more effective than focused ones. We, third, then allow for the possibility of an attacker being present already during the training phase. When we allow for this possibility, we find that citizens can learn to adapt to the presence of an attacker and as a result the accuracy of information aggregation improves. Lastly, we study the case where an attacker knows the true state of the world and then chooses an action to convince

citizens from the opposite. Always choosing the action opposite to the state or choosing a random action is easily learned by the citizens. However, if we allow the attacker to be a separate deep Q-learner, the resulting attack is far more effective.

While the simulations provide us with a rich set of predictions, they alone are not sufficient to decide whether independent Q-learning is a realistic model of human behaviour. We, therefore, test part of the predictions from our simulations in a laboratory experiment with six treatments and a total of 560 participants. The participants are grouped into independent networks of eight and tasked to guess the state of the world correctly in each of several periods. As in our model, participants receive private information about the state of the world and observe the previous actions of those other participants they are connected with. We employ a 2×3 factorial between-subjects design where the two treatment dimensions are two different types of networks and three different types of attacks. By and large, the results from the human-subject experiment are in agreement with the predictions derived with our model.

Our work relates to several strands of literature. First, it relates to models of information aggregation in social networks. The standard information aggregation game, outlined in Section 2, involves agents (repeatedly) guessing an unknown state of the world taking into account an informative private signal and observations of actions of their neighbors on the social network in the previous period. Note that this model involves *no fake news*. There are two main approaches to studying this game. In the rational agent approach, agents' decision are best responses and the analysis focuses on the Nash equilibria of the game. In an important contribution, [Mossel et al. \(2015\)](#) show that under certain conditions on the social network, in the limit $T \rightarrow \infty$ (number of periods) and $n \rightarrow \infty$ (number of agents), agents will converge and will agree on the correct state of the world. This approach is not tractable for finite T and n and therefore does not allow, for example, the study of transient behavior. An alternative approach to studying the game of information aggregation is heuristic (see [Golub and Sadler, 2016](#), for a review of many important contributions). While heuristic approaches are more flexible to changes in the model setup and allow the study of transient behavior, agent strategies can be too simplistic. For example, agents may treat repeated observations of a neighbor's guess as conveying additional information, even when this is not the case (this occurs for example when this neighbor's only information is his private signal). Furthermore, heuristic behavior does not adapt to changing circumstances and would therefore not allow us to study how agent behavior changes in the presence of fake news. Our deep multi-agent reinforcement learning approach combines both worlds. It is flexible since it does not require the model to be

analytically tractable and finds near-optimal policies through reinforcement learning.¹

Second, our work relates to the literature on deep multi-agent reinforcement learning (DMARL) as a tool to solve games. DMARL has been applied to solve games of different structure. Policies or value functions are approximated by deep neural nets and trained via backpropagation. Some recent examples are [Lerer and Peysakhovich \(2017\)](#); [Letcher et al. \(2018\)](#); [Foerster et al. \(2018\)](#); [Balduzzi et al. \(2018\)](#). Often the focus is on methods that allow agents to converge on a particular, socially desirable policy, such as cooperation in social dilemmas without changing the payout structure. Since this is not an issue for most of our analyses, we make use of Independent Q-learning ([Tampuu et al., 2017](#)).

Third, our paper is related to the experimental literature about human interaction on social networks. This literature analyzes several ways in which people can interact on networks, from cooperation and public good provision on networks ([van Leeuwen et al., 2019](#); [van Leeuwen et al., 2020](#)) to the acquisition and spread of information in networks ([Centola, 2010](#); [Goyal et al., 2017](#)). However, there are hardly any human-subject experiments on fake news or disinformation. A notable exception is [Stewart et al. \(2019\)](#), who study a network voting game with bots in some treatments. Our setting and analyses differ significantly from theirs.²

To the best of our knowledge, we are the first to develop a flexible computational model of fake news on social networks in which agents act according to learned best response functions.³ We provide evidence that predictions of this model are realistic in the sense that they correspond to findings from a human-subject experiment.

Our simulations and the experiment help us to answer three main questions. (i) What determines the effectiveness of the attacker in reducing the accuracy of information aggregation? (ii) How well can agents learn adapt to the presence of an attacker? (iii) When attacker and agents learn simultaneously, how does behavior evolve over time?

Concerning the first question, we find that fake news spread more easily in networks with separate clusters than in more "balanced" networks. Furthermore, a fake news attack is more effective if the attacker spreads less careful disinformation over more agents than if he invests his efforts in more thoroughly convincing fewer people. Fake news attacks are

¹Our work is also related to studies of competitive information diffusion ([Alon et al., 2010](#); [Etesami and Başar, 2014](#)). However, while the literature on competitive information diffusion focuses on the strategic seeding of a pre-determined diffusion process, our work focuses on modeling the information diffusion process itself through the agents' learned actions.

²In addition, we are not studying voting outcomes but information aggregation more generally. We also link our experiment to the predictions of a very flexible reinforcement-learning model.

³In addition, we contribute to the MARL literature by providing an important and concrete real-world application of methods that to date have found limited practical application.

also more effective if agents with more connections on the network are attacked and if agents are attacked who have weaker private information. Concerning the second question, we find that agents can indeed learn to adapt to the presence of fake news. However, learning to adapt to fake news comes at a cost – if agents are trained in the presence of fake news, the accuracy when no fake news are present is lower than in the case that no fake news are possible at any stage. Regarding the third question, when agents and attackers learn simultaneously, behavior does not necessarily converge, but interesting irregular cyclical dynamics may evolve. Our findings have important implications for the fight against disinformation, which we discuss in the concluding section.

2 Modelling the spread of fake news as a network game

We consider a sequential game of information aggregation over T periods indexed by $t = 1, \dots, T$ with a set of agents N (for details, see [Mossel et al., 2015](#); [Golub and Sadler, 2016](#)). Agents interact via a fixed, possibly directed, graph $\mathcal{G}(N, E)$, where E denotes the set of edges connecting the nodes N . For concreteness, one can think of the agents as users of a social network, such as Twitter. A directed link from agent A to agent B would then imply that A follows B on Twitter. Figure 1 illustrates the networks used in this paper.

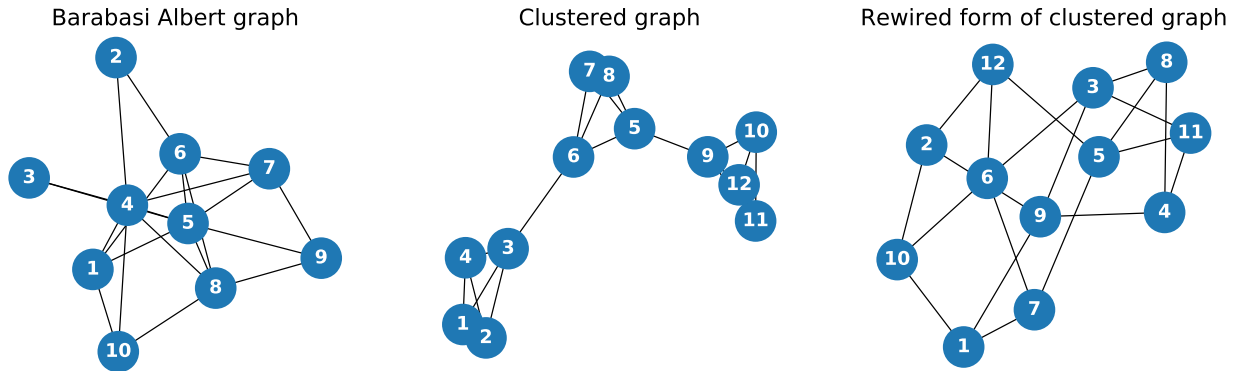


Figure 1: Left: Instance of a Barabasi-Albert random network with $|N| = 10$ that was used in our analyses (unless otherwise stated). Middle: stereotypical clustered network with $|N| = 12$ with three ($n_c = 3$) fully connected clusters of four nodes ($s_c = 4$) each. Right: the clustered network in balanced form produced by random rewiring of link pairs (preserves degree distribution). These networks were used in our analysis of the effectiveness of information aggregation and attack (i) in clustered vs. balanced networks and (ii) for spread out vs. focused attack strategies.

Agents are presented with a claim that is either true or false. Again, for concreteness, one can think of the claim as a factual statement encountered in a piece of news that has been shared among the agents on the social network. We restrict our analysis to the case when claims are factual statements that can be objectively verified as true or false. While this prevents us from studying the dynamics of opinions, our focus is on the spread of fake news. True claims are denoted as $\theta = 1$ and false claims as $\theta = 0$. The unconditional probability of a claim being true or false is $\mathbb{P}(\theta = 0) = \mathbb{P}(\theta = 1) = 1/2$. This choice simplifies our analysis without affecting any of our qualitative results. Agents receive private signals $s^i \sim F_\theta$ once at the beginning of an episode. In the baseline F_θ is such that private signals are always informative but noisy. In each period agents choose a binary action $a_t^i \in \{0, 1\}$. Agents observe the actions of their neighbors on the graph from the previous period. At the end of the episode, agents are rewarded for actions that matched the veracity of the claim presented. For a given claim θ and sequence of actions $\{a_t^i\}$, an agent’s total discounted reward is $r_i = \sum_t^T \gamma^t \mathbb{1}\{a_t^i = \theta\}$.

Note that in this game, for $t < T$, it is not necessarily optimal to choose the action that matches the maximum of the posterior likelihood of θ . This is because an agent wants to prevent others from simply copying its action, since that reveals no information about the others’ private signal. Injecting noise into actions may therefore be beneficial. Despite these concerns, it can be shown that, under certain conditions on the network, in the limit $T \rightarrow \infty$ (number of periods) and $n \rightarrow \infty$ (number of agents), agents will converge and will agree on the correct θ (Mossel et al., 2015).

