When is Society Susceptible to Manipulation?

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 the social norms and network structures that 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 develop a new centrality measure and describe how our model offers insights into designing networks that are resistant to manipulation.

1 Introduction

People's beliefs can directly impact their actions. These beliefs are usually formed through a combination of individual and social learning, and a large literature details conditions under which learning aggregates beliefs in a way that leads agents to correctly learn an underlying state of the world.

An ability to shape beliefs implies an ability to steer agents towards taking specific actions. In this paper, we consider a social learning environment where agents try to learn an underlying state in order to make a one-time choice between different actions. Agents receive private signals about the state and use these signals in addition to the information they obtain from their neighbors to update their beliefs. A strategic principal is interested in having agents take a certain action, and can try to influence the beliefs in the network by sending costly (and misleading) signals to some of the agents. Agents cannot differentiate whether a signal they are receiving is 'organic' or coming from the principal. Social learning therefore provides a positive externality as agents spread organic news, but also a negative externality as agents unknowingly spread misinformation from the principal. Some agents are *stubborn* – they are endowed with knowledge of the true state and only spread correct information. Being connected to these agents can therefore offer some protection from the influence of the principal.

We say that an agent is manipulated if her beliefs converge to the true state and she takes the correct action in the absence of interference from the principal, but chooses the wrong action due to incorrect beliefs when the principal interferes with the learning process. Our interest is in characterizing the conditions under which the principal can use social learning to his advantage in order to manipulate the agents, and in understanding the social norms and network structures that help or hinder this spread of misinformation in society.

Contribution and Overview of Results. We provide a classification of networks that describes when manipulation is possible. Agents in our model use DeGroot updating to aggregate the beliefs of their neighbors with their own signals in a linear fashion. Theorem 1 provides a tight characterization of the beliefs of agents under any interference strategy by the principal, and Proposition 2 proves that these beliefs are related to a novel centrality measure that we call DeGroot centrality. We then show that depending on how agents weigh their own signals, a substantial fraction of the population can be tricked into believing that the underlying state is different from the actual state. Theorem 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 moderate society whose members use their own beliefs as well as other agents' opinions is the society that is most susceptible to this kind of manipulation.

For these moderate societies, the stubborn agents can help spread the truth about the underlying state, but their ability to do so is limited by the network structure. We classify networks into dense and sparse topologies, and show in Theorem 4 that dense networks are highly resistant to manipulation: even as the size of the network grows, the presence of a *constant* number of stubborn agents *anywhere* in the network is enough to guarantee imperviousness. By contrast, sparse networks are more susceptible to manipulation, and both the number *and* location of stubborn agents are important for the network to be impervious. In particular, the number of stubborn agents required may grow with the size of the network. If there are not enough stubborn agents, or if there is a sufficient number that 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.

We use the above results to provide a characterization of manipulation in networks that can be represented as a combination of sparse and dense networks, and show the existence of a *phase transition* in Theorem 5: as the network gets sufficiently dense, all opportunities for manipulation suddenly vanish. Proposition 7 shows that agents being skeptical about their news source does not necessarily

lead to better learning. We then extend our results on several dimensions in Section 6 and, for the interested reader, provide a numerical study in the appendix that examines the concepts in the paper on data from the advice network described in Jackson et al. (2012).

Related Literature. The agents in our model use DeGroot learning to update their beliefs. 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. Jadbabaie et al. (2012) consider agents that update their own information in a Bayesian fashion and aggregate the information of their neighbors in a DeGroot fashion, and their particular formulation of DeGroot agents is closely related to the one we consider in this paper. Bohren and Hauser (2017) examine learning when agents have a misspecified model of the world.

Our model also includes stubborn agents who hold correct beliefs about the state of the world. 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 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 beliefs, and the presence of stubborn agents who know the state is always useful for everyone in the network. Nevertheless, as we discuss, even with the positive contribution that these agents provide to the learning process, manipulation might still be unavoidable.

The proliferation of false news on social networks has been the central focus of some recent work. Candogan and Drakopoulos (2020) and Papanastasiou (2020) examine how (Bayesian) agents exchange information on a social network and show 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 misinformation into the network, and consequently there is no notion of manipulation. The idea that a principal can use social learning to manipulate agents towards taking a certain action has been studied in the context of replicator dynamics in Mostagir (2010). Unlike this work, we examine richer learning dynamics in an environment where some agents consistently spread correct information while others spread their beliefs without critical reasoning, which as Pennycook and Rand (2018) show in recent experimental work, might be one of the primary mechanisms through which misinformation spreads in social networks.

2 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 parameter $\lambda > 0$ for each agent i. We refer to this process as *organic news*. For simplicity, we assume agents digest news at the same times $\tau = 1, 2, \ldots$, which correspond to the arrivals of a single Poisson process, but might correspond to different articles (i.e., different messages) for different agents. For all $\tau \in \{1, 2, \ldots\}$, the organic news for agent i generates a signal $s_{i,\tau} \in \{S, R\}$ according to the distribution:

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

i.e., the signal is correlated with the underlying truth. All organic news' signals for agent i are independent across time and across other agents. 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 allow for the possibility that $p_i = 1/2$, so that agent i faces an identification problem and cannot rely on her organic news alone, but instead must rely on others in order to learn the true state.

Principal In addition to the organic news process, there is a strategic principal who may also generate news of his own. We assume, without loss of generality, that the true state is $y = S$ and the principal wants to convince agents of state R.[1] The principal picks an influence strategy $x_i \in \{0, 1\}$ for each agent i in the network. The principal then generates news of his own, which is always signal R, and the influence strategy indicates which agents receive these signals. 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^* which is received by all agents with $x_i = 1$.[2] The principal incurs an upfront investment cost $\varepsilon > 0$ for each agent with $x_i = 1$.

Once again, for simplicity we assume agents digest news at the same rates, so an agent i with $x_i = 1$ receives organic news at time τ with probability $\lambda/(\lambda + \lambda^*)$ and receives the principal's news with probability $\lambda^*/(\lambda + \lambda^*)$, but is unable to differentiate between the nature of the news. On the other hand, an agent i with $x_i = 0$ always receives organic news. An organic message always follows the distribution in (1), whereas a message from the principal always gives a signal of R, i.e., it is

[1] This is without loss because if the underlying state is indeed R, then as we establish in Proposition 1, agents will learn that state without interference from the principal, and therefore he will elect not to intervene.

[2] To simplify our setup, we do not allow the principal to send S messages, vary his influence strategy, or change the intensity of his messages over time. We explore how this affects our results in Section 6.4.

misinformation.

The principal can be one of two types. He can either be a strategic type $\mathcal{S}$ or a truthful type $\mathcal{T}$. 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.

Agents There are two types of agents in the model. *DeGroot* agents learn about the state by combining both (i) what they read in the news and (ii) what their friends believe about the state. *Stubborn* agents are endowed with knowledge of the true state y at $t = 0$, through being well-educated or knowledgeable about the subject. Stubborn agents will not change their beliefs over time. We denote the set of DeGroots as D and the set of knowledgeable stubborn agents as K. The total population in society is denoted by n, with m denoting the number of stubborn agents in that society.

Unlike stubborn agents, DeGroot agents start with prior q about the state y at $t = 0$, and must use their own signals combined with social learning to try and learn the state. Specifically, every DeGroot agent:

(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 with intensity λ and all signals are independent over time with $\mathbb{P}(s_{i,\tau} = y) = p_i$ (i.e., takes the news at face value).

The implicit assumption in the DeGroot learning process is that DeGroots are not aware of a principal who might be tampering with this process and sending misinformation. DeGroots absorb *all* news as if it is coming from organic sources. We relax this in Section 6.1, where agents try to simultaneously learn how trustworthy their news sources are, and can appropriately discount their own news if they suspect it is interfered with.

Formally, let $\pi_{i,t} \in \Delta(\{S, R\})$ represent the belief of agent i about the underlying state at time t. Given history $h_{i,t} = (s_{i,1}, s_{i,2}, \ldots, s_{i,\tau^*})$ up until time t (where τ^* is the last message received before 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 (where $z_{i,t}^S + z_{i,t}^R = \tau^*$); then DeGroot agent i has direct "personal experience":

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

and $\text{BU}(R|h_{i,t}) = 1 - \text{BU}(S|h_{i,t})$. The experience function BU 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 with no knowledge of principal interference.

DeGroot agents also form beliefs by talking to (and exchanging beliefs with) their neighbors after every unit of time. For all agents i, there are weights θ_i, α_{ij} such that agent i holds belief $\pi_{i,t}$ for all $t \in \{1, 2, \ldots\}$:

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

where $\theta_i + \sum_{j=1}^{n} \alpha_{ij} = 1$. As convention, we assume the link $i \to j$ indicates that agent j is a neighbor of i (i.e. i listens to j) but not necessarily vice versa. We refer to this as the *DeGroot update* (DU) process.

Network Structure Each agent i has a neighborhood $N(i) \subset \{1, \ldots, n\}$ that consists of other agents she listens to in every period (i.e., her "friends"). Note that because stubborn agents do not change their beliefs over time, their neighborhoods are immaterial. On the other hand, each DeGroot agent i's neighborhood is specified by her weights $(\theta_i, \{\alpha_{ij}\}_{j=1}^{n})$, with larger weights representing stronger connections. In matrix notation, we can represent the *influence* of the social network as:

$$\mathbf{W} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{DK} & \mathbf{A}_{DD} \end{pmatrix}$$

where $\mathbf{A}_{DK}$ is the DeGroot by stubborn agent weight matrix, given by entries $\{\alpha_{ij}\}_{i \in D,\, j \in K}$, and $\mathbf{A}_{DD}$ is the DeGroot by DeGroot agent weight matrix, given by entries $\{\alpha_{ij}\}_{i,j \in D}$. We refer to this social network as $\mathbf{G}$, and denote the weights w_{ij} from matrix $\mathbf{W}$.

It is also sometimes insightful to look at the unweighted representation of $\mathbf{G}$, which we denote by social network $\mathbf{G}^*$. The unweighted social network is a binary relation between pairs of agents representing whether agent i listens to agent j *at all*. By convention, a link $i \to j$ exists in the undirected social network $\mathbf{G}^*$ if and only if j is in i's neighborhood, i.e., $j \in N(i)$, because $\alpha_{ij} > 0$.

Payoffs At time $t = T$, each agent chooses an action $a_i \in \{S, R\}$. Payoffs for the strategic principal and agent are given in Table 1.[3] 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 her action a_i with the state y if she knows the state with certainty. Otherwise, the parameter b captures any asymmetry in the payoffs between

[3] An example to help visualize this payoff table is the following: the states of nature S and R can be mapped to whether a particular vaccine is safe (state S) or risky (state R). Similarly, an agent's actions can be thought of as analogous to "vaccinate" (action S) and "not vaccinate" (action R). 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.

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

Table 1. Terminal payoffs when the strategic principal wants agents to take action R. The parameter b is in $(-1, 1)$.

the two states.[4] Recall that, on the other hand, the principal always prefers that agents take action R instead of action S, and so has an incentive to convince agents of $y = R$ (when the state is in fact $y = S$). 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 (which 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 $c(\mathbf{x}) = \sum_{i=1}^n \varepsilon x_i$ the cost of the principal for implementing the network influence strategy $\mathbf{x}$ at $t = 0$. 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, $u^p(\mathbf{a}) - c(\mathbf{x})$. Agents simply choose an action that maximizes their utility $u_i(y, a_i)$ given belief of the state $\pi_{i,T}$.

Note that the action of the stubborn agent is always S and yields her a payoff of 1. On the other hand, a DeGroot agent will take action S if and only if her terminal belief about state S, $\pi_{i,T}(S)$, exceeds the threshold $(1 + b)/2$, because then action S gives her more (expected) utility than action R. The principal chooses his optimal influence strategy, denoted $\mathbf{x}^*$, based on the expectation of his utility from the actions $\mathbf{a}$ that the agents take at time T. Observe that the optimal influence strategy $\mathbf{x}^*$ may or may not be unique.

3 Learning Dynamics and Centrality

In this section, we characterize the learning dynamics and terminal beliefs of agents in the presence of the principal's interference. Our key insight is the relationship between the limiting beliefs of the agents and a novel centrality measure that we call *DeGroot centrality*. This measure captures an agent's susceptibility to misinformation by computing her centrality amongst other (DeGroot) agents who update their beliefs using possibly misinformative signals sent by the principal.

