When is Society Susceptible to Manipulation?

Mohamed Mostagir* Asuman Ozdaglar† James Siderius‡

Abstract

We consider a social learning model where agents learn about an underlying state of the world from individual observations as well as from exchanging information with each other. A principal (e.g. a firm or a government) interferes with the learning process in order to manipulate the beliefs of the agents. By utilizing the same forces that give rise to the "wisdom of the crowd" phenomenon, the principal can get the agents to take an action that is not necessarily optimal for them but is in the principal's best interest. We characterize which networks are susceptible to this kind of manipulation and derive conditions under which a social network is impervious and cannot be manipulated. In the process, we generalize some known centrality measures and describe how our model offers insights into designing networks that are resistant to manipulation.

1 Introduction

In a recent emergency report, the World Health Organization lists "Vaccine Hesitancy" —defined as "the reluctance or refusal to vaccinate despite the availability of vaccines"— as one of the top ten global health threats in 2019.[1] This hesitancy is believed to be one of the main factors behind the resurgence of several health crises, including the recent increase in measles cases by more than 30% worldwide. The reasons for choosing not to vaccinate are varied and complex, but one primary driver is the belief that vaccines are unsafe and have serious adverse effects.

People hold beliefs about all kinds of different issues, e.g. whether a particular vaccine is safe or whether burning fossil fuels contribute to global warming. In these and many other examples, there is usually a ground truth – an underlying state of the world – that agents are trying to learn. In the case of the measles vaccine, the state can be that the vaccine is "safe" or "risky".[2] Agents form

*University of Michigan, Ross Business School
†Massachusetts Institute of Technology
‡Massachusetts Institute of Technology

[1] https://www.who.int/emergencies/ten-threats-to-global-health-in-2019

[2] Vaccines, like any medication, may have side effects, and therefore safety here is understood in the statistical sense. The Center for Disease Control lists several groups who should not be vaccinated, like immuno-compromised individuals or pregnant women. Setting aside these groups and focusing on healthy individuals, the MMR vaccine, for example, has a 1 in a million chance of causing a severe allergic reaction (https://www.cdc.gov/vaccines/hcp/vis/vis-statements/mmr.html), and hence the vaccine is deemed safe enough and approved for use by the Food and Drug Administration. Agents however, do not have to believe that this information is correct, and may have to be convinced (or not) of its accuracy.

beliefs about this underlying state through receiving private signals (for example, by doing their own research on the issue) as well as communicating and exchanging opinions with their neighbors, and a large literature studies conditions under which social learning aggregates beliefs in a way that leads agents to learning the correct state of the world.

In many instances, the beliefs of the agents directly impact their actions. In the example above, an agent would choose to vaccinate if she believes that the state is "safe" and would choose not to vaccinate otherwise, and an aspect that is often ignored in the social learning literature is that there is usually an entity, for e.g. a business, a lobbying group, or a government, that can tamper with the learning process in order to influence these beliefs and steer agents towards a particular action. For example, Broniatowski et al. (2018) provide evidence that Russian bots spread anti-vaccination propaganda online, and Newsweek magazine reports that "most of the new measles cases are in Eastern European and Central Asian countries frequently targeted by Russian disinformation."[3] Similarly, a recent episode of the show Planet Money reports how firms like Cambridge Analytica selectively pushes certain stories and not others in order to "create a fake view of the world with real stories", i.e. the content itself does not even have to be false; it is enough for it to be biased enough in order to influence the beliefs of the receiver.[4] Less malicious examples exist of course – a firm may simply try to influence the beliefs of consumers in order to make them buy a product, or a public health campaign may try to convince the population to adopt certain hygiene practices that can be useful in reducing the risk of communicable diseases.

Building on the above, we consider a social learning environment where a *principal* tries to manipulate the learning process of the agents. Agents in our paper are heterogeneous on multiple dimensions. In addition to their different network locations and how well-connected they are, they can also vary in how they interpret their own signals and how they use the information they obtained from their friends or colleagues to update their opinions. Some agents may choose to aggregate the opinions of their peers without conducting more thorough research or without considering how these opinions were reached. Others may be more discerning, choosing instead to try and determine how a peer reached a particular conclusion before blindly incorporating it into their own opinion. This heterogeneity allows us to study manipulation in the context of the two most common social learning models in the literature – Bayesian and DeGroot learning. Importantly, the recent experimental and empirical work of Chandrasekhar et al. (2015) shows that societies are indeed composed of a mixture of Bayesian and DeGroot learning types, and that the proportion of types can be different from one

[3] https://www.newsweek.com/russian-trolls-promoted-anti-vaccination-propaganda-measles-outbreak-1332016
[4] https://www.npr.org/2019/05/24/726536757/episode-915-how-to-meddle-in-an-election

society to the next. As shown in that work, about 10% of the sample of Indian villagers considered in the paper behave in a way that is consistent with Bayesian updating, while the remaining agents behave in a DeGroot fashion. In contrast, the proportion of Bayesian to DeGroot agents is roughly equal in the sample of college students studied in the paper. Our model thus captures a realistic aspect of social networks by incorporating this learning diversity and —as we show later— demonstrates how the proportions of learning types in a population, among other factors, determine whether a society is susceptible to manipulation.

To summarize, this paper builds an opinion dynamics model with the following three components. First, opinions are formed as a result of both individual and social learning. Second, agents are heterogeneous in how they incorporate their peers' opinions into their own beliefs. In particular, they have varying levels of sophistication in how they treat these opinions. Third, there is a strategic principal who can utilize the social aspect of opinion formation in order to manipulate the agents' beliefs to his benefit. These three aspects combine to give a novel model that provides analytical insights into how beliefs spread in these heterogeneous environments, as well as practical implications to the design of these networks in order to make them impervious to manipulation.

Contribution and Overview of Results. The primary contribution of this paper is to examine a rich mixed-learning environment where the learning process of the agents is manipulated by a strategic principal. With few exceptions, previous literature has traditionally eschewed such heterogeneity and considered information aggregation by either DeGroot agents or Bayesian agents. More importantly, none of that literature considers the case where a principal tries to influence the learning process. Our model combines Bayesian agents and a more general formulation of DeGroot agents to answer the following questions: 1. Can a strategic principal consistently manipulate the beliefs of some agents in the network in order to make them take certain actions? And 2. What are the driving factors that make some networks amenable to such manipulation while other networks are more resistant?

We answer the above questions by providing a classification of networks that describes when such manipulation is possible. In our model, agents try to learn the true state of the world in order to make a one-time choice between different actions. In the example given earlier, the possible actions are **vaccinate** or **not vaccinate** and the state can be whether the vaccine is **safe** or **risky**. Agents receive signals about the underlying state – for example, they might read news stories or examine research articles about vaccination– and they use these signals in addition to the information they obtain from their neighbors to update their beliefs and eventually uncover the state. The principal has an unknown type: he can either be truthful or strategic. A truthful principal does not interfere

with the learning process, but a strategic principal can choose to send costly signals to the agents. These signals do not have to be tied to the state and can be intentionally misleading. Agents do not know the type of the principal and cannot differentiate whether a signal they are receiving is *organic* or coming from a strategic principal.

Agents try to take an action that matches the state, so in the example above they would like to choose **vaccinate** if the state is **safe** or **not vaccinate** if the state is **risky**. The principal is interested in having the agents take a specific action, for example, the action **not vaccinate**, regardless of what the state actually is. We say that an agent is manipulated if her beliefs converge to the true state and she takes the correct action when the principal is the truthful type, but chooses the wrong action due to incorrect beliefs when the principal is the strategic type (this corresponds, in this example, to taking the action **not vaccinate** when the state is **safe**). These dynamic environments often admit a multiplicity of equilibria, which complicates their analysis. We first provide a few technical results that show that for a long-enough horizon, an equilibrium always exists and is essentially unique, in the sense that the degree of manipulation in society does not depend on which equilibrium is selected. We then use these results to show in Theorem 1 that Bayesian agents are never manipulated, but that depending on parameters related to the network structure and how agents weigh their own signals, a substantial fraction of DeGroot agents can be tricked into believing that the underlying state is different from the actual state. Proposition 3 shows that under mild conditions, extreme societies that are inclined towards herding (agents discount their own signals and put their faith in what other agents think) *or* towards individuality and narcissism (agents discount everything except their own signals) are basically impossible to manipulate. On the other hand, a well-tempered society whose members use their own beliefs as well as other agents' opinions is the society that is most prone to this kind of manipulation.

For these well-tempered societies, the Bayesian agents can help spread the truth about the underlying state, but their ability to do so is limited by the network structure. We provide a characterization of which network topologies are manipulable in terms of a centrality measure that we call DeGroot Centrality, and we use this measure to classify networks into dense and sparse topologies. Theorem 2 shows that dense networks are highly resistant to manipulation: even as the size of the network grows, the presence of a *constant* number of Bayesian agents *anywhere* in the network is enough to guarantee imperviousness. On the other hand, sparse networks are more susceptible to manipulation, and both the number of the Bayesian agents as well as where these agents are located are important for the network to be impervious. In particular, the number of Bayesians required may grow with the size of the network. If there are not enough Bayesians, or if there is a sufficient number of Bayesians but

they are not well-located, then the principal can manipulate almost the entire population by targeting only a fraction of the agents, i.e. it becomes cheaper and easier for the principal to manipulate.

Finally, we apply our results to the network topologies commonly studied in the literature and use DeGroot centrality and the dense/sparse classification to determine which of these topologies are easier to manipulate. In an effort to bring our results closer to real-world networks, we further apply our results to data from an advice network in an Indian village, obtained from Jackson et al. (2012). The data provides an actual network topology from the village but no information about which agents are Bayesian. We analyze different scenarios of Bayesian placements in this network to highlight the concepts introduced in the paper. Ultimately, we believe that the work in Chandrasekhar et al. (2015) —which identifies which agents learn in a Bayesian vs. DeGroot fashion— and the methodological approach introduced in this paper jointly provide a complete framework for studying manipulation in these heterogeneous real-world networks.

Related Literature. Our model combines both DeGroot and Bayesian agents. DeGroot learning has been extensively studied in several literatures. For example, Golub and Jackson (2010) give conditions under which beliefs converge to the true state of the world. There is also a rich literature (e.g. Acemoglu et al. (2011) and Bikhchandani et al. (1992)) that looks at when agents who learn in a Bayesian fashion can correctly aggregate information. Others, such as Jadbabaie et al. (2012), consider agents that are somewhere between DeGroot and Bayesian agents in how they update their beliefs about the state of the world, and their particular formulation of DeGroot agents is the one we consider in this paper. Some recent work looks at a mixed learning environment. Mueller-Frank (see Mueller-Frank (2014)) examines how a network of DeGroot agents and a single Bayesian agent aggregates information, and Chandrasekhar et al. (2015) experimentally examine learning in an environment where some agents are designated as Bayesian and others are not. One major differentiating factor of our work compared to this literature is the presence of a principal who can intentionally confound learning, and we examine the conditions under which this may or may not be possible.

The Bayesian Persuasion literature initiated by Kamenica and Gentzkow (2011) considers a principal who sends messages to agents in order to make them take a certain action. In the standard setup, everyone is strategic, there is no state uncertainty or learning from the environment, there is no notion of organic and strategic messages, and most importantly, there is no ambiguity over the type of the principal. In our paper, agents do not know the type of the principal and cannot tell whether the signals they receive originate from a strategic principal or are more organic. This uncertainty about the principal's type relates our work to that of Morris (2001) and more generally, to the litera-

ture on reputation formation, which started with the work of Kreps and Wilson (1982) and Milgrom and Roberts (1982). This literature considers short-lived Bayesian agents that interact sequentially with a principal. In contrast, our paper examines a setup where there is a principal interacting simultaneously with a collection of agents who are connected on a social network and who update their beliefs through the signals they receive as well as from the social interactions amongst themselves.

As we mentioned earlier, our first result shows that the Bayesian agents in our model eventually figure out the true state of the world. Once this happens, they become somewhat similar to stubborn agents, in the sense that their (correct) opinion about the state remains unchanged. Opinion dynamics with stubborn agents have been studied in Acemoglu et al. (2013) and Yildiz et al. (2013) among others. The primary differences between our work and these papers is the presence of a strategic principal, which fundamentally changes the role that these stubborn agents play. In the cited literature, the presence of stubborn agents leads to divergence of opinions and generally hinders learning about the true state of the world. In contrast, the learning difficulty in our model comes from the strategic principal who tries to manipulate the agents, and in that sense the presence of stubborn agents who realize the principal's type is always useful for everyone in the network, i.e unlike the work above, the stubborn agents can only help society discover the true state of the world. Nevertheless, as we discuss, even with the positive contribution that these agents provide to the learning process, manipulation might still be unavoidable.

The recent proliferation of false news on social networks, while not a primary focus of our paper, provides a current application of our work. Recent theoretical work in Candogan and Drakopoulos (2017) and Papanastasiou (2018) examines how (Bayesian) agents exchange information on a social network and shows how misinformation can spread in these models and what the platform (over which the agents are communicating) can do about it. The existence of fake news in these models is exogenous, i.e. unlike our model, there is no principal or news provider that strategically injects such misinformation into the network, and consequently there is no notion of manipulation. In addition, we examine a mixed learning environment with varying degrees of sophistication, which, as Pennycook and Rand (2018) show in recent experimental work, might be one of the primary reasons why misinformation propagates in social networks.

2 Model

We first provide an informal description of how agents learn in our model. Agents continuously receive news about a specific topic, for example by scrolling through the stories that appear in their

news feed. In the absence of interference from the principal, the news that agents receive is *organic*, and, together with communicating with other agents, is enough for them to update their beliefs and figure out the state of the world correctly. The principal may however interfere with the news generation process for some of the agents, so that these agents see both organic and fake stories as they scroll through their feed. Agents cannot differentiate which stories are correct and which are not, and so they update their beliefs using both types of stories. As mentioned in the introduction, the stories that the principal provides do not even have to be false, but can simply be correct stories that are curated in a way that leaves a specific impression. For simplicity however, we will refer to the stories that the principal provides as fake news. Once enough time has elapsed and agents have learned the state of the world, they take an action based on their belief of what the state is.

2.1 Formal Model

We consider a directed social network with n agents trying to learn a binary state of the world $y \in \{S, R\}$ over time. Time is continuous and agents learn over a finite horizon, $t \in [0, T)$. At time $t = 0$, the underlying state $y \in \{S, R\}$ is drawn, with $\mathbb{P}(y = S) = q \in (0, 1)$.

Organic News News is generated according to a Poisson process with unknown parameter $\lambda_i > 0$ for each agent i; for simplicity, we assume that λ_i has atomless support over $(\underline{\lambda}, \infty)$ with $\underline{\lambda} > 0$. We refer to this process as *organic news*. Let us denote by $(t_1^{(i)}, t_2^{(i)}, \ldots)$ the times at which news occurs for agent i. For all $\tau \in \{1, 2, \ldots\}$, the organic news for agent i generates a signal $s_{t_\tau^{(i)}} \in \{S, R\}$ according to the distribution:

$$\mathbb{P}\left(s_{t_\tau^{(i)}} = S \middle| y = S\right) = \mathbb{P}\left(s_{t_\tau^{(i)}} = R \middle| y = R\right) = p_i \in [1/2, 1)$$

i.e., the signal is correlated with the underlying truth. The value of p_i indicates the richness of agent i's signal, and can be interpreted as her ability to deduce the true state from the facts presented in the organic news. We assume that p_i may be equal to $1/2$, in which case the organic news serves only as noise for agent i, who cannot infer the true state simply from this news.

Principal In addition to the organic news process, there is a principal who may also generate news of his own. At $t = 0$, the principal picks an influence state $\hat{y} \in \{R, S\}$. This is the state that the principal would like agents to believe, regardless of what the true state actually is. The principal then picks an influence strategy $x_i \in \{0, 1\}$ for each agent i in the network. The influence state $\hat{y}$ corresponds to the signal the principal sends to (some) agents, and the influence strategy indicates which agents the principal wants to send the signal to. If the principal chooses $x_i = 1$ for any agent i, then he

(principal) generates news according to an independent Poisson process with intensity λ_i^* which is received by all agents where $x_i = 1$. We assume the principal commits to sending signals at this intensity, which may not exceed some (exogenous) threshold $\overline{\lambda}$.[5] We denote by $\hat{t}_1^{(i)}, \hat{t}_2^{(i)}, \ldots$ the arrival times of *all* news, either from organic sources or from the principal, for agent i. At each time $\hat{t}_\tau^{(i)}$, if the news is organic, the agent gets a signal according to the above distribution, whereas if the news is sent from the principal, she gets a signal of $\hat{y}$. The principal incurs an upfront investment cost $\varepsilon > 0$ for each agent with $x_i = 1$.

The principal can be one of two types. He can either be a strategic type $\mathcal{S}$ or a truthful type $\mathcal{T}$. The type of the principal, which is denoted by ω, is drawn at $t = 0$ with $\mathbb{P}(\omega = \mathcal{T}) = \mu_0 \in (0, 1)$ and does not change over time. If the principal's type is $\omega = \mathcal{T}$, we assume he is committed to implementing $x_i = 0$ for all agents; that is, he does not interfere with the learning process. On the other hand, the $\omega = \mathcal{S}$ type of the principal may play any influence strategy $\mathbf{x} \equiv \{x_i\}_{i=1}^n$ over the network (and may randomize over network strategies). Specifically, he may choose $x_i = 1$ for some agent i, with influence state $\hat{y} \neq y$, to spread misinformation. The uncertainty of the principal's type generates uncertainty for agent i about the true nature of her signal distribution.

Agents Agents have different degrees of sophistication. We think of these sophistication levels as separate from whether the agent has skill in distinguishing the state y from the news alone (i.e., her p_i or λ_i). Specifically, sophistication in our model refers to how an agent uses the beliefs in her network to form her own belief about the state. Each agent is either Bayesian (B) or DeGroot (D), and the sophistication type of each agent is common knowledge and consistent across time. DeGroot agents differ from Bayesians in that DeGroot agent i:

(a) Uses a simple learning heuristic to update beliefs about the underlying state from other agents.

(b) Believes all signals arrive according to a Poisson process and all signals are independent over time with $\mathbb{P}\left(s_{i,\hat{t}_\tau} = y\right) = p_i$ (i.e., takes the news at face value).

Each agent perfectly observes her signals but does not observe the signals received by any other agent. All agents have perfect recall. We let $\mathcal{H}_{i,t}$ denote the set of possible private histories of signals at agent i up until time t, and $h_{i,t} \in \mathcal{H}_{i,t}$ a particular history realization. Let $\pi_{i,t} \in \Delta(\{R, S\})$ represent the belief of agent i about the underlying state at time t.

DeGroot Update: DeGroot agents form their opinions about the state both through their own experience (i.e. the signals they receive) and by talking to their neighbors. Given history $h_{i,t} =$

[5] One can interpret $\overline{\lambda}$ as the maximum capacity that the principal can send his messages. The principal may elect $\lambda_i^* < \overline{\lambda}$ if $\overline{\lambda}$ when is large because choosing $\lambda_i^* = \overline{\lambda}$ would make the evidence of $\hat{y}$ so overwhelming that the agent would realize $x_i = 1$ (i.e., in other words, the bias in agent i's signals would become obvious to her).