In principle, reinforcement learning (RL) can be used to learn the agents’ policy (or response functions) in any game. Many methods have been proposed to learn reasonable best response functions in various multi-agent reinforcement learning (MARL) settings (Foerster et al., 2016; Letcher et al., 2018; Lerer and Peysakhovich, 2017). In this paper, the agents’ response functions will be learned via independent Q-learning (IQL) and Q functions will be approximated by deep recurrent neural networks (Tampuu et al., 2017; Foerster et al., 2016). We relegate a discussion of the RL and IQL framework to Appendix A.1.

For the simulations, we consider three types of *undirected* graphs. An instance of a Barabasi-Albert random graph with $|N| = 10$ (Barabási and Albert, 1999); a clustered graph with three clusters ($n_c = 3$) of fully connected nodes with four nodes each ($s_c = 4$) such that $|N| = 12$; and a balanced graph ($|N| = 12$) that is obtained by randomly rewiring the clustered graph. Random rewiring involves repeatedly selecting two edges uniformly at random and swapping the terminal nodes of these edges. This preserves the degree sequence,

that is the number of neighbors of each node, but it removes the clusters, so that the rewired graph is “balanced”. The Barabasi-Albert random graph is often used to model social networks and is therefore a natural choice for our application. The clustered graph is a stereotypical example of a social network which is comprised of cliques that are weakly connected. It thereby represents a more “fragmented” or “polarized” social network. The balanced graph removes these clusters from the graph while maintaining the degree sequence and can thus help isolate the effect of clusters on information aggregation. We analyze different types of networks, but the network remains fixed during each game.

3 Fake news and information aggregation

3.1 Fake news

We extend the model of information aggregation by introducing an adversary – the so called attacker – who wants to reduce the other agents’ total reward $\sum_i r_i$. To disambiguate between the attacker and other agents, we often refer to other agents as citizens. The attacker wants to persuade citizens to support false claims and dissuade them from supporting true claims. The attacker can achieve this in one of two ways. First, in a *biased signal attack*, he can distribute a fixed budget of bias $\beta = \sum_i \beta^i$ across a subset of the agents such that citizen i ’s private signal is drawn from $s^i \sim F_\theta(\beta_i)$. Biased signals can be thought of as targeted ads to users of the social network. Indeed, it has been shown that fake news attacks have taken exactly this form (Chiou and Tucker, 2018). A fixed bias budget corresponds roughly to an attacker with a fixed budget for targeted social media ads. We choose $s^i \sim \mathcal{N}(\theta + \beta^i(1 - 2\theta), \sigma^2)$. We set $\sigma^2 = 1$ throughout. In this case the choice of $\{\beta_i\}_{i \in N}$ is *hand-tuned*, that is set by the researcher and not optimized. Second, in the *agent takeover attack*, the attacker can directly take over an agent and act on his behalf. In this case the attacker optimizes his attack policy by independent Q-learning.

At the beginning of each episode θ all private signals are drawn at random according to the distributions given above. For *citizens* we train a single Deep Q-Network (DQN). In period t , the citizen observes the episode private signal s^i , his and the actions of his neighbors in the previous period $\{a_{t-1}^j\}_{j \in B_i}$ (B_i is the set of neighbors of agent i augmented by i) and the agent id i which identifies the agent’s network position. To maintain a constant length input vector, unobserved actions are encoded as -1 . Thus, the corresponding DQN is a function $Q : \mathbb{R} \times \{-1, 0, 1\}^{|N|} \times N \times \{0, 1\} \mapsto \mathbb{R}$.

In the *agent takeover attack*, the attacker is represented by a Deep Q-Network, which is distinct from the citizen DQN. At the beginning of each episode the attacker chooses a citizen uniformly at random from N and acts on his behalf. In period t , the attacker observes θ without noise, his and the actions of his neighbors in the previous period $\{a_{t-1}^j\}_{j \in B_i}$ and the agent id i . Thus, the corresponding DQN is a function $Q: \{0, 1\} \times \{-1, 0, 1\}^{|N|} \times N \times \{0, 1\} \mapsto \mathbb{R}$. In the *biased signal attack*, the attacker is *not* represented by a DQN. Instead, the biases are hand-tuned. At the beginning of each episode the attacker chooses a set of citizens uniformly at random from N and delivers biased signals to them as outlined above.

3.2 Measuring information aggregation

Intuitively, agents should be able to learn to extract information from their private signals and their neighbors' actions, such that at the end of an episode their expected reward ($\mathbb{E}[r_T^i]$) should be significantly higher than at the beginning of an episode ($\mathbb{E}[r_0^i]$). That is, agents should aggregate information over time. In the following, we give a formal definition of information aggregation for finite N and T that is useful in our case.

Definition 1 (Information aggregation). Let $\mathcal{F} = \{s^i\}_{i \in N}$ denote the set of all realized signals in a given episode. Let $\hat{\theta} = \operatorname{argmax}_{\theta} \mathbb{P}[\theta \mid \mathcal{F}]$ denote the maximum-a-posteriori estimator of θ given $\mathcal{F}$. Then information is aggregated if for all $i \in N$, we have $a_T^i = \hat{\theta}$.

That is, under this definition information is aggregated if agents act as if they had seen all private signals. In reality, it is not reasonable to expect perfect information aggregation with actions learned via independent Q-learning. We therefore define the following measure of information aggregation.

Definition 2 (Accuracy of information aggregation). The accuracy of information aggregation at time step t is $A_t = \mathbb{E}[\mathbb{1}\{a_t^i = \theta\}]$.

A useful benchmark to compare A_t to is of course $\mathbb{E}[\mathbb{1}\{\hat{\theta} = \theta\}]$. Another useful benchmark is $\mathbb{E}[\mathbb{1}\{\tilde{\theta} = \theta\}]$, where $\tilde{\theta} = \operatorname{argmax}_{\theta} \mathbb{P}[\theta \mid s^i]$ is the MAP estimator given only a single private signal. A_T can never exceed $\mathbb{E}[\mathbb{1}\{\hat{\theta} = \theta\}]$ and A_1 can never exceed $\mathbb{E}[\mathbb{1}\{\tilde{\theta} = \theta\}]$ and of course $\mathbb{E}[\mathbb{1}\{\hat{\theta} = \theta\}] > \mathbb{E}[\mathbb{1}\{\tilde{\theta} = \theta\}]$ provided $|N| > 1$. Once we have trained the agents via Q-learning it is easy to estimate A_t by evaluating a set of batches of games B and computing $\hat{A}_t = 1/|B|1/|N| \sum_{b \in B, i \in N} \mathbb{1}\{a_{t,b}^i = \theta\}$, where $a_{t,b}^i$ is the action taken by agent i in batch $b \in B$. In the following we will simply refer to $\hat{A}_t$ as the accuracy.

Lastly, we compare the accuracy of the DQN against the DeGroot heuristic (DeGroot, 1974). Under this heuristic information aggregation rule at $t = 0$, we set $a_0^i = \tilde{\theta}(s^i)$. In

each subsequent period actions are set to $a_t = Ua_{t-1}$, where a_t is the vector of actions taken by citizens and $U = D^{-1}(I + M)$, where I is the identity matrix, M is the graph adjacency matrix and D is a diagonal matrix with entries D_{ii} being the number of neighbors of node i plus one (i.e. a self loop). Information is therefore aggregated through repeated averaging of past actions.

4 Simulating the spread of fake news using independent Q-learning

Our objective is to improve our understanding of the propagation of fake news in social networks and discover potential interventions to curb its effectiveness. This raises three main questions: (i) What determines the effectiveness of an attacker in reducing the accuracy of information aggregation? In particular, how does the network structure and the choice of the attacked agent affect attack effectiveness? (ii) How well can agents learn to adapt to the presence of an attacker? (iii) When attacker and citizens learn simultaneously (agent takeover attack), how do policies change over time? To study these questions we set up a series of training and testing scenarios.

4.1 Training scenarios

We implement the following training scenarios. (i) **Baseline**: We train citizens in the absence of any attack, i.e. all citizens receive unbiased signals and no citizen is taken over by the attacker. This scenario establishes a benchmark for the ability of IQL agents to learn to aggregate information in the absence of attack. (ii) **Biased signal attack**: We train citizens in the presence of a biased signal attack where one or more randomly chosen citizens receive a biased signal. This scenario allows us to evaluate whether citizens can learn to adapt to the presence of a biased signal attack and thereby mitigate the adverse effects of the attack. (iii) **Agent takeover attack**: We train citizens and attacker simultaneously. This scenario allows us to understand the effectiveness of attack as it co-evolves with the citizen’s effort to mitigate the attacker’s ability to manipulate citizens. Note that IQL may not converge to stable policies in such a game. It is interesting, however, to understand the action dynamics that emerge under IQL as they can help us understand the co-evolution of fake news attack and defense in real social networks. (iv) **Random action attack**: We train citizens in the presence of an attacker that selects a single citizen uniformly at random and then, in each period, picks an action uniformly at random. This scenario serves as

a benchmark for the agent takeover attack. The accuracy under the random action attack should exceed the accuracy under the agent takeover attack.

4.2 Testing scenarios

Given the models obtained for these different training scenarios we conduct a number of testing scenarios. First, for each attack training scenario, we evaluate accuracy, as defined above, both in the presence of the corresponding attack (as trained) and in the absence of any attack. For the baseline scenario we consider two testing scenarios in addition to testing in the absence of any attack (as trained). First, we evaluate accuracy in the baseline model for the Barabasi-Albert graph in the presence of a biased signal attack with a single attacked agent and $\beta = 3$. Second, we evaluate accuracy in the baseline model for the clustered and balanced graphs under two biased signal attack scenarios. In the *focused* scenario, a single, randomly chosen citizen receives a biased signal with $\beta^i = 3$. In the *spread* scenario, two, randomly chosen citizens receive biased signals with $\beta^i = 1.5$ each.