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

3.1 Learning

We aim to understand the asymptotic learning dynamics that emerge for a given (arbitrary) network strategy $\mathbf{x}^*$ of the principal. We provide a closed-form expression for DeGroot terminal beliefs as a function of the chosen influence vector $\mathbf{x}^*$. These terminal beliefs induce actions for each DeGroot agent i at T, which in turn provides an expression for whether agent i mislearns the state under $\mathbf{x}^*$.

When the network consists entirely of stubborn agents, the principal is unable to get anyone to take the incorrect action. On the other hand, when the network consists of all DeGroot agents, generally it will be possible to convince DeGroot agents of the wrong state, as long as the influence cost ε is not too large. The interesting case happens in the mixed environment where both DeGroot and stubborn agents co-exist. In this setting, there are two opposing forces: (i) the stubborn agents who know and can communicate the correct state information, and (ii) the DeGroot agents who may confound the learning process through simple learning heuristics. Our interest in in whether the principal can effectively utilize the second force to his benefit, despite the presence of the first.

We make some assumptions about the rate of information arrival, the informativeness of organic signals, and the network structure:

Assumption 1. Each of the following hold:

(i) *Amenability to mislearning*: For all agents i, $p_i < \frac{\lambda^*+\lambda}{2\lambda}$.

(ii) *Strong connectedness*: For every two agents i, j in unweighted social network $\mathbf{G}^*$, there exists both a directed path from i to j and from j to i.

(iii) *Irrelevance of noise*: For every DeGroot agent i, θ_i is positive if and only if $p_i > 1/2$.

(iv) *Identifiability*: There exists some agent i where $p_i > 1/2$ (i.e., state R and state S can be identified by at least one agent from solely organic news).

The first part of the assumption ensures that 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 own signals (and ignoring others). However, if the agent utilizes social learning, she may be able to learn the true state. The second part of the assumption requires that the beliefs of any one agent can reach (or influence) any other agent, albeit indirectly through others. The third part requires that all agents in the network listen to the news they receive if and only if their organic news is believed to be meaningful. Lastly, we assume the organic news contains valuable information for at least one agent, otherwise learning would be impossible with all DeGroot agents, even without principal interference.

To understand the role of the network structure in the principal's problem, we need to characterize asymptotic learning for the DeGroot agents. Let y' denote an arbitrary state. We write $h_{i,t}(\mathbf{x}^*)$ as the (random) history of news (both organic and inorganic) up until time t induced by the principal's action $\mathbf{x}^*$ (which, naturally, depends on his type). We first establish that the personal experience component of all agents converges almost surely for a long learning horizon:

Lemma 1. *The personal-experience Bayesian update term BU$(S|h_{i,t})$ converges almost surely to a constant BU$(S|h_{i,\infty}(\mathbf{x}^*)) \in \{0, q, 1\}$ as $T \to \infty$.*

Given Lemma 1, we observe that in the limiting case (i.e., large t), the beliefs of the DeGroot agents approximately follow:

$$\boldsymbol{\pi}_{t+1}(y') = \text{BU}(\mathbf{h}_\infty(\mathbf{x}^*))(y') \odot \boldsymbol{\theta} + \mathbf{W}\boldsymbol{\pi}_t(y')$$

where the matrix $\mathbf{W}$ is the influence matrix from Section 2, $\boldsymbol{\theta} = (\mathbf{1}_K, \theta_{m+1}, \ldots, \theta_n)$, $\odot$ denotes the element-by-element product, and $\text{BU}(\mathbf{h}_\infty(\mathbf{x}^*))(y')$ is the vector of converged personal-experience beliefs of state y', per Lemma 1 (with the convention that $\text{BU}(S|h_{i,\infty}) = 1$ for all stubborn agents). Given this formulation, we have the following asymptotic result for the beliefs of DeGroot agents:

Theorem 1. *Under Assumption 1, the beliefs of the agents about state y' converge almost surely to:*

$$\boldsymbol{\pi}_t(y') \overset{a.s.}{\to} (\mathbf{I} - \mathbf{W})^{-1}(\text{BU}\left(\mathbf{h}_\infty(\mathbf{x}^*)\right)(y') \odot \boldsymbol{\theta})$$

for any principal action $\mathbf{x}^*$.

First, we look to characterize beliefs in the baseline case of a truthful principal (i.e., $\omega = \mathcal{T}$), i.e., $\mathbf{x}^* = \mathbf{0}$. We obtain the following result, which is similar to the findings in Jadbabaie et al. (2012):

Proposition 1. *If Assumption 1 holds, then all agents learn the true state almost surely (i.e., $\pi_{i,t}(S) \overset{a.s.}{\to} 1$ for all i) when the principal is truthful.*

Without a strategic principal, learning occurs despite the fact that DeGroot agents are only updating their beliefs using naive learning heuristics. We now turn our attention to whether it is possible for (some) agents to mislearn the state when the strategic principal plays $\mathbf{x}^* \neq \mathbf{0}$. This is the main focus of the remainder of the paper.

Under Assumption 1, we know by Lemma 1 that $\text{BU}(\mathbf{h}_\infty(\mathbf{x}^*))(R) \odot \boldsymbol{\theta}$ converges almost surely to the vector:

$$\left(\boldsymbol{\gamma} \equiv \begin{pmatrix} \mathbf{0}_K \\ \mathbf{x}^*_D \end{pmatrix}\right) \odot \boldsymbol{\theta}$$

where the subscripts K and D denote the intervention vector $\mathbf{x}$ associated with those types of agents. In other words, if an agent i is DeGroot and receives misinformation signals from the principal (i.e., $x_i = 1$), then we write $\gamma_i = 1$ and otherwise we write $\gamma_i = 0$. The (limit) beliefs of stubborn agents are naturally a point-mass on the true state (i.e., zero belief on $y' = R$). This allows a succinct representation of the limiting beliefs of the incorrect state $y' = R$ for all agents:

$$\boldsymbol{\pi}_t(\mathbf{x}^*) \to (\mathbf{I} - \mathbf{W})^{-1}(\boldsymbol{\gamma}(\mathbf{x}^*) \odot \boldsymbol{\theta}) \equiv \boldsymbol{\pi}_\infty(\mathbf{x}^*) \tag{2}$$

Equation (2) provides a closed-form expression for the beliefs of the agents in state R when T is large. We note that this expression depends on the network structure, the a priori knowledge of the agents about the state (i.e., agent types), the personal-experience weights $\boldsymbol{\theta}$, and the network action $\mathbf{x}$ of the principal (captured through $\boldsymbol{\gamma}$). We discuss each of these factors in more detail below.

1. *Personal-Experience Weights*: Each agent i's personal-experience belief update propagates to the beliefs of other agents in society precisely through her weight θ_i, which factors into the expression $\boldsymbol{\theta} \odot \boldsymbol{\gamma}$. In Section 4.1, we show the nuances of how increases in θ_i can either help or hurt the spread of misinformation, as a function of other agents' θ_j for all $j \neq i$.

2. *Network Structure*: Recall from Section 2 that $\mathbf{W}$ represents the influence matrix. The term $(\mathbf{I} - \mathbf{W})^{-1}_{ij}$ represents the entire accumulation of (direct or indirect) influence j has over i.[5] In Section 5, we focus on how the topology of the social network $\mathbf{G}$ shapes how influence propagates through this term.

3. *Agent Types*: The type of the agent (i.e., replacing a DeGroot agent with a stubborn one) has an impact on both $(\mathbf{I} - \mathbf{W})^{-1}$ and $\boldsymbol{\theta} \odot \boldsymbol{\gamma}$. Through the expression $(\mathbf{I} - \mathbf{W})^{-1}$, one can see that a DeGroot agent may not take the incorrect actrion herself but still spreads some of the misinformation she observes from the beliefs of her friends (or in her news). A stubborn agent on the other hand does not propagate nor succumb to misinformation, which limits the influence the principal can have in the population.

The set of agents taking the incorrect action is entirely determined by the principal's optimal choice of $\mathbf{x}^*$ and the resulting limiting beliefs $\boldsymbol{\pi}_\infty$. For some stylized settings, we can characterize the principal's optimal strategy $\mathbf{x}^*$, as well as the set of agents who will take the incorrect action because of $\mathbf{x}^*$, as a function of these model parameters (see Section 4 and Section 5). In Appendix A, we

[5] To see this, note the term $(\mathbf{I} - \mathbf{W})^{-1}$ can be written in expanded form as $\sum_{\ell=0}^{\infty} \mathbf{W}^\ell$, where each $\mathbf{W}^\ell$ represents how the beliefs of agent i propagate to agent j who is ℓ hops away in the social network.

present the general optimization problem of the principal for arbitrary network parameters, as well as some technical results of interest. For an illustration of how these techniques can be visualized in the context of a real-world social network, we encourage the reader to look at Appendix C.

3.2 Manipulation

A main focus of our paper is characterizing the conditions under which an agent chooses the terminal action that matches the underlying state (and therefore maximizes her ex-post payoff) given her belief at time T. Recall from the previous section that the beliefs of all agents converge almost surely to some limit belief, based on which she takes her terminal action. To this end, we define what it means for agent i to be manipulated:

Definition 1 (Manipulation)**.** Let $\mathbf{x}^*$ be the optimal network influence strategy for the strategic principal. We say that agent i is *manipulated* under the network influence strategy $\mathbf{x}$ if:

1. Agent i's terminal action a_i *does not* match the underlying state when the principal's type is $\omega = \mathcal{S}$ and $\mathbf{x} = \mathbf{x}^*$, almost surely.

2. Agent i's' terminal action a_i *does* match the underlying state when the principal's type is $\omega = \mathcal{T}$ and $\mathbf{x} = \mathbf{0}$, almost surely.

In other words, manipulation of agent i implies that a strategic principal interferes with the learning process, and this causes the agent to mislearn the true state that she would have correctly learned in the absence of such interference (by Proposition 1). Agents whose beliefs of state R converge to a value higher than $(1-b)/2$ are necessarily manipulated.

To be able to speak about the extent of manipulation (i.e., how many agents are manipulated) when the principal acts optimally, it is important to consider the entire set of optimal strategies for the principal. Let $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ be two optimal strategies for the principal and let k_1 and k_2 denote the corresponding number of manipulated agents at time T. We say the principal's optimal influence strategies are *manipulation-invariant* if for all $\kappa > 0$, there exists T^* such that for all $T > T^*$, the manipulation at horizon T satisfies:

$$\mathbb{P}_{(k_1,k_2)\sim(\mathbf{x}_1^*,\mathbf{x}_2^*)}\left[k_1 = k_2\right] \geq 1-\kappa$$

for any two optimal strategies $\mathbf{x}_1^*, \mathbf{x}_2^*$. In other words, manipulation is the same under all optimal strategies for the principal if these strategies are manipulation-invariant. We can then state:

Theorem 2. *There exists a set $\mathcal{P} \subset \mathbb{R}^2_+ \times (-1,1)$ of measure zero,[6] such that for all $(\varepsilon, \lambda, b) \notin \mathcal{P}$, the principal's optimal strategies are manipulation-invariant.*

Manipulation-invariance of the principal's strategies guarantees that, with high probability, our welfare analysis (i.e., the number of agents who mislearn the state) does not depend on the strategy the principal chooses or the realization of the signals or actions during the learning process, as $T \to \infty$. Because of Theorem 2, we can refer without ambiguity to the "number of manipulated agents" in the principal's optimal strategy. Note that it may be possible that *different* agents are manipulated under different optimal principal strategies but the *total number* of manipulated agents remains unchanged.

3.3 DeGroot Centrality

We characterize manipulation in an arbitrary network by developing a centrality measure called DeGroot centrality, which is closely related to the familiar eigenvector, Katz-Bonacich, and PageRank centrality measures from the social learning literature. A definition that allows for a simple visualization of DeGroot centrality is based on weighted walks: fix the weighted social network $\mathbf{G}$, its matrix representation $\mathbf{W}$, and its unweighted counterpart network $\mathbf{G}^*$ (see Section 2). A *walk* between agents i and j is any directed path $W = i \to v_1 \to v_2 \to \ldots \to v_k \to j$ such that all links exist in the unweighted social network $\mathbf{G}^*$.[7] The *weight* of a walk W is given by:

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

We say that a walk W is stubborn-avoiding if none of the agents along the walk are stubborn. Let $\mathcal{W}_{ij}$ be the (countable) set of all stubborn-avoiding walks (of any length) between agents i and j.[8] Recall the vector $\boldsymbol{\gamma}$ from Section 3.1; we will refer to this as the *influence* parameter for the principal, which directly depends on the choice of interference $\mathbf{x}^*$. We now define our key centrality measure:

[6] The condition that $(\varepsilon, \lambda, b)$ must lie outside $\mathcal{P}$ is often referred to as a *genericity condition*, where we say that $(\varepsilon, \lambda, b)$ are *generic* if they satisfy this property. Its purpose is to eliminate knife-edge cases where specifically chosen parameters may make some agents indifferent between multiple actions, but if perturbed just slightly, the agent prefers a unique action. Another interpretation of the genericity condition is that provided $(\varepsilon, \lambda, b)$ are drawn randomly from a smooth distribution over some subset of $\mathbb{R}^2_+ \times (-1,1)$, then with probability 1 the principal's optimal strategies will be manipulation-invariant.