$\left(s_{i,\hat{t}_1^{(i)}}, s_{i,\hat{t}_2^{(i)}}, \ldots, s_{i,\hat{t}_{\tau_i}^{(i)}}\right)$ up until time t with $\tau_i = \max\{\tau : \hat{t}_\tau^{(i)} \leq t\}$, each agent forms a personal belief about the state according to Bayes' rule. Let $z_{i,t}^S$ and $z_{i,t}^R$ denote the number of S and R signals, respectively, that agent i received by time t; then the DeGroot agent has direct "personal experience":

$$g_{i,t}(S|h_{i,t}) = \frac{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R}q}{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R}q + p_i^{z_{i,t}^R}(1-p_i)^{z_{i,t}^S}(1-q)}$$

and $g_{i,t}(R|h_{i,t}) = 1 - g_{i,t}(S|h_{i,t})$. The experience function $g_{i,t}$ represents the direct contribution of the observed signals into agent i's belief, and is related to the personal Bayesian update in Jadbabaie et al. (2012). It is the belief any fully Bayesian agent would hold about the state y *in isolation* and without principal interference. DeGroot agents also form beliefs by talking to their neighbors every time interval of length $\Delta > 0$ small.[6] For all agents i, there are weights θ_i, α_{ij} such that agent i holds belief $\pi_{i,t}$ for all $k\Delta < t \leq (k+1)\Delta$ according to:

$$\pi_{i,t} = \theta_i g_{i,t}(h_{i,t}) + \sum_{i=1}^{n} \alpha_{ij}\pi_{j,k\Delta}$$

for all $k \in \mathbb{N}$, where $\theta_i + \sum_{j=1}^n \alpha_{ij} = 1$ (we have suppressed dependence on y). As convention, we assume the link $i \to j$ suggests that i listens to j. We refer to this as the *DeGroot update* (DU) process.

Bayesian Update: We assume it is common knowledge for Bayesian agents that there are n agents arranged in a given social network $\mathbf{G}$, with signal structures $\{p_i\}_{i=1}^n$. Furthermore, agent i observes the history of beliefs in her neighborhood $\mathcal{N}_i$, given by $\Pi_{i,t} = \bigtimes_{t'=0}^t \bigtimes_{j\in\mathcal{N}_i} \pi_{j,t'}$. Given the private history of signals and history of neighborhood beliefs, the belief map ϕ_t at time t of a Bayesian agent is of the form:

$$\phi_t : (h_{i,t}, \Pi_{i,t}) \mapsto \pi_{i,t+dt}$$

and pinned down by Bayes' rule. We will say the Bayesian is *truthful* if she reports belief $\pi_{i,t}$ to all agents in her out-neighborhood is the belief given by $\phi_t(h_{i,t}, \Pi_{i,t})$. We will assume throughout this paper that all Bayesian agents are truthful.[7] Notice that Bayesian agents may be oblivious (i.e., receiving no signals at all about the state), in which case they have to rely on the network to learn what the state of the world is.

At the same time, Bayesians hold (private) beliefs about the type of the principal (and whether

[6] In particular, we assume Δ is arbitrarily small so the probability that any agent has two signals within an interval of length Δ is close to zero.

[7] This is contrast to previous papers (such as Rosenberg et al. (2009)) where Bayesian agents may experiment with reporting false beliefs to better learn about the information of other agents in the network. In light of Theorem 1, when T is large, Bayesians learn the correct state, so even if Bayesians strategically report beliefs in the network, reporting truthfully is a best-response to other Bayesians reporting truthfully as well.

		Agent **R**	Agent **S**
State y **R**	**R**	$1, 1+b$	$0, 0$
	S	$1, b$	$0, 1$

Table 1. Terminal Game.

signals are corrupted by the principal's influence). We will denote the belief (that the principal is truthful type) of a Bayesian i about the principal's type at time t as $\mu_{i,t}$, which is unobservable to other agents in the network, including i's neighbors. Such beliefs are updated using Bayes' rule whenever possible, as in a perfect Bayesian equilibrium (see Fudenberg and Tirole (1991)).

Payoffs At time $t = T$, each agent chooses an action $a_i \in \{S, R\}$.[8] Payoffs for the principal and agent are given in Table 1. The first entry in a cell is the principal's payoff while the second is the agent's payoff (so for example, the top-left cell corresponds to the case when the state is R and the agent chooses action R. This gives the principal a payoff of 1 and the agent a payoff of $(1 + b)$).

We assume that $b \in (-1, 1)$ so that agent i would match its action a_i with the state y if it were known with certainty. Otherwise, the parameter b captures any asymmetry in the payoffs between the two states.[9] Note that, on the other hand, the principal always prefers agents take action R instead of action S, and so has an incentive to convince agents of $y = R$ even when $y = S$. Let $u_i(y, a_i)$ denote the payoff of agent i when the state is y and she takes action a_i; $u_i^p(a_i)$ is the payoff for the principal at agent i (and only depends on that agent's action). The total payoff for the principal is given by $u^p(\mathbf{a}) = \sum_{i=1}^n u_i^p(a_i)$, which is the summation of the payoffs from period-T actions of all n agents (where $\mathbf{a} \equiv \{a_i\}_{i=1}^n$). We denote by $c(\mathbf{x}) = \sum_{i=1}^n \varepsilon \mathbf{1}_{x_i=1}$ the cost of the principal for implementing the network influence strategy $(\hat{y}, \mathbf{x})$ at $t = 0$

Each agent chooses a mixed strategy σ_i mapping terminal beliefs, $\pi_{i,T}$, to a distribution over actions, $\Delta(\{S, R\})$. Similarly, the principal chooses a mixed network influence strategy σ^p mapping his type ω and the current state y to a distribution over network influence, $\Delta(\hat{y}, \mathbf{x})$, with the restriction that the truthful principal type $\omega = \mathcal{T}$ always plays a pure network-influence strategy of $\mathbf{x} = \mathbf{0}$, i.e. does not interfere with the organic signals. We assume that the principal has total payoff given by the difference between her future utility (via the actions of the agents) and the cost of the network influence, $u^p(\mathbf{a}) - c(\mathbf{x})$.

[8] The example given in the introduction can be modeled using this payoff table as follows: the states of nature S and R can be mapped to whether a vaccine is safe (state S) or risky (state R). Similarly, the actions can be thought of as analogous to the "vaccinate" (action S) and "not vaccinate" (action R) actions. In this sense, a player wants to match her action to the state, e.g. taking action S when the state is S indicates vaccinating when the vaccine is safe.

[9] For instance, it may be more costly to vaccinate your child if vaccines do have averse effects than it is to not vaccinate even if they are safe.

3 Equilibrium and Learning

In this section, we present a brief summary of the solution concept and the learning dynamics which follow. These results are completely technical, and so we elected to delegate them to the appendix in order to preserve the flow of the paper and focus on the structural results. We refer the reader to Appendix A and Appendix B, respectively, for a more formal treatment.

Equilibrium We informally describe our equilibrium concept and provide some basic results. The relevant details are given in Appendix A. By definition, DeGroot agents update beliefs mechanically and simply take all news received at face value. At time $t = T$, each DeGroot agent chooses an action which maximizes her payoff given her belief about the state. All of this is common knowledge to both the Bayesian agents and the principal. In addition, Bayesian agents observe their neighbors' beliefs over time, know the network structure, and know the principal's type is drawn at $t = 0$ such that he is truthful ($\omega = \mathcal{T}$) with probability μ_0 and strategic ($\omega = \mathcal{S}$) with probability $1-\mu_0$. The principal and the Bayesian agents play a *perfect Bayesian equilibrium*, i.e. the Bayesians update beliefs about the type of the principal and the underlying state y simultaneously, taking as given the strategy of the strategic principal in equilibrium. Then, at $t = T$, each Bayesian agent chooses an action which maximizes her payoff given her belief about the state. Similarly, the principal chooses his network influence strategy to maximize his payoff taking as given how agents learn and ultimately select terminal actions. In all of this, we require in equilibrium that strategies in fact be best-responses for both the principal and the Bayesians, as standard.

We will say that an agent is *manipulated* if she learns the correct state (i.e., takes the correct action) when the principal is truthful, but takes the incorrect action when he is strategic. That is, the principal's interference successfully tricks some agent into taking a suboptimal action she *would not* have taken without the interference. Our first main result, presented in Appendix A, is that an equilibrium always exists, which does not follow immediately from standard existence results. Second, we show that as the learning horizon becomes long (i.e., $T \to \infty$), for almost all[10] parameters given in the problem, the number of manipulated agents is the same under any equilibrium, almost surely. This allows us to refer to the "number of manipulated agents" in equilibrium without ambiguity, even if the identity of those agents may be different under different equilibria. Throughout the paper, we will refer to this property as *essential uniqueness*.

Learning We provide a full characterization of limit beliefs as $T \to \infty$ in Appendix B. Based on that

[10] One can interpret "almost all" as meaning that if some parameters (ε, b) generate multiple equilibria, then by perturbing one or both of them by some small amount, the equilibrium becomes (essentially) unique.

characterization, we prove the following result.

Theorem 1. *Under standard connectivity and organic signal distribution assumptions (detailed in Assumptions 1 and 2(c) in the appendix), no Bayesian agent is manipulated almost surely as* $T \to \infty$.

Theorem 1 states that Bayesian agents are never manipulated, so in equilibrium any attempts to thwart learning are eventually detected. By knowing the network structure, Bayesian agents can infer the signal distributions of others in the network, which in turn informs them of whether the principal is attempting to manipulate *anyone*. This implies that for large enough T^*, for all $T > T^*$, Bayesian agents can be treated as *stubborn agents* who hold strong beliefs about the true state.

Since DeGroots operate mechanically, this allows us to characterize their beliefs as $T \to \infty$ as in standard in the social learning literature. Let us define $\boldsymbol{\gamma}$ as the limiting personal experience vector given by:

$$\boldsymbol{\gamma} = \begin{pmatrix} \mathbf{0}_B \\ \mathbf{x}_D \end{pmatrix}$$

where we have implicitly assumed the first m agents are Bayesian, without loss of generality, and where the subscripts denote the vector for those type of agent. In other words, if an agent is DeGroot and receives fake signals from the principal (i.e., $x_i = 1$), then we write $\gamma_i = 1$ and otherwise we write $\gamma_i = 0$. We also replace all of the limit beliefs of Bayesian agents by a point-mass on the true state i.e., zero belief on $\hat{y}$). Then for a suitable influence matrix $\mathbf{A}$, we can represent the limit-beliefs using the familiar Leontif inverse form:

$$\boldsymbol{\pi} \to (\mathbf{I} - \mathbf{A})^{-1}(\boldsymbol{\gamma} \otimes \boldsymbol{\theta}) \tag{1}$$

This provides a closed-form expression for the beliefs of the agents for large T. We note that this expression depends on the network structure, sophistication of the agents, personal-experience weight, and the network action $\mathbf{x}$ of the principal (captured through $\boldsymbol{\gamma}$).

4 Manipulation and Network Topology

In this section, we consider fixed networks of size n with m Bayesian agents. We use the term "network structure" to collectively refer to the neighborhoods of the Bayesian agents $\{\mathcal{N}_i\}_{i\in\mathcal{B}}$, and the DeGroot influence matrix and personal-experience weight vector, $(\mathbf{A}, \boldsymbol{\theta})$. We address the central question of our paper: when is a population susceptible to manipulation? Towards this, we make the following definition:

Definition 1. A network is *impervious* to manipulation if no agents are manipulated (in equilibrium); otherwise it is *susceptible*.

Our first step is to define the principal's optimization problem of choosing his (mixed) network strategy when T is large. We present these findings in Appendix C. Generally, this optimization problem is computationally intractable, but we can still provide a characterization of manipulation in terms of a novel centrality measure that we call *DeGroot Centrality*. This measure captures how the network structure propagates the principal's injected fake signals to any specific agent. Loosely, it corresponds to how much influence other DeGroots (who receive these signals) have on agent i's own belief. Under appropriate normalization, this centrality measure is exactly equal to an agent's belief of the incorrect state given in Equation (1). Despite the equivalence, we find that the interpretation of beliefs in terms of centrality to be meaningful for our results, and so will reference it when appropriate. We refer the reader to Appendix C for a more complete discussion of the network-relevant details.

Using the concept of DeGroot centrality, we show that dense networks (in a sense to be made precise) are always impervious to manipulation as long as there is a *constant* number of Bayesian agents located *anywhere* in the network. That is, the number of Bayesians needed does not scale as n gets large and the agents need not have particular network positions. On the contrary, when network is sparse, it can be the case that anything less than a *linear* number of Bayesians will lead to manipulation, and moreover these Bayesians must be situated in specific network locations for manipulation not to happen. Finally, we use the personal-experience weights $\boldsymbol{\theta}$ as a proxy for societal norms, and provide comparative statics on how these norms affect manipulation.

4.1 Dense Networks

For compactness, let us represent the network $\mathbf{G}$ as the concatenation of the Bayesian adjacency matrix and the DeGroot influence matrix $\mathbf{A}$, given by the (i,j)-elements:

$$\mathcal{A}_{ij} = \begin{cases} \alpha_{ij}, \text{ if } i \in \mathcal{D} \\ 1, \text{ if } i \in \mathcal{B} \text{ and } j \in \mathcal{N}_i \end{cases}$$

In other words, let $\mathbf{G}$ be a directed, weighted network where the weight w_{ij} of the link from $i \to j$ is equal to 1 if either i is a Bayesian and j is in i's neighborhood, and otherwise it is equal to α_{ij}. We can define a *walk*, W_{ij}, between agent i and agent j to be a sequence of arcs, $i \to u_1 \to u_2 \to \cdots \to u_n \to j$, starting with i and ending with j. We let $\mathcal{W}_{ij}$ be the (countable) set of all walks between agents i and j in $\mathbf{G}$ of any length. Finally, define the *log-diameter* of the network $\mathbf{G}$ to be:

$$d_{\mathbf{G}} \equiv \max_{i,j} \min_{W_{ij} \in \mathcal{W}_{ij}} \sum_{(k \to \ell) \in W_{ij}} -\log(w_{k\ell})$$

Using this, we can define the *density* of a network as follows:

Definition 2 (Dense Networks)**.** We say that network $\mathbf{G}$ is δ-*dense* if has a log-diameter of at most $\log(n+\delta)$.

Theorem 2 (Constant Bayesians)**.** *For every δ, there exists a universal constant $m^*(\delta)$ such that every network $\mathbf{G}$ which is δ-dense and contains at least $m^*(\delta)$ Bayesians is impervious to manipulation.*

We make a few comments about Theorem 2. First, the number of Bayesians needed to make the network impervious is constant and does not scale as n gets large, as long as the network diameter does not grow too quickly with n. This implies that even with a vanishingly small fraction of Bayesians in the population, the principal will be unable to manipulate beliefs. Second, the location of the Bayesians is irrelevant for the result to hold. Even if an adversary chooses the network position of the Bayesians, only m are needed in any network of any size to make it impervious. Finally, we note that just because the shortest path between an agent i and every Bayesian is less than $\log(n+\delta)$ does not imply agent i will not be manipulated. This needs to hold *uniformly* across all DeGroot agents. One can easily construct an example where a DeGroot is close to all the Bayesians, but because she talks to other DeGroots who only talk to each other, echo chambers drive beliefs away from the truth. In that case, the network does not satisfy the small log-diameter condition.

We also point out that Theorem 2 should be viewed as a *worst-case* bound for imperviousness. First, the result should not be interpreted as the location of Bayesians in the network does not matter. If the Bayesians are in better network positions, it may be the case that even for $m \ll m^*(\delta)$, a given δ-dense network is impervious. Second, the bound does not suggest that the worst-case number of Bayesians needed for imperviousness is monotone in log-diameter. In other words, if network $\mathbf{G}'$ has a bigger log-diameter than $\mathbf{G}$ (for the same n), this does not imply that $\mathbf{G}'$ requires more Bayesians than $\mathbf{G}$ to avoid manipulation, *even if* the Bayesians are chosen in a worst-case way. Rather, all it guarantees is that if the number of Bayesians meets the threshold $m^*(\delta)$ in a δ-dense network, there will never be manipulation. We perform a numerical study of how Bayesian placement and number affect manipulation in Section 5 in an Indian social network.

We conclude this section by briefly mentioning a couple of examples of interest where one can easily apply the result.

Example 1 (Complete Network)**.** Consider the complete network on n vertices. We suppose that, for simplicity, $\theta_i = \alpha_{ij} = 1/(n+1)$ for all DeGroot agents i and agents j (of any kind). This corresponds to each agent weighing each source of opinion (each neighbor, plus their own news) equally. The log-diameter of this network is no exactly $\log(n+1)$ for any $n \geq 2$. Therefore, only a constant number

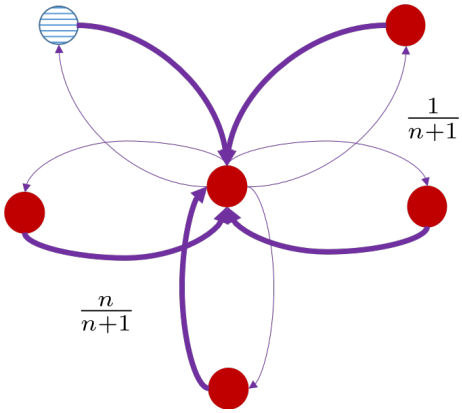

Figure 1. Influential Star Network. A weighted directed arrow from node i to node j indicates that i puts that much weight on j's belief. Shaded node represents a Bayesian agent.

of Bayesian agents are needed by Theorem 2 (applying the result for $\delta = 1$), and in particular, one can show that $m \geq (1+b)/(1-b)$ are required for the complete network of size n. □

Example 2 (Influential Star Network)**.** Consider Figure 1 which shows one type of star network. We suppose that, for simplicity, $\theta_i = 1/(n+1)$ for all agents; that is, each agent weighs its own news as if it were in the complete network. Let agent 1 be the central agent of the star and agents $\{2,\ldots,n\}$ be on the periphery. For agent $i \in \{2,\ldots,n\}$, we have $\alpha_{i1} = n/(n+1)$ and $\alpha_{ij} = 0$ for all other j. For agent 1, we have $\alpha_{1j} = 1/(n+1)$ for all agents j. In other words, the central agent is *highly influential*, as all peripheral agents are influenced much more by this agent than their own news.

Once again, for any $n \geq 2$, the log-diameter of the network is at most $\log(n+3)$; between any two agents on the periphery, we have $\log((n+1)^2/n) = \log(n+2+1/n) \leq \log(n+3)$. In fact, if the number of Bayesians satisfies $m \geq 2(1+b)/(1-b)$, the network is impervious. This is true even when all of the Bayesians are on the periphery. So, in a seemingly very asymmetric network, still only a constant number are needed. This does not imply, however, that fewer Bayesians would make the network susceptible. For instance, in this example, a single Bayesian in the center of the star *always* makes the network impervious when n is large enough. □

We briefly mention a counterexample for a network that has a log-diameter that grows faster than $\log(n+\delta)$ for any constant δ, *despite having short paths between any two agents*. Consider some undirected network $\mathbf{T}$, and let $\mathbf{G}$ be the corresponding weighted network where all DeGroots take $\theta = 1/n$ and place equal weight on all of its neighbors (and 0 elsewhere). If $\mathbf{T}$ has a constant diameter, does $\mathbf{G}$ satisfy the small log-diameter condition to be considered dense? In general, no. As an extreme example, consider two cliques of size $n/2$, and a single connection between them. This network has diameter 3 for all n. However, the log-diameter is $\log\left(\frac{n^4}{2(n-1)^2}\right) \approx \log(n^2)$. Thus, the small

log-diameter condition is helpful for showing that some *asymmetric networks* are impervious, as in the star, but not networks of extreme homophily, despite having small "undirected diameter."