5 Simulation Results

Our analyses yield four main results, which we will discuss in more detail below.

5.1 Baseline information aggregation

In the absence of an attack, citizens learn to aggregate information with an accuracy close to the optimal benchmarks $\mathbb{E}[\mathbb{1}\{\hat{\theta} = \theta\}]$ as $t \rightarrow T$ and $\mathbb{E}[\mathbb{1}\{\tilde{\theta} = \theta\}]$ for $t = 0$. Our method’s accuracy also substantially outperforms the accuracy achieved by the DeGroot information aggregation heuristic. In the presence of an attack, information aggregation is severely disrupted. These results are illustrated in Figure 2 (A).

5.2 Determinants of the effectiveness of an attack

Our simulations allow us to study the following determinants of attack effectiveness: network position of the attacked agent, signal strength of the attacked agent, network structure (clustered vs balanced) and distribution of bias across agents (focus vs spread). Let us consider each in turn.

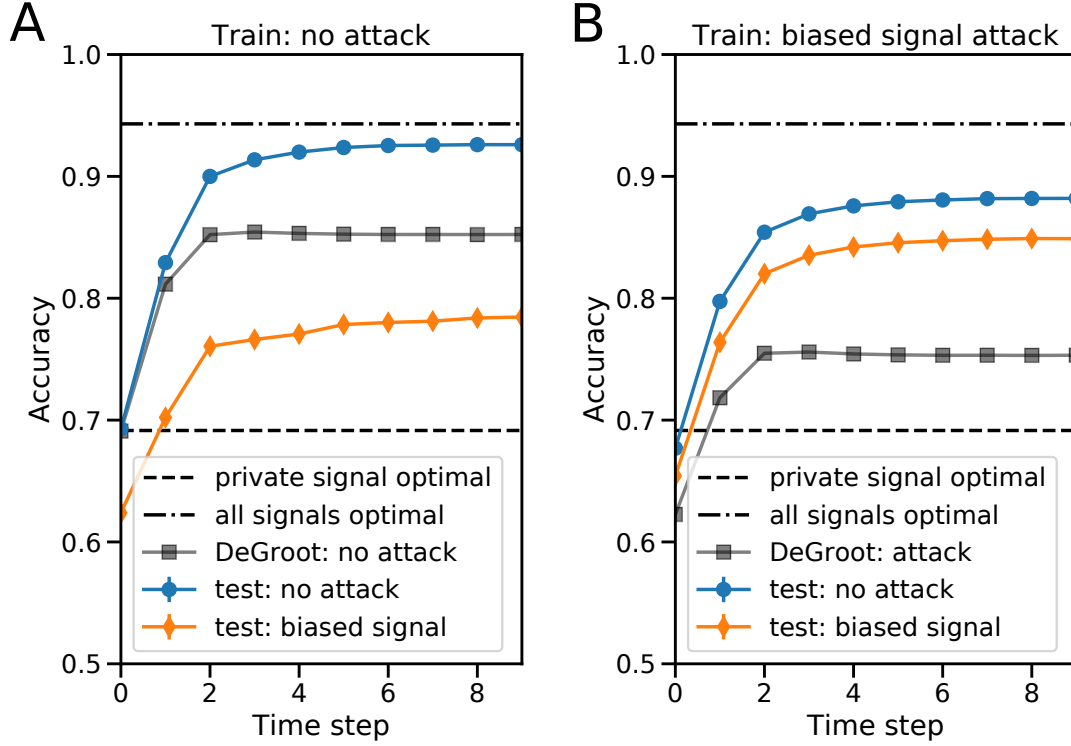


Figure 2: (A-B): Information aggregation over time in Barabasi-Albert graph. (A) Baseline scenario. Under the attack test scenario, a single agent receives a biased signal. (B) Biased signal attack scenario. The upper and lower dashed lines corresponds to the benchmarks $\mathbb{E}[\mathbb{1}\{\hat{\theta} = \theta\}]$ and $\mathbb{E}[\mathbb{1}\{\tilde{\theta} = \theta\}]$ respectively. Square markers correspond to the accuracy achieved by the De Groot information aggregation heuristic in the baseline and biased signal attack scenarios respectively.

To study the effect of network position, we compute the accuracy in the final time step ($t = T$) conditional on the network position of the attacked agent when evaluating the baseline model under the biased signal attack in the Barabasi-Albert network. Denote this conditional accuracy by $\hat{A}_T(i)$, where i is the attacked agent. Let $\hat{A}_T$ denote the baseline accuracy in the absence of attack. We then define the decline in accuracy relative to baseline as $\Delta A(i) = \hat{A}_T - \hat{A}_T(i)$. We plot $\Delta A(i)$ as a function of i in Figure 3 (A). It is clear that there is substantial heterogeneity in the attack effectiveness. As the tight correlation between $\Delta A(i)$ and the degree of i in Figure 3 (A) shows, the heterogeneity in $\Delta A(i)$ arises from the network structure which implicitly gives more influence to those nodes with more neighbors (higher degree). If an attacker is able to target the biased signal to such highly connected nodes, his attack will be more effective.

Next, we study the effect of the signal strength. Again, we restrict ourselves to evaluating the baseline model under the biased signal attack in the Barabasi-Albert network. A natural measure of a signal’s strength, or informativeness, is the absolute value of the log likelihood ratio of the two states of θ conditional on the signal. Let $f(\theta | s)$ denote the posterior distribution over θ given the signal s . Then, the signal strength is given by $L(s) = |\log[f(\theta = 1 | s)/f(\theta = 0 | s)]|$. Let L_k denote the k th decile of the empirical distribution of signal strengths. We define the conditional accuracy $\hat{A}(L_k)$ as the average accuracy in the final time step conditional on the attacked agent’s signal strength lying in $[L_k, L_{k+1}]$. $\Delta A(L_k)$ is defined analogously to $\Delta A(i)$. We plot $\Delta A(L_k)$ against L_k in Figure 3 (B). There exists a strong negative correlation between attack effectiveness as measured by $\Delta A(L_k)$ and the attacked agent’s signal strength. This is intuitive. If an agent has received a strong private signal in support of some value of θ , a larger bias will be required to convince him of the contrary. Thus, if an attacker can target the biased signal to agents with weak private signals, his attack will be more effective.

Lastly, we consider the case of clustered vs rewired/balanced networks and spread our vs focused attacks. This analysis is done evaluating the baseline model under the biased signal attack in the clustered and balanced networks. A number of results are worth noting. As can be seen in Figure 3 (C), for each scenario (baseline/no attack, focus and spread attacks) accuracy in the balanced network exceeds accuracy in the clustered network. This suggests that information aggregation is more effective in the balanced network. This is intuitive as the balanced networks has a shorter maximum path length between any two nodes thereby allowing information to “propagate” faster between nodes. In Figure 3 (C), we can also see that variation in the accuracy between training runs with different seeds (each run corresponds to one marker for a particular scenario) is smaller for the balanced network. This suggests that in the balanced network learning good policies is an easier

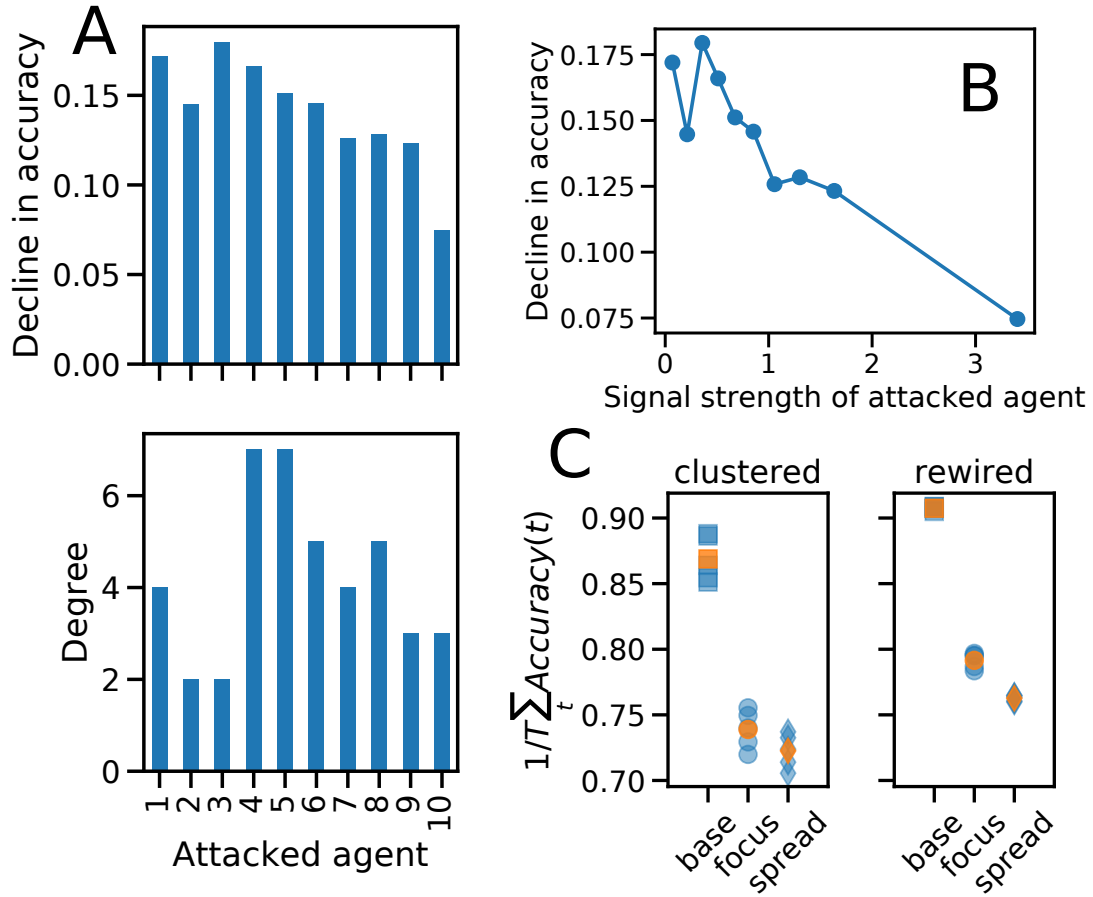


Figure 3: (A-B) Agents are trained in the absence of any attacker on a Barabasi-Albert graph. (A) Top: Average decline in accuracy conditional on agent id which determines network position. Bottom: degree of attacked agent. (B) Effectiveness of conditioning on attacked agent signal strength. (C) Agents trained in the absence of any attacker for clustered (left) and rewired / balanced (right). Base: no attack. Focus: a single, randomly chosen agent receives a strong biased signal ($\beta^i = 3$). Spread: two, randomly chosen agents receive weak biased signals ($\beta^i = 1.5$). Blue markers represent training runs with different seeds for a given network and attack scenario. Orange markers are averages over the different training runs. Error bars computed over outcomes with different random seeds but fixed neural network weights are too small to be visible on the graphs.

task than in the clustered network. We can also see from Figure 3 (C) that spread attacks are more effective both in the clustered and balanced networks.