[7] Note that by "directed path" we allow for the possibility that $v_i = v_j$ for $i \neq j$ along the walk (i.e., repeated vertices).

[8] By convention, any walk containing a stubborn agent will necessarily have weight zero, so taking $\mathcal{W}_{ij}$ to be the set of all walks gives identical results, but makes it easier to misapply the result (by including walks that pass through stubborn agents, and failing to zero the weight of the entire walk).

Definition 2. The *DeGroot centrality* of agent i is equal to:

$$\mathcal{D}_i(\boldsymbol{\gamma}) = \sum_{j=1}^{n} \left(\theta_j \gamma_j \sum_{W \in \mathcal{W}_{ij}} w_W \right)$$

Our centrality measures captures the *level of influence* that other DeGroot agents have on agent i's own belief. The next proposition shows that this centrality measure, applied to any agent i is exactly equal to that agent's belief of the incorrect state:

Proposition 2. *The DeGroot centrality of agent i is equal to her limiting belief $\pi_{i,\infty}(R)$ of the incorrect state R when the principal exerts influence $\boldsymbol{\gamma}$, i.e., $\mathcal{D}(\boldsymbol{\gamma}) = (\mathbf{I} - \mathbf{W})^{-1}(\boldsymbol{\gamma} \odot \boldsymbol{\theta})$.*

Proposition 2 therefore establishes that the DeGroot centrality of agent i, defined as the weighted-sum of all stubborn-avoiding walks to other DeGroot agents, corresponds precisely to her belief in the incorrect state R, as given by Theorem 1. Moreover, the DeGroot centrality of an agent i can also be related to the centralities of her neighbors via the following recursive relationship:

$$\mathcal{D}_i(\boldsymbol{\gamma}) = \theta_i \gamma_i + \sum_{j=1}^{n} w_{ij} \mathcal{D}_j(\boldsymbol{\gamma})$$

where by definition the DeGroot centrality of a stubborn agent is 0. We provide an example of how to apply both the weighted-walk and recursive definitions of DeGroot centrality, and their equivalence to beliefs of the incorrect state, in Example 2 located in Appendix B.1.

Discussion. As we mentioned before, our definition of centrality shares some similarities with other measures in the literature. There are three key parts of the DeGroot centrality definition: (i) longer walks are discounted more than shorter walks (i.e., closer friends have more impact than those further away), (ii) there is differentiation between agents who are influenced by the principal (targeted DeGroots) and those who are not (stubborn agents and non-targeted DeGroots), and (iii) there is a normalization of the weights so that more neighbors means less influence per neighbor (i.e., the sum of influence weights is always 1). Table 2 illustrate which of these properties are shared in eigenvector, Katz-Bonacich, and PageRank centrality.

In particular, none of the centrality measures capture property (ii). This property highlights the fact that the other measures solely describe graph or network properties, whereas DeGroot centrality captures both network properties *and* the principal's strategy: DeGroot agents who are targeted by the principal contribute towards the centrality of an agent, but those who are not targeted do not. In that sense, different nodes in the network exert different types of influence on the centrality measure

Centrality	Discounted Walks	Asymmetric Influence	Normalized
Eigenvector	✗	✗	✗
Katz-Bonacich	✓	✗	✗
PageRank	✓	✗	✓
DeGroot	✓	✓	✓

Table 2. Comparison of Centrality Measures.

of other nodes, and that type is in turn dependent on the targeting strategy. For instance, PageRank centralities do not change if the network remains the same; on the other hand, if the principal plays some network strategy $\mathbf{x}_1$ instead of $\mathbf{x}_2$, the DeGroot centralities *will* change, even if the underlying network itself does not. Thus, the defining feature of DeGroot centrality is its ability to not only capture network structure, but also capture how the strategic provision of information shapes centrality.

Finally, we note that property (i) is also more general in DeGroot centrality than in the typical Katz-Bonacich or PageRank centrality sense: Because every node represents an agent with a heterogeneous θ_i, the discount (or dampening factor) applied in each step of the walk is different from one node to the next, and thus provides some additional subtlety which is explored in Section 4.1.

4 Principal's Optimal Influence

Throughout Section 3, we have characterized the learning dynamics of the population, holding fixed the influence of a possibly strategic principal. A key determinant of manipulation, however, is *how* a strategic principal chooses his influence to maximize his own payoff. Holding the network topology fixed, we consider how changes in the environment affect the principal's strategy, agents' beliefs, and the persistence of manipulation. In the next section (Section 5), we study how these elements are affected by the network topology.

We will say that society is *impervious* to manipulation if no agents are manipulated in the principal's optimal strategy; otherwise it is *susceptible*.[9] Our first comparative static analyzes the effects from changing cultural norms relating to information assimilation and social learning. The second demonstrates how influence costs and the relative payoffs from taking the incorrect vs. correct action shape the principal's intervention, sometimes in counterintuitive ways. Underlying both of these is the tension between a principal trying to spread misinformation, and stubborn agents spreading knowledge of the true state.

[9]This implies that manipulation is a binary property of the network: it either exists or not. Section 6.3 extends this definition to consider *how many* agents are manipulated.

4.1 Comparative Statics on Personal Experience: Cultural Norms

This section examines the effect that the personal experience term 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 little 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, 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 $(\mathbf{G}', \boldsymbol{\theta}')$ is a *network preservation* of $(\mathbf{G}, \boldsymbol{\theta})$ if $w'_{ij} = w_{ij}(1-\theta'_i)/(1-\theta_i)$ for all DeGroot agents i.

A network preservation corresponds to a shifting of weights between an agent's own experience and that of her neighbors' opinions, while preserving the relative proportions of the network weights. We call this network preservation *homogeneous* 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 before and after). The homogeneous network preservation corresponds to a unilateral shift in attitudes about the importance of one's own perceptions. Most naturally, in a homogeneous network, θ can be thought of an attitude parameter tuned to the cultural norms of the population.

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

Theorem 3. *Let $\mathbf{G}$ be an arbitrary network with homogenous θ, and let $\mathbf{G}_{\theta'}$ denote the network preservation of $\mathbf{G}$ where all agents have θ'. There exist cutoffs $0 < \underline{\theta} < \theta^* < \overline{\theta} < 1$ such that:*

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

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

(c) If $b > 1/2$,[10] there exists $\varepsilon > 0$ such that when $\theta' \in (\theta^*, \overline{\theta})$ the network $\mathbf{G}_{\theta'}$ is susceptible, but when $\theta \in (\overline{\theta}, 1)$ the network $\mathbf{G}_{\theta'}$ is impervious.

[10] 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.

Theorem 3 shows that the comparative statics on manipulation are *non-monotone* in θ. A society that supports an intermediate amount of weight on each agent's own experience is the society that is most susceptible to manipulation. While social learning can be both helpful and detrimental to uncovering truth, it is most harmful (in the presence of strategic interventions) when used in moderation.

Manipulation becomes impossible when a society is more inclined towards herding (i.e., very small θ), as it relies entirely on social learning. If the community has at least one stubborn 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, but guarantees agents will eventually find the truth. Conversely, social learning plays little role in a culture that supports strong individuality and narcissism (i.e., very large θ). Thus, 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), which is costly. With intermediate θ, however, agents both incorporate their own experience *and* employ social learning, allowing the principal to leverage social forces to spread his message without getting completely drowned out by the stubborn agents.

The next result considers heterogeneous settings and stands in contrast to Theorem 3. Let us consider a set D_1 of DeGroot agents with θ_1 and a set D_2 of DeGroot agents with θ_2.

Proposition 3. *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 $b < 1$ there exists $\bar{\theta}$ such that for all $\theta < \bar{\theta}$, the network is impervious if there is at least one stubborn agent in the network, regardless of ε.*

In heterogeneous settings (where there might be more than one value of θ in society), even if some agents have a small θ and discount all news from the principal, they can still mislearn the state if other agents hold high θ. Those agents discounting their own experiences are now manipulated because they listen mostly to the experiences of others, who may be voicing the beliefs of stubborn agents, but may also be voicing the misinformation they receive from the principal.

Network Formation Considerations. The above observations highlight a novel channel for the formation of social networks as a means for avoiding misinformation. We briefly detour to consider how agents in a society might choose the weights to assign to their personal experiences so as to maximize their chances of learning the correct state. Formally, if an agent learning through DeGroot-style

heuristics could choose her θ_i, how should she do so?

Note that if other agents are relying strongly on social learning over personal experience (i.e., low θ_j), then agent i can benefit by setting $\theta_i \approx 0$ as well to receive influence from only those agents who know the truth with certainty. However, as other agents increase their θ_j, agent i will conform to the beliefs of her immediate peers who, as observed in Proposition 3, may or may not be more amenable to misinformation. In a more individualistic culture, if agent i believes she is more discerning of the news relative to her peers, she is better off picking a larger θ_i herself when the values of θ_j are large.

This defines a coordination game where agents try to arrive at a cultural standard for θ by matching others' choices. When agent i does not match this cultural norm, she risks making a naive decision while ignoring (smart) stubborn agents in the population (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). Loosely, the equilibria of this game correspond to the entire spectrum of homogenous θ. But as we saw in Theorem 3, some equilibria can be more socially inefficient than others. For instance, when agents choose an intermediate θ that splits learning between personal experience and social forces, manipulation is generally worst for society.

4.2 Influence Costs and Payoff Asymmetry

We conclude this section by considering how manipulation is affected by the cost of sending misinformation and the payoff asymmetry under the two different actions S and R. Recall that ε captures the per-agent cost of sending misinformation,[11] whereas b parametrizes the agent's natural affinity toward one action or the other.[12] We obtain the following comparative static:

Proposition 4. *When ε increases, the number of manipulated agents never increases (but may decrease). Similarly, if the network is susceptible with payoff asymmetry b, it is still susceptible when increasing b.*

Perhaps counterintuitively, however, increases in b can create incentives for the principal to target fewer agents and decrease manipulation as a whole, as seen in the following example:

Example 1 (Targeting "Low Hanging" Fruit)**.** Consider the social network consisting of three DeGroots and one stubborn agent arranged along a bidirectional line as in Figure 1. Let $\varepsilon \in (1, 3/2)$ throughout. All DeGroot agents weigh their neighbors and themselves according to $\theta_i = \alpha_{ij} = 1/(1 + |N(i)|)$. First, suppose that $b = 0.3$, so the belief cutoff to take action R is given by $\pi_{\text{cutoff}}(R) = 0.35$.

[11] More general cost structures are considered in Section 6.2.

[12] For instance, if consuming a risky product provides more disutility than a safe product provides utility, then $b > 0$, indicating that the agent must be (much) more confident in the product's safety to choose action S.

Figure 1. Network for Example 1 (Solid = DeGroot, Shaded = Stubborn).

Then it can be shown that the principal manipulates all three agents by sending misinformation to agents 1 and 3, which yields payoff $3-2\varepsilon > 0$.[13] Now suppose b increases to $b = 0.6$, so the belief cutoff to take action R is given by only $\pi_{\text{cutoff}}(R) = 0.2$. Then, one can show the principal manipulates only two of the three agents by sending misinformation to only one agent (for instance, by manipulating agents 1 and 2 through sending signals to agent 1), which yields a payoff of $2-\varepsilon > 3-2\varepsilon$. Thus, while manipulation became "easier" because the cutoff required to take the wrong action had decreased, the number of manipulated agents also decreases from 3 to 2. □

5 Network Topology

We now turn our attention to examine how the topology of the network and the placement of stubborn agents affect manipulation. We provide a classification of networks into dense and sparse topologies, and compare these structures in terms of what is required to make them impervious. The upshot is that a *constant* number of these stubborn agents located *anywhere* in a dense network makes it impervious. Sparse networks on the other hand require more resources (number of stubborn agents needed) and planning (where to place these agents). Even when the number of stubborn agents is enough to make the network impervious, the location of these agents have to be carefully chosen. Further, it is possible that there are scenarios where the number of stubborn agents required grows with the size of the network, making imperviousness more difficult to achieve.