4.2 Susceptible Networks

We now consider networks that are *sparse*, in the sense that they have a large log-diameter. For this, consider the directed ring network as a prototypical example with $\theta_i = \theta_i^{(n)}$ (any function of population size n) for all DeGroots i, and $\alpha_{ij} = 1 - \theta_n$ for $j = i - 1$, and $\alpha_{ij} = 0$ for all other j. Under this assumption, each DeGroot listens to her own news and the opinion of one other agent, who in turn listens to only one other agent, and so on. For now, let us assume that m Bayesian agents form a continuous chain in the ring. This would arise in a setting where Bayesians only talk to each other; for example, a subpopulation of educated students who have little interaction with students who are less educated. For concreteness, in the section will assume $\theta_i = 1/(n+1)$ for all DeGroot i and that $m \ll n$,[11] but (as we show) the results are applicable across a wide range of personal-experience weights $\theta_i^{(n)}$.

The following is an illustration that shows that the principal can manipulate in the above setup. Consider a heuristic optimization problem where the principal maximizes only along a single dimension. We make the following restriction on the heuristic problem: (i) the principal can only influence a continuous arc in the ring, and then does not exert influence for the remaining agents, and (ii) he wants to induce the maximal number of agents to believe the false state.[12] Note that the principal's network strategy in equilibrium may be different from the strategy we describe here, but we use this to show that *some* strategy beats $\mathbf{x} = \mathbf{0}$, and therefore no intervention is not a best-response.

Therefore, the principal selects some τ so that for all agents on the arc before τ receive fake and organic news (i.e., $x_i = 1$) and all agents after τ receive only organic news ($x_i = 0$). We can solve this problem by directly characterizing the limit beliefs of the DeGroot agents:

$$(\mathbf{I} - \mathbf{A})^{-1} = \begin{pmatrix} \mathbf{I}_B & \mathbf{0}_{B,D} \\ \mathbf{X}_{D,B} & \mathbf{X}_{D,D} \end{pmatrix}$$

where

$$\mathbf{X}_{D,B} = \begin{pmatrix} n/(n+1) & 0 & \dots & 0 \\ n^2/(n+1)^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ n^{n-m}/(n+1)^{n-m} & 0 & \dots & 0 \end{pmatrix}$$

[11] Formally, we assume $m = \omega(n)$; that is $m \leq \eta n$ eventually for all constants $\eta > 0$.

[12] That is, of all feasible network strategies σ^p, the principal maximizes the number of agents taking actions $\mathbf{R}$ when the state is $\mathbf{S}$.

and

$$\mathbf{X}_{D,D} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ n/(n+1) & 1 & 0 & \dots & 0 \\ n^2/(n+1)^2 & n/(n+1) & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ n^{n-m}/(n+1)^{n-m} & n^{n-m-1}/(n+1)^{n-m-1} & n^{n-m-2}/(n+1)^{n-m-2} & \dots & 1 \end{pmatrix}$$

and $\gamma \otimes \theta$ is equal to:

$$(\boldsymbol{\gamma} \otimes \boldsymbol{\theta}) \sim \begin{pmatrix} \mathbf{1}_B \\ 1/(n+1) \cdot \boldsymbol{\gamma} \end{pmatrix}$$

where $\boldsymbol{\gamma}$ is the influence vector defined at the end of Section 3.

Consider the DeGroot agent at location τ away from the last Bayesian agent. write her belief in terms of her DeGroot centrality $\mathcal{D}$, a function of $\boldsymbol{\gamma}$:

$$\mathcal{D}(\boldsymbol{\gamma}) \sim \sum_{j=0}^{\tau-1} \frac{n^j}{(n+1)^{j+1}} \gamma_{\tau-j}$$

If the principal has chosen $\gamma_i = 1$ for all previous agents, then the above reduces to:

$$\mathcal{D}_\tau(\boldsymbol{\gamma}) \sim 1 - \left(\frac{n}{n+1}\right)^\tau$$

when τ is sublinear, $\mathcal{D}_\tau(\boldsymbol{\gamma}) \to 0$, whereas when $\tau = \alpha n$, we get that $\mathcal{D}_\tau(\boldsymbol{\gamma}) \to 1 - e^{-\alpha}$. Recalling that agents with $\mathcal{D}_\tau(\boldsymbol{\gamma}) > (1-b)/2$ will choose the incorrect action, we get that all but $\log\left(\frac{2}{1+b}\right)$ proportion of the DeGroot agents are manipulated (when $b \geq (2-e)/e$).

Now consider the principal choosing to not exert influence for all agents after some threshold τ^*. Then we obtain:

$$\begin{aligned} \mathcal{D}_\tau(\boldsymbol{\gamma}) &\sim \sum_{j=0}^{\tau-\tau^*-1} \frac{n^j}{(n+1)^{j+1}} \gamma_{\tau-j} + \sum_{j=\tau-\tau^*}^{\tau} \frac{n^j}{(n+1)^{j+1}} \gamma_{\tau-j} \\ &\sim \left(\frac{n}{n+1}\right)^{\tau-\tau^*} \cdot \left[1 - \left(\frac{n}{n+1}\right)^{\tau^*}\right] \end{aligned}$$

As the principal wants to maximize the number of agents who believe the false state, we pick:

$$\tau^*(n,b) = \inf\left\{\tau : \left[1 - \left(\frac{n}{n+1}\right)^\tau\right] \cdot \left(\frac{n}{n+1}\right)^{n-m-\tau} > \frac{1-b}{2}\right\}$$

When n is large, $\tau^*(n,b) \approx n \log\left(\frac{2+e(1-b)}{2}\right)$. Therefore, we can write the cost curve for the principal,

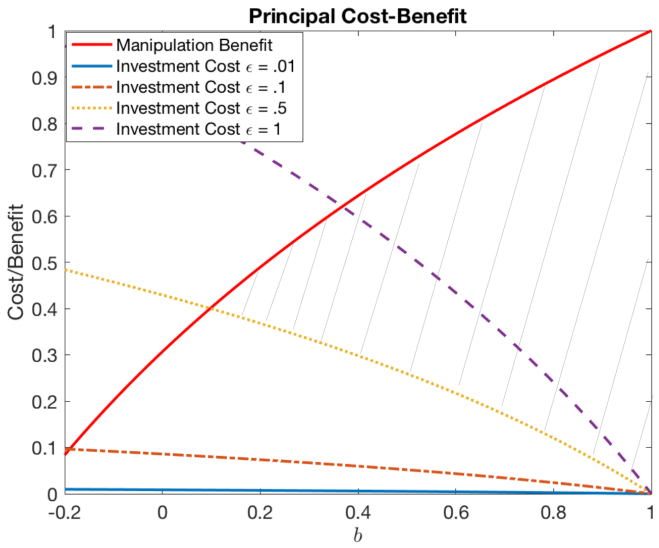

Figure 2. Cost-Benefit Curves for Principal. The shaded region indicates the profitable region for the principal when the cost ϵ is equal to 0.5.

denoted $C(\varepsilon, b)$, and the benefit curve, $B(b)$, under this strategy:

$$C(\varepsilon, b) = \varepsilon \log\left(\frac{2 + e(1-b)}{2}\right)$$
$$B(b) = \max\left\{1 - \log\left(\frac{2}{1+b}\right), 0\right\}$$

These are plotted for different values of ε in Figure 2. The cost-curve denotes the per-agent cost of the τ^*-cutoff network influence strategy, whereas the benefit-curve denotes the fraction of the population manipulated (also the utility of the principal) for the same strategy. Whenever $C(\varepsilon, b) < B(b)$, the principal strictly prefers the cutoff strategy to no influence (i.e., $\mathbf{x} = \mathbf{0}$).

For illustration, consider the case of $b = 0$, where the agent simply picks her action corresponding to the state she believes is more likely. In this case, $C(\varepsilon, 1) = \varepsilon(1 - \log(2e/(2+e))) \approx 0.86\varepsilon$ and $B(1) = 1 - \log(2) \approx 0.307$. This implies that almost 31% of the population is manipulated under this strategy, and it is profitable for the principal whenever $\varepsilon < \varepsilon^*$, where $\varepsilon^* \equiv B(1)/C(1,1)$. This holds for any number of Bayesian agents at the beginning of the ring, holding constant the continuous ring of DeGroot agents. A graphical depiction of this is seen in Figure 3.

This discussion illustrates that agents at the end of the ring might be at-risk of being manipulated. However, the principal For instance, the principal may want to stagger the agents receiving fake news in order to expend less cost while preventing a long string of agents who only receive organic news. While the exact optimal strategy is only possible via computation (see Theorem 4), we do know that

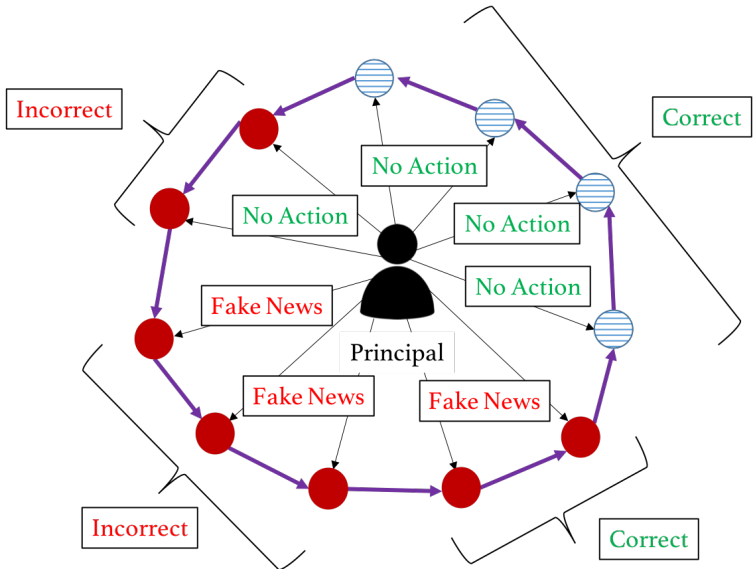

Figure 3. Beliefs in the Ring Network. A directed arrow from node i to node j indicates that i listens to j. Shaded nodes represent Bayesian agents.

the principal *will* manipulate for any $b \in (b^*, 1)$ where $b^* \equiv (2-e)/e \approx -0.246$, for a long enough ring and sufficiently small cost ε. Formally, we have the following result for any $\theta_i^{(n)}$:

Proposition 1. *Suppose there exists $\beta > 0$ such that $\theta_i^{(n)} \geq \beta/(n-m)$ for all i. Then there exists a non-empty region $\mathcal{R}$ for (ε, b) such that the ring network with many DeGroot agents and any number of Bayesian agents (in a chain) is susceptible; moreover, a constant fraction of DeGroots are manipulated in equilibrium.*

The main issue here is that the Bayesians form a continuous arc, and so the network is fundamentally equivalent to one where the arc is replaced by a single Bayesian. Moreover, the long arc of DeGroot agents, who receive fake news drowns out the beliefs of the Bayesians who know the true state. This holds even when DeGroot agents largely discount their own experience, thereby making the influence of the principal less effective. On the other hand, we obtain imperviousness if the Bayesian agents are "sprinkled" throughout the ring, which depends directly on each agent's propensity to listen entirely to her own signals:

Proposition 2. *If $\theta_i^{(n)} = \beta_i f(n)$ for some function $f : \mathbb{N} \to [0,1]$ and $\beta_i \in [\underline{\beta}, \overline{\beta}]$ for all i, then under conditions on (ε, b):*[13]

(a) There exists a placement of $n/f(n)$ Bayesians (up to a constant) such that the network is impervious for a sufficiently large population,

[13] A sufficient condition is that $\varepsilon < 1$ and either $\lim_{n\to\infty} f(n) = 0$ or $b \leq 1 - 2\overline{\beta} \limsup f(n)$.

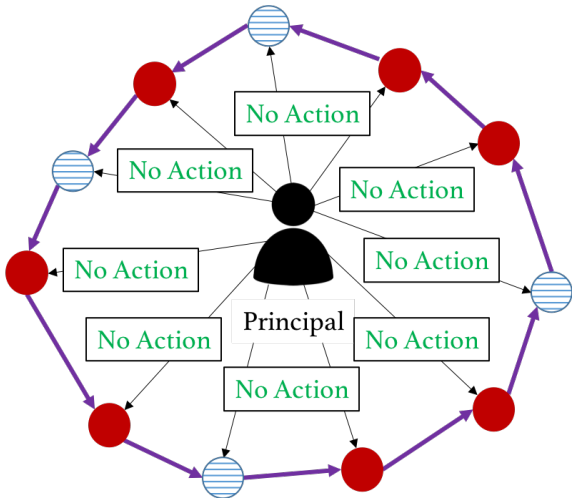

Figure 4. An illustration of Proposition 2 and Corollary 1

(b) Any placement of fewer than $n/f(n)$ Bayesians (up to a constant) makes the network susceptible for all sufficiently large populations.

In other words, Proposition 2 states that $\Theta(n/f(n))$ Bayesians are both *necessary and sufficient* for imperviousness in the ring network. Let $N(i)$ denote the neighborhood of agent i in an undirected network $\mathbf{T}$; then we obtain the following corollary:

Corollary 1. *Suppose that $\theta_i^{(n)} = \alpha_{ij} = 1/(1+|N(i)|)$ for DeGroots i and all $j \in N(i)$. If $\varepsilon < 1$ and $b < 0$, then $\Theta(n)$ optimally-placed Bayesians are necessary and sufficient for imperviousness in the ring. On the other hand, only $\Theta(1)$ Bayesians anywhere in the complete network are necessary and sufficient for imperviousness for this region of (ε, b).*

Corollary 1 gives a characterization of imperviousness for the ring network, which does not satisfy the log-diameter condition in Theorem 2. However, this imperviousness comes with more stringent requirements on resources and planning. First, as the network grows in size, the number of Bayesian agents must also grow in proportion, so that a constant fraction of the population is still Bayesian. This is in contrast to Theorem 2, where only a constant number are needed for large n. Second, the location of the Bayesian agents is paramount to preventing manipulation in networks like the ring, whereas specific placement of Bayesians do not matter in dense networks.

Finally, we conclude with an example of the *balanced star network*, where agents are aligned in a star network but weigh their personal experiences according to Corollary 1. We show that despite the seemingly added symmetry, as compared to Example 2, the network fails to satisfy the log-diameter condition, and so introduces unique vulnerabilities not present in the complete network.

Example 3 (Balanced Star Network)**.** Consider the balanced star network of Figure 5. Suppose that for agents on the periphery $\theta_i = \alpha_{i1} = 1/2$ whereas the core agent 1 updates as in Example 2, $\theta_1 = \alpha_{1j} = 1/(n+1)$. The log-diameter condition is unsatisfied because the log-diameter grows as $\approx \log(2n)$.

When the central agent is Bayesian, then either all of the agents are manipulated (if $b < 0$ and $\varepsilon < 1$) or none of them are (otherwise), i.e., the network is impervious. If Bayesians are only on the periphery, then if $m \leq \beta n$ for all $\beta > 0$ as n grows large (i.e., the number of peripheral Bayesians is sublinear), Bayesians have a vanishing fraction of influence in the network. The DeGroot centrality of the core agent converges to $\mathcal{D}_1(\boldsymbol{\gamma}) = ||\boldsymbol{\gamma}||_1/n$, whereas the DeGroot centrality of the peripheral agent i converges to $\mathcal{D}_i(\boldsymbol{\gamma}) = \frac{1}{2}\boldsymbol{\gamma}_i + \frac{1}{2}||\boldsymbol{\gamma}||_1/n$. In other words, for peripheral agents, their belief is half of the average news experience and half of their own experience, whereas the core agent's belief is simply an average of all experiences.

Given a sublinear number of Bayesians, the network is impervious if and only if $\varepsilon < \max\{1/(1-b), 1\}$ for large n; otherwise, a *linear* number of Bayesians on the periphery are required to prevent manipulation. If $b > 0$, then the principal targets $(1-b)$ fraction of the population; if $b < 0$, the principal targets all agents in the network, except the central agent. We note that the principal *targets the core agent last*, in contrast to the influential star network of Example 2, where the principal should target this agent first. While the balanced star network is more symmetric in that no agent has disproportionate influence on the population, it also prevents the central agent from acting as a spokesperson for the knowledgable Bayesians on the periphery. □

4.3 Comparative Statics on Personal Experience: Cultural Norms

We now consider the effect that θ has on manipulation. The way agents take into account their own experience relative to the opinions of others can vary substantially. An agent might put a small weight on her own experience relative to what she hears from her friends (because, for example, she believes she is not well-informed about the topic at hand). Conversely, an agent might weigh her own experience much higher compared to the information she obtains from her friends, or she can simply weigh her experience similarly to her friends' beliefs. As we show, all of these variations lead to substantial differences when it comes to manipulation. In what follows, we study what happens for a fixed network structure as the vector of experience weights $\boldsymbol{\theta}$ changes.

Definition 3 (Network Preservation)**.** We say $(\mathbf{A}', \boldsymbol{\theta}')$ is a *network preservation* of $(\mathbf{A}, \boldsymbol{\theta})$ if $\alpha'_{ij} = \alpha_{ij}(1-\theta'_i)/(1-\theta_i)$ for all DeGroot agents i.

A network preservation corresponds to a shifting of weights between an agent's own experience

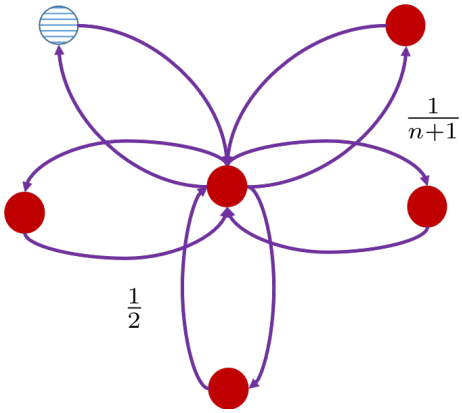

Figure 5. Balanced Star Network

and that of her neighbor's opinions, while preserving the relative proportions of the network weights. The balanced star network (Figure 5) is a network preservation of the influential star network (Figure 1) and vice-versa. We call this network preservation *homogenous* if it is a network-preservation with $\boldsymbol{\theta} = \theta\mathbf{1}$ and $\boldsymbol{\theta}' = \theta'\mathbf{1}$ (i.e., all agents have the same experience weights both before and after). The homogenous network-preservation corresponds to a unilateral shift in attitudes about the importance of one's own perceptions. Most naturally, in a homogenous network, θ can be thought of an attitude parameter tuned to the cultural norms of the population.

For the following result, we fix b and the homogenous network $\mathbf{A}$ with an arbitrary self-experience weight $\boldsymbol{\theta} = \theta\mathbf{1}$. For simplicity, we make the additional assumptions: (i) there exists at least one Bayesian agent in the population, and (ii) there is at least one DeGroot not adjacent to a Bayesian.