5.3 Adapting to attacks

So far we have evaluated models trained in the absence of an attack. However, one can expect that users of social networks will adapt over time to the presence of fake news. We therefore investigate to what extent citizens can learn to mitigate the effect of an attacker. For this purpose we train citizens in the presence of a biased signal attacker, a random attacker and an agent takeover attacker for the Barabasi-Albert network. We then evaluate the accuracy of each of these models in the absence of an attack and in the presence of the attack under which they were trained. We summarize our results in Table 1.

Table 1: Test accuracy of agents in the final time step ($t = T = 10$) for different training and testing scenarios. All scenarios were run with the Barabasi-Albert graph. At test time without attack all agents receive unbiased signals and no agent is attacked. At test time with attack agents are exposed to the attack scenario they were trained under (except in the baseline case where attack is biased signal with $\beta = 3$).

Training scenario	Testing accuracy ($t = T$)	
	Without attack	With Attack
Baseline (train: no attack, test: hand-tuned $\beta = 3$)	0.926	0.784
Biased signal attack ($\beta = 3$)	0.882	0.849
Random action attack	0.909	0.908
Agent takeover attack	0.863	0.844

Let us first contrast the accuracy of the baseline model (trained without attack) under a biased signal attacker with the accuracy of a model trained under this attack scenario. When trained in the presence of an attacker the test accuracy under attack increases from 0.78 to 0.84 relative to the baseline. This can also be seen in Figure 2 (B). We conclude that agents can indeed learn to adapt to the presence of fake news. However, this adaptation comes at a cost. When trained in the presence of an attacker the test accuracy without attack decreases from 0.92 to 0.88 relative to the baseline. We conjecture that this is because agents learn to trust strong private signals less and are less likely to follow their neighbors' actions. A similar picture can be seen for the random attacker and the agent

takeover attacker. To summarize, agents can learn to adapt to the presence of an attacker which reduces attack effectiveness. However, adaptation comes at a cost. If trained under attack and tested in the absence of attack, accuracy is lower than in the baseline which is trained without attack.

5.4 Dynamics of attack and defense

We now take a closer look at the agent takeover attack. Recall that in this scenario, a separate DQN agent is trained to minimize the total reward of the citizens by acting on behalf of a randomly chosen agent. This attacker knows θ and must then choose an action to convince the citizens of the opposite. Simply choosing the opposite action would be a naive strategy that is likely to be detected quickly by the citizens as they learn in the presence of this strategy. Always choosing a random action in every period, as the random attacker does, is also quite easily detectable by the citizens. This intuition is confirmed by results reported in Table 1 and Figure 4.

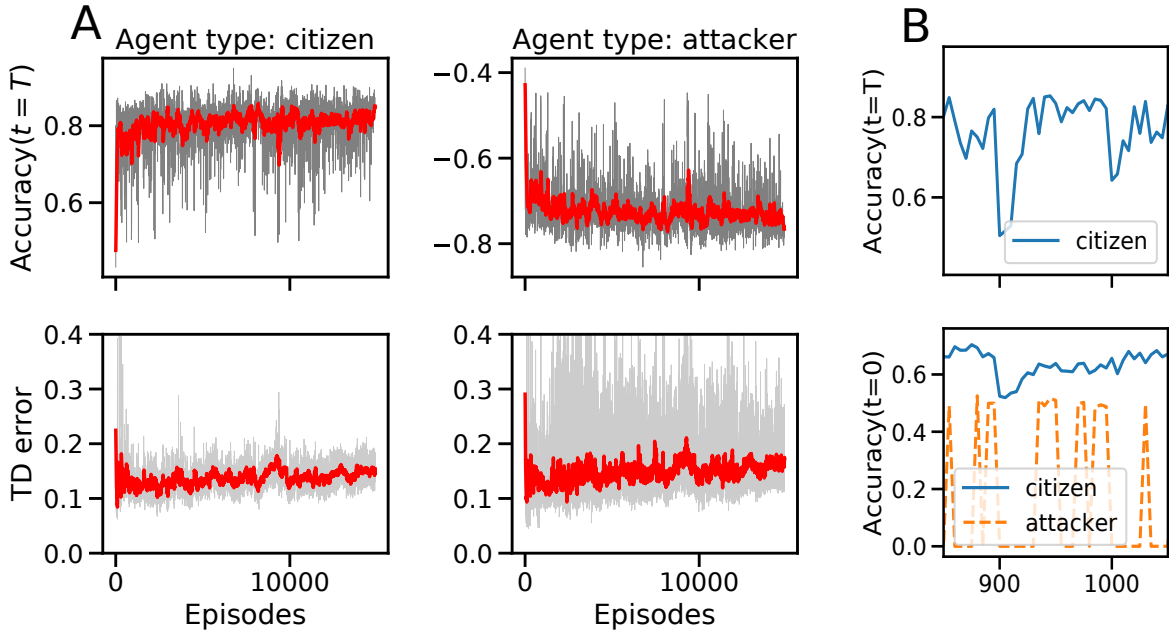


Figure 4: Dynamics of co-evolving attacker and citizen agents on a Barabasi-Albert graph. (A) Red lines are exponentially weighted moving averages. Top: accuracy in the last time step ($t = T = 10$) over episodes. Bottom: TD error of Q learning. (B) Close up of accuracy and actions of citizens and attacker early on during training averaged over multiple batches. Top: citizen accuracy in the last time step ($t = T = 10$). Bottom: citizen and attacker accuracy in the first time step ($t = T = 0$).

From Table 1 we see that the trained attacker clearly “outperforms” (i.e. more effective attack) the random action attack benchmark. The bottom panel of Figure 4 (B) shows the accuracy of the attacker in the first time step as a simple summary statistic of its strategy. The accuracy is either 0.5 indicating that the attacker “randomizes” or 0 indicating that the attacker “lies”. While Q-learning only allows for deterministic policies the attacker can learn to lie if the attacked agent is in one partition of N and be truthful if the attacked agent is in another partition. Since the attacked agent is chosen at random, this strategy amounts to a randomized attack strategy.

Note further, that neither the attacker’s nor the citizens’ policy converges as can be seen in the highly variable reward and unstable TD-error in Figure 4 (A). Indeed, as can be seen in the bottom panel of Figure 4 (B), the attacker’s policy fluctuates between randomization and lying between episodes. These fluctuations can be seen as irregular “cycles” in the policy space of the attacker and the citizen. One such cycle in Figure 4 (B) shows the attacker switching to a lying strategy, while the citizens lose “trust” in their private signal (they nearly ignore it in the first time step). The result of these two events coming together can lead to substantial, temporary declines in accuracy.

We don’t claim that these dynamics are accurate representations of the learning dynamics between users of social networks and pushers of fake news. However, we consider the failure of IQL to converge and the irregular dynamics that emerge more of a feature than a bug. Indeed these dynamics suggest two interesting, qualitative take aways. First, attackers and citizens are unlikely to converge to stable policies in real life. We should expect a continuing cycle of adaptation with periods of a high prevalence of fake news followed by periods of low prevalence. Second, the co-evolution of attackers and citizens can have unexpected effects on the citizens’ ability to aggregate information in the social network. As we observe in Figure 4 (B), the presence of neighbors that do not act in accordance with their private signals (i.e. the attacked agents) can lead citizens to learn to mistrust their own private signals.

6 Experimental Design and Procedures

To see whether the results obtained from the reinforcement learning model tell us something about actual human behavior in networks, we conduct a human-subject experiment. In this experiment, we test comparative statics obtained with the reinforcement learning model.

The experiment was pre-registered before the data collection began. The pre-registration is available at <https://aspredicted.org/blind.php?x=rf557t>. Data collection and data analysis were entirely separated.

6.1 Course of Events

In the beginning of the experiment, participants read the instructions. Before they are able to proceed to the first round, they also have to answer a set of comprehension test questions correctly. The instructions and comprehension test questions are reproduced in Appendix B and C.

Participants are randomly assigned to groups of eight participants. The eight participants of one group are connected on a network (the type of network that the participants are connected on depends on the treatment and will be shown in Section 6.2 below). The group composition remains the same during the whole experiment, and participants only interact with participants of their own group (directly or indirectly).

The experiment consists of six rounds. Each round consists of ten periods. At the beginning of each round, the true state of the world is randomly determined to be either -1 or 1 , each with probability one half.⁴ The state of the world remains the same throughout all periods of a round.