5.1 Dense Networks

We start by defining what it means for a network to be dense. Recall that $\mathcal{W}_{ij}$ is the set of all stubborn-avoiding walks between i and j. To this end, we define the *log-diameter* of the network $\mathbf{G}$ as:

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

where the weights w_{ij} are from the matrix-representation $\mathbf{W}$ of $\mathbf{G}$ (see Section 2). Using this, we can define the *density* of a network as follows:

[13] To see this, observe that sending more signals cannot improve the principal's payoff, and targeting only a single agent leads to only that agent being manipulated when $b = 0.3$, which is worse than targeting no one since $\varepsilon > 1$. A more formal proof is given in Appendix D.

Definition 4 (Dense Networks)**.** We say that network $\mathbf{G}$ is δ-*dense* if it has a log-diameter of at most $\log(n+\delta)$.

We can then utilize this definition to give the following result:

Theorem 4 (Constant Number of Stubborn Agents)**.** *For every δ, there exists a universal constant $m^*(\delta)$ such that every network $\mathbf{G}$ which is δ-dense and contains at least $m^*(\delta)$ stubborn agents is impervious to manipulation.*[14]

Theorem 4 implies that a vanishingly small fraction of stubborn agents in the population is all that is required to make the principal unable to manipulate beliefs. Moreover, if the positions of those stubborn agents were to be chosen by an adversary, manipulation will still not be possible as long as the network $\mathbf{G}$ satisfies the log-diameter condition for every placement of $m^*(\delta)$ stubborn agents. Finally, note that the shortest path between agent i and every stubborn agent being less than $\log(n+\delta)$ does not automatically imply that agent i will not be manipulated. This needs to hold *uniformly* across all DeGroot agents, otherwise the log-diameter condition is not satisfied. Example 5 in Appendix B demonstrates how a DeGroot agent that is directly connected to a stubborn agent can still believe the incorrect state as a result of her living in a DeGroot bubble where echo chamber effects are rampant.

Theorem 4 guarantees that if the number of stubborn agents meets the threshold $m^*(\delta)$ in a δ-dense network then there will never be manipulation, but this bound may not be tight. In Appendix C, we perform a numerical study of how the number of stubborn agents and their placements might affect manipulation in practice, as compared to the log-diameter bound provided in Theorem 4.

We conclude this section by mentioning a few examples of interest where one can easily apply the result of Theorem 4, along with one cautionary example where the result cannot be utilized. These examples are worked out in detail in Appendix B.2.

(i) *The complete network*: The complete network is the most dense network, and is impervious provided the number of stubborn agents satisfies $m \geq (1+b)/(1-b)$.

(ii) *Influential star network*: In the influential star network, most agents listen to a single (central) agent. This network can be impervious even if the central agent herself is DeGroot. This occurs because the network is sufficiently dense, as it is possible to get from one agent to another by passing through that central agent. We can then apply Theorem 4 and show that $m \geq 2(1+b)/(1-b)$ stubborn agents are sufficient for imperviousness, irrespective of their location.

[14] For the interested reader, Appendix A offers a stronger version of Theorem 4 that requires satisfying a weaker notion of local density everywhere in the network.

(iii) *Echo chamber network*: An echo-chamber network is a network where DeGroot agents communicate almost-exclusively with other DeGroot agents. For instance, consider two cliques of size $n/2$, one of all DeGroots and one of all stubborn agents, with a single connection between them. The unweighted network $\mathbf{G}^*$ has diameter 3 for all n, but admits the same manipulation as in Example 5, despite a *linear* number of stubborn agents. It is easy to check the shortest path between most DeGroots and a stubborn agent is roughly $\log(n^2/2)$, so does not satisfy the log-diameter condition of Theorem 4 for any δ.

5.2 Susceptible Networks: An Example

As a prelude to our discussion of sparse networks, we demonstrate the traits that make such networks susceptible to manipulation by considering the directed ring network as a prototypical example. We follow this up with a more general characterization in Section 5.3.

In an episode of the show Planet Money on NPR, the political consultant David Goldstein discusses how firms like Cambridge Analytica interfere in elections by targeting agents with messages in order to push them towards specific actions, and how this strategy can be profitable even if it fails to sway most agents who receive such messages:[15]

> "You might be immune and the guy next to you might be immune, and the guy next to that person might be immune, but if I only need to change [the minds of] 3% of people in order to affect a given result, then I can go 97 people down and not have an effect but as long as I have an effect on 1, 2, and 3, then I can literally change the world."

Goldstein's quote was made in a different context from the one we consider, but it perfectly encapsulates the example of the ring network in Figure 2. In this example, the principal targets several agents with misinformation despite knowing that some of these agents will not be directly affected and will still figure out the correct state of the world. This targeting however reverberates through the network, and allows the principal to manipulate agents at the end of the ring without sending them any messages, leading to an overall lower cost of manipulation. We now discuss this example in detail.

Consider a ring network with homogenous $\theta_i = \theta$. Network weights are given by $\alpha_{ij} = 1 - \theta$ 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 her immediate neighbor, who in turn listens to her immediate neighbor, etc. Consider the following stubborn agent placements:

1. *Continuous chain*: Assume the first m agents on the ring network are all stubborn (i.e., the stubborn agents talk *mostly* with other stubborn agents), and the remaining agents are DeGroots.

[15] This quote comes at the 16:30 minute mark at https://www.npr.org/2019/05/24/726536757/episode-915-how-to-meddle-in-an-election

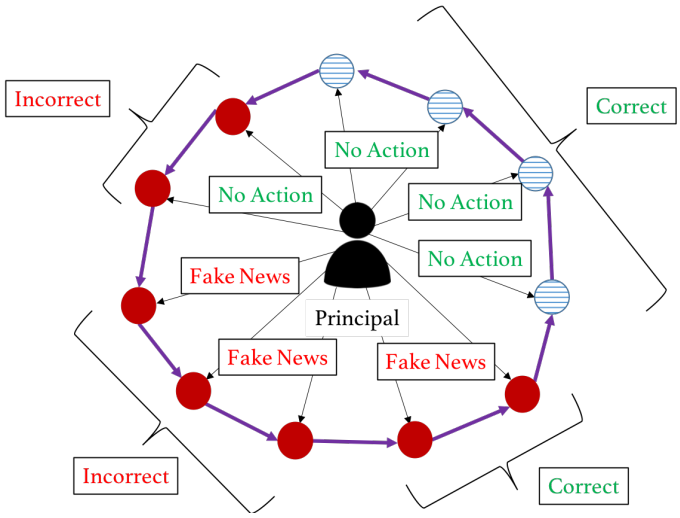

Figure 2. Beliefs in the Ring Network. A directed arrow from node i to node j indicates that i listens to j. Shaded nodes represent stubborn agents.

2. *Sprinkled*: The stubborn agents are "sprinkled" throughout the network so that the distance of any DeGroot agent i to the nearest stubborn agent is minimized.

Continuous Stubborn Agent Chain. For illustration, consider the case of a continuous chain of stubborn agents with $\theta = \frac{1}{n+1}$ and $m \ll n$, so that there are fewer stubborn agents than DeGroots. Suppose the principal solves an easier influence problem along only a single dimension: (i) he exerts influence along 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 manipulated agents. This is not necessarily the principal's optimal network strategy, but we use this to show that there is *some* strategy that beats $\mathbf{x} = \mathbf{0}$, and therefore there must be some manipulation by Corollary 2. An illustration of this strategy is given in Figure 2.

Consider DeGroot agent i at location τ away from the last stubborn agent. We can write her belief in terms of her DeGroot centrality $\mathcal{D}_i(\boldsymbol{\gamma})$, a function of $\boldsymbol{\gamma}$:

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

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

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

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

the DeGroot agents are manipulated when $b \geq (2-e)/e$. Therefore, *even with a growing population of stubborn agents, a linear number of DeGroot agents are manipulated.*

When stubborn agents form a continuous chain, the network is fundamentally equivalent to one where the chain is replaced by a single stubborn agent who knows the truth at the end of this chain. The long chain of DeGroot agents who receive misinformation drowns out the beliefs of the DeGroots at the beginning of the ring.

Sprinkled Stubborn Agents. We now consider the effects of stubborn agent placement by "sprinkling" them throughout the ring. Unlike with the continuous chain, in this case, we obtain the following result:

Proposition 5. *There exists a constant m^* such that if there are $m > m^*$* sprinkled *stubborn agents in the ring network, it is impervious to manipulation for any n.*

In contrast to Theorem 4, Proposition 5 imposes firm restrictions on the placement of stubborn agents, but similar to that theorem, it shows that only a constant number are needed. Recall in dense networks, neither the *placement* nor *the number* of stubborn agents have to meet particularly demanding conditions to guarantee imperviousness. However, in the ring network, placement becomes crucial, despite still only requiring a small fraction of stubborn agents to avoid manipulation. Although the ring network is sparse, because agents largely discount their own experiences ($\theta \sim 1/n$) and echo chambers are limited,[16] a few optimally-placed stubborn agents limit the spread of misinformation.

Recall that because $\theta = 1/(n+1)$, as the network gets large, agents mostly dismiss their own experiences. A more natural assumption is to suppose agents weigh all social influences equally with their own experience. Formally, we consider the *equal-influence* weighting given by:

$$\theta_i = \alpha_{ij} = \frac{1}{1+|N(i)|} \tag{3}$$

for DeGroots i and all $j \in N(i)$. While in the complete network this corresponds to setting $\theta_i = 1/(n+1)$ as before, in the ring this instead admits weighting $\theta_i = \theta = \alpha_{i(i-1)} = 1/2$. In contrast to Proposition 5, when DeGroots listen more to their own news, we obtain a result in stark contrast to the case of dense networks:

[16] Because opinions only flow in one direction, echo chambers are not too strong. Here, sparsity is the main driver of manipulation, and thus requires special placement to avoid it. In the case of other sparse networks, such as the bidirectional ring or cliques of all DeGroots (e.g., Example 5), echo chambers can be much worse and drive beliefs even farther away from truth.

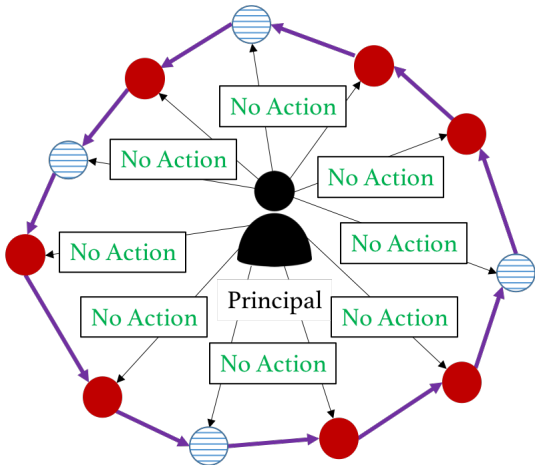

Figure 3. An illustration of Proposition 5 with "sprinkled" stubborn agents.

Proposition 6. *Consider the ring network with equal-influence weighting and $\varepsilon < 1$. Then there exists a constant $c > 0$,*[17] *such that the network is impervious with $c \cdot n$ sprinkled stubborn agents agents, but is susceptible if there are fewer than $c \cdot n$ stubborn agents* or *their configuration is not sprinkled.*

In the equal-influence ring network, imperviousness comes with stringent requirements on *both* resources (number of stubborn agents) and planning (their placement). First, as the network grows in size, the number of stubborn agents must also grow in proportion, so that a constant fraction of the population is still stubborn. Second, the location of the stubborn agents is paramount to preventing manipulation. We summarize these findings in Table 3.

Network	Resources	Planning	# Manipulated (when possible)
Dense Network	$\Theta(1)$	Anywhere	$\Omega(1)$
Ring Network	$\Theta(1)$	Sprinkled	$\Omega(n)$
Equal-Influence Ring	$\Theta(n)$	Sprinkled	$\Omega(n)$

Table 3. Properties based on Network Density.

The principal's ability to manipulate in networks that do not satisfy the log-diameter condition of Theorem 4 is not unique to the ring network. In addition to the more general characterization of sparse networks in the next section, Example 7 in Appendix B.3 provides another demonstration of susceptibility on the star network with equal-influence weighting.

[17] Here, and throughout the entire paper, by *constant* we mean $c \in \Theta(1)$, so there exist $\underline{\beta}, \overline{\beta}$ independent of n such that $\underline{\beta} \leq c \leq \overline{\beta}$.