Proposition 3. *There exist* $0 < \underline{\theta} < \theta^* < \overline{\theta} < 1$ *such that:*

(a) If $\theta' \in (0, \underline{\theta})$*, the network* $\mathbf{A}_{\theta'}$ *is impervious for any* $\varepsilon > 0$*.*

(b) The network $\mathbf{A}_{\theta'}$ *is impervious for* $\theta' \in (\theta^*, \overline{\theta})$ *only if it is impervious for* $\theta' \in (\theta^*, 1)$ *for any* $\varepsilon > 0$*.*

(c) If $b > 1/2$,[14] *there exists* ε^* *such that when* $\theta' \in (\theta^*, \overline{\theta})$ *the network* $\mathbf{A}_{\theta'}$ *is susceptible, but when* $\theta \in (\overline{\theta}, 1)$ *the network* $\mathbf{A}_{\theta'}$ *is impervious.*

Proposition 3 shows that the comparative statics on manipulation are *non-monotone* in θ. A society that support an intermediate amount of weight on each agent's own experience is the society that is most susceptible to manipulation. On the other hand, when a society is more inclined towards herding (i.e., very small θ), then manipulation is impossible. This is because agents ignore their own

[14] When b is small, the network can exhibit no manipulation for any θ' or a "phase transition" instead: there exists θ^{**} such that $\theta' < \theta^{**}$ is impervious but $\theta' > \theta^{**}$ is susceptible.

experience and instead rely entirely on social learning. If the community has at least one sophisticated agent, then the beliefs of that agent spread throughout the network. This may come at the cost of agents dismissing accurate information from organic news sources and thus learning more slowly. Therefore, θ can also be thought of as the direct influence that *all* news (containing a possible mixture of biases or propaganda) has on a representative agent in society.

On the other hand, a culture that supports strong individuality and narcissism (i.e., very large θ) is more difficult to manipulate compared to when θ is intermediate, but easier compared to when θ is small. This is because social influence plays little role with high θ, and the principal cannot exploit social network effects to propagate his message, i.e. the principal is no longer able to reach a large population by only targeting a small subset of agents, and instead has to reach all agents directly (e.g., door-to-door campaigning). When this is the case, spreading false ideas can cease to be profitable and manipulation becomes more difficult. However, for small enough investment costs ε, manipulation may still be possible even when θ is high.

The next result considers heterogeneous settings and stands in contrast to Proposition 3. In heterogeneous settings, even if agents discount all news from the principal by having a small θ, they can still mislearn the state if other agents have high θ. To demonstrate why learning breaks down in the presence of heterogeneity, suppose that we have a set D_1 of DeGroot agents with θ_1 and a set D_2 of DeGroot agents with θ_2. Regardless of the network structure, heterogeneity leads to those agents discounting their own experiences to be manipulated, as they listen mostly to the experiences of others who still incorporate misinformation into their beliefs. This is formalized in the following proposition.

Proposition 4. *Suppose agents in D_1 are strongly connected and there exists at least one link from D_2 to D_1. For fixed θ_2, there exists $\bar{b}$ such that for all $b > \bar{b}$, even as $\theta_1 \to 0$, all DeGroot agents (including those in D_1) are manipulated for sufficiently small ε. On the other hand, if $\theta_1 = \theta_2 = \theta$, for every $\bar{b} < 1$ there exists $\bar{\theta}$ such that for all $\theta < \bar{\theta}$, the network is impervious if there is at least one Bayesian in the network regardless of ε.*

We briefly detour to consider how agents in a society might choose θ. Consider the problem of a boundedly-rational agent who wants to avoid manipulation. This agent learns using DeGroot-style heuristics, but tries to choose her network weights θ_i,$\{\alpha_{ij}\}_j$ in a clever way. This means that θ_i is agent i's best-response to other agents' choices of θ_j. In particular, agent i is incentivized to conform to some cultural standard for θ by matching others choices of θ_j. To see this, note that if other agents are herding, then the agent can avoid manipulation by also choosing her θ_i close to 0 (i.e., ignore

the news). This is true regardless of how rich agent i's signal structure is, as it can still be flooded with (undetectable) misinformation. Instead, the agent can rely on the truth emerging from social communication, knowing that sophisticated agents will discover the ground truth and spread it.

However, as other agents increase their θ_j's, agent i's beliefs may start to incorporate misinformation that it receives from those agents who are not necessarily sophisticated. In particular, if agent i believes she is more able to discern the state from *accurate* news compared to her peers, then she would be better off picking a large θ_i herself. In a more individualistic culture, if agent i chooses low θ_i, she will come to believe ideas observed by her neighbors and to some extent her neighbors' neighbors, but not many more. If agent i senses these nearby agents might be amenable to believing falsehoods, then it is in agent i's best-interest to choose a higher θ_i as well. In other words, agents would listen mostly to their own ideas and take friends' opinions with a grain of salt.

In this way, θ_i can be seen as a *cultural norm* that is plausibly consistent (either high, medium, or low) across agents in the population. When agent i does not match this cultural norm, she risks making a naive decision while ignoring her informed peers (picking θ_i high when others pick low) or risks listening to bad advice when knowing better herself (picking θ_i low when others pick high). On the other hand, when the population as a whole settles on intermediate choices for θ_i, the principal can leverage social externalities while minimizing the influence of informed agents to his biggest advantage.

5 Numerical Experiments

The previous section and Examples 1, 2, and 3 show how our results can be applied to the network topologies commonly studied in the literature. In this section, we examine these results in the context of real-world network data coming from Jackson et al. (2012). The network we consider represents an advice network in an Indian village, and consists of 144 nodes and 320 edges, where an edge between nodes i and j represents undirected communication between these two agents. In the following we look at different placement of Bayesian agents in this network in order to further demonstrate the concepts introduced throughout the paper.

Similar to the setup we have so far, the principal tries to manipulate a subset of agents in the population by sending messages to some agents (not necessarily the same set of agents he is trying to manipulate) in the network. We compute the optimal strategy for the principal given the network topology (and we assume for simplicity that all weights θ are fixed at $\frac{1}{n}$).[15] corresponds to assigning

[15] A weight of $\theta = 1/n$ is used to directly compare to dense networks of Section 4, such as the complete network.

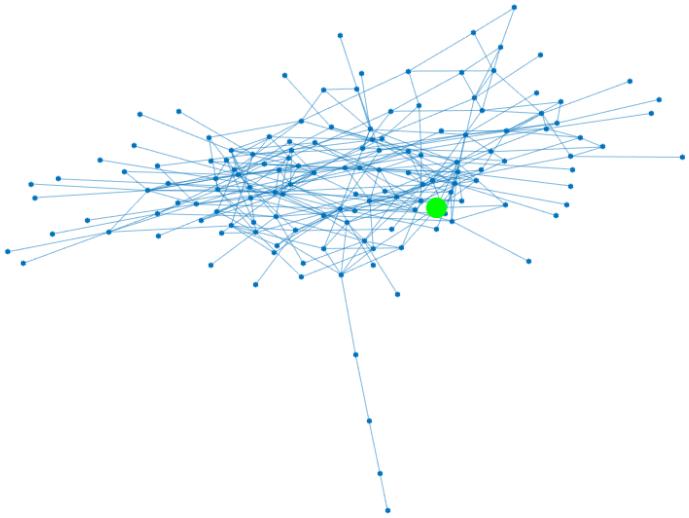

Figure 6. Central Bayesian, $b = 0$.

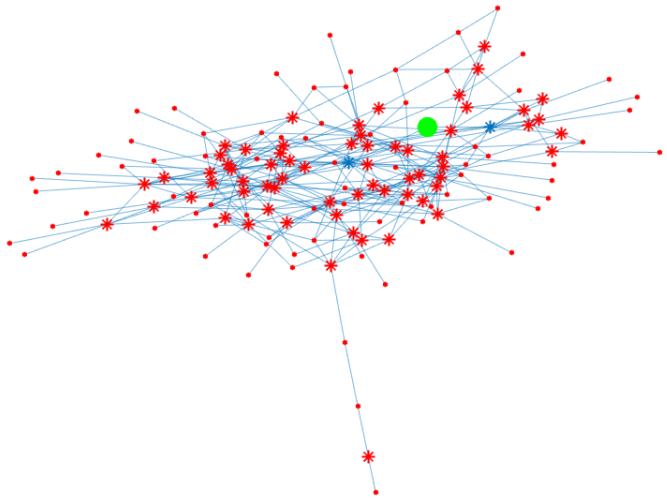

Figure 7. Peripheral Bayesian, $b = 0$.

roughly equal influence from personal signals and from signals of the rest of the population. We start with Figure 6 as an illustration that shows the network with only a single Bayesian agent. Throughout the figures in this section, green nodes represent Bayesian agents, and nodes represented with an asterisk indicate agents directly targeted by the principal. Non-Bayesian agents are colored either blue or red, to indicate whether under the principal's optimal strategy the agent is manipulated (red) or not (blue). Thus, a network of all-blue and green agents means that this particular placement of the Bayesian agents results in a network that is impervious to manipulation.

Throughout we fix $\varepsilon = 1/2$ (recall ε is the cost of sending messages to a single agent). For our first two examples, we consider the game in Table 1 and assume that $b = 0$, i.e. that agents' terminal actions reflect whichever state they believe is more likely. We focus on two particular agents, referred to in the data as Agent 70 and Agent 59. In Figure 6, Agent 70 (with degree 7 and eigenvector centrality 0.0121) is a Bayesian agent whose location results in the DeGroot centralities of all agents in the network being equal to zero in equilibrium, so the strategic principal does not interfere with learning. In other words, the principal has no profitable strategy with which he can manipulate even a single member of the population.

On the other hand, Agent 59 is much more peripheral in the network, with a degree of 2 and eigenvector centrality of 0.0044. If Agent 59 is the Bayesian agent, as is the case in Figure 7, then the average DeGroot centrality (and terminal belief in equilibrium) is $\bar{\pi} = 0.529$ and manipulation is inevitable and quite severe.

These two cases are summarized in Figure 8. Each dot in this graph represents the DeGroot Cen-

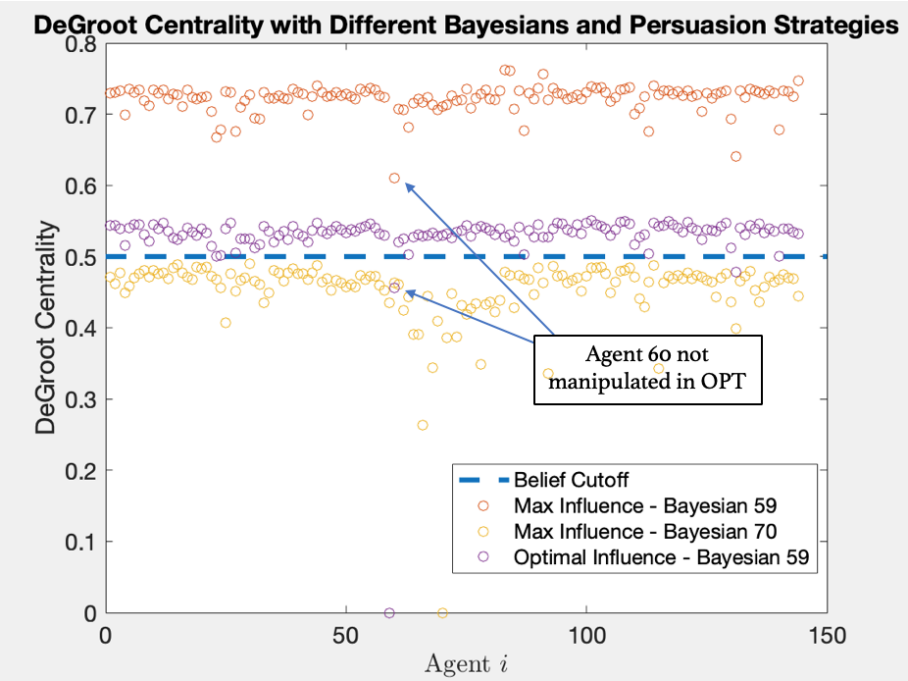

Figure 8. DeGroot Centrality for Single Bayesian, $b = 0$.

trality of the corresponding agent in the network under one of the two placements of the Bayesian agent and under a particular strategy for the principal. Agents whose DeGroot Centrality are above 0.5 are manipulated. Yellow dots correspond to the DeGroot centality of the agents in Figure 6 (with Agent 70) when the principal *targets the entire population* (i.e. when he sends messages to every single DeGroot agent). Notice that all the yellow dots are below the threshold of 0.5, and hence no agent is manipulated despite the efforts of the principal. On the other hand, if the principal applies the same strategy (targeting everyone) to the network in Figure 7 (with Agent 59) then, as can be seen from the red dots, every single DeGroot agent is manipulated since all DeGroot centralities lie above the cutoff.

Most importantly in Figure 8 however are the purple dots lying just above the dotted cutoff line. These dots represent the DeGroot centralities of the agents in Figure 7 when the principal applies the equilibrium targeting strategy depicted in the figure. Note that despite targeting 67 agents (46% of the population) instead of the entire population, the principal is able to obtain almost the maximum manipulation possible at a fraction of the cost (expends less than 50% of the cost), with only three agents (e.g., Agent 60 in the figure) escaping manipulation ($< 2\%$ of the population). This is in contrast to the complete network, where exactly one Bayesian anywhere in the network is sufficient.

The rest of the figures examine the situation for different values of b. We have seen that when b is equal to zero, manipulation is very sensitive to the placement of the *single* Bayesian agent. As b

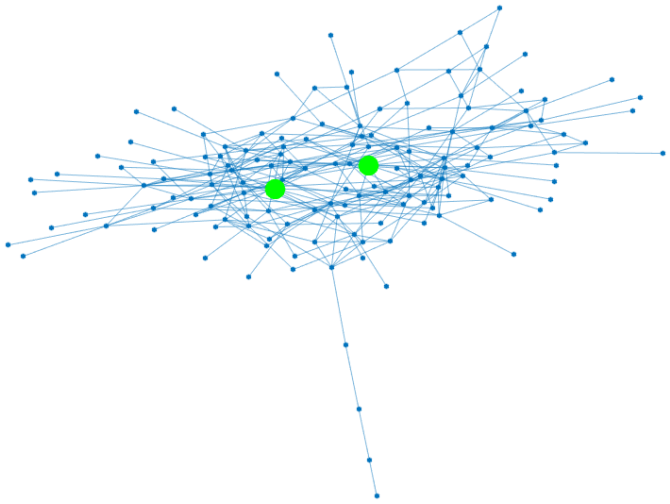

Figure 9. Two Well-Placed Bayesians, $b = 0.5$.

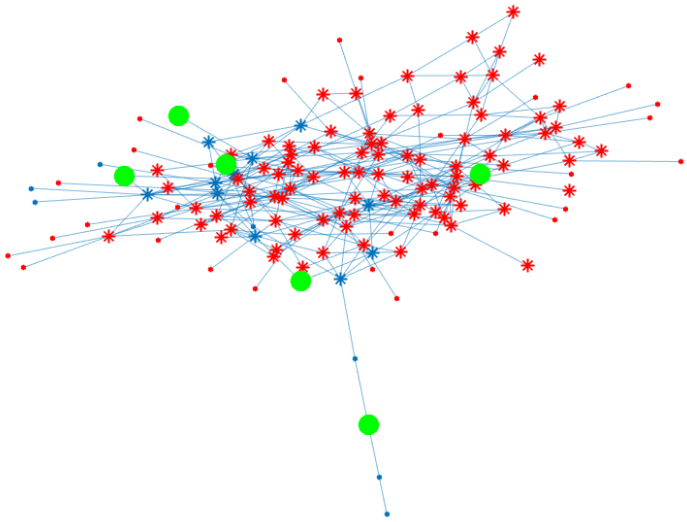

Figure 10. Five Poorly-Placed Bayesians, $b = 0.5$.

becomes lower and the cost of taking the risky action and mismatching the state increases, manipulation becomes exceedingly difficult. Similarly, as b increases, it becomes less costly for the agents to take the risky action, and hence it becomes easier to manipulate them. Figure 9 shows that with $b = 0.5$, two Bayesian agents (instead of one) are now required to prevent manipulation, provided they occupy network positions that again lead to low DeGroot centralities (across all γ) for the other agents. Similar to the ring network studied earlier, both the number and location of the Bayesians matter. Figure 10 shows that even with five Bayesian agents, large-scale manipulation is possible because these agents occupy less central positions. In the case of the complete network, three Bayesians are both necessary and sufficient for imperviousness when $b = 0.5$; in other words, the best-case placement in this network is better (requires only two Bayesians) but the worst-case placement in this network is also worse (requires at least six Bayesians). Similar conclusions are reached when b has a higher value (0.8), as can be seen in Figure 12 and Figure 13 in Appendix E.

6 Conclusion

In this paper, we embed the classic social learning problem in a principal-agent(s) setting and examine what conditions allow a principal to interfere with the learning process of the agents in order to shape their beliefs. These interactions are common in marketing, public health, politics, and many other contexts, and we provide a model that allows us to study these environments in a formal setup. In an effort to bring our model closer to real-world networks, we employ a diverse population that

possess different degrees of sophistication, which we model by considering a mixed-learning environment. We find that in this more general setup, the ability of a self-interested principal to manipulate a population depends on the learning mechanisms employed by the agents, the network structure, and the social norms in the network (as modeled by how much agents are willing to incorporate their friends' opinions into their own beliefs). We show that manipulation or lack thereof can be quite sensitive to these factors. In particular, we develop a centrality measure that we call DeGroot Centrality, which we use to classify networks into dense and sparse topologies. DeGroot Centrality is a measure that can be used to quickly identify which agents in the population are at risk of being manipulated. We demonstrate the use of this measure by studying manipulation in several common network topologies as well as an actual topology from an advice network in an Indian village. We show how some networks can be resilient with the presence of a small number of Bayesian agents, whereas others continue to be susceptible to manipulation unless the number and location of Bayesian agents meet certain criteria.

Our work can be extended on several fronts. For example, the principal can choose to vary the intensity of his messages over time, and/or can choose different intensities for different agents (as opposed to the fixed rate we use throughout the paper). One can also consider scenarios with more than two states, which will require further assumptions on the signal structure in our model. Another possibility are cases where the state that is preferred by the principal is a priori unknown to the Bayesian agents, which complicates their inference problem. We have studied the dynamics of our learning model in the limit, and characterizing the strategies played by the principal and the Bayesian agents in the short-term is also a relevant but challenging problem to solve.

Finally, and as we mention at several points in the paper, experimental investigations of these mixed learning environments is an emerging area (see the aforementioned Chandrasekhar et al. (2015)), and our framework can be utilized to provide several testable hypotheses about how agents behave in these principal-agents settings. Understanding how behavior departs from our theoretical findings can be used to enrich the theoretical framework as well as provide a bedrock for a deeper understanding of these networks and how they interact with social learning and manipulation in practice.

Appendix

A Formal Solution Concept

We define our equilibrium concept and prove existence and essential uniqueness. In particular, we first show existence of an equilibrium for any horizon T, so our solution concept is always well-defined. Secondly, while there may be many equilibria in general, we prove our equilibrium is *essentially unique* in the sense that, as $T \to \infty$, all equilibria are outcome-equivalent under generic conditions. By uniqueness of the equilibrium, we establish that learning dynamics, asymptotic beliefs, and realized payoffs are an inherent property of the network (and other primitives) rather than a specific equilibrium we choose.