The participants' task is to guess the state of the world correctly in each of the ten periods of this round (the state of the world is framed neutrally as a coin flip in the experimental instructions). Thus, in each period, they have to make a binary choice, guessing -1 or 1 .⁵ The true state of the world in a round is not revealed to participants until the end of the experiment. Participants can draw on two types of information to make their guesses. First, they can use their own private signal. This private signal is drawn once for a round and is communicated to participants before the round starts. The signal is drawn from a normal distribution with standard deviation one and is independent across participants and rounds (the mean of a participant's signal depends on the treatment; the average of this mean across all participants of a group is always negative if the true state of the world is -1 and positive if the state of the world is $+1$, so that the signals are on average always informative). The second type of information participants can use are the past actions of

⁴We choose -1 and 1 in the experiment as states of the world instead of 0 and 1 , so that negative signals correspond to signals pointing to one state of the world and positive signals to the other state.

⁵In the first round, there is a time limit of 60 seconds per decision, in the other rounds of 30 seconds per decision. However, the time limit is not strict, after that time a message pops up asking participants to make their choice.

all participants that they are connected to on the network. For instance, when making the guess for the fifth period of a round, a participant connected to two other participants on the network, knows the guesses of these two other players from periods one to four (in addition to her own private signal).

In each period, a participant who guessed the state of the world correctly receives 100 points. At the end of the experiment, one period per round is selected randomly to be actually paid out (this is clearly communicated to participants in the instructions; only one period per round is paid out, so that participants cannot hedge). Points are exchanged into euros at an exchange rate of 40 points to 1 EUR. In addition to the payoffs from the tasks, each participant receives a show-up fee of five euros.

6.2 Treatments

The experiment makes use of a 2×3 -factorial between-subjects design. The two treatment dimensions are the type of network (rewired network or clustered network) and the type of fake-news attack represented by the existence and type of biased signals (no attack, spread-out attack, concentrated attack).

Figure 5 shows the networks used in the experiment. In the treatments without attack, all private signals are unbiased (thus, all signals are drawn from a normal distribution with the mean equal to the true state of the world). In the treatments with a spread-out attack, four members of the group receive an unbiased signal and four members receive a biased signal, where the bias is 1.75 (that is, the four attacked members with the biased signal receive a draw from a normal distribution with mean 0.75 if the state of the world is -1 and mean -0.75 if the state of the world is 1). In the treatments with a concentrated attack, six group members receive an unbiased signal and two receive a biased signal with a bias of 3.5 (thus, the two attacked members receive a signal from a normal distribution with mean 2.5 if the state of the world is -1 and -2.5 if the state of the world is 1).

The instructions contain the information that there may be biased signals – there is no deception in the experiment. However, we do not tell participants whether they are in a treatment with or without biased signals or how many biased signals of which strength there are.

Table 2 summarizes the design. We number treatments consecutively: T1 is *rewired_no-attack*, T2 is *rewired_spread*, T3 is *rewired_concentrated*, T4 is *clustered_no-attack*, T5 is *clustered_spread*, and T6 is *clustered_concentrated*. We consider the treatments without

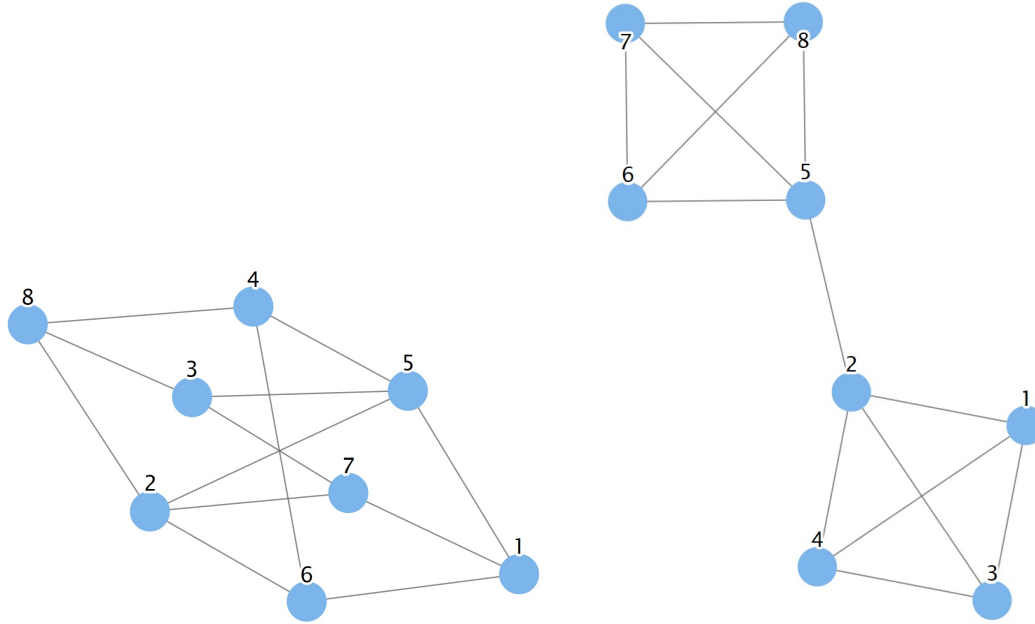


Figure 5: Rewired network (left) and clustered network (right), as used in the experiment.

attack, T1 and T4, control treatments. The number of observations corresponds to the number announced in the pre-registration.

Table 2: Design Summary. This table shows the treatment abbreviations and the numbers of participants and networks (in parentheses) per treatment.

	No attack	Spread-out attack	Concentrated attack
Rewired network	T1, 56 (7)	T2, 112 (14)	T3, 112 (14)
Clustered network	T4, 56 (7)	T5, 112 (14)	T6, 112 (14)

6.3 Hypotheses

Our outcome variable of interest is the accuracy of participants' guesses of the state of the world. The hypotheses are derived with the multi-agent reinforcement model. The hypotheses are as in the pre-registration and can be grouped into three parts (a to c below). The predictions by the model are our hypotheses, thus in statistical terms these are the alternative hypotheses.

- a) Accuracy is greater in the control treatments than in the treatments with fake news for the same network type: $\text{acc}(T1) > \text{acc}(T2)$, $\text{acc}(T1) > \text{acc}(T3)$, $\text{acc}(T4) >$

$\text{acc}(\text{T5})$, and $\text{acc}(\text{T4}) > \text{acc}(\text{T6})$.

- b) Accuracy is greater in the treatment with concentrated attacks than in those with spread-out attacks for the same network type: $\text{acc}(\text{T3}) > \text{acc}(\text{T2})$ and $\text{acc}(\text{T6}) > \text{acc}(\text{T5})$.
- c) Accuracy is greater in rewired networks than in clustered networks for the same type of attack: $\text{acc}(\text{T1}) > \text{acc}(\text{T4})$, $\text{acc}(\text{T2}) > \text{acc}(\text{T5})$ and $\text{acc}(\text{T3}) > \text{acc}(\text{T6})$.

Note that for hypothesis c1, $\text{acc}(\text{T1}) > \text{acc}(\text{T4})$, our model indicates that the difference between the two accuracies is so small that we did not expect to find statistically significant results.⁶

The hypotheses vary in terms of how likely it is that they will be confirmed. The hypothesis set a) comparing the control treatments to the treatments with fake news – that is, biased signals – are likely to be confirmed in experiments. These hypotheses do not require a high degree of rationality by participants to hold. Hypothesis sets b) and c) are less obvious and it is easy to come up with theories pointing towards the opposite. For instance, if those citizens with particularly strong signals are most likely to stick with their signals, while other citizens are more likely to react to others' actions, this could lead to concentrated attacks being more successful (i.e., leading to lower accuracy) than spread-out attack, in contrast to our predictions. It would also be conceivable that attacks are more successful in rewired networks than in clustered networks (in contrast to our predictions), as the biased signals could spread to a larger set of participants immediately instead of propagating first within a cluster.

6.4 Procedures

We report data from a total of 560 participants. 56 experimental subjects (seven independent networks of eight) participated in each of the two control treatments. 112 subjects (14 independent networks) participated in each of the four fake-news treatments.⁷ The experiment was conducted at the LINEEX laboratory in Valencia. The experiment was carried out in Spanish (instructions and comprehension test questions in the appendix are the English versions; the translation to Spanish was carefully verified). After the experiment,

⁶This is written in the pre-registration as follows: *"These are the predictions/hypotheses from the model. Note, however, that we expect the treatment difference "c1. $\text{acc}(\text{T1}) > \text{acc}(\text{T4})$ " to be so small that we do not expect to observe statistically significant results (this is a reason for choosing a lower number of observations in the control treatments than in the other treatments..."*

⁷The treatments were balanced across the ten experimental sessions, so that in each session there were zero or one control treatment networks and one or two fake-news treatment networks (as fewer data points were collected from the control treatments).

participants filled out a short questionnaire asking for a few demographic variables. Approximately half of the participants are female. The average age of participants is about 21 years. About half of them are students of economics, finance, or business. More than 90% of participants are Spanish.

7 Experimental Results

The key outcome variable we use is the mean accuracy in a network, i.e. the mean across participants on the network across all rounds and periods. As participants only interact with participants in the same network, these network-level variables are statistically independent. We rely on Wilcoxon rank-sum tests for hypothesis testing, using one-sided tests because of the directedness of our pre-registered hypotheses. Robustness checks making use of regressions can be found in Appendix D. The statistical analysis generally follows the description in the pre-registration.⁸

Figure 6 illustrates the experimental results, depicting the mean accuracy per treatment. Table 3 shows the hypotheses discussed above, i.e. the relationships predicted by the reinforcement learning model, together with the relationship observed in the data and the p -values of Wilcoxon rank-sum tests.