5.3 Sparse Networks

We now generalize the insights of Section 5.2 to a wide array of sparse networks. First, we consider a continuum of networks that are parametrized by a sparsity parameter η, and proceed to give a characterization of manipulation for all η. Second, we provide a sufficient condition for imperviousness in symmetric networks. Because symmetric networks may either be dense (e.g., complete) or sparse (e.g., ring), this result provides a more complete understanding of imperviousness across different levels of sparsity. Our main findings highlight how the requirements on resources and planning become more demanding as the network becomes more sparse, corroborating the results from Section 5.2.

Convex Combination of Ring and Complete Networks Let us fix the stubborn and DeGroot agents in the population and consider two different network structures $\mathbf{G}_c = (\boldsymbol{\theta}^c, \mathbf{W}^c)$ and $\mathbf{G}_r = (\boldsymbol{\theta}^r, \mathbf{W}^r)$, corresponding to the complete network and the ring network with equal-influence weighting. By Theorem 4 and Example 3, we know $\mathbf{G}_c$ is impervious to manipulation with a constant number of stubborn agents (located anywhere). On the other hand, in $\mathbf{G}_r$, an unbounded (in n) number of agents are manipulated whenever the number of stubborn agents is sublinear *or* these agents are not in specific network positions.

Now consider parameter $\eta \in [0,1]$ and define the network $\mathbf{G}_\eta$ as $\theta_i = \eta \cdot \theta_i^c + (1-\eta) \cdot \theta_i^r$, and $\alpha_{ij} = \eta \cdot \alpha_{ij}^d + (1-\eta) \cdot \alpha_{ij}^s$ for all i, j. Note that as η varies from 0 to 1, the network becomes more dense and, by construction, the network is susceptible at $\eta = 0$ but impervious at $\eta = 1$. The following result provides the full characterization for intermediate η:

Theorem 5. *Suppose there are either $o(n)$ stubborn agents or that these agents form a continuous chain in the ring. There exists η^* such that: (i) if $\eta < \eta^*$, $\mathbf{G}_\eta$ the number of manipulated agents grows unboundedly in the size of the network n, whereas (ii) if $\eta > \eta^*$, $\mathbf{G}_\eta$ is impervious to manipulation.*

Theorem 5 shows that a *phase transition* exists between dense and sparse networks: when the network gets sufficiently dense, all opportunities for manipulation suddenly vanish. Similarly, it shows the results of Section 5.2 are fairly robust: qualitative conclusions for the ring network generalize to sparse networks even as they become slightly more dense.

Symmetric Networks of Degree k We now consider *symmetric networks*, where any two agents have identical network positions. In particular, for any directed unweighted network $\mathbf{G}^*$, we say $\mathbf{G}^*$ is symmetric if and only if for every pair of vertices i, j, there exists a function $f : \{1, \ldots, n\} \to \{1, \ldots, n\}$

such that $f(i) = j$ and $k \to \ell$ exists in $\mathbf{G}^*$ if and only if $f(k) \to f(\ell)$ exists in $\mathbf{G}^*$.[18] We say a network $\mathbf{G}$ is *symmetric* if the unweighted analog, $\mathbf{G}^*$, is symmetric and $\{\alpha_{ij}, \theta_i\}$ satisfy the equal-influence weighting (Equation 3) for all links $i \to j$ that exist in $\mathbf{G}^*$ (i.e., $j \in N(i)$ whenever $i \to j$ exists in $\mathbf{G}^*$, and $\alpha_{ij} = 0$ if $j \notin N(i)$).

In other words, a symmetric network $\mathbf{G}$ is one where all agents are symmetric in the unweighted sense, and employ equal-influence weighting. When the network is strongly connected, k-regularity (i.e., each agent has k neighbors) is a necessary (but not sufficient) condition for symmetry, so in particular we have $\theta_i = \alpha_{ij} = 1/(1+k)$ for all agents i whenever there exists a link $i \to j$. Therefore, symmetric networks can be partitioned into classes of "degree-k" symmetric networks. We will also say K is a *symmetric placement* of stubborn agents if the induced subgraph $\mathbf{G}^* \backslash K$ is symmetric.

Suppose that a fraction ϕ of all the links going into stubborn agents are links between stubborn and DeGroot agents. Then, within the class of symmetric networks, we obtain the following characterization:

Theorem 6. *Suppose $\mathbf{G}$ is a degree-k symmetric network with a symmetric placement of $m = |K|$ stubborn agents. Then the network is impervious to manipulation if $\phi k m/(n-m) = \phi k |K|/|D| \geq (1+b)/(1-b)$.*

Theorem 6 further demonstrates how sparsity tends to make manipulation easier. Here the degree of the agents, k, functions as a measure of sparsity, and along with the ratio of stubborn agents to DeGroots in the population, $m/(n-m) = |K|/|D|$, determines a sufficiency condition for imperviousness. Moreover, Theorem 6 highlights how stubborn agent placement becomes more demanding with sparsity, on two fronts: first, the placement must be *symmetric*, and cannot be arbitrarily chosen, and second, the stubborn agents ought to be placed in such a way that the links going from these agents to DeGroots (and vice-versa) are maximized. Both of these requirements become more difficult as the network becomes sparser.

The bound in Theorem 6 is in fact tight in many common network topologies. In the case of the complete network, we have $\phi k = |D|$, so $m \geq (1+b)/(1-b)$, which is the exact bound we saw in Example 3. In this case, any placement of stubborn agents is symmetric and has the same ϕ, so the restriction in Theorem 6 is immaterial. The result is also tight in the ring network, where $k = 1$, so m needs to be linear in n to avoid manipulation (as in Proposition 6). Here, the symmetric placement is more challenging and requires careful planning; unsurprisingly, the symmetric placement corresponds precisely to the "sprinkling" arrangement of Section 5.2.

[18] This is simply the definition of a graph automorphism.

6 Extensions

In an effort to illustrate how social learning changes in the presence of strategic interventions, we have presented a parsimonious framework with a number of simplifying assumptions. In this section, we consider how the results and conclusions change in the face of additional complications. While these extensions offer further areas of exploration and more detailed analyses, they also demonstrate how our simplified framework can be applied without much loss of generality.

6.1 Learning the Principal's Type

In Section 2, we have assumed that agents share their beliefs about the state with their neighbors, but not about the type of the principal. We now endow the DeGroot agents with some degree of skepticism. Agents are aware of the possibility of a strategic principal, and in addition to learning about the state, they also update their beliefs on whether the news they receive is organic or strategic. Does this skepticism always decrease manipulation?

To provide an answer to this, we introduce a coupled belief dynamics process for DeGroot agent who may be aware of possible misinformation. In addition to sharing beliefs $\pi_{i,t}$ about the state, we assume agents also share $\mu_{i,t}$, their belief that the principal is truthful instead of strategic. Every DeGroot agent has prior $\mu_{i,0}$ about whether their news source is entirely organic, and personal-experience weight $\tilde{\theta}_i$ about this prior. Moreover, we assume that DeGroot agents exchange beliefs about the principal's type according to the influence matrix $\tilde{\mathbf{W}}$ (where $\tilde{\mathbf{W}}$ is not restricted to be equal to $\mathbf{W}$).

The coupled dynamics process occurs as follows. Agents endogenously choose how much weight to put on the belief they form from reading the news. This weight is directly proportional to how trustworthy they believe the news source is. If an agent believes much of the news they receive is misinformation sent by the principal, then the agent puts much more weight on social learning and largely dismisses the news she observes. Thus, instead of putting a constant weight θ_i on their own news, DeGroot agents put $\mu_{i,t}\theta_i$ weight on their personal news update. Formally, the belief update process obeys the following law of motion:

$$\mu_{i,t+1} = \tilde{\theta}_i \mu_{i,0} + \sum_{j=1}^{n} \tilde{\alpha}_{ij} \mu_{j,t}$$

$$\pi_{i,t+1} = \mu_{i,t}\theta_i \cdot \text{BU}(h_{i,t}) + \frac{1-\theta_i\mu_{i,t}}{1-\theta_i} \sum_{j=1}^{n} \alpha_{ij} \pi_{j,t}$$

Note that as $\mu_{i,t} \to 1$, we recover the baseline model from Section 2, whereas when $\mu_{i,t} \to 0$, agents dismiss their personal experience entirely. With this formulation, the next result shows how we can

reduce this belief process to the baseline model:

Proposition 7. *The coupled-belief dynamics process is equivalent to the baseline model where agents use personal-experience weights* $\boldsymbol{\theta}'$ *given by:*

$$\boldsymbol{\theta}' = \boldsymbol{\theta} \odot (\mathbf{I} - \tilde{\mathbf{W}})^{-1}(\boldsymbol{\mu}_0 \odot \tilde{\boldsymbol{\theta}})$$

and the corresponding network preservation on $\mathbf{W}$.

Proposition 7 shows how the personal weights of the belief update process can arise endogenously when agents engage in a coupled belief update that considers the trustworthiness of their own news. One can apply similar comparative statics as in Section 4.1 to understand how DeGroot skepticism affects limit beliefs about the state:

(a) Uniformly more skepticism about the veracity of information (i.e., lower $\boldsymbol{\mu}_0$) does not necessarily lead to better outcomes (i.e., less manipulation): $\boldsymbol{\theta}'$ is increasing in $\tilde{\boldsymbol{\theta}}$, and by Theorem 3(c), it is possible for manipulation to *increase* when $\boldsymbol{\theta}$ decreases. That being said, sufficient skepticism across all agents leads to imperviousness by Theorem 3(a), when there is at least one stubborn agent in the population.

(b) For the same reason as (a), the baseline model may protect more agents from manipulation than this revised model where agents take into account the possibility of misinformation.

(c) Extreme skepticism (as opposed to just *more* skepticism) about the veracity of information does not necessarily protect a given agent. By Proposition 3, if other agents are less skeptical, then this agent can still be manipulated by absorbing misinformation acquired from social learning.

(d) However, when $\varepsilon \approx 0$, additional skepticism about the accuracy of news always improves the beliefs of DeGroot agents. These can be seen directly through the DeGroot centrality expression, $(\mathbf{I} - \mathbf{W})^{-1}(\mathbf{1}_D \odot \boldsymbol{\theta})$, and given that $(\mathbf{I} - \mathbf{W})^{-1}$ contains all non-negative entries, it is monotone in $\boldsymbol{\theta}$.

6.2 Alternative Cost Functions

We have assumed throughout that the principal's cost function follows the form $c(\mathbf{x}) = \sum_{i=1}^{n} \varepsilon x_i$. In particular, we have assumed costs are *linear* and *homogenous* across agents. We consider two variants of this:

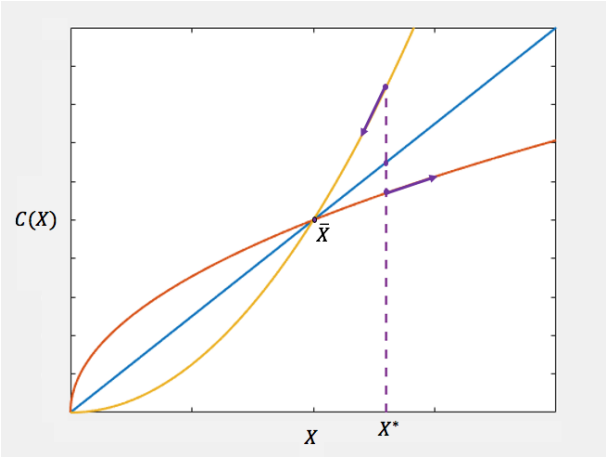

Figure 4. Concave vs. Convex Costs.

1. *Non-linear specification*: Suppose that $c(\mathbf{x}) = C\left(\sum_{i=1}^{n} x_i\right)$, but that c may not scale directly with $X \equiv \sum_{i=1}^{n} x_i$. In particular, there may be *concave costs* with the intervention: we assume $C' > 0$ but $C'' < 0$, with $C'(0) > \varepsilon$ and $\lim_{X\to\infty} C'(X) = 0$. Similarly, there may be *convex costs* with the intervention: we assume $C' > 0$ and $C'' > 0$, with $C'(0) < \varepsilon$ and $\lim_{X\to\infty} C'(X) = \infty$.

2. *Heterogenous costs*: Certain agents may be more expensive to target than others, such as celebrities or those who do not use social media, etc. Thus, we assume there is a vector of costs $\boldsymbol{\varepsilon} = \{\varepsilon_i\}_{i=1}^{n}$ so that $c(\mathbf{x}) = \boldsymbol{\varepsilon} \bullet \mathbf{x}$.