Recall that agents choose strategies $\{\sigma_i\}_{i=1}^n$ over terminal actions and the principal chooses a network-influence strategy σ^p. Each agent i tries to maximize her expected utility given her belief $\pi_{i,T}$ by solving:

$$\sigma_i \in \arg\max_{\sigma_i'} \mathbb{E}^{\pi_{i,T}}_{a_i \sim \sigma_i'}[u_i(y, a_i)] \tag{*}$$

Additionally, each Bayesian agent i has belief $\mu_{i,t}$ at time t over the type of the principal $\omega \in \{\mathcal{S}, \mathcal{T}\}$ which she updates continuously according to Bayes' rule given $(h_{i,t}, \Pi_{i,t})$, taking as given the on-path equilibrium play σ^p of the principal. The principal of type $\mathcal{S}$ solves:

$$\sigma^p(\mathcal{S}) \in \arg\max_{\sigma^{p'}} \mathbb{E}_{[(\hat{y},\mathbf{x}),\pi_{i,T}]\sim\sigma^{p'}}\left[u^p(\mathbf{a}) - c(\mathbf{x})\right] \tag{**}$$

where we recall the dependence of terminal beliefs $\pi_{i,T}$ on the choice of $(\hat{y}, \mathbf{x})$. Then we define:

Definition 4 (Equilibrium)**.** We say $\sigma \equiv (\{\sigma_i\}_{i=1}^n, \sigma^p)$ is an *equilibrium* if DeGroot agents solve (*), Bayesian agents solve (**) taking σ^p as given, and the principal solves (**) taking $\{\sigma_i\}_{i=1}^n$ as given.

To obtain existence of an equilibrium, we reduce our setup to that of a reputation game with incomplete information, see Fudenberg and Levine (1989). There are a few important differences, however, because information is incomplete on two dimensions: the state of the world and the type of the principal. First, with respect to beliefs about the state, all agents operate mechanically given the principal's strategy of σ^p. This determines the state information observed by the agents, their (random) terminal beliefs $\boldsymbol{\pi}_T$, and their random terminal actions $\mathbf{a}$. Second, with respect to beliefs about the type of the principal, the principal and the Bayesian agents play a perfect Bayesian equilibrium (PBE), so the principal may be concerned about his reputation when playing some strategy σ^p. On the other hand, DeGroot agents do not doubt the veracity of their signals but interpret all news at face value. This makes the solution to the fixed-point problem slightly nuanced; nonetheless we obtain:

Theorem 3. *For every learning horizon T, there exists an equilibrium σ.*

A main focus of our paper will be characterizing under what conditions an agent chooses the terminal action which maximizes his or her payoff (i.e., matches the underlying state) given her belief at time T. To this end, we define what it means for agent i to be manipulated in equilibrium σ:

Definition 5 (Manipulation)**.** We say that agent i is *manipulated* under a realization $(\hat{y}, \mathbf{x}, \mathbf{a})$ of equilibrium σ if:

1. Her terminal action a_i *does not* match the underlying state y when the principal's type is $\omega = \mathcal{S}$.
2. Her terminal action a_i *does* match the underlying state y when the principal's type is $\omega = \mathcal{T}$.

In other words, manipulation of agent i implies that a strategic principal interferes with the learning process, and this actually causes agent i to mislearn the true state. An agent may be manipulated under some realizations of an equilibrium but not others, and moreover since σ may not be unique, the set of manipulated agents can differ depending on the equilibrium and realization of this equilibrium we analyze. This multiplicity motivates us toward a concept of *outcome equivalence*, where the number of manipulated agents under any equilibrium realization is the same with high probability. For this to hold, we need to allow agents a long time to learn; therefore, as $T \to \infty$, if all equilibria become outcome-equivalent, we say that the equilibrium is *essentially unique*.

For fixed (large) T, consider two equilibria $\sigma^{(1)}$ and $\sigma^{(2)}$ (which may be the same); we say these equilibria are κ-*outcome-equivalent* if the number of manipulated agents, $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, is the same with probability at least $1-\kappa$:

$$\mathbb{P}_{(\mathcal{M}^{(1)},\mathcal{M}^{(2)})\sim(\sigma^{(1)},\sigma^{(2)})}\left[\mathcal{M}^{(1)}=\mathcal{M}^{(2)}\right]\geq 1-\kappa$$

for any $\kappa > 0$. In other words, the equilibria are outcome equivalent (up to tolerance κ) for learning horizon T if the number of agents being manipulated is the same across any two equilibria with probability at least $1-\kappa$. This motivates our definition of essential uniqueness:

Definition 6. We say that the equilibria are *essentially unique* if for all $\kappa > 0$, there exists $T^*(\kappa)$ such that for all $T > T^*(\kappa)$, the equilibria for horizon T are κ-outcome-equivalent.

Proposition 5. *For generic parameters (ϵ, b), the equilibrium is essentially unique.*

Essential uniqueness of the equilibrium guarantees that, with high probability, our welfare analysis (i.e., the number of agents who mislearn the state) does not depend on the equilibrium we choose, or the realization of the signals or actions from that equilibrium, as $T \to \infty$. We do not rule out the possibility of multiple equilibria, or different realizations of the same equilibrium, yielding substantive differences when the learning horizon T is small. Likewise, even for large T, it may be possible that *different* agents are manipulated under different equilibria but the *total number* of these manipulated agents remains unchanged. For this reason, the focus of the paper will be on which network structures lead to manipulation for a non-empty subset of agents in *some* equilibrium as $T \to \infty$, noting that the identity of these agents may be different under different equilibria, but the welfare properties are invariant.

B Limit Beliefs and Asymptotic Learning

One can think of the principal as an adversarial designer who picks his network strategy in a way that maximizes his own payoffs. As is typical in design problems, it is easiest to first consider the equilibrium actions of the agents holding fixed the strategy $\sigma^p(\mathcal{S})$. In particular, in this section we aim to understand the asymptotic learning dynamics that emerge for a given network strategy of the strategic principal. DeGroot agents have one-dimensional identification problem of learning the true state y. In contrast, Bayesian agents have a two-dimensional identification problem. In addition to learning the state of the world, they also learn about the type of the principal, and whether he is interfering in the learning process.

We show that for large T and under mild conditions, the Bayesians always learn the true state of the world regardless of any efforts the principal may exert to thwart learning. On the other hand, we provide a closed-form expression for DeGroot terminal beliefs, as a function of the chosen strategy σ^p. These terminal beliefs induce random terminal actions for each DeGroot agent i at T, which in turn provide an expression for the probability that agent i is manipulated under σ^p.

For notational purposes, assume there are $\{1,\dots,m\}$ Bayesian agents and $\{m+1,\dots,n\}$ DeGroot agents, as usual. Throughout this paper, we will make standard network connectedness assumptions, reminiscent of those from Jadbabaie et al. (2012) and other social learning models:

Assumption 1. The network defined by $\mathcal{A}$ is *strongly connected*[16] and the personal-experience weight, θ_i, is positive if and only if $p_i > 1/2$.

The first part of the assumption requires that the beliefs of any one agent can reach (or influence) any other agent, albeit indirectly through others. It can be relaxed in the event there are distinct components (entirely separated), where the analysis presented here can be applied to each component independently. The second part requires that all agents in the network listen to the news they receive if and only if the organic signals are informative of the true state. Agents whose organic signals provide only noise instead form their beliefs entirely from social influences. We also introduce the following assumption about the signals received by agents in the population:

Assumption 2. Let $\lambda_{\max}$ be the largest (realized) λ_i (i.e., $\max_i \lambda_i$). Then:

(a) No agent receives organic news faster than $\underline{\lambda} + \overline{\lambda}/2$; that is, $\lambda_{\max} < \underline{\lambda} + \overline{\lambda}/2$.

(b) Every agent in isolation is susceptible to mislearning; that is, for all agents i:

$$p_i < \frac{\overline{\lambda} + \lambda_{\max}}{2\lambda_{\max}}$$

(c) There exists some agent i (DeGroot or Bayesian) whose signal is reasonably informative:

$$p_i > \frac{\overline{\lambda} + \underline{\lambda}}{2\underline{\lambda} + \overline{\lambda}}$$

Note that assumption (a) guarantees that both (b) and (c) are possible. Condition (b) ensures that agents use the social network as a way to protect themselves against possible manipulation. If agents are left in isolation, and the principal attempts to corrupt their signals, then it is impossible for agent i to uncover the truth simply from performing Bayesian updating on her signals. This allows us to isolate the impact of social learning on preventing a strategic principal from gaining widespread influence. Condition (c), however, ensures that some agent in the network gets a strongly informative signal. This is necessary to ensure the principal's influence cannot entirely corrupt the ground truth by disguising the signal generating process as purely organic news under a different (false) state. Finally, note that while we impose (a) and (b) hold for simplicity of analysis, only (c) must be common knowledge (for the principal and Bayesians).

B.1 Bayesian Learning

Recall that the probability that the principal is the truthful type $\mathcal{T}$ is given by ζ. Consider the following properties (**) about asymptotic learning for any fixed σ^p:

1. If σ^p is a pure strategy, then if $\mathbf{x} = \mathbf{0}$, $\lim_{T\to\infty} \mu_{i,T} = \mu_{t,0}$ for all Bayesians i; however, if $\mathbf{x} \neq \mathbf{0}$ and $\hat{y} \neq y$, then $\lim_{T\to\infty} \mu_{i,T} = 0$ for some Bayesian i.

[16] Formally, the network is strongly connected if there exists a directed path between any two agents (i.e., $\exists k$ such that $\mathcal{A}^k_{ij} > 0$ for all i, j).

2. If σ^p is a mixed strategy with support $\nu > 0$ on any $\mathbf{x} \neq \mathbf{0}$ and $\hat{y} \neq y$, then $\lim_{T\to\infty} \mu_{i,T} = 0$, if the principal takes action $\mathbf{x} \neq \mathbf{0}$, for some Bayesian i.

In particular, (1) states that if in equilibrium, the principal commits to a pure strategy that mimics the truthful type, then all Bayesians will be unable to differentiate between this type and the strategic one. On the other hand, if the principal interferes *anywhere* in the network by sending false messages, the at least one Bayesian will recognize he is interfering. If the principal plays a mixed strategy, then (2) guarantees as $T \to \infty$, the amount of influence the principal can have without detection goes to zero.

Theorem 1. Under Assumption 1 and 2(c), (**) holds. In particular, no Bayesian agent is manipulated almost surely as $T \to \infty$.

Informally, this implies that the Bayesian agents "figure everything out" about the play of the principal when they are aware he may be strategic (i.e., $\mu_0 < 1$), and can therefore remove his influence and identify the correct state. We require that there is at least an agent with rich enough signals; otherwise, the principal can simply fool the entire community. However, this assumption is fairly weak. We do not impose that this "expert" be a Bayesian agent, and in fact it may be possible that all Bayesian agents are unable to decipher to the news altogether (i.e., $p_i = 1/2$ for every Bayesian). Bayesian agents are able to infer the presence of fake signals in the network even if the principal does not send these signals directly to him or her.

As Bayesian agents become more convinced of the true type of the principal, they are able to make correct inferences about the underlying state. Moreover, these agents can then communicate their conclusions, through their communicating their beliefs, to the rest of the network. Therefore, Bayesian agents provide a positive informational externality, which assists all agents in getting accurate information about the true type (and future play) of the principal. In this way, *in the limit*, Bayesian agents become stubborn agents. However, unlike much of the previous literature on stubborn agents, their presence reduces the amount of misinformation which can persist in the population.

When the network consists entirely of Bayesian agents, the principal is unable to manipulate. On the other hand, when the network consists of all DeGroot agents, manipulation will always be possible (in general) when the influence cost ε is not too large. The interesting case will come in the mixed learning environment, where there are both DeGroot and Bayesian agents. In this setting, there are two opposing forces: (1) the Bayesian agents who can accurately deduce the state information and communicate this over the network, and (2) the DeGroot agents who may confound the learning process through simple learning heuristics. We will study whether the principal can effectively use (2) to his benefit, despite the presence of (1).

B.2 DeGroot Learning and Network Structure

To understand the role of the network structure in the principal's problem, we need to characterize *asymptotic* learning for DeGroot agents. Recall we denote by y the realized state and let us write y' as an arbitrary state. For large enough t, the beliefs of the DeGroot agents evolve approximately according to the law of motion:

$$\pi_{i,t+1}(y'|y,\omega) = \theta_i g_i(h_{i,\infty}|y,\omega) + \sum_{i=1}^{n} \alpha_{ij}\pi_{j,t}(y'|y,\omega)$$

By Theorem 1, every Bayesian agent i's limit belief, denoted by $\pi^B_\infty(y'|y,\omega)$, is approximately given by $\pi^B_{i,\infty}(y|y,\omega) = 1$. In matrix notation, we can write this as

$$\boldsymbol{\pi}_{t+1}(y'|y,\omega) = \mathbf{A}\boldsymbol{\pi}_t(y'|y,\omega) + \mathbf{g}(\mathbf{h}_\infty(y'|y,\omega)) \otimes \boldsymbol{\theta}$$

where the matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{D,B} & \mathbf{A}_{D,D} \end{pmatrix}$$

and $\mathbf{A}_{D,B}$ is the DeGroot by Bayesian agent weight matrix and $\mathbf{A}_{D,D}$ is the DeGroot by DeGroot agent weight matrix, and $\otimes$ is the element-by-element product. The random vector $\mathbf{g}(\mathbf{h}_\infty(y'|y,\omega))$ has distribution given by:

$$\mathbf{g}^*(y') \equiv \mathbf{g}(\mathbf{h}_\infty(y'|y,\omega)) \sim \begin{pmatrix} \pi^B_\infty(y'|y,\omega)\mathbf{1}_B \\ g(h_{m+1,\infty}(y'|y,\omega)) \\ g(h_{m+2,\infty}(y'|y,\omega)) \\ \cdots \\ g(h_{m+n,\infty}(y'|y,\omega)) \end{pmatrix}$$

where the $h_{i,\infty}$ is the random history of news (both organic and fake) induced by the principal's strategy $\sigma^p(\omega)$. The deterministic vector $\boldsymbol{\theta}$ is then given by $\boldsymbol{\theta} = (\mathbf{1}_B, \theta_{m+1}, \theta_{m+2}, \ldots, \theta_n)'$. Given this formulation as classical DeGroot learning, we present the following asymptotic result for the beliefs of the DeGroot agents:

Proposition 6. *For principal type ω, as $t \to \infty$, the beliefs of the DeGroot agents in the network, $\boldsymbol{\pi}^D$, about the state converge almost surely to:*

$$\boldsymbol{\pi}^D_t \overset{a.s.}{\to} (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{g}^* \otimes \boldsymbol{\theta}) \equiv \boldsymbol{\pi}^D_\infty$$

where $\mathbf{G}$ depends on the true state y^* and the play of the principal of type ω, $\sigma^p(\omega)$.

This result builds on the standard belief characterization from the social learning literature, with a couple of caveats. First, agents receive information "externally" from idiosyncractic news which they incorporate into their own personal belief. For this reason, the expression for asymptotic beliefs resembles the steady-state Leontif input-output economy. The other key difference, however, is that Bayesian agents are absorbing states in the population. In particular, they are not sensitive to the choice of σ^p, although DeGroot agents are through the vector $\mathbf{g}^*$.

Lastly, we comment that when $\omega = \mathcal{T}$, Proposition 6 guarantees that for T large, all agents (Bayesian or DeGroot) learn the true state when the connectivity conditions of Assumption 1 and the organic signal conditions of Assumption 2 are satisfied. This can be seen from the fact that $\mathbf{g}^*$ approaches the vector of all 1's at $y' = y$ given that $\sigma^p(\mathcal{T})$ chooses $\mathbf{x} = \mathbf{0}$ with probability 1. Therefore, without a strategic principal, learning occurs despite the fact that DeGroot agents are only updating their beliefs using the heuristic. Characterizing manipulation (when T is large) simply reduces to asking which agents mislearn the state when the strategic principal plays some $\sigma^p(\mathcal{S})$. This is the main focus for the remainder of the paper.

C General Characterization and DeGroot Centrality

We can give a characterization of manipulation in an arbitrary network by observing that it is closely related to a centrality measure that resembles eigenvector centrality and Katz-Bonacich centrality in the social learning literature. Consider some vector $\boldsymbol{\gamma}$, which we will call the *influence* parameter for

the principal, of dimension $(n-m)\times 1$. Toward defining our centrality measure, let us define the characteristic-vector, parametrized by $\boldsymbol{\gamma}$, to be:

$$\boldsymbol{\xi}(\boldsymbol{\gamma}) \equiv \begin{pmatrix} \mathbf{0}_m \\ \boldsymbol{\theta} \otimes \boldsymbol{\gamma} \end{pmatrix}$$

We define the *DeGroot centrality* vector to be:

$$\mathcal{D}(\boldsymbol{\gamma}) \equiv [(\mathbf{I}-\mathbf{A})^{-1}]\boldsymbol{\xi}(\boldsymbol{\gamma})$$

It measures the *level of influence* the other DeGroot agents have on the agent's own belief. One way to interpret the term $[(\mathbf{I}-\mathbf{A})^{-1}]_{ij}$ is the number of *weighted* walks between i and j that do no pass through a Bayesian. That is, define the weight of a walk $W = i \to v_1 \to v_2 \to \ldots \to v_n \to j$ to be:

$$w_W = \prod_{(v_i \to v_{i+1}) \in W} \alpha_{v_i, v_{i+1}}$$

If $\mathcal{W}_{ij}$ is the set of walks between i and j that do not pass through a Bayesian, it can be shown that:

$$[(\mathbf{I}-\mathbf{A})^{-1}]_{ij} = \sum_{W \in \mathcal{W}_{ij}} w_W < \infty$$

Our definition of centrality is a *generalization* of Bonacich centrality. We can think of the θ_i terms as being the discount factors that are *node-dependent* and reflect the level of stubbornness or influence of that agent. Bayesian agents can be taken as stubborn agents that communicate the truth and have an effective $\theta_i = 1$. For instance, in a setting where $\boldsymbol{\theta} = (1-\beta)\mathbf{1}$ and all the agents are DeGroot, we recover β-Bonacich centrality by setting $\boldsymbol{\gamma} = \mathbf{1}$.

The characteristic vector for agent i is the vector of state opinions in a world where each agent i receives an experience of deterministic value γ_i in every period. Each DeGroot agent employs a cutoff strategy; if at time T her terminal belief exceeds $(1-b)/2$ she chooses **R**, and otherwise she chooses **S**. We can then write the principal's problem as

$$\begin{aligned} \boldsymbol{\Gamma}^* = \arg\max & \sum_{i=m+1}^{n} z_i - \varepsilon\gamma_i \\ \text{s.t. } \forall i:\ & z_i \le \mathcal{D}_i(\boldsymbol{\gamma}) + (1+b)/2 \\ \forall i:\ & \gamma_i, z_i \in \{0,1\} \end{aligned}$$

Theorem 4. *Given investment cost $\varepsilon > 0$ and a solution $\boldsymbol{\Gamma}^*$ to the principal's problem, a network is impervious if $\mathbf{0} \in \boldsymbol{\Gamma}^*$; otherwise it is susceptible.*

The intuition for the result is as follows. The principal can choose to either send fake news ($\gamma_i = 1$) or not ($\gamma_i = 0$) for each agent. The choice of $\boldsymbol{\gamma}$ impacts the principal's payoffs in two ways: (i) a direct, separable cost ε for each $\gamma_i = 1$ and (ii) a network impact captured in the DeGroot centrality (i.e., how the experiences of DeGroot agents impact the beliefs of others) from the aggregate vector $\boldsymbol{\gamma}$. As the principal tries to manipulate more agents, the greater the DeGroot centralities are and the more likely he is to convince other DeGroots of the incorrect state. Importantly, these *network externalities* can help or hurt the principal's objective. On one hand, the impact of $\gamma_i = 1$ diffuses throughout the network and is not sufficient on its own to convince even agent i of the wrong state. However, when the influence vector consists of many agents, it can serve to convince both these agents *and others* of the wrong state, even if these other agents are not directly influenced by the principal (i.e., $\gamma_i = 0$).