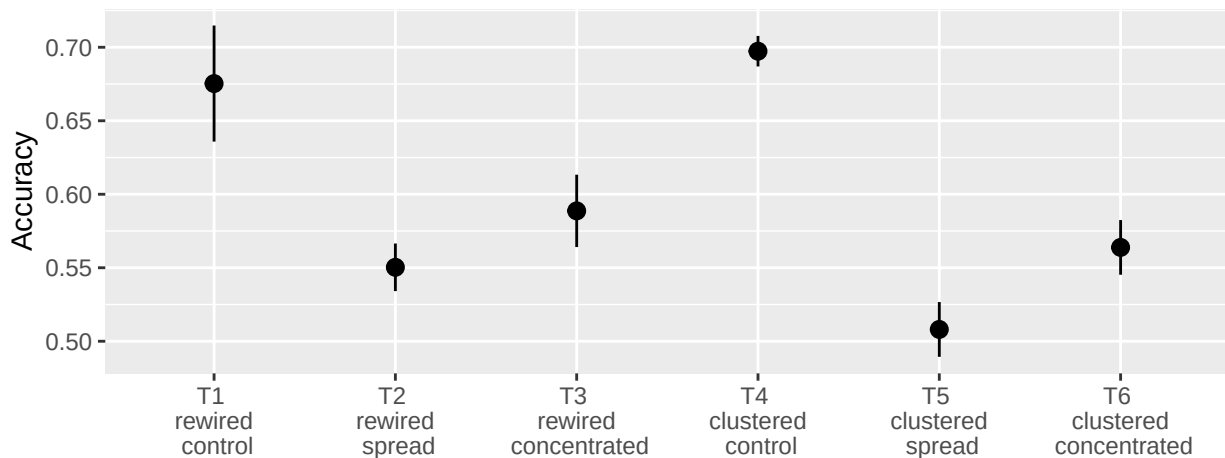


Figure 6: Mean accuracy in the treatments of the experiment.

The directions of treatment effects from the experimental data are always as predicted, with the exception of the comparison of the two control treatments where the results are

⁸The pre-registration mentions the mean accuracy across all periods and rounds as first outcome measure. In addition, the pre-registration mentions the mean across all periods of the last round only as a second outcome measure, a measure that is more noisy than the first measure and that we have discarded.

Table 3: Comparative statics in predictions and data and p -values. This table shows the comparative statics of the predictions, i.e. the hypotheses, jointly with the directions of treatment effects, jointly with the p -values of the one-sided Wilcoxon rank-sum tests. * For this hypothesis, no significant differences were expected ex ante, as discussed in Section 6.3 and the pre-registration.

	Hypothesis	Data	p -value
Control (left) vs. fake news (right)			
a1. acc(T1) vs. acc(T2)	>	>	0.003
a2. acc(T1) vs. acc(T3)	>	>	0.078
a3. acc(T4) vs. acc(T5)	>	>	0.000
a4. acc(T4) vs. acc(T6)	>	>	0.001
Concentrated (left) vs. spread (right)			
b1. acc(T3) vs. acc(T2)	>	>	0.135
b2. acc(T6) vs. acc(T5)	>	>	0.020
Rewired (left) vs. clustered (right)			
c1. acc(T1) vs. acc(T4)*	>	<	0.896
c2. acc(T2) vs. acc (T5)	>	>	0.047
c3. acc(T3) vs. acc(T6)	>	>	0.275

very close together and statistically highly insignificant. This is the hypothesis for which we already mentioned in the pre-registration that we do not expect to find support for it, as also discussed in Section 6.3. Note that for this comparison, there is only data from 14 groups in total, less than for the other comparisons. Considering the three blocks of hypotheses, there is strong evidence that accuracy is lower in the fake-news treatments than in the control treatments. Three of the four tests are strongly significant while the fourth is marginally significant. There is also support that accuracy is lower when the fake-news attack is spread-out across more different participants with lower intensity than when it is concentrated on fewer participants. This comparison is statistically significant at the 5%-level for the comparison when the network is clustered, while the p -value is slightly above the level for marginal significance in the rewired-network comparison. There is also some evidence that accuracy is lower in clustered networks than in rewired networks in the presence of a fake-news attack. When the fake-news attack is more effective, i.e. when it is spread-out, this difference is statistically significant. When this fake-news attack is less effective, i.e. when it is concentrated, the direction of the treatment effect points in the same direction, while the treatment difference is statistically insignificant.

In sum, our experimental results support the comparative statics obtained with the reinforcement learning model in Section 4. Not all hypothesized treatment differences are statistically significant at the 10% level. However, basically all treatment directions are as hypothesized and most of the treatment differences are statistically significant.

8 Concluding Remarks

The deliberate manipulation of the public’s perception of facts via fake news, in particular on social networks, has become a growing concern for policy makers and technology companies alike. Despite its importance, little is understood about the spread of fake news on social networks. This is in large part due to the technical difficulties involved in studying complex decision making behavior on social networks. We develop a framework for the study of fake news on social networks that is theoretically grounded yet computationally tractable and flexible. We achieve this by first extending a standard game of information aggregation to accommodate fake news and then applying state-of-the-art deep multi-agent reinforcement learning to solve the game. We use predictions of the model to form hypotheses that we can test in experiments with human subjects. In these experiments we find evidence that our model of human behaviour is indeed useful to understand the spread of fake news in social networks.

Our findings suggest a number of interventions that can contribute to making fake news attacks in social networks less effective. First, social media companies could adjust their algorithms suggesting users which other users to connect to (such algorithms exist for many social media platforms, including Facebook, LinkedIn and Twitter). These algorithms could give preference to connections that are in different clusters of the network, thereby making the network more balanced. Second, keeping some information on social networks – such as the number of connections – private can make it harder for attackers to condition their attacks on the network position. In addition, simply making agents aware of the potential presence of fake news will already contribute to their adaption and mitigate the effect of disinformation.

References

- Alon, N., Feldman, M., Procaccia, A. D., and Tennenholtz, M. (2010). A note on competitive diffusion through social networks. *Information Processing Letters*, 110(6):221–225.
- Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. (2018). The mechanics of n-player differentiable games. *arXiv preprint arXiv:1802.05642*.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197.
- Chiou, L. and Tucker, C. (2018). Fake news and advertising on social media: A study of the anti-vaccination movement. Technical report, National Bureau of Economic Research.
- Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.
- DeGroot, M. H. (1974). Reaching a consensus. *Journal of American Statistical Association*, 69:118–121.
- Etesami, S. R. and Başar, T. (2014). Complexity of equilibrium in diffusion games on social networks. In *2014 American Control Conference*, pages 2065–2070. IEEE.

- Foerster, J., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145.
- Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems.
- Golub, B. and Sadler, E. (2016). Learning in Social Networks. *The Oxford Handbook of the Economics of Networks*.
- Goyal, S., Rosenkranz, S., Weitzel, U., and Buskens, V. (2017). Information acquisition and exchange in social networks. *The Economic Journal*, 127(606):2302–2331.
- Hausknecht, M. and Stone, P. (2015). Deep recurrent q-learning for partially observable mdps. In *Sequential Decision Making for Intelligent Agents – Papers from the AAAI 2015 Fall Symposium*.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.
- Lerer, A. and Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.
- Letcher, A., Foerster, J., Balduzzi, D., Rocktäschel, T., and Whiteson, S. (2018). Stable opponent shaping in differentiable games. *arXiv preprint arXiv:1811.08469*.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518:529–533.
- Mossel, E., Sly, A., and Tamuz, O. (2015). Strategic learning and the topology of social networks. *Econometrica*, 83(5):1755–1794.
- Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G., and Plotkin, J. B. (2019). Information gerrymandering and undemocratic decisions. *Nature*, 573(7772):117–121.

- Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., and Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4):e0172395.
- van Leeuwen, B., Offerman, T., and Schram, A. (2020). Competition for status creates superstars: An experiment on public good provision and network formation. *Journal of the European Economic Association*, 18(2):666–707.
- van Leeuwen, B., Ramalingam, A., Rojo Arjona, D., and Schram, A. (2019). Centrality and cooperation in networks. *Experimental Economics*, 22(1):178–196.

Online Appendix to "Fake News in Social Networks: Reinforcement Learning and Experimental Evidence"

Christoph Aymanns

Jakob Foerster

Co-Pierre Georg

Matthias Weber

This online appendix contains material in addition to the main text. Section [A](#) contains details about the implementation of the deep multi-agent reinforcement learning. Section [B](#) contains the experimental instructions. Section [C](#) reproduces the comprehension test questions used in the experiment. Section [D](#) contains additional statistical analyses of the experimental data as robustness check.

A Deep Multi-Agent Reinforcement Learning

A.1 Multi-agent reinforcement learning

A.2 Single-Agent Reinforcement Learning

Consider an MDP specified by the following tuple: $\langle S, A, P, r, \gamma \rangle$. In the fully observable case the environment has an observed true state, $s \in S$, and at each time step the agent chooses an action $a \in A$, which induces a probabilistic transition to the next state via the state transition function $P(s'|s, a) : S \times A \times S \rightarrow [0, 1]$. The reward function assigns an instantaneous reward to each state-action pair, $r(s, a) : S \times A \rightarrow \mathbb{R}$, and $\gamma \in [0, 1)$ is a discount factor.

In single agent RL the task of the agent is to maximize the discounted reward per episode, $R_t = \sum_{l=0}^{\infty} \gamma^l r_{t+l}$. The agent's policy, $\pi(a|s) : S \times A \rightarrow [0, 1]$, induces a state-action-value function (Q-function), $Q^\pi(s_t, a_t) = \mathbb{E}_{s_{t+1:\infty}, a_{t+1:\infty}} [R_t | s_t, a_t]$. Q-learning aims to estimate the optimal action-value function, $Q^*(x, a) = \max_{\pi} Q^\pi(x, a)$, via an estimated Q-function, $Q(s, a)$.