A number of results are completely unaffected by these changes, for example, Theorem 4 for dense networks (Section 5.1), the characterization of manipulation in symmetric networks (Section 5.3), and the comparative statics on θ (Section 4.1). This follows from the fact these sufficiency conditions establish an upper bound on the DeGroot centralities of the agents in the population, and thus hold independently of the principal's costs for intervention.

More generally, however, changes in the cost function will change the scope of manipulation (e.g., in the ring). Let $\mathbf{x}^*$ be the optimal principal intervention with cost function $c(\mathbf{x}) = \sum_{i=1}^{n} \varepsilon x_i$ and $X^* = \sum_{i=1}^{n} x_i^*$. We provide the following comparative result relating these more general cost functions to the one provided in Section 2:

Proposition 8. *Let $\bar{X}$ denote the (unique) crossing point of $C(X)$ and εX. If there is concave cost and $X^* \geq \bar{X}$ then manipulation never decreases; if there is convex cost and $X^* \geq \bar{X}$, then manipulation never increases, whereas if $0 < X^* < \bar{X}$, the network is always susceptible.*

The intuition for the result can be seen in Figure 4. When $X^* \geq \bar{X}$, concave costs encourage the principal to expend more resources at lower marginal cost than in the linear case, increasing

manipulation; on the other hand, convex costs entice the principal to slow her influence and save on higher marginal costs (similar to Example 1).

Lastly, we note the optimization problem admitting an exact characterization of the optimal strategy (given in Appendix A) can be modified easily to account for heterogeneous ε_i without affecting the nature of the problem. In stylized examples such as the ring, star, or complete network, the analysis can be applied as is by considering the average costs of sending signals (i.e., $\frac{1}{n-m}\sum_{i=m+1}^{n}\varepsilon_i$) in place of ε. As such, this generalization does not affect the qualitative findings of this paper.

6.3 Extent of Manipulation

Throughout the paper, we have focused on conditions where manipulation occurs for at least one agent (or none at all). In many contexts, a more appropriate metric is the number of manipulated agents, possibly relative to the population size. While we provide some characterization of the number of manipulated agents throughout (see Table 3), we present here a technical reduction that shows how imperviousness can be easily generalized to this problem.

Definition 5. We say a network is k*-impervious* if there are k or fewer manipulated agents. Similarly, a k*-cut subnetwork* of a network $\mathbf{G}$ is a network obtained from coalescing a set $\mathcal{K}$ of (at most) k vertices from the network (i.e., replacing all vertices in $\mathcal{K}$ by a single vertex u) and setting $\alpha_{iu}=\sum_{j\in\mathcal{K}}\alpha_{ij}$ for all agents i, with $\theta_u=1$.

With this transformation, we get the following reduction:

Proposition 9. *If there exists a k-cut subnetwork that is impervious to manipulation (with the exception of u) when $\varepsilon_u=0$ (and $\varepsilon_i=\varepsilon$ for all other agents), then the original network is k-impervious.*

Therefore, having a complete understanding of k-imperviousness in networks reduces to understanding imperviousness and finding "clever" k-cuts. As an immediate corollary to Proposition 9, we get a log-diameter condition that generalizes Theorem 4 for k-imperviousness:

Corollary 1. *Consider a k-cut subnetwork with the k-cut vertex u removed.*[19] *If the log-diameter of this network does not exceed $\log(n-k+\delta)$, the network is k-impervious if it has $m>m^*(\delta)$ stubborn agents (where $m^*(\delta)$ is the same as in Theorem 4.)*

In Example 8 of Appendix B, we show how in a core-periphery network, Theorem 4 cannot be applied for any value of δ. Yet, Corollary 1 establishes the network is k-impervious for a small value of k agents (agents on the periphery), so a vanishing fraction of agents are manipulated.

[19] Note that this network is not a "valid" subnetwork in the sense that some agents have $\theta_i+\sum_j\alpha_{ij}<1$ after the removal of u. However, DeGroot centrality (and log-diameter) are still well-defined provided that sub-stochasticity is satisfied: $\boldsymbol{\theta}+\mathbf{W}\mathbf{1}\leq\mathbf{1}$.

6.4 Dynamic Targeting Policies

In Section 2, we assume the principal chooses to send each agent with $x_i = 1$ a signal of intensity λ^*, with message $\hat{y}$, which costs him ε. Here, we relax this specification in the following ways:

(a) The principal may send different messages (call these messages $\hat{y}_i \in \{S, R\}$) and/or apply different intensities to different agents, i.e., λ_i^*.

(b) The principal may vary its message and/or intensity throughout time, i.e., $\lambda_i^*(t)$, $\hat{y}_i(t)$.

(c) The principal pays a larger cost for greater intensity messages; that is, the principal pays $\frac{1}{t}\int_0^t \tilde{\varepsilon}(\lambda_i^*(t'))\ dt'$, where $\tilde{\varepsilon}$ is an increasing, convex, and continuous function.

While this relaxation provides many more decision variables for the principal, the outcomes can be analyzed in nearly the exact same way. The following result makes that clear:

Proposition 10. *Consider the model of Section 6.2 with heterogenous (but linear) costs* $\varepsilon_j = \tilde{\varepsilon}(\lambda(2p_j - 1))$ *for each agent* j*. Every agent manipulated in this model is manipulated when the principal is allowed to use dynamic targeting policies, and vice-versa.*

The result of Proposition 10 should not be interpreted as dynamic targeting policies not helping the principal, but rather, that the problem can be analyzed in a static setting using an alternative cost formulation. In this setting, we see that the principal must pay higher costs to send signals to agents whose organic signals are more informative, and that those who are more skilled at interpreting organic news are more difficult to manipulate through their direct personal experience, unlike in the baseline model.

7 Conclusion

In this paper, we consider a classic social learning setup when some of the information in the network is injected by a strategic principal, and we identify conditions that allow this 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,[20] and we provide a model that allows us to study them in a formal setup. We employ a diverse population that possess different degrees of knowledge about the state, which we model by using classical DeGroot agents and knowledgeable

[20] For example, Allon and Zhang (2017) examine a model where agents learn about service quality from their experience as well as what they hear from their friends, and ask how the firm should incorporate this learning process into its decisions about which service levels to offer.

stubborn agents. We find that in this setup, the ability of a self-interested principal to manipulate a population depends on 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 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, and show that sparse topologies are typically more susceptible than dense ones. We demonstrate how some networks can be resilient with the presence of a small number of these stubborn agents, whereas others continue to be susceptible to manipulation unless the *number* and *location* of these agents meet certain demanding criteria.

Our work can be extended on several fronts. We have studied the dynamics of our learning model in the limit, and characterizing the strategies played by the principal in the short-term is also an important but challenging problem. Relatedly, when agents have imperfect recall (e.g. because of costly information acquisition as in Liu (2011) or recency bias, these short-term dynamics become especially relevant, even when the learning horizon is long. Finally, as discussed in Section 4.1, agents can use their social network as a way to protect themselves against potential misinformation. Understanding how agents form their social circles to acquire accurate information is an unexplored avenue for models of social network formation in the presence of misinformation, and provides yet another area of potential future work.

Appendix

A Technical Details

Local Density. We provide a generalization of Theorem 4, which is often more useful in practice, especially when stubborn agents "disconnect" the network (i.e., there exist DeGroots i, j with the only directed walks between them containing stubborn agents). In Example 4 (see Appendix B), we apply the result to one variant of the star network.

Definition 6. The log-distance between i and j is:

$$d_{ij} = \min_{W_{ij} \in \mathcal{W}_{ij}} \sum_{(i' \to j') \in W_{ij}} -\log(w_{i'j'})$$

We say that network $\mathbf{G}$ is δ-locally dense if there exist subsets $I_1, \ldots, I_k$ of agents in $\mathbf{G}$ such that: (i) $\cup_{\ell=1}^{k} I_\ell = \{1, \ldots, n\}$ (i.e., the subsets cover $\mathbf{G}$) and (ii) the log-distance between every two agents $i, j \in I_\ell$ is at most $\log(|I_\ell| + \delta)$.

Proposition 11. *If the network $\mathbf{G}$ is δ-locally dense and contains $m^*(\delta)$ stubborn agents (from Theorem 4) in each set I_ℓ, the network is impervious.*

It is easy to see Theorem 4 is a special case of Proposition 11 by taking $I_1 = \{1, \ldots, n\}$ and checking the log-distance between every two agents in $\mathbf{G}$ (i.e., log-diameter) is at most $\log(|I_1| + \delta) = \log(n + \delta)$.

Full Characterization of Principal's Problem. We can write the principal's problem as the following integer (binary) program:

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

Theorem 7. *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 principal can choose to either send misinformation ($\gamma_i = 1$) 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 $\boldsymbol{\gamma}$. In Appendix C, we use this problem to solve explicitly for the optimal strategy in a real-world social network.

Note that $\mathcal{D}_i(\boldsymbol{\gamma})$ is *linear* in $\boldsymbol{\gamma}$, which makes the problem an integer program (IP) for any network $\mathbf{G}$. 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 7:

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 $\mathcal{K} \neq \varnothing$ 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 sends signals to all of the DeGroot agents, the influence from the stubborn agents will still dominate (i.e., ensure DeGroots take the correct action). 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 $\boldsymbol{\gamma} = \mathbf{0}$.

B Worked Examples

B.1 Demonstration of DeGroot Centrality

Example 2 (Illustration of DeGroot Centrality)**.** Consider the triangle network in Figure 5, with one stubborn agent and two DeGroot agents all talking to each other. Suppose the DeGroot agents listen to themselves and their friends equally so that $\theta_i = \alpha_{ij} = 1/3$ for $j \neq i$, as shown by the solid lines. Using Theorem 1, we can characterize the limiting beliefs of the DeGroots about the incorrect state $y' \neq y$:

$$\begin{aligned}\boldsymbol{\pi}(y') &\overset{a.s.}{\to} \left(\mathbf{I} - \begin{pmatrix} 0 & 0 & 0 \\ 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 0 \end{pmatrix}\right)^{-1} \left(\begin{pmatrix} 0 \\ x_2 \\ x_3 \end{pmatrix} \odot \begin{pmatrix} 1 \\ 1/3 \\ 1/3 \end{pmatrix}\right) \\ &= \begin{pmatrix} 0 \\ \frac{3}{8}x_2 + \frac{1}{8}x_3 \\ \frac{1}{8}x_2 + \frac{3}{8}x_2 \end{pmatrix}\end{aligned}$$

To measure DeGroot centrality, let us first consider the stubborn-avoiding weighted walks from a

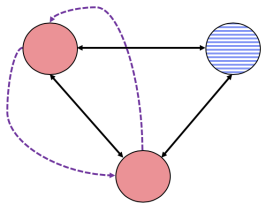

Figure 5. Triangle Network (shaded agent = Stubborn; solid agents = DeGroot). Solid lines represent social network connections while dashed lines represent weighted walks that avoid stubborn agents.