The trade-off between these two effects depends on the underlying network structure.

Note that $\mathcal{D}_i(\boldsymbol{\gamma})$ is *linear* in $\boldsymbol{\gamma}$; despite this, such an optimization problem is generally intractable. However, we can provide sufficient conditions for showing that a network is either impervious or susceptible to manipulation. These conditions, for most networks in practice, tend to be much more useful than direct application of this optimization problem. For notation purposes, for a subset $\mathcal{K} \subset D$ of DeGroot agents let $\mathbf{1}_{\mathcal{K}}$ denote the vector given by:

$$[\mathbf{1}_{\mathcal{K}}]_i = \begin{cases} 1, \text{ if } i \in \mathcal{K} \\ 0, \text{ otherwise} \end{cases}$$

Then we obtain the following corollary to Theorem 4:

Corollary 2. *Fix some $\varepsilon > 0$; then the network is:*

(a) Impervious *to manipulation if $\mathcal{D}_i(\mathbf{1}_D) < (1-b)/2$ for every DeGroot agent i, or*

(b) Susceptible *to manipulation if there exists a subset $\mathcal{K}$ of DeGroot agents such that:*

$$\sum_{i=m+1}^{n} 1_{\mathcal{D}_i(\mathbf{1}_{\mathcal{K}}) > (1-b)/2} > \varepsilon |\mathcal{K}|$$

Note that the condition on imperviousness is sufficient but not necessary. It simply states that if the principal attempts to send fake signals to all of the DeGroot agents, this is still not enough to convince them of the false state. We see this result holds regardless of the cost of investment ε; in particular, it becomes a necessary condition as well when $\varepsilon \to 0$. However, a necessary *and* sufficient condition for susceptibility is given by (b). While it is challenging to verify that there exists no subset $\mathcal{K}$ that is profitable for the principal to manipulate, it is often easy to simply check that some subset $\mathcal{K}$ does better than $\boldsymbol{\gamma} = \mathbf{0}$.

D Proofs

D.1 Main Body

Proof of Theorem 1. Note by Assumption 2(c), there is some agent i with "effective" probability of signal R:

$$\tilde{p}_i \equiv \frac{\lambda_i^*}{\lambda_i + \lambda_i^*} + (1-p_i)\frac{\lambda_i}{\lambda_i + \lambda_i^*} \leq \frac{\overline{\lambda}}{\underline{\lambda} + \overline{\lambda}} + (1-p_i)\frac{\underline{\lambda}}{\underline{\lambda} + \overline{\lambda}} < p_i$$

Therefore, the probability that agent i gets signal R is strictly less than p_i if and only if $y = S$, by Lemma 1. If the agent is a Bayesian, then as $T \to \infty$ the agent will hold belief $\pi_{i,T}(y) \to 1$. On the other hand, if the agent is DeGroot, then $g_{i,t}(h_{i,t})$ encodes the difference $z_{i,t}^R - z_{i,t}^S$, and because Δ is small, any observer of $g_{i,t}(h_{i,t})$ for all t knows $z_{i,t}^R + z_{i,t}^S$ by counting the number of changes to $g_{i,t}(h_{i,t})$ for all $t' \leq t$, since $p_i > 1/2$. Thus, an observer of $g_{i,t}(h_{i,t})$ can identify $\lambda_i + \lambda_i^*$ (i.e., the arrival rate of all signals) as $T \to \infty$ from $z_{i,t}^R + z_{i,t}^S$. Moreover, as $T \to \infty$, this observer can deduce the quantity $(2\tilde{p}_i - 1)(\lambda_i + \lambda_i^*)$ from $z_{i,t}^R - z_{i,t}^S$, where again $\tilde{p}_i$ is the effective probability of getting signal R (from both organic and principal signals). Since $\tilde{p}_i < p_i$ when $y = S$, regardless of x_i, an observer of $g_{i,t}(h_{i,t})$ can deduce the correct state y when it is either R or S. We refer to the agent i that satisfies Assumption 2(c) as the "special agent," which we denote i^* from here on.

Now we use this fact to show every Bayesian learns the state y. There is a walk from every Bayesian i to this special agent, which may pass through other Bayesians, because the network is strongly connected. Denote the (maximal) sequence of Bayesians along such a walk as $i \to j_1 \to \cdots \to j_k \to i^*$. By Lemma 2, Bayesian j_1 is not manipulated because agent j_1 can effectively observe $g_{i^*}(h_{i^*,t'})$ (or $\pi_{i^*,t'}$ if i^* is a Bayesian) for all $t' < t - \text{dist}(j_1, i^*) \cdot \Delta$, which implies that the belief of Bayesian j_1, π_{j_1}, converges to a point-mass on the correct state. However, by identical reasoning using Lemma 2, this shows that agent j_2 is not manipulated (who can indirectly observe the belief of agent j_1), and so on, to suggest that Bayesian j_2 will not be manipulated. Straightforward induction proves that Bayesian agent i will not be manipulated.

We know that playing $\mathbf{x}$ where $x_i = 1$ for either an agent i with $\theta_i = 0$ is dominated by the strategy $\mathbf{x}'$ where $x'_j = x_j$ for all $j \neq i$ and $x'_i = 0$, so will not be played in equilibrium. If σ^p is a pure action of $\mathbf{x} = \mathbf{0}$, then in equilibrium the signals provide no information about the type of the principal (as both types play $\mathbf{x} = \mathbf{0}$), and so no agent updates her prior μ_0. Otherwise if $\mathbf{x}^* \neq \mathbf{0}$ when $\hat{y} \neq y$, we know that $x_i = 1$ for some agent i. Some Bayesian will be connected to agent i through a chain of DeGroots, and aware that $g_{i,t}(h_{i,t})$ is converging to the incorrect state by Lemma 2, and by the logic from the paragraph above, will know the true state, and hence realize the principal is strategic and attempting to manipulate.

Now suppose the principal plays some $\mathbf{x}^* \neq \mathbf{0}$ when $\hat{y} \neq y$ only with probability $\nu > 0$. If the principal does play $\mathbf{x}^*$, then given history $h_{i,t}$ we have for sufficiently large t:

$$\mu_{i,t} \leq \frac{\mathbb{P}_{i,t}[h_{i,t}|\mathbf{x} = \mathbf{0}]\mu_0}{\mathbb{P}_{i,t}[h_{i,t}|\mathbf{x} = \mathbf{0}](\mu_0 + (1-\eta)(1-\mu_0)) + \mathbb{P}_{i,t}[h_{i,t}|\mathbf{x} = \mathbf{x}^*]\eta(1-\mu_0)}$$

with high probability, where $\mathbb{P}_{i,t}[h|\mathbf{x} = \mathbf{x}']$ is the conditional probability of observing history h given (pure) strategy $\mathbf{x}'$, at time t for Bayesian agent i. For fixed η, since $\mathbb{P}_{i,t}[h_{i,t}|\mathbf{x} = \mathbf{0}]/\mathbb{P}_{i,t}(h_{i,t}|\mathbf{x} = \mathbf{x}') \to 0$ as $t \to \infty$, we see that still $\lim_{t\to\infty} \mu_{i,t} = 0$ for a Bayesian connected through a chain of DeGroots to an agent with $x_i = 1$. $\square$

Proof of Theorem 2. It is sufficient to prove that the sum of weighted walks passing through Bayesian agent is bounded below by a constant $\rho(\delta, m)$ which only depends on the log-diameter δ of $\mathbf{G}$ and the number of Bayesian agents m, with $\lim_{m\to\infty} \rho(\delta, m) = 1$ for all δ. To see this, note that $\mathcal{D}_i(\mathbf{1}_D) \leq (1 - \rho(\delta, m))$ for all i, so we can construct $m^*(\delta)$ from:

$$m^*(\delta) = \inf\{m : \rho(\delta, m) \geq (1+b)/2\}$$

which exists because the set above is non-empty if the limit of ρ converges to 0. By Corollary 2 (and noting $1 - \rho(\delta, m) \leq (1-b)/2$ is an equivalent condition) this implies the network is impervious to manipulation.

Let w_i^B be the sum of weighted walks that *end* with a Bayesian agent, which clearly a lower bound on $\rho(\delta, m)$. Since the log-diameter of the network is less than δ, we know that between any two agents i and j, there exists a walk $W_{ij}^* = i \to u_1 \to u_2 \to \cdots \to u_k \to j$ such that:

$$-\log(\alpha_{iu_1}) - \sum_{\ell=1}^{k-1} \log(\alpha_{u_\ell u_{\ell+1}}) - \log(\alpha_{u_k j}) = -\log\left(\alpha_{iu_1} \cdot \alpha_{u_k j} \cdot \prod_{\ell=1}^{k-1} \alpha_{u_\ell u_{\ell+1}}\right) \leq \log(n+\delta)$$

$$\implies \alpha_{iu_1} \cdot \alpha_{u_k j} \prod_{\ell=1}^{k-1} \alpha_{u_\ell u_{\ell+1}} = W_{ij}^* \geq \frac{1}{n+\delta}$$

Let us define an *intermediate walk* to be a walk of weight at least $1/(n+\delta)$ between two DeGroot

agents i, j. Additionally, let us say a k-weighted walk from DeGroot i ending at some Bayesian j is the concatenation of k intermediate walks; in other words, the ending vertex of one intermediate walk is the starting vertex of the next. If we let $\mathcal{I}_k$ denote the set of k-weighted walks starting at i, then we observe:

$$w_i^B \geq \sum_{k=1}^{\infty} \sum_{W \in \mathcal{I}_k} w_W$$

Observe there are at least $(n-m-2)^{k-1}$ k-weighted walks from i to any given Bayesian j. To see this, note that because i has a weighted walk of weight $1/(n+\delta)$ to every other vertex v in $\mathbf{G}$ (by the above inequality), the number of k-weighted walks is the number of intermediate vertices between i and j which do not include Bayesians (or i, j themselves). Moreover, we note that by the previous inequality:

$$\sum_{W \in \mathcal{I}_k} w_W \geq m \cdot (n-m-2)^{k-1} \cdot \left(\frac{1}{n+\delta}\right)^k$$

Putting the pieces together, we have that:

$$\begin{aligned}
\rho(\delta, m) \geq w_i^B &\geq \sum_{k=1}^{\infty} m \cdot (n-m-2)^{k-1} \cdot \left(\frac{1}{n+\delta}\right)^k \\
&= \frac{m}{n+\delta} \sum_{k=1}^{\infty} \left(\frac{n-m-2}{n+\delta}\right)^k \\
&= \frac{m}{n+\delta} \frac{1}{1-\frac{n-m-2}{n+\delta}} \\
&= \frac{m}{m+\delta+2}
\end{aligned}$$

Finally, noting that $\lim_{m\to\infty} m/(m+\delta+2) = 1$ completes the proof. □

Proof of Proposition 1. In Section 4.2, we computed DeGroot centrality by matrix inversion; here we will employ the walk approach. If $j > i$, the weighted walk from agent i to agent j is simply the influence multiplied over the length of the walk, $(j-i)$, that is, $\prod_{\eta=i+1}^{j}(1-\theta_\eta^{(n)})$. If $j < i$, the influence is zero because every walk passes through a Bayesian. Therefore, the DeGroot centrality of agent j is given by:

$$\mathcal{D}_j(\boldsymbol{\gamma}) = \sum_{\kappa=m+1}^{j} \theta_\kappa^{(n)} \left(\prod_{\eta=\kappa+1}^{j} 1-\theta_\eta^{(n)}\right) \gamma_\kappa = 1 - \prod_{\eta=m+1}^{j}(1-\theta_\eta^{(n)}) - \sum_{\kappa=m+1}^{j} \theta_\kappa^{(n)} \left(\prod_{\eta=\kappa+1}^{j} 1-\theta_\eta^{(n)}\right)(1-\gamma_\kappa)$$

where the equality follows from noting all weighted walks sum to 1, and subtracting the influence from the Bayesians and the influence from those DeGroots κ with $x_\kappa = 0$. If $\gamma_\kappa = 1$ for all previous $\kappa < j$, then for some $\beta > 0$:

$$\mathcal{D}_j(\mathbf{1}_D) \geq 1 - \prod_{\eta=m+1}^{j} \left(1 - \frac{\beta}{n-m}\right) = 1 - \left(\frac{n-m-\beta}{n-m}\right)^{j-m}$$

On the other hand, if $\gamma_i = 1$ for all $\tau \leq \tau^*$, then for $j \leq \tau^*$:

$$\mathcal{D}_j(\boldsymbol{\gamma}) \geq 1 - \left(\frac{n-m-\beta}{n-m}\right)^{j-m}$$

whereas for $j \geq \tau^*$:

$$\begin{aligned}\mathcal{D}_j(\boldsymbol{\gamma}) &= 1 - \prod_{\eta=m+1}^{j}(1-\theta_\eta^{(n)}) - \sum_{\kappa=\tau^*+1}^{j}\theta_\kappa^{(n)}\left(\prod_{\eta=\kappa+1}^{j}1-\theta_\eta^{(n)}\right) \\ &\geq 1 - \left(\frac{n-m-\beta}{n-m}\right)^{j-m} - \sum_{\kappa=\tau^*+1}^{j}\theta_\kappa^{(n)}\left(\prod_{\eta=\kappa+1}^{j}1-\theta_\eta^{(n)}\right)\end{aligned}$$

Suppose that $\tau^* = n$ (i.e., all DeGroot agents receive the principal's signals). Then $\lim_{n\to\infty}((n-m-\beta)/(n-m))^{n-m} = e^{-\beta}$, so $\lim_{n\to\infty}\mathcal{D}_n(\boldsymbol{\gamma}) = 1 - e^{-\beta}$. If the principal wants to manipulate the most people, then it is clear that we want $\mathcal{D}_j(\mathbf{x}_{\text{cutoff}}(\tau^*)) > (1-b)/2$ for agent $j = n$ at the end of the ring. Therefore, we solve for τ^*:

$$\tau^* = \inf\{\tau : \mathcal{D}_j(\mathbf{x}_{\text{cutoff}}(\tau^*)) > (1-b)/2\}$$

where the infimum is well-defined for some b because $1-e^{-\beta} > 0$ for all $\beta > 0$. If the principal chooses strategy $\mathbf{x}(\tau^*)$, then he guarantees that all DeGroot agents who are manipulatable are manipulated, as the only agents not manipulated are those at the beginning of the ring who are not manipulated for any $\boldsymbol{\gamma}$. Therefore, the cost of this strategy (and noting that the principal never exerts effort for the Bayesians) is $\tilde{C}(\varepsilon, b) = \varepsilon\tau^* \leq \varepsilon(n-m)$.

Next, we compute the number of agents not manipulated under this influence strategy. We note that these agents consist of an arc of length approximately $n \cdot \ell$, with:

$$1 - \left(\frac{n-m-\beta}{n-m}\right)^{\ell(n-m)} \geq (1-b)/2$$

which as n grows large is equivalent to:

$$1 - e^{-\beta\ell} \geq (1-b)/2 \Longrightarrow \ell \leq \frac{1}{\beta}\log\left(\frac{2}{1+b}\right)$$

Therefore, for n large, the total benefit from the network influence strategy with cutoff τ^* is given by $\tilde{B}(b) \geq (n-m)\left(1 - \frac{1}{\beta}\log\left(\frac{2}{1+b}\right)\right) - \epsilon_b$ with $\lim_{n\to\infty}\epsilon_b = 0$. Consider the region $\underline{\mathcal{R}}$ is given by:

$$\underline{\mathcal{R}} \equiv \left\{(\varepsilon, b) \in \mathbb{R}^2_{++} \,:\, b > 1 - 2e^{-\beta} \cap \varepsilon < \left(1 - \frac{1}{\beta}\log\left(\frac{2}{1+b}\right)\right)\right\} \neq \varnothing$$

For any $(\varepsilon, b) \in \underline{\mathcal{R}}$, the network influence strategy $\mathbf{x}^*$ with cutoff τ^* does strictly better than $\mathbf{x} = \mathbf{0}$, which does strictly better than any $\mathbf{x}' \neq \mathbf{0}$ where no agents are manipulated. Therefore, the network is susceptible for all $(\varepsilon, b) \in \mathcal{R} \supset \underline{\mathcal{R}}$, with $\mathcal{R} \neq \varnothing$.

Finally, to see that at least $\Omega(n)$ DeGroots are manipulated, we note that $\mathbf{x}^*$ does better than any network influence strategy where $\omega(n)$ DeGroots are manipulated. If $\mathbf{x}^{**}$ were such a strategy, then the cost of $\mathbf{x}^{**}$ would be bounded below by 0, but the benefit of $\mathbf{x}^{**}$ would be bounded above by a sequence $\{\tilde{B}_n^{**}\}$ such that $\lim_{n\to\infty}\tilde{B}^{**}/(n-m) = 0$. Thus, $\mathbf{x}^*$ outperforms $\mathbf{x}^{**}$, which is a contradiction.

□

Proof of Proposition 2. Fix b. We show that $\Theta(n/f(n))$ Bayesians are both necessary and sufficient for imperviousness:

1. Necessary: Suppose there are fewer than $A \cdot n/f(n)$ Bayesians for any constant A. By the pigeonhole principle, there must exist a DeGroot agent i^* with at least $\rho_n f(n)$ DeGroots before him after the previous Bayesian agent j^*, with $\lim_{n\to\infty} \rho_n = \infty$. Suppose the principal manipulates all of these DeGroot agents (i.e., $\gamma_i = 1$ for all $i < i^*$ after the previous Bayesian j^*). Then,

$$\begin{aligned}
\mathcal{D}_{i^*}(\mathbf{1}_D) &= 1 - \prod_{\kappa=j^*+1}^{i^*} (1-\theta_\kappa^{(n)}) \\
&= 1 - \prod_{\kappa=j^*+1}^{i^*} (1-\beta_\kappa f(n)) \\
&\geq 1 - (1-\underline{\beta} f(n))^{i^*-j^*} \\
&\geq 1 - (1-\underline{\beta} f(n))^{\rho_n f(n)}
\end{aligned}$$

   We consider separately the cases that $\limsup_{n\to\infty} f(n) > 0$ and $\limsup_{n\to\infty} f(n) = 0$. In the former case, we have an infinite subsequence of networks such that for some $\eta > 0$, $\mathcal{D}_{i^*}(\mathbf{1}_D) \geq 1-(1-\underline{\beta}\eta)^{\rho_n \eta} \to 1$. In the latter case, we note that $\lim_{n\to\infty} f(n) = 0$, so $\lim_{n\to\infty}(1-\underline{\beta} f(n))^{\rho_n f(n)} \sim 1 - e^{-\rho_n \underline{\beta}} \to 1$. Finally, consider the last $\nu f(n)$ DeGroots along this chain for any constant $\nu \in (0,1)$. For each of these DeGroots (denoted i), the fomer case becomes $\mathcal{D}_i(\mathbf{1}_D) \geq 1 - (1 - \underline{\beta}\eta)^{(1-\nu)\rho_n \eta} \to 1$ and the latter case becomes $\mathcal{D}_i(\mathbf{1}_D) \geq 1 - e^{-(1-\nu)\rho_n \underline{\beta}} \to 1$ for n sufficiently large. Thus, a vanishingly small fraction of the population (less than $1/(1+\rho_n)$) are not manipulated along this chain, therefore we see the network is susceptible for all $\varepsilon < 1$.