Training is carried out by collecting samples from the environment in order to obtain Monte-Carlo-Estimates of the expected return. For any sampled transition, the current estimate, $Q(s, a)$, is compared to a greedy one-step lookahead using the Bellman optimality

operator, $\mathcal{T}Q(s, a) = \mathbb{E}_{s'}[r + \gamma \max_{a'} Q(s', a')]$. We can use samples to evaluate the Bellman update:

$$Q(s, a)_{k+1} = Q(s, a)_k + \alpha \left(r + \gamma \max_{a'} Q(s', a')_k - Q(s, a)_k \right)$$

Here k is the iteration number, α is the learning rate and $r + \gamma \max_{a'} Q(s', a')_k - Q(s, a)_k$ is commonly referred to as the temporal-difference or TD-error. In the tabular case the Q -values for each state-action pair are maintained separately and the Bellman update is a contraction mapping. As a consequence at convergence this iterative process results in the optimal Q -function, $Q(s, a)_\infty = Q^*(s, a)$. Finally, $Q^*(s, a)$ trivially defines the optimal policy $\pi^*(x, a) = \delta(\arg \max_{a'} Q^*(x, a') - a)$, where $\delta(\cdot)$ is the Dirac-delta function. In contrast to the tabular case, DQN (Mnih et al., 2015) uses a neural network parametrized by a large number of weights, ϕ , to represent the Q -function.

In order to reduce the variance of the update process the average square of the TD-error across a large number of transitions (the *batch*) is used as the DQN-loss. Using backpropagation the parameters of the neural network are updated to minimize the magnitude of the DQN error $\mathcal{L}(\phi) = \sum_{j=1}^b [(y_j^{DQN} - Q(x_j, a_j; \phi))^2]$. Here $y_j^{DQN} = u_j + \gamma \max_{a'_j} Q(x'_j, a'_j; \phi^-)$, is the target function and ϕ^- is the target network, which contains a stale copy of the parameters. This target network helps to stabilize the training.

So far we have assumed that the agent has access to the Markov state, s , of the system. In a partially observable settings we augment the MDP with an observation space, Z , and observation function $O(s)$. In particular the observation $z \in Z$ is produced by the observation function $O(s) : S \rightarrow Z$. We further define an action-observation history $\tau \in T \equiv (Z \times U)^*$, which is used to condition a stochastic policy $\pi(a|\tau) : T \times U \rightarrow [0, 1]$. In recurrent deep RL (Hausknecht and Stone, 2015), this can be achieved using recurrent neural networks, such as LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014) which we use here.

A.3 Independent Q-Learning

In multi-agent reinforcement learning, each agent $i \in N$ receives a private observation $O(x, i)$, where i is the agent index and O is the observation function. The agents also receive individual reward r_t^i and take actions a_t^i . Furthermore the state-transition conditions on the joined action $\mathbf{a} \in \mathbf{A} \equiv A^n$.

In independent Q-learning each agent further estimates a Q -function $Q^i(\tau^i, a^i)$, treating the other agents and their policies as part of a non-stationary environment. A combination

DQN with IQL commonly uses parameter-sharing across agents combined with an agents specific index in the observation function in order to accelerate learning while still allowing for specialization of policies.

A.4 Implementation details

Our code is available on github. Our neural network architecture implements the IQL architecture in [Foerster et al. \(2016\)](#).

B Experimental Instructions

We first reproduce the complete instructions of the treatments with rewired networks, T1-T3 in Section [B.1](#). We then describe where the instructions of the treatments with clustered networks, T4-T6, differ from the reproduced instructions. These differences are described in Section [B.2](#).

B.1 Instructions in the Treatments with Rewired Networks

Instructions

Please read these instructions carefully as they explain how you earn money from the decisions that you make. You will be paid privately at the end, after all participants have finished the experiment.

During the experiment you are not allowed to use the internet, your mobile phone, tablet, or notebook. You are also not allowed to communicate with other participants.

If you have a question at any time, please raise your hand and someone will come to your desk to answer your question in private.

The experiment consists of 6 rounds. Each round consists of 10 periods. Your decision in each period and round may affect your final payoff. **At the end, only your earnings from one randomly selected period per round will be paid out to you.** This means that for each round, one period will be randomly selected, and the payment from the experiment will be the sum of these 6 randomly selected periods (one per round). The points from the selected period will be exchanged into EUR at the exchange rate 40 points = 1 EUR. In addition, you will receive 5 EUR for your participation.

You have been randomly assigned to a **group of 8 participants/players** (including you). Your group will not change during the experiment. You will not know the identity of any other group member nor will they know your identity even after the experiment is over.

The following describes what you will be doing in **each** of the 6 rounds.

General information

In the experiment, all members of one group are connected in a network. Below, you can see this network. The network will be the same throughout the whole experiment. Within each group, players are assigned numbers from 1 to 8. As shown in the graph below, the numbers determine the positions in the network. Each player will get to know their player number during the experiment. This player number and the position in the network will remain the same throughout the experiment. The lines in the graph indicate which players are connected. A connection between two players means that the players can observe each other's past decisions in the current round (what decisions players can make will be explained later).

[Figure B.1 appears here in the experimental instructions.]

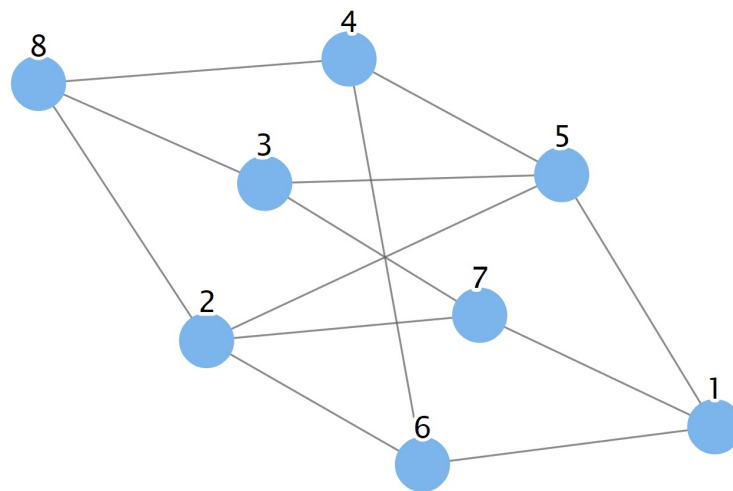


Figure B.1: Rewired network (not labeled in the experimental instructions).

In this network, player 7 can for example observe the past decisions of players 1, 2, and 3 but not the decisions of other players. As another example, player 2 can observe the past decisions of players 5, 6, 7, and 8.

Setting, decisions, payments

At the beginning of each round, we flip a fair coin (to be precise, the computer's random number generator "flips the coin"). On one side of the coin, the number 1 is written. On the other side of the coin, the number -1 is written. That is, 1 and -1 appear with equal probability. Each player's task is to correctly guess the outcome of the coin flip (later, we will explain which hints/information you will receive that may be relevant for your guesses).

In each period (of each round), you need to state whether your guess of the outcome of the coin flip is -1 or 1. If your guess of the number in a period was correct, you receive 100 points, if the guess was incorrect you receive nothing for that round (remember that only one period of each round will later be randomly selected for payment). Note that the outcome of the coin flip for any round will not be revealed to you until the very end of the experiment when the payments are determined.

In the first round, you have up to 60 seconds for each decision. In the other rounds, you have up to 30 seconds. When the time elapses, a notification pops up (you will still be able to make a decision after this notification, but we ask you to respond within the given time in order not to prolong the experiment).

Signals and information

At the beginning of each round (after the coin has been flipped), each player receives a signal about the outcome of the coin flip. This signal has a random component in it.

Two types of signals exist: unbiased signals and biased signals. The mean (that is, the long-term average) of the unbiased signal is equal to the outcome of the coin flip. The mean of the biased signal is not equal to the outcome of the coin flip. Both unbiased signals and biased signals follow a so-called normal distribution with standard deviation 1. The signals are drawn independently for all players (this means that the random disturbance added to the mean in one player's signal does not influence the random disturbance added to the mean of another player's signal).

At least half of the players of a network receive an unbiased signal. That is, at least four players in a network receive a signal that is on average correct. You do not know whether your signal is biased or unbiased. Note that the average of the mean of the 8 signals in one group is always closer to the outcome of the coin flip than to the number that did not come up in the coin flip. In simple terms, this means that you can always expect that all of the group's signals considered jointly are informative about the outcome of the coin flip.

The next graph shows how likely it is that such a signal falls into a certain range. In this example, the outcome of the coin flip (that players cannot observe) is 1. The probability that the random signal of any player who receives an unbiased signal falls into a certain range of numbers corresponds to the area under the curve. For example, the gray shaded area shows how likely it is that the outcome of this random signal falls into the range from 1.5 to 2.5.

[Figure B.2 appears here in the experimental instructions.]

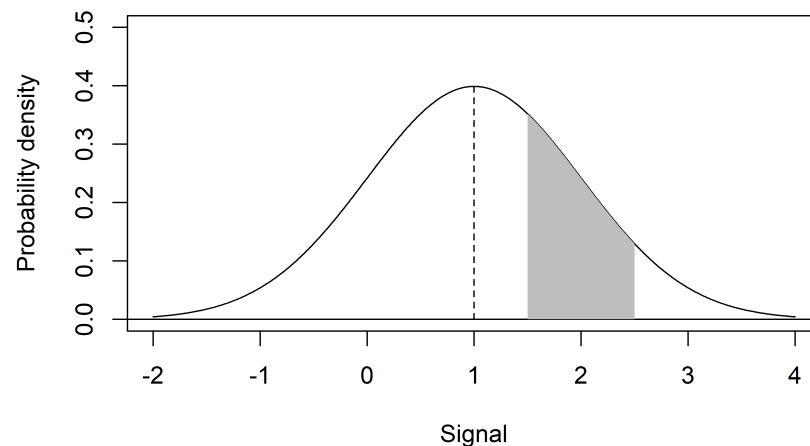


Figure B.2: Density of a normal distribution (not labeled in the experimental instructions).