DeGroot agent i back to itself. There is a unique such walk of length $2r$ for $r = 0, 1, 2, \ldots$ from i to i, with weight $(1/3)^{2r}$ (by simply pinging back and forth between the two DeGroots, as in the dashed lines). Therefore, the total weight of stubborn-avoiding walks from i to i is $\sum_{r=0}^{\infty}(1/3)^{2r} = \frac{9}{8}$. Similarly,

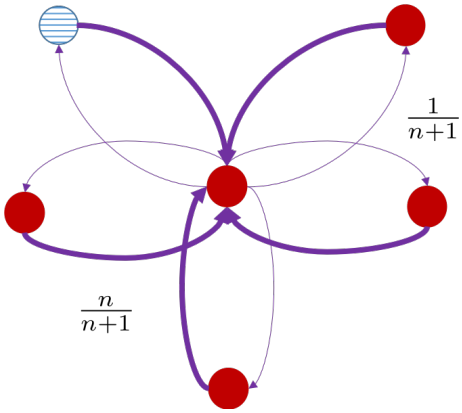

Figure 6. Influential Star Network. A weighted directed arrow from node i to node j indicates that i puts that much weight on j's belief. Shaded node represents stubborn agents.

there is a unique stubborn-avoiding walk of length $2r+1$ for $r = 0, 1, 2, \ldots$ from i to $j \neq i$, with weight $(1/3)^{2r+1}$. Therefore, the total weight of walks from i to j is $\sum_{r=0}^{\infty}(1/3)^{2r+1} = \frac{3}{8}$. Using Definition 2, we see that for DeGroot i:

$$\begin{aligned}\mathcal{D}_i(\boldsymbol{\gamma}) &= \frac{9}{8}\theta_i\gamma_i + \frac{3}{8}\theta_j\gamma_j \\ &= \frac{3}{8}x_i + \frac{1}{8}x_j\end{aligned}$$

which is equal to her belief of the incorrect state (as anticipated by Proposition 2). Thus in the above network, if the principal targets both DeGroots, their common belief in the incorrect state will be equal to $1/2$. Note that we get the same results if we instead use the recursive definition of DeGroot centrality. When $x_2 = x_3 = x$, we obtain by symmetry:

$$\begin{aligned}\mathcal{D}_i(\boldsymbol{\gamma}) &= \frac{1}{3}\cdot x + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot \mathcal{D}_i(\boldsymbol{\gamma}) \\ \implies \frac{2}{3}\mathcal{D}_i(\boldsymbol{\gamma}) &= \frac{1}{3}x \\ \implies \mathcal{D}_i(\boldsymbol{\gamma}) &= \frac{1}{2}x\end{aligned}$$

which coincides with the previous calculation when $x_2 = x_3 = x \in \{0, 1\}$. □

B.2 Applications of Theorem 4

Example 3 (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 exactly $\log(n+1)$ for any $n \geq 2$. Therefore, only a constant number of stubborn agents are needed by Theorem 4 (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 4 (Influential Star Network)**.** Consider Figure 6 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

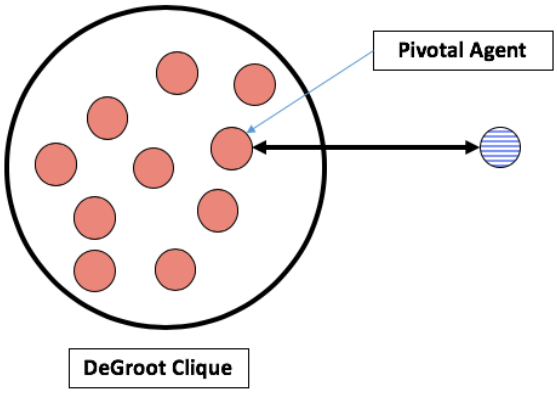

Figure 7. A clique of DeGroot agents with a single connection to a stubborn agent.

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 stubborn agents satisfies $m \geq 2(1+b)/(1-b)$, the network is impervious. This is true even when all of the stubborn agents 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 stubborn agents would not be sufficient to make the network impervious, if placed in better positions. For instance, a single stubborn agent in the center of the star *always* makes the network impervious when n is large enough. To see this, we can apply the local density result of Proposition 11,[21] by considering subsets $I_\ell = \{1, \ell\}$ for $\ell = 2, \ldots, n$. The log-distance of each I_ℓ is given by $\log(1+1/n) = \log(|I_\ell| + 1/n - 1) \leq \log(|I_\ell| + 1)$. Thus, when $b = 0$ and applying the bound in Example 3, we see that if the stubborn agent is the central agent, the network is impervious (whereas we would require $m \geq 2$ on the periphery). □

Example 5 (Echo chambers)**.** Consider Figure 7, with a clique of size $n-1$ consisting entirely of DeGroot agents and a single stubborn agent. We assume the clique and the stubborn agents are joined by just a single (bidirectional) link connecting the stubborn agents with one DeGroot in the clique, and with no other connections going between the islands. We call this DeGroot agent the *pivotal agent*. This defines an undirected social network $\mathbf{G}^*$. For simplicity, let $\mathbf{G}$ have the weights given by $\theta_i = \alpha_{ij} = 1/(1+|N(i)|)$ whenever $i \to j$ in $\mathbf{G}^*$.

We see that the pivotal agent can reach any other agent with a path of log-weight at most $-\log\left(\frac{1}{n}\right) = \log(n)$ and therefore satisfies the density condition with $\delta = 0$. Notice, however, that if the principal targets all agents in the DeGroot clique, then when n is large, the pivotal agent will still have an arbitrarily incorrect belief, i.e., $\pi_{i,T}(R) \to 1$ as $n \to \infty$. To see this, note that as $n \to \infty$, almost every walk (of any length) from the pivotal agent ends up at another DeGroot agent.[22] Because DeGroots only

[21] We cannot apply Theorem 4 here because the stubborn agents disconnects all the DeGroots from each other, so the log-diameter is $+\infty$.

[22] To be precise, $\mathbf{G}$ has weights that represent a random walk for all DeGroot agents, where the agent chooses a link uniformly at random. The probability the walk ever reaches a stubborn agent is $\frac{1}{n}\left(1 + \left(\frac{n-1}{n}\right)\left(\frac{1}{n-1}\right)\sum_{k=0}^{\infty}\left(\frac{n-2}{n-1}\right)^k\right) = \frac{1}{n}\left(\frac{n^2+n-1}{n^2}\right)$, and as $n \to \infty$, tends toward 0.

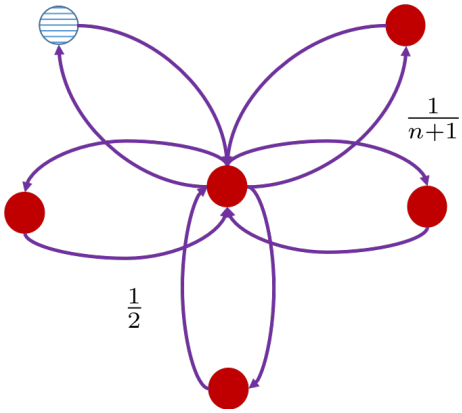

Figure 8. Balanced Star Network

talk amongst themselves, there is an *echo chamber* whereby the misinformation sent by the principal circulates within the DeGroot island and the beliefs of the stubborn agents never propagate. Therefore, while the pivotal agent is close to the stubborn agent, the fact that most of her friends, and friends of friends are not, almost all of the influence exerted on the pivotal agent comes from others exposed to misinformation.

Compare this to the case where *every* DeGroot agent is pivotal, which now satisfies the log-diameter condition for $\delta = 1$. Even though DeGroot agents are friends almost exclusively with other DeGroot agents, who receive possible misinformation, Theorem 4 guarantees the network is impervious. This is precisely because an echo chamber effect no longer amplifies incorrect beliefs of the agents, simply because each DeGroot agent is friends with at least *one* stubborn agent, limiting the principal's influence. □

Finally, we expand on Example 5 to show how Theorem 4 applies only to *log-diameter*, which may not coincide with the notion of diameter in undirected networks:

Example 6 (Echo chambers, revisited)**.** Consider a variant of the unweighted social network $\mathbf{G}^*$ of the "echo chambers" network from Example 5, but now where there are two cliques of size $n/2$, one clique which is all DeGroot and one clique which is all stubborn, with a single connection between them. Note that $\mathbf{G}^*$ has a diameter of 3, since it is possible to get from any agent in one clique to any other agent in the other clique with a walk that has no more than 3 steps. Because the diameter of the network stays constant with n, it is natural to classify this as a "small diameter" network.

Yet, straightforward computation reveals that the log-diameter of $\mathbf{G}$ is $\log\left(\frac{n^4}{2(n-1)^2}\right) \approx \log(n^2/2)$, which does not satisfy the conditions of Theorem 4 for any δ. In fact, as we saw before in Example 5, no constant number of stubborn agents are guaranteed to make this network impervious. Therefore, having a small diameter in $\mathbf{G}^*$, even as n grows, does not necessarily imply the conditions of Theorem 4 will be satisfied for small log-diameter. □

B.3 Sparse Networks

Finally, the last sparse example is the *balanced star network*, where agents are aligned in a star network but employ equal-influence weighting. We show that despite the seemingly added symmetry, as compared to Example 4, the network fails to satisfy the log-diameter condition, and so introduces unique vulnerabilities not present in the asymmetric star network of Example 4.

Example 7 (Balanced Star Network)**.** Consider the balanced star network of Figure 8. Suppose that for agents on the periphery $\theta_i = \alpha_{i1} = 1/2$ whereas the core agent 1 updates as in Example 4, $\theta_1 = \alpha_{1j} = 1/(n+1)$. The log-diameter condition is unsatisfied because the log-diameter grows as $\approx \log(2n)$.

When the central agent is stubborn, 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 stubborn agents 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 stubborn agents is sublinear), Stubborn agents 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 $\mathcal{D}_i(\boldsymbol{\gamma}) = \frac{1}{2}\gamma_i + \frac{1}{2}||\boldsymbol{\gamma}||_1/n$. 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 stubborn agents, the network is impervious if and only if $\varepsilon < \max\{1/(1-b), 1\}$ for large n; otherwise, a *linear* number of stubborn agents 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 4, 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 stubborn agents on the periphery. □

To conclude, we present an application of the results in Section 6.3 which allow us to characterize when a network is almost impervious but still has a few agents manipulated:

B.4 k-imperviousness

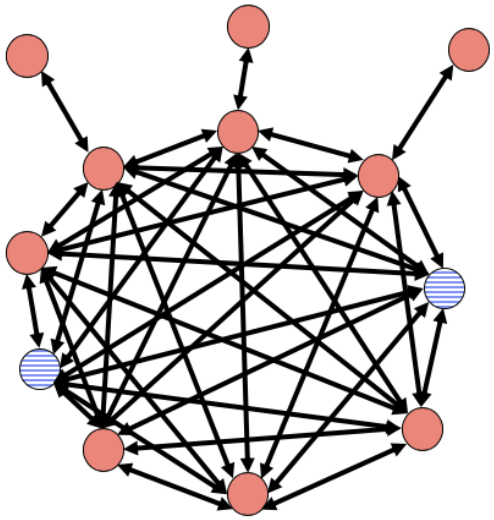

Figure 9. Core-Periphery Network.

Example 8. Consider the (unweighted) core-periphery network $\mathbf{G}^*$ shown in Figure 9, with $n-k$ agents in the core and k agents on the periphery who listen to only one agent in the core (in Figure 9, $k = 3$). Suppose the weights are given by the equal-influence weighting scheme. Fixing k, the log-diameter of the network is bounded below by $\log(3n)$ for sufficiently large n, which does not satisfy

the conditions of Theorem 4 for any value of δ as n grows. On the other hand, there is an obvious k-cut which leaves the complete network as k-cut subnetwork (and after removing u), which has a log-diameter bounded above by $\log(n+1)$ (for sufficiently large n), and therefore is k-impervious with at least $m^*(k+1)$ stubborn agents located in the core via Example 3 (the complete network is dense) and Corollary 1 (density condition for k-impervious). □

C Numerical Experiments

The previous examples 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 stubborn 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 $\frac{1}{n}$). We start with Figure 10 as an illustration that shows the network with only a single knowledgeable stubborn agent. Throughout the figures in this section, green nodes represent stubborn agents, and nodes represented with an asterisk indicate agents directly targeted by the principal (according to his *optimal* strategy). Conversely, DeGroot agents are colored either blue or red, to indicate whether under the principal's 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 stubborn 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 10, Agent 70 (with degree 7 and eigenvector centrality 0.0121) is a stubborn agent whose location results in no manipulation, because the principal has no profitable strategy with which he can manipulate even a single member of the population. Naturally, all agents have a DeGroot centrality of 0 when the principal chooses not to exert any influence.

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 stubborn agent, as is the case in Figure 11, then the average DeGroot centrality (and terminal belief under the principal's optimal strategy) is $\bar{\pi} = 0.529 > \pi_{\text{cutoff}} \equiv 0.5$ and manipulation is inevitable and quite severe.

These two cases are summarized in Figure 12. Each dot in this graph represents the DeGroot centrality of the corresponding agent in the network under one of the these two stubborn agent placements, and under a particular strategy for the principal:

1. *Optimal influence*: corresponds to the DeGroot centrality of the agents when the principal exerts the influence he would in his optimal strategy.

2. *Max influence*: corresponds to the DeGroot centrality of the agents when the principal targets every DeGroot agent, even though such influence may be "overkill" or ineffective.