2. Sufficient: Once again, we divide this into two cases. First, consider $\lim_{n\to\infty} f(n) = 0$. Suppose we sprinkle $A \cdot n/f(n)$ Bayesians such that $\lceil f(n)/A \rceil$ is the farthest distance between any two "adjacent" Bayesian agents along the ring. Then for all DeGroots i, letting $j^*(i)$ be the nearest Bayesian:

$$\begin{aligned}
\mathcal{D}_i(\mathbf{1}) &= 1 - \prod_{\kappa=j^*(i)+1}^{i} (1-\theta_\kappa^{(n)}) \\
&= 1 - \prod_{\kappa=j^*(i)+1}^{i} (1-\beta_\kappa f(n)) \\
&= 1 - (1-\overline{\beta} f(n))^{i-j^*(i)} \\
&\leq 1 - (1-\overline{\beta} f(n))^{\lceil f(n)/A \rceil}
\end{aligned}$$

   Then $\mathcal{D}_i(\mathbf{1}) \to 1 - e^{-\beta/A}$, which is also less than $(1-b)/2$ for sufficiently large A. Therefore, by Corollary 2, the network is impervious to manipulation.

   Now suppose that $\limsup_{n\to\infty} f(n) > 0$. Then there is an infinite subsequence of networks where $f(n)$ is uniformly bounded away from 0, and so $\Theta(n/f(n))$ Bayesians implies that we can stagger the Bayesians so that each DeGroot only neighbors a Bayesian. Thus, provided that $\theta_i^{(n)} \leq (1-b)/2$, DeGroot agent i is not manipulated even if $\gamma_i = 1$, which is implied by $b \leq$

$1 - 2\overline{\beta} \limsup f(n)$ given in Proposition 2.

□

Proof of Corollary 1. In the ring network, we have $\theta_i^{(n)} = 1/2$, so $f(n)$ is a constant with $b < 1-2(1/2) = 0$ with $\varepsilon < 1$. We can apply Proposition 2 to note that $\Theta(n)$ Bayesians are necessary and sufficient. The second part of the statement follows from Theorem 2 and Example 1, and of course noting that beliefs in the network of all DeGroot agents simply converges to $||\boldsymbol{\gamma}||_1/n$, so is susceptible to manipulation without at least one Bayesian. □

Proof of Proposition 3. For part (a), we note that by Proposition 6, limiting DeGroot beliefs of the true state y^* for $\boldsymbol{\theta} = \theta'\mathbf{1}$ are given by:

$$\boldsymbol{\pi}_\infty^D = (\mathbf{I} - \mathbf{A}_{\theta'}^{-1})(\mathbf{g}^* \otimes \boldsymbol{\theta}') \leq (\mathbf{I} - \mathbf{A}_{\theta'}^{-1})(\boldsymbol{\gamma}^* \otimes [\mathbf{1}_B \ \mathbf{0}_D])$$

We first prove that the asymptotic bound for DeGroot beliefs is continuous in θ' around $\theta' = 0$. Clearly the network preservation of $\mathbf{A}_{\theta'}$ is continuous in θ', so it is sufficient to prove that as $\theta' \to 0$, $\mathbf{I} - \mathbf{A}_{\theta'}$ is non-singular. First note that λ is an eigenvalue of $\mathbf{I} - \mathbf{A}_{\theta'}$ if and only if $1 - \lambda$ is an eigenvalue of $\mathbf{A}_{\theta'}$. Thus, it suffices to show that the eigenvalue of $\mathbf{A}_{\theta'}$ are uniformly bounded away from the unit circle as $\theta' \to 0$. We note that:

$$\lim_{\theta' \to 0} \mathbf{A}_{\theta'} = \begin{pmatrix} \mathbf{0}_{\mathcal{B}} \\ \mathbf{S} \end{pmatrix}$$

for some row-stochastic matrix $\mathbf{S}$. Then, for any vector $\mathbf{v}$ such that $||\mathbf{v}||_2 = 1$, note that

$$\mathbf{v}' \equiv \begin{pmatrix} \mathbf{0}_{\mathcal{B}} \\ \mathbf{S} \end{pmatrix} \mathbf{v} = \begin{pmatrix} \mathbf{0}_{\mathcal{B}} \\ \mathbf{S}\mathbf{v} \end{pmatrix}$$

Note that $||\mathbf{v}'||_2 \leq ||\mathbf{v}||_2$, which holds with equality only if $\mathbf{v}_{\mathcal{B}} = \mathbf{0}_{\mathcal{B}}$. However, this is a contradiction by definition of our RHS vector. Thus, $(\mathbf{I} - \mathbf{A}_{\theta'})^{-1}$ is a continuous operation at $\theta' = 0$. But notice that when we substitute $\theta' = 0$, applying DeGroot centrality and noting the characteristic-vector $\boldsymbol{\gamma} \to \mathbf{0}$ shows that DeGroot centrality tends to 0, so beliefs of the correct state tend toward 1. Then applying continuity yields the claim in (a).

Because $\lim_{\theta' \to 1} \mathbf{A}_{\theta'} = \mathbf{0}$, it is obvious that beliefs are continuous at $\theta' = 1$. Moreover, when $\theta' = 1$, an amendable DeGroot agent i is manipulated if and only if $\gamma_i = 1$, which is profitable if and only if $\varepsilon < 1$. Call the strategy of targeting all amenable DeGroots as $\mathbf{x}_{\text{amen}}$, which has a net utility of $(1-\varepsilon)(n-m)$. If $b < 1/2$, then (c) holds vacuously; to show (b), we just note by continuity that there exists some θ^{**} such that the network with $\theta' \in (\theta^{**}, 1)$ is either impervious (if $\varepsilon < 1$) or susceptible (if $\varepsilon > 1$) independent of θ'. Setting $\theta^* = \theta^{**}$ and $\overline{\theta} = (1 + \theta^{**})/2$ gives us (b).

Now consider $b > 1/2$ and let $\theta^* = 1/2$. Suppose the principal chooses $\mathbf{x}_{\text{amen}}$ with the only difference being that he does not target the DeGroot agent not adjacent to any Bayesians; call this $\mathbf{x}_{\text{spec}}$. By just considering first-order walks, we see that the DeGroot centrality of this agent is at least $(1-\theta^*)\theta^* = 1/4$, so this agent is still manipulated under $\mathbf{x}_{\text{spec}}$. Similarly since all other DeGroot agents *are* targeted and have $\theta = 1/2$, these agents are also manipulated. Therefore the net utility of strategy $\mathbf{x}_{\text{spec}}$ is $(1-\varepsilon)(n-m) + \varepsilon$, which beats $\mathbf{x}_{\text{amen}}$. Let $\overline{\theta}$ be the infimum of all $\theta > 1/2$ where agent i is manipulated if and only if $\gamma_i = 1$ for all i (call this proprty **Independence**); we know such an infimum exists because independence holds at $\theta' = 1$. We claim that for all $\theta' \in (\overline{\theta}, 1)$, independence holds. To see this, it is sufficient to show that if independence holds with some θ_1', then independence holds for any $\theta_2' > \theta_1'$. By way of contradiction, consider some strategy $\mathbf{x}_2$ which violates independence with θ_2. This implies that for some DeGroot i^*, the sum of weighted walks to other DeGroots j with $\gamma_j = 1$

exceeds $(1-b)/2$ with θ_2. However, the sum of weighted walks with θ_1 is *necessarily* larger, because $\alpha_{ij,1} > \alpha_{ij,2}$ for all i, j. Thus, $\mathbf{x}_2$ violates independence under θ_1', a contradiction.

By construction, there exists some ε^* such that $\theta' \in (\theta^*, \overline{\theta})$ is susceptible (because $\mathbf{x}_{\text{spec}}$ dominates $\mathbf{0}$) but where $\mathbf{x}_{\text{amen}}$ is dominated by $\mathbf{0}$. Also by our previous observation, for $\theta' \in (\overline{\theta}, 1)$, an agent is manipulated if and only if $\gamma_i = 1$, so the network is impervious if and only if $\varepsilon > 1$, which holds for ε^*. Therefore, these $\theta^*, \overline{\theta}$ satisfy (b) and (c). □

Proof of Proposition 4. We will appeal to the first part of Corollary 2. Let $j_2^* \in D_2$ be the agent in D_2 adjacent to an agent $j_1^* \in D_1$. Now consider an arbirary agent $j \in D_1$. Since D_1 is strongly connected, there exists a walk between j and j_1^*, which implies there is also a walk from j to j_2^*; let us denote this walk by $W_{jj_2^*} = j \to v_1 \to \cdots \to v_k \to j_1^* \to j_2^*$. Suppose $\theta_1 \in [0, \overline{\theta})$ for some $\overline{\theta} < 1$. Let us write the weight of this walk explicitly as:

$$w_{jj_2^*} = \theta_2 \prod_{(v_i \to v_{i+1}) \in W_{jj_2^*}} (1-\theta_1)\alpha_{v_i, v_{i+1}} > C_{jj_2^*} > 0$$

where the constant $C_{jj_2^*}$ does not depend on θ_1. If we take $\overline{b} = 1 - 2\min_{j \in D_1} C_{jj_2^*} < 1$, then we see that for all $b > \overline{b}$, all $j \in D_1$ have DeGroot centrality $\mathcal{D}_j(\mathbf{1}_D) \geq w_{jj_2^*} \geq C_{jj_2^*} \geq (1-b)/2$. Thus, all agents in D_1 are manipulated when ε is sufficiently small, regardless of their θ_1, and in particular as $\theta_1 \to 0$. On the other hand, all agents in D_2 have $\theta_2 \geq \min_{j \in D_1} C_{jj_2^*}$, so by the same argument agents in D_2 are manipulated.

The second result is just a rephrasing of Proposition 3(a). □

D.2 Supplementary

Lemma 1. *In every equilibrium, the principal chooses* $\mathbf{x} = \mathbf{0}$ *when* $y = R$.

Proof of Lemma 1. By the same arguments as in the reputation literature (see Fudenberg and Levine (1989)), the principal can do no worse than mimicking the committed truthful type who implements $\mathbf{x} = \mathbf{0}$ regardless of y, as $T \to \infty$. Of course, when $y = R$ and $\mathbf{x} = \mathbf{0}$, all agents j learn the true state and take action $a_j = R$ at time T. This yields a payoff for the principal of $n \cdot 1 - \varepsilon \cdot 0$, which is maximal. Any other pure strategy cannot exceed a payoff of $n - \varepsilon$, and therefore $\mathbf{x} = \mathbf{0}$ is the unique equilibrium outcome when $y = R$. □

Lemma 2. *For generic* $\mathbf{A}$ *and all finite* t*, every Bayesian agent discovers:*

(a) $\{g_{j,\Delta\tau}(h_{j,\Delta\tau})\}_{\tau'=1}^{t/\Delta}$ *for all DeGroots with* $\theta_j > 0$*,*

(b) The terminal belief $\pi_{j,t}$ *for* <u>all</u> *agents* j*,*

connected to her through a path of only DeGroots and with $\theta_j > 0$*. In particular, it can deduce* $\pi_{j,\infty}$ *for all agents* j *connected to her through a path of only DeGroots.*

Proof of Lemma 2. For simplicity of exposition, let time t be discrete and denote the increments on length Δ where DeGroots exchange information (i.e., consider $t = 0, \Delta, 2\Delta, \ldots, \tau\Delta, \ldots$).

Part 1. First, we show that $g_{j,t}(h_{j,t})$ for DeGroots j and $\pi_{j,t}$ for Bayesians j can take on at most countably many values. We note that the difference between $z_{j,t}^S$ and $z_{j,t}^R$ is a sufficient statistic for $g_{j,t}(h_{j,t})$. Since this difference is an integer, there are at most countable values for $g_{j,t}(h_{j,t})$. To show Bayesian beliefs must come from a finite set at each timestamp t, note that an equivalent learning model is one where all Bayesians update beliefs continuously at $t - dt$, observe DeGroot beliefs, then

reupdate at $t+dt$ (with the first and last learning "sub-periods" ending in finite time). Since only the total number of $z^S_{j,t}$ and $z^R_{j,t}$ matter for any agent j in the beliefs of Bayesian agent i, we conclude as well that there are at most countably many beliefs for $\pi_{i,t}$ for any Bayesian i at time t.

Part 2. We take an approach similar to the one in Mueller-Frank (2014), but applied to our more complicated setting;. Fix some agent i and denote by $\mathcal{N}^k(i)$ the be the k-order neighborhood of i passing through DeGroots (i.e., the agents k "hops" away through a path of DeGroots). We let the k-action neighborhood of agent i, denoted by $\mathcal{Q}_{i,k}$ be $\{g_j(h_{j,1}), g_j(h_{j,2}), \ldots, g_j(h_{j,t-k}\}$ for any DeGroot $j \in \mathcal{N}^k(i)$ and $\{\pi_{j,1}, \pi_{j,2}, \ldots, \pi_{j,t-k}\}$ for any Bayesian agent $j \in \mathcal{N}^k(i)$. We show that in each period t, there exists a generic set of $\mathbf{A}_{n-m,m}$ network weights so that any Bayesian agent i knows with certainty its k-action neighborhood. In other words, the Bayesian can deduce the personal-experience belief of every DeGroot and the overall beliefs of every Bayesian in its k-order neighborhood, for all intervals of time except the last k, for all t. We prove this by induction on t, by noting that $\pi_{j,t}$ is a polynomial of degree t in the weights of $\mathbf{A}_{n-m,m}$, and a function of j's $(t-1)$-action neighborhood. When $t=1$, the Bayesian can observe her own neighborhood, so knows $\pi_{j,1}$ for every Bayesian j and knows that:

$$\pi_{j,1} = \theta_j g_j(h_{j,1}) + \sum_{\ell=1}^{n} \alpha_{j\ell} q \Longrightarrow g_j(h_1) = \frac{\pi_{j,1} - (1-\theta_j)q}{\theta_j}$$

and of course the above is a linear function (i.e., polynomial of degree 1). Now suppose the statement is true for t; we want to show it holds for $t+1$. We can clearly know $\pi_{j,t+1}$ for any Bayesian agents in our first-order neighborhood. For any DeGroot j in i's neighborhood, we write:

$$\pi_{j,t+1} = \theta_j g_j(h_{j,t+1}) + \sum_{\ell=1}^{n} \alpha_{j\ell} \pi_{\ell,t}$$

By the inductive hypothesis, we can express $\pi_{\ell,t}$ as a polynomial of degree t in the network weights $\mathbf{A}_{m-n,n}$ and as a function of its t-action neighborhood. Let us write this polynomial as $\Gamma^t_j(\mathbf{A}_{m-n,n}, \mathcal{Q}_{j,t})$; then:

$$\pi_{j,t+1} = \theta_j g_j(h_{j,t+1}) + \sum_{\ell=1}^{n} \alpha_{j\ell} \Gamma^t_\ell(\mathbf{A}_{m-n,n}, \mathcal{Q}_{\ell,t-1})$$

which shows the first-part of the inductive hypothesis, which is that $\pi_{j,t+1}$ can be expressed as a $t+1$-order polynomial in the network weights $\mathbf{A}_{m-n,n}$ and as a function of j's t-action neighborhood, $\pi_{j,t+1} = \Gamma^{t+1}_j(\mathbf{A}_{m-n,n}, \mathcal{Q}_{j,t})$. Consider any two distinct t-action neighborhoods for agent j, $\mathcal{Q}_{j,t}$ and $\mathcal{Q}'_{j,t}$. Then define:

$$\mathcal{L}^t_j(\mathbf{A}_{m-n}, n) = \Gamma^{t+1}_j(\mathbf{A}_{m-n,n}, \mathcal{Q}_{j,t}) - \Gamma^{t+1}_j(\mathbf{A}_{m-n,n}, \mathcal{Q}'_{j,t})$$

Note that $\mathcal{L}^t_j(\mathbf{A}_{m-n}, n)$ is a polynomial (of degree $t+1$), and by the same Lemma as in Mueller-Frank (2014), the set of weights $\mathbf{A}_{m-n,n}$ which make the above expression vanish has measure zero. Therefore, for any *generic* set of $\mathbf{A}_{m-n,n}$, the two t-action neighborhoods for agent j, $\mathcal{Q}_{j,t}$ and $\mathcal{Q}'_{j,t}$, are distinguishable entirely by $\pi_{j,t+1}$. By the fact that the number of agents in the network is finite and the set of possible k-action neighborhoods is at most countable implies the set of network weights which allow i to distinguish between *any two* t-action neighborhoods for j is also generic. Because agent i has a finite number of neighbors, this implies that at time $t+1$, agent i can construct his $\mathcal{Q}_{i,t+1}$ for a generic set of $\mathbf{A}_{m-n,n}$, completing the inductive step. Finally, the set of $\mathbf{A}_{m-n,n}$ which allow all agents to distinguish between t-action neighborhoods for j is generic; taking an intersection over all t preserves genericity because it is the complement of a countable union of countable sets. Thus, each Bayesian agent knows its t-action neighborhood at time t.

Part 3. Finally, we show that if a Bayesian knows its t-action neighborhood at time t, it can deduce

limit beliefs $\pi_{j,t}$ eventually for all j connected through a path of DeGroots. Since the diameter of the network is finite (because there are a finite number of agents), as $T \to \infty$, each Bayesian observes an unlimited sequence of $\{g_{j,t}(h_{j,t})\}$ for DeGroots j and an unlimited sequence of beliefs $\{\pi_{j,t}\}$ for Bayesians j, *given they are connected to* i *through other DeGroots*. Clearly, the Bayesian learns the asymptotic belief of other Bayesians because these beliefs are observed directly in $\mathcal{Q}_{i,t}$, and the beliefs must converge by the martingale convergence theorem. Similarly, by the MCT, it must be the case that $g_{j,t}(h_{j,t}) \to g_j^*(h_{j,\infty})$ for all DeGroot agents, because $g_{j,t}$ is a Bayesian update on j's own signals obtained by time t. Finally, we note that we can partition the network $\mathbf{G}$ into components of DeGroot agents who are not connected to each other if all Bayesians were removed from the network. It is clear from Proposition 6 that taking as given the beliefs of the Bayesians who separate the components, and the g_j^* for all DeGroots j in the component, the asymptotic belief for each agent j can be computed. Moreover, if Bayesian i has a path through DeGroots to agent j, then Bayesian i knows all of this in its limiting t-action neighborhood, $\lim_{t\to\infty} \mathcal{Q}_{i,t}$. □

Proof of Theorem 3. We rewrite the game defined in Section 2 as an extensive form Bayesian game. Existence thus follows from Theorem 8.5 in Fudenberg and Tirole (1991) for dynamic, finite games of incomplete information. The game is depicted in Figure 11, consisting of three time periods and $m+1$ players (where m is the number of Bayesians). At time $t=0$, nature draws the state of nature y and the type of player 1 (the principal) which is truthful ($\mathcal{T}$) with probability μ_0 and strategic ($\mathcal{S}$) with probability $1-\mu_0$. While player 1 observes y, none of the other m players do. At time $t=1$, the principal chooses a strategy σ^p over his action set of network influence strategies $\mathbf{x}$; the action set of the truthful principal is a singleton, $\{\mathbf{0}\}$, whereas the action set of the strategic principal is $\{0,1\}^n$. At time $t=2$, Bayesian agents receive multi-dimensional signals $s_i \in ([0,1])^{|\mathcal{N}_i|\times[0,T)}$ from a joint distribution G_i, conditional on $\mathbf{x}$. This joint distribution reflects the observation of beliefs in the Bayesians' neighborhood. Then Bayesian agents play a strategy over their action set $\{\mathbf{R},\mathbf{S}\}$.