On your screen, you will always see your signal for the ongoing round, your past decisions in that round, and the past decisions of the players connected to you. The following overview shows the course of events in **each of the 6 rounds**:

- The outcome of the coin flip is determined (but not revealed to any of the players)
- Each player receives a signal.
- Each player enters his/her guess (-1 or 1) for period 1.
- Each player receives the information what players connected to him/her guessed previously. Each player enters his/her guess (-1 or 1) for period 2.
- The previous step is repeated until the guess for period 10 has been entered. Then the round terminates.

B.2 Differences in Instructions in the Treatments with Rewired Networks

The instructions of the treatments with clustered networks, T4-T6, are identical to the instructions reproduced in Section B.1, except for the graph of the network and its description. The graph from treatments T4-T6 is shown in Figure B.3.

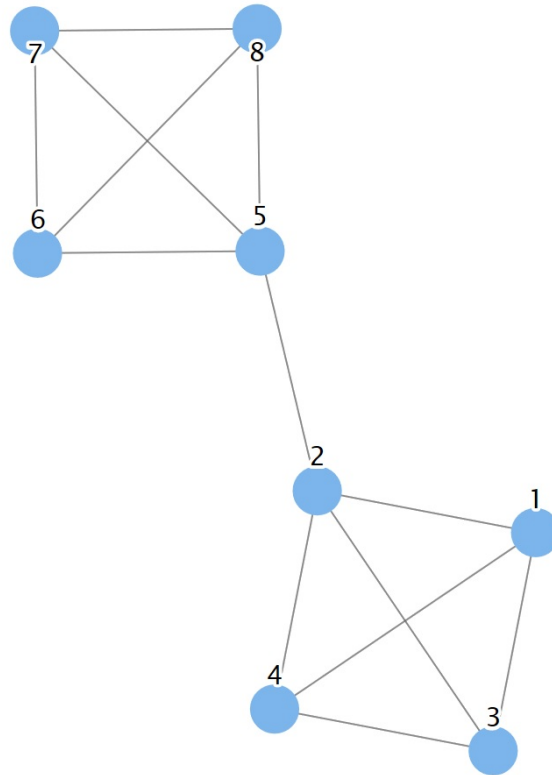


Figure B.3: Clustered network (not labeled in the experimental instructions).

The description of the graph in the instructions of the clustered treatments is as follows: “In this network, player 7 can for example observe the past decisions of players 5, 6, and 8 but not the decisions of other players. As another example, player 2 can observe the past decisions of players 1, 3, 4, and 5.”

C Comprehension Test Questions

Below, we reproduce the comprehension test questions that all participants had to answer correctly before starting the experiments. The questions were the same in all treatments.

We indicate the correct answer with a check mark.

Choose the correct option.

1. At the end of the experiment, one period of each round will be randomly selected and your payment then depends on...

- a) your decisions in these 6 selected periods and the outcomes of the coin flip (to be precise, your payment only depends on whether your guesses in these 6 selected periods were correct). ✓
- b) your decisions and others' decision, not on the outcomes of the coin flips.
- c) luck. Your payment does not depend in any way on your decisions in the experiment.

2.

- a) You can observe the past decisions of all other players in your network.
- b) You can observe the past decisions of the players that you are connected to on the network, but you do not know who can observe your past decisions.
- c) You can observe the past decisions of the players that you are connected to on the network, and exactly those players can observe your past decisions. ✓

3.

- a) The outcome of the coin flip is always the same for all players in the network. However, it can be different from period to period and from round to round.
- b) The outcome of the coin flip is always the same for all players in the network. It is also the same in all periods of a round. However, it can be different from round to round. ✓
- c) The outcome of the coin flip may be different for different players in the network. However, it is the same for each player across periods and rounds.

4. The signal about the outcome of the coin flip is...

- a) on average correct for at least four players in the network (where "on average" refers to the theoretical mean of the probability distribution). This may be any number from four to eight throughout the whole experiment. The number may also vary from round to round, as nothing is written about this in the instructions. ✓
- b) on average correct for at least four players in the network (where "on average" refers to the theoretical mean of the probability distribution). This number is necessarily different from round to round.
- c) never less than -1 and never greater than 1.

5. If a player receives a biased signal,...

- a) the mean of this signal is -1 if the outcome of the coin flip was 1 and 1 if the outcome of the coin flip was -1.
 - b) the mean of this signal is unknown to you, as the instructions do not provide an exact number for it. ✓
 - c) at least four other players' signals are also biased in this round.
6. The instructions contain the following sentence: "Note that the average of the mean of the 8 signals in one group is always closer to the outcome of the coin flip than to the number that did not come up in the coin flip." What does this mean in simple terms?
- a) You can always expect that the average of all signals in a group equals either 1 or -1. Whether you can expect this average to equal 1 or -1 depends on the outcome of the coin flip.
 - b) You can always expect that all of the group's signals considered jointly are informative about the outcome of the coin flip. ✓
 - c) All 8 signals are necessarily between -1 and 1.

D Robustness Check: Regression Analyses of the Experimental Data

As a robustness check to the statistical analysis in Section 7, we present here the results from regression analyses. In a specification without demographic control variables, we estimate the following equation with OLS:

$$ACCURACY_i = \beta_0 + \beta_1 CLUST_i + \beta_2 FAKE_NEWS_i + \beta_3 SPREAD_i + \beta_4 CLUST_i * FAKE_NEWS_i + \beta_5 CLUST_i * SPREAD_i + \varepsilon_i \quad (D.1)$$

The outcome variable *ACCURACY* is the average accuracy in a network (across all periods and rounds, as in the analysis in Section 7). There is thus only one observation per network and observations are statistically independent. *CLUST* is a binary variable taking values 0 (in the treatments with the rewired network) or 1 (in the treatments with the clustered network). *FAKE_NEWS* is 0 in the control treatments, otherwise 1. *SPREAD* equals 1 in the two treatments with spread-out attacks, otherwise 0.

In a specification including demographic control variables, the right side of Equation equation D.1 additionally includes the term $+\beta_6 GENDER + \beta_7 AGE + \beta_8 STUDIES$, where *GENDER*

represents the fraction of females in a network, *AGE* the average age, and *STUDIES* the fraction of students of economics, finance, or business.

We report the coefficient estimates of these two regressions in Table D.1. Table D.2 is similar to Table 3 in the main text. In Table D.2, we reiterate our hypotheses and describe to what sign of sums of regression coefficient these hypotheses lead in the regressions (in the table, we use the variable names rather than the estimated coefficients, because this makes clearer what is meant; e.g., we write *FAKE_NEWS* + *SPREAD* instead of $\beta_2 + \beta_3$). The column “Prediction” states what needs to hold for the sum of these coefficients if the prediction/hypothesis holds. The following column contains the relation for the data/regressions (one column for both regression specifications, because it turns out that these observed directions of treatment effects are always identical for both specifications). Showing the statistical significance of the obtained relationships, the last two columns contain the *p*-values of (one-sided) tests of the relationships, separately for the specifications without and with demographic control variables.

Table D.1: Regression coefficient estimates. This table reports the coefficient estimates from OLS regressions without (left) and with (right) demographic control variables. The outcome variable is the average accuracy in a network across periods and rounds.

	w/o dem. controls	w dem. controls
INTERCEPT	0.675	0.579
CLUST	0.022	0.017
FAKE_NEWS	−0.087	−0.094
SPREAD	−0.038	−0.034
CLUST * FAKE_NEWS	−0.047	−0.041
CLUST * SPREAD	−0.017	−0.012
GENDER		−0.078
AGE		0.007
STUDIES		0.004
R^2	0.413	0.442
N	70	70

The tests concerning the first four hypotheses are strongly significant, the level of significance of the tests of the other hypotheses is a bit lower and depends on the chosen regressions specification. By and large, the regression analysis confirms the findings discussed in the main text.

Table D.2: Comparative statics in predictions (hypotheses) and tests on regressions coefficients

Hypothesis	Regression coefficients	Prediction	Data w/o & w/ dem. controls	<i>p</i> -value w/o dem. controls	<i>p</i> -value w/ dem. controls
Control (left) vs. fake news (right)					
a1. acc(T1) > acc(T2)	FAKE_NEWS + SPREAD	< 0	< 0	0.000	0.000
a2. acc(T1) > acc(T3)	FAKE_NEWS	< 0	< 0	0.007	0.004
a3. acc(T4) > acc(T5)	FAKE_NEWS + SPREAD + CLUST * FAKE_NEWS + CLUST * SPREAD	< 0	< 0	0.000	0.000
a4. acc(T4) > acc(T6)	FAKE_NEWS + CLUST * FAKE_NEWS	< 0	< 0	0.000	0.000
Concentrated (left) vs. spread (right)					
b1. acc(T3) > acc(T2)	SPREAD	< 0	< 0	0.088	0.115
b2. acc(T6) > acc(T5)	SPREAD + CLUST * SPREAD	< 0	< 0	0.026	0.056
Rewired (left) vs. clustered (right)					
c1. acc(T1) > acc(T4)*	CLUST	< 0	> 0	0.709	0.666
c2. acc(T2) > acc (T5)	CLUST + CLUST * FAKE_NEWS + CLUST * SPREAD	< 0	< 0	0.069	0.108
c3. acc(T3) > acc(T6)	CLUST + CLUST * FAKE_NEWS	< 0	< 0	0.190	0.202

This table shows the predictions (i.e., the hypotheses), formulated in terms of the regression specification, jointly with the results of these sums of coefficient estimates in the data and with (one-sided) *p*-values from the regressions (both without and with demographic control variables).

* For this hypothesis, no significant differences were expected ex ante, as discussed in Section 6.3 and the pre-registration.

Swiss Finance Institute

Swiss Finance Institute (SFI) is the national center for fundamental research, doctoral training, knowledge exchange, and continuing education in the fields of banking and finance. SFI's mission is to grow knowledge capital for the Swiss financial marketplace. Created in 2006 as a public-private partnership, SFI is a common initiative of the Swiss finance industry, leading Swiss universities, and the Swiss Confederation.