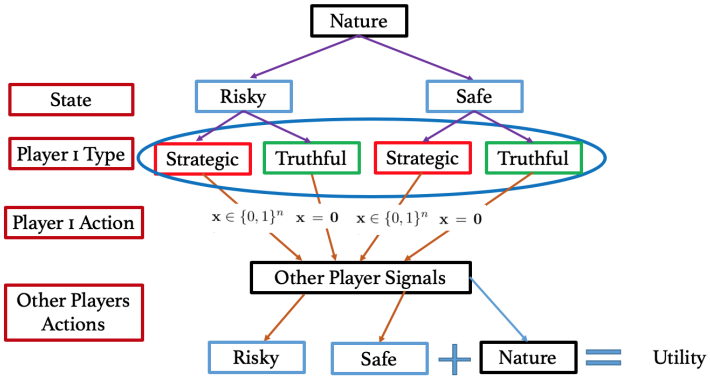

Figure 11. Extensive-Form Representation.

Payoffs are given as follows. For the Bayesian agents, their utilities are given by Table 1. The principal receives two payoffs, one determined by nature and one determined by the play of the Bayesians, and always pays the cost of the network influence $\varepsilon||\mathbf{x}||_1$. The payoff from the Bayesians is given directly by Table 1, and is additive across Bayesians. For the other payoff, nature generates a signal $\tilde{s} \in [0,1]^{n-m}$, according to a distribution $\tilde{G}$ corresponding to the random beliefs of the DeGroot agents in the learning process, which again is conditional on $\mathbf{x}$ *and* may depend on the realization of $\{s_i\}_{i\in\mathcal{B}}$. The nature payoff is then given by $\sum_{j=1}^{n-m} 1_{\tilde{s}^{(i)}\geq(1-b)/2}$, where $\tilde{s}^{(i)}$ is the i-th component of $\tilde{s}$.

Proof of Proposition 5. Suppose that agent i has belief $\pi_i(\mathbf{R})$ at time T. Agent i's best-response is the action $a_i = \mathbf{R}$ if $\pi_i(\mathbf{R}) > (1-b)/2$, $a_i = \mathbf{S}$ if $\pi_i(\mathbf{R}) < (1-b)/2$, or any strategy in the simplex $\Delta(\mathbf{R}, \mathbf{S})$ if $\pi_i(\mathbf{R}) = (1-b)/2$. Therefore, the equilibrium play of the agents in the terminal stage is pinned-down as a function of terminal beliefs.

By Theorem 1 and Proposition 6, as $T \to \infty$, the beliefs of all agents converge almost surely to some $\boldsymbol{\pi}_\infty$, given a network action $\mathbf{x}$. We can construct a set B^* which consists of all the values of b where some agent i has a limit belief $\lim_{t\to\infty} \pi_{i,t} \to (1-b)/2$, for some network action $\mathbf{x}$. Because there are finitely many agents and finitely many network actions, the set B^* is finite, so has measure zero, implying that $(-1,1)\backslash B^*$ has full measure.

Finally, consider fixing some $\mathbf{x}$ and any $b \in (-1,1)\backslash B^*$. Given fixed λ, for every $\kappa > 0$, there exists T^* such that for all $T > T^*$, the probability that all beliefs at time T are within λ of their limits is at least $1-\kappa$:

$$\mathbb{P}[||\boldsymbol{\pi}_T - \boldsymbol{\pi}_\infty||_\infty < \lambda] \geq 1 - \kappa$$

by Theorem 1 and Proposition 6. Since the set of B^* contains no b's with an agent holding $\pi_{i,\infty} = (1-b)/2$, we can pick T^* large enough such that λ is small enough whereby each agent i plays a known action a_i with probability at least $1-\kappa$ at time T. Choosing action $\mathbf{x}$ gives the principal a known net payoff of $\mathcal{M}(\mathbf{x}) - \varepsilon||\mathbf{x}||_1$ with probability $1-\kappa$ (which we deem the "likely payoff") and some other payoff with probability κ.

Now suppose two network strategies $\mathbf{x}_1, \mathbf{x}_2$ have a different number of manipulated agents. If $\mathbf{x}_1$ and $\mathbf{x}_2$ give the same likely payoff, this implies that $\mathcal{M}_1(\mathbf{x}_1) - \varepsilon||\mathbf{x}_1||_1 = \mathcal{M}(\mathbf{x}_2) - \varepsilon||\mathbf{x}_2||_1$, which implies that:

$$\varepsilon = \frac{\mathcal{M}(\mathbf{x}_1) - \mathcal{M}(\mathbf{x}_2)}{||\mathbf{x}_1||_1 - ||\mathbf{x}_2||_1}$$

because $||\mathbf{x}_1||_1 \neq ||\mathbf{x}_2||_1$. Noting that both the numerator and denominator are integers, we see that by taking the generic set of irrational ε, we guarantee that whenever $\mathbf{x}_1$ and $\mathbf{x}_2$ have a different number of manipulated agents, the principal has a strictly higher likely payoff under one. Since we took κ to be arbitrary, we can choose κ small (by increasing T) such that the principal prefers action $\mathbf{x}_1$ to $\mathbf{x}_2$ if he prefers the likely payoff of $\mathbf{x}_1$ to the likely payoff of $\mathbf{x}_2$ (as the payoff of any "unlikely" outcome is bounded above by n). Thus, for the set of irrational ε and $b \in (-1,1)\backslash B^*$, the principal plays a randomized strategy over network actions which induces the "likely" outcome of that network action with probability at least $1-\kappa$. Each of the network actions in the support of this randomized strategy must have the same number of of manipulated agents. This holds for arbitrary $\kappa > 0$ as T grows large. □

Proof of Proposition 6. By convention, let the first m agents in the network be Bayesian. Each DeGroot agent updates its belief according to the law of motion:

$$\boldsymbol{\pi}^D_{t+1} = \boldsymbol{\theta} \otimes \mathbf{g}(\mathbf{h}_t) + \mathbf{A}_{DB}\boldsymbol{\pi}^B_t + \mathbf{A}_{DD}\boldsymbol{\pi}^D_t$$

By the martingale convergence theorem, we know that $\boldsymbol{\pi}^B_t$ converges almost surely to some $\boldsymbol{\pi}^B_\infty$. Moreover we know that $\boldsymbol{\theta} \otimes \mathbf{g}^*$ is eventually a constant almost surely. To see this, consider the following four cases for each DeGroot agent i:

1. $x_i = 0$: By Assumption 1 either $p_i > 1/2$, so Bayesian update will converge to a point mass on the true state y, or $\theta_i = 0$ and so is identically zero for all t.

2. It is impossible that $\lambda_i \geq \lambda_i^*/(2p_i - 1)$ by Assumption 2(b) for any agent i.

3. If $\lambda_i < \lambda_i^*/(2p_i - 1)$ and $x_i = 1$, the Bayesian update will converge to a point mass on the princi-

pal's influence state $\hat{y}$.

Therefore, for any $\kappa > 0$, we can write for sufficiently large T:

$$\begin{aligned}\boldsymbol{\pi}^D_{t+1} - \boldsymbol{\pi}^D_t &= \boldsymbol{\theta} \otimes (\mathbf{g}(\mathbf{h}_{t+1}) - \mathbf{g}(\mathbf{h}_t)) + \mathbf{A}_{DB}(\boldsymbol{\pi}^B_t - \boldsymbol{\pi}^B_{t-1}) + \mathbf{A}_{DD}(\boldsymbol{\pi}^D_t - \boldsymbol{\pi}^D_{t-1}) \\ &\leq \kappa \mathbf{1} + \mathbf{A}_{DD}(\boldsymbol{\pi}^D_{t+1} - \boldsymbol{\pi}^D_t)\end{aligned}$$

for all $t > T$. As an aside, we prove that $(\mathbf{I} - \mathbf{A}_{DD})$ is invertible. It suffices to prove that all the eigenvalues of $\mathbf{A}_{DD}$ lie strictly within the unit circle, in which case all eigenvalues of $(\mathbf{I} - \mathbf{A}_{DD})$ are bounded away from zero. Denote by $Q_i = 1 - \sum_{j=m+1}^{n} \alpha_{ij}$ the amount of weight placed on one's experience and the Bayesian agents, combined, which by the assumption that each $\theta_i > 0$, is strictly positive. Take the matrix:

$$\mathbf{Q} = \begin{pmatrix} 1/Q_{m+1} & 0 & \cdots & 0 \\ 0 & 1/Q_{m+2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/Q_n \end{pmatrix}$$

Then we note that $\mathbf{A}_{DD}\mathbf{Q}$ is row-stochastic, so by the Perron-Frobenius theorem all eigenvalues lie strictly within the unit circle except for the largest, which is exactly equal to 1. Consider any arbitrary vector $\mathbf{v} \in \mathbb{R}^{n-m}$:

$$||\mathbf{A}_{DD}\mathbf{v}||_2 < ||\mathbf{A}_{DD}\mathbf{Q}\mathbf{v}||_2 \leq ||\mathbf{v}||_2$$

where the strict inequality follows from the fact that all eigenvalues of $\mathbf{Q}$ are strictly greater than 1. Back to the original claim, for any $\kappa > 0$, there exists T sufficiently large such that:

$$\boldsymbol{\pi}^D_{t+1} - \boldsymbol{\pi}^D_t \leq \kappa(\mathbf{I} - \mathbf{A}_{DD})^{-1}\mathbf{1}$$

which implies that $\boldsymbol{\pi}^D_t$ must converge almost surely to some $\boldsymbol{\pi}^D_\infty$. This implies that $\boldsymbol{\pi}_\infty$ must solve the fixed-point problem:

$$\boldsymbol{\pi}^D_\infty = \boldsymbol{\theta} \otimes \mathbf{g}^* + \mathbf{A}_{DB}\boldsymbol{\pi}^B_\infty + \mathbf{A}_{DD}\boldsymbol{\pi}^D_\infty$$

If not, then the difference between the left-hand side and right-hand side is always some positive amount η for at least one ω, and so every iteration of belief updating changes the belief of type ω by at least η, contradicting convergence. By setting $g_i(\sigma^p(\omega))$ to $\pi^B_{i,\infty}$ and $\theta_i = 1$ for all Bayesian agents i (which is the correct belief of the Bayesians by Theorem 1), we can reduce this expression to:

$$\boldsymbol{\pi}_\infty = \boldsymbol{\theta} \otimes \mathbf{g}^*(\mathbf{a}^p) + \mathbf{A}\boldsymbol{\pi}_\infty$$

Note that $\mathbf{A}_{DD}$ has all eigenvalues lying (strictly) in the unit circle if and only if $\mathbf{A}$ does. Therefore, we can solve this fixed-point problem explicitly:

$$\boldsymbol{\pi}_\infty = (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{g}^* \otimes \boldsymbol{\theta})$$

which proves the claim of Proposition 6. □

Proof of Theorem 4. As we saw in Theorem 3 and Proposition 5, the equilibrium play of the DeGroots is pinned-down by their beliefs, which when T is large is high probability close to its limit. By Theorem 1 and Proposition 6, the DeGroot centrality $\mathcal{D}(\boldsymbol{\gamma})$ is equivalent to the belief $\boldsymbol{\pi}_\infty(R)$ when $y = S$ and $\mathbf{g}^* = \boldsymbol{\gamma}$. We let z_i denote an agent i who is manipulated at the limit (and thus for large T). Recall that by Theorem 1, no Bayesian agent is manipulated so we can set $z_i = 0$ for all $i \in \{1, \ldots, m\}$. Similarly, we suppose the principal can "elect" to manipulate agent i only if its DeGroot centrality is

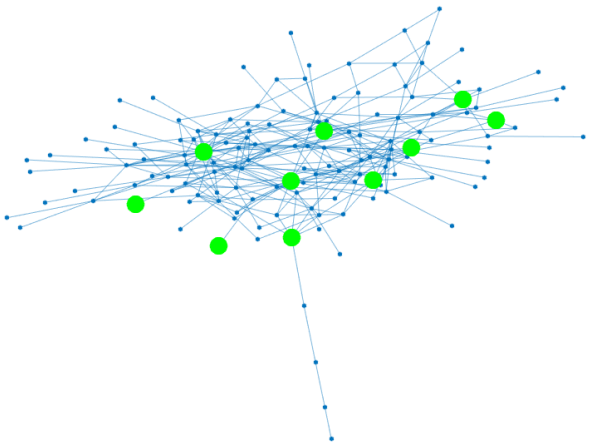

Figure 12. Ten Well-Placed Bayesians, $b = .8$.

above $(1-b)/2$; in other words:

$$z_i \leq 1 + \mathcal{D}_i(\boldsymbol{\gamma}) - (1-b)/2 = \mathcal{D}_i(\boldsymbol{\gamma}) + (1+b)/2$$
$$z_i \in \{0,1\}$$

Finally, note that the principal gains an additional payoff of 1 for each manipulated agent and pays a cost of ε for each γ_i. Combining these we get the integer program in Theorem 4. Every $\mathbf{x}$ except those that try to target Bayesians can be represented as $(\mathbf{z}, \boldsymbol{\gamma})$, but such $\mathbf{x}$ are dominated by another network action because of Theorem 1. Similarly, each feasible $(\mathbf{z}, \boldsymbol{\gamma})$ corresponds to some network action $\mathbf{x}$, as given in Section 3. □

Proof of Corollary 2. Consider any set $\mathcal{K}$ of amenable DeGroot agents. Because $(\mathbf{I} - \mathbf{A})^{-1}$ consists of all nonnegative entries, we know that $\mathcal{D}(\mathbf{1}_\mathcal{K}) < \mathcal{D}(\mathbf{1}_\mathcal{D})$. Under (a), every feasible solution requires that $\mathbf{z} = \mathbf{0}$. Therefore, the IP objective is maximized if and only if $\boldsymbol{\gamma} = \mathbf{0}$, which implies the network is impervious. On the other hand, suppose (b) holds. Then $\boldsymbol{\gamma} = \mathbf{1}_\mathcal{K}$ and $\mathbf{z} = 1_{\mathcal{D}_i(\mathbf{1}_\mathcal{K}) > (1-b)/2}$ is a feasible solution to the IP, and the objective yields $||\mathbf{z}||_1 - \varepsilon||\boldsymbol{\gamma}||_1 > 0$ by the assumption in (b). Thus, the feasible solution $(\boldsymbol{\gamma}, \mathbf{z}) = \mathbf{0}$ does not maximize the IP as it gives an objective of 0, so $\mathbf{0} \notin \boldsymbol{\Gamma}^*$, and thus the network is susceptible. □

E Additional Figures

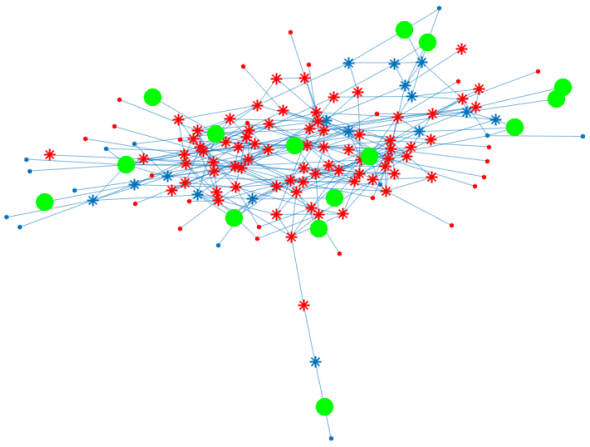

Figure 13. Fifteen Poorly-Placed Bayesians, $b = .8$.

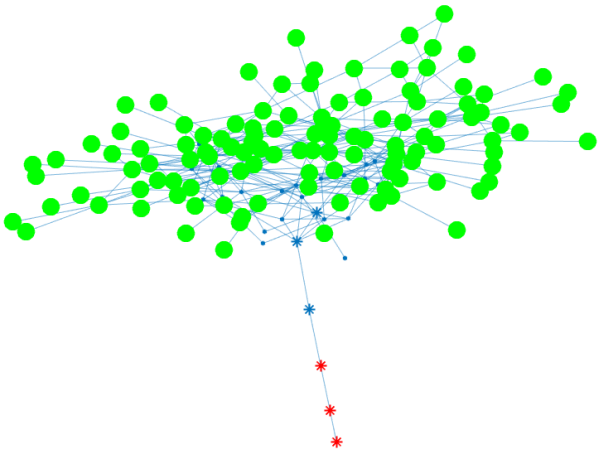

Figure 14. With 116 Bayesians, $b = .8$.

References

Acemoglu, Daron, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar (2013), "Opinion fluctuations and disagreement in social networks." *Mathematics of Operations Research*, 38, 1–27.

Acemoglu, Daron, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar (2011), "Bayesian learning in social networks." *The Review of Economic Studies*, 78, 1201–1236.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992), "A theory of fads, fashion, custom, and cultural change as informational cascades." *Journal of political Economy*, 100, 992–1026.

Broniatowski, David A, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze (2018), "Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate." *American journal of public health*, 108, 1378–1384.

Candogan, Ozan and Kimon Drakopoulos (2017), "Optimal signaling of content accuracy: Engagement vs. misinformation."

Chandrasekhar, Arun G, Horacio Larreguy, and Juan Pablo Xandri (2015), "Testing models of social learning on networks: Evidence from a lab experiment in the field." Technical report, National Bureau of Economic Research.

Fudenberg, Drew and David Levine (1989), "Reputation and equilibrium selection in games with a patient player." *Econometrica*, 57, 759–78.

Fudenberg, Drew and Jean Tirole (1991), *Game Theory*. MIT Press.

Golub, Benjamin and Matthew O Jackson (2010), "Naive learning in social networks and the wisdom of crowds." *American Economic Journal: Microeconomics*, 2, 112–49.

Jackson, Matthew O., Tomas Rodriguez-Barraquer, and Xu Tan (2012), "Social Capital and Social Quilts: Network Patterns of Favor Exchange." *American Economic Review*, 102, 1857–1897, URL https://www.aeaweb.org/articles?id=10.1257/aer.102.5.1857.

Jadbabaie, Ali, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi (2012), "Non-bayesian social learning." *Games and Economic Behavior*, 76, 210–225.

Kamenica, Emir and Matthew Gentzkow (2011), "Bayesian persuasion." *American Economic Review*, 101, 2590–2615.

Kreps, David M and Robert Wilson (1982), "Reputation and imperfect information." *Journal of economic theory*, 27, 253–279.

Milgrom, Paul and John Roberts (1982), "Predation, reputation, and entry deterrence." *Journal of economic theory*, 27, 280–312.

Morris, Stephen (2001), "Political correctness." *Journal of political Economy*, 109, 231–265.

Mueller-Frank, Manuel (2014), "Does one bayesian make a difference?" *Journal of Economic Theory*, 154, 423–452.

Papanastasiou, Yiangos (2018), "Fake news propagation and detection: A sequential model."

Pennycook, Gordon and David G Rand (2018), "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition*.

Rosenberg, Dinah, Eilon Solan, and Nicolas Vieille (2009), "Informational externalities and emergence of consensus." *Games and Economic Behavior*, 66, 979–994, URL https://ideas.repec.org/a/eee/gamebe/v66y2009i2p979-994.html.

Yildiz, Ercan, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione (2013), "Binary opinion dynamics with stubborn agents." *ACM Transactions on Economics and Computation (TEAC)*, 1, 19.