Agents whose DeGroot centralities are above $\pi_{\text{cutoff}} = 0.5$ are manipulated. Yellow dots correspond to the DeGroot centality of the agents in Figure 10 (with Agent 70), but when the principal employs max influence. Notice that all the yellow dots are below the threshold of π_{cutoff}, and hence no agent

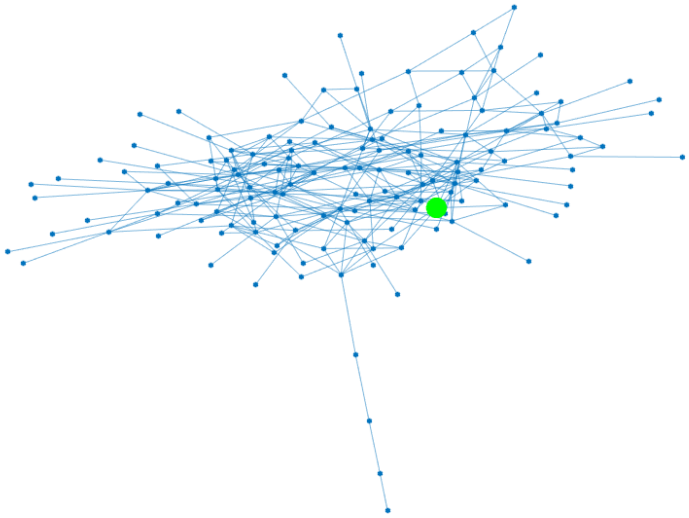

Figure 10. Central Stubborn Agent, $b = 0$.

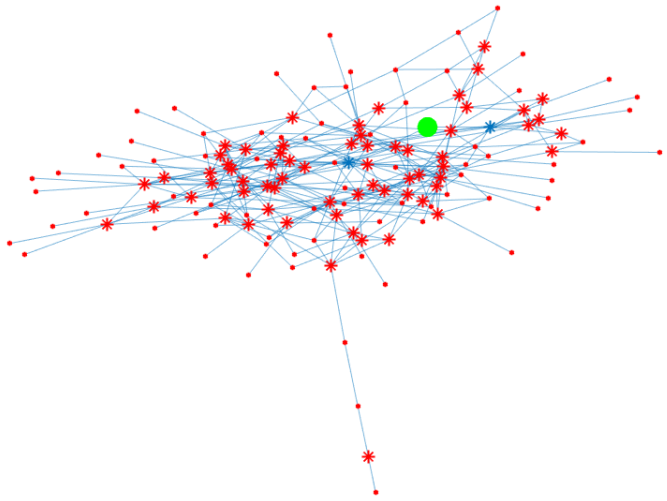

Figure 11. Peripheral Stubborn Agent, $b = 0$.

is manipulated despite the most intensive efforts of the principal. Thus, the stubborn agent communicates the truth effectively, and the principal cannot interfere. On the other hand, if the principal applies the same max-influence strategy to the network in Figure 11 (with Agent 59 as the stubborn agent) 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 12 however are the purple dots lying just above the dotted cutoff line, corresponding the principal's optimal strategy. These dots represent the DeGroot centralities of the agents in Figure 11 when the principal applies the optimal 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 (such as Agent 60 in the figure) escaping manipulation ($< 2\%$ of the population).

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* stubborn agent. As b 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 13 shows that with $b = 0.5$, two stubborn 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 stubborn agents matter. Figure 14 shows that even with five stubborn agents agents, large-scale manipulation is possible because these agents occupy less central positions. In the case of the complete network, three stubborn agents are both necessary and sufficient for imperviousness when $b = 0.5$; in other words, the best-case placement in this network is better than in the complete network (requires only two stubborn agents) but the worst-case placement in this network is also worse than the complete network (requires at least six stubborn agents).

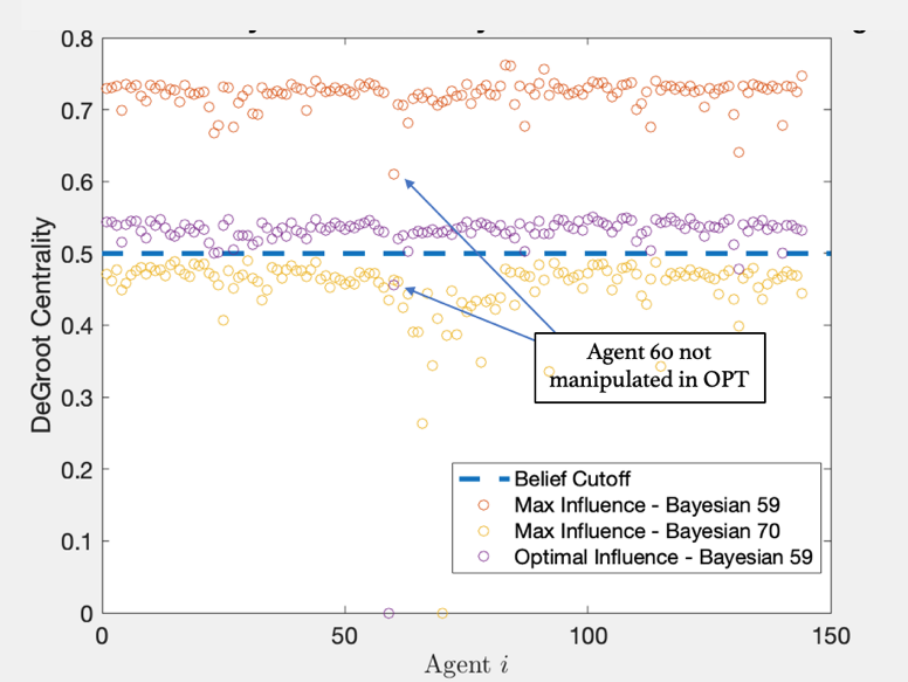

Figure 12. DeGroot Centrality for Single Stubborn Agent, $b = 0$.

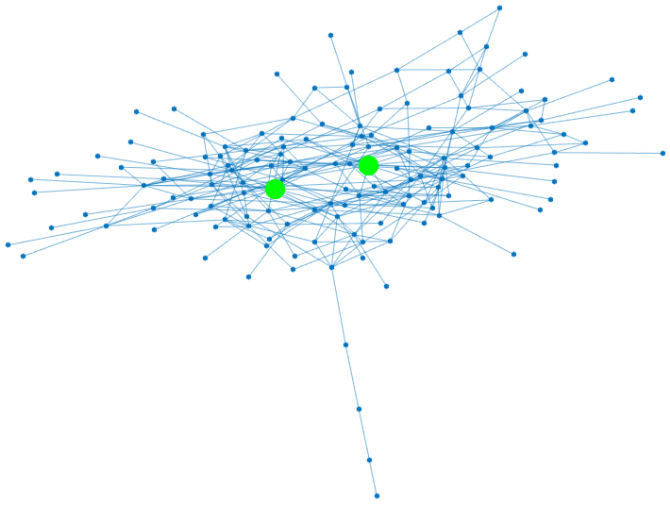

Figure 13. Two Well-Placed Knowledgeable Stubborn Agents, $b = 0.5$.

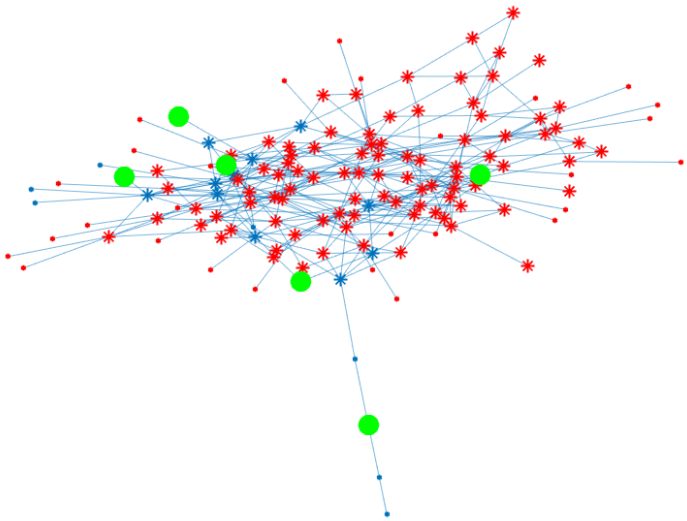

Figure 14. Five Poorly-Placed Knowledgeable Stubborn Agents, $b = 0.5$.

D Proofs

D.1 Section 3

Proof of Lemma 1. We first prove that $\text{BU}(S|h_{i,t})$ is a martingale. Consider the filtration with respect to the history $h_{i,t}$. Then:

$$\begin{aligned}\mathbb{E}\left[\text{BU}(S|h_{i,t+1})|h_{i,t}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\theta=S}|h_{i,t+1}\right]|h_{i,t}\right] \\ &= \mathbb{E}[\mathbf{1}_{\theta=S}|h_{i,t}] \\ &= \text{BU}(S|h_{i,t})\end{aligned}$$

where the second to last inequality follows from the law of iterated expectations. Because the Bayesian update term is a belief and bounded between 0 and 1, we know by the martingale convergence theorem that $\text{BU}(S|h_{i,t})$ converges almost surely to a random variable X. We next prove that X is a constant almost surely. If $p_i = 1/2$, then $\text{BU}(S|h_{i,t}) = \text{BU}(S|h_{i,0}) = q$ for all t and so trivially converges to constant q. Otherwise, we know that if $\gamma_i = 1$ then DeGroot agent i receives signal R with probability $\frac{\lambda^*}{\lambda+\lambda^*} + \frac{\lambda}{\lambda+\lambda^*}(1-p_i) > 1/2$ by Assumption 1. We show that $\text{BU}(S|h_{i,t})$ converges almost surely to 0. Consider the biased random walk $z^{\Delta}_{i,t} = z^R_{i,t} - z^S_{i,t}$. For all t we can write:

$$\begin{aligned}\text{BU}(S|h_{i,t}) &= \frac{p_i^{z^S_{i,t}}(1-p_i)^{z^R_{i,t}}q}{p_i^{z^S_{i,t}}(1-p_i)^{z^R_{i,t}}q + p_i^{z^R_{i,t}}(1-p_i)^{z^S_{i,t}}(1-q)} \\ &= \frac{q}{q + \left(\frac{p_i}{1-p_i}\right)^{z^{\Delta}_{i,t}}(1-q)} \\ &\overset{a.s.}{\to} 0\end{aligned}$$

because for a biased random walk with the probability of R greater than $1/2$, we know that $z^{\Delta}_{i,t} \overset{a.s.}{\to} \infty$, and $p_i > 1/2$.

Similarly, if $\gamma_i = 0$, then DeGroot agent i receives signal S with probability $p_i > 1/2$. We show that $\text{BU}(S|h_{i,t})$ converges almost surely to 1. Consider the same biased random walk; then for all t we can write:

$$\begin{aligned}\text{BU}(S|h_{i,t}) &= \frac{p_i^{z^S_{i,t}}(1-p_i)^{z^R_{i,t}}q}{p_i^{z^S_{i,t}}(1-p_i)^{z^R_{i,t}}q + p_i^{z^R_{i,t}}(1-p_i)^{z^S_{i,t}}(1-q)} \\ &= \frac{q}{q + \left(\frac{1-p_i}{p_i}\right)^{-z^{\Delta}_{i,t}}(1-q)} \\ &\overset{a.s.}{\to} 1\end{aligned}$$

because for a biased random walk with the probability of S greater than $1/2$, we know that $-z^{\Delta}_{i,t} \overset{a.s.}{\to} \infty$, and $p_i > 1/2$. □

Lemma 2. *The spectral radius of matrix* $\mathbf{W}$ *is strictly less than 1.*

Proof. It is equivalent to prove that all the eigenvalues of $\mathbf{W}$ lie strictly within the unit circle. For stubborn agents or DeGroot agents with $\theta_i = 1$, these agents have $\alpha_{ij} = 0$ for all j, so $\mathbf{W}_i$ is the zero vector. Thus, these agents introduce an additional eigenvalue of 0, which of course lies within the unit circle, without affecting the rest of the eigenvalues. Therefore, it is without loss of generality to

consider DeGroot agents with $\theta_i < 1$ for all agents i (where we assign arbitrary θ values for those with θ equal to 1 (as we have already identified this as irrelevant.) Then let us define the diagonal matrix:

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{1-\theta_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{1-\theta_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{1-\theta_n} \end{pmatrix}$$

Then we note that $\mathbf{QW}$ 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. Further, because there is at least one agent with $p_i > 1/2$, by Assumption 1, this agent is either stubborn or DeGroot with $\theta_i > 0$, so $\mathbf{Q}$ has at least one eigenvalue strictly greater than 1, with corresponding eigenvector $\mathbf{v}^*$. Moreover, none of the eigenvalues of $\mathbf{Q}$ are less than or equal to 1.

Consider any arbitrary vector $\mathbf{v} \in \mathbb{R}^{n-m}$. By Assumption 1 (strong connectedness), we know there exists k such that $\mathbf{QW}^k\mathbf{v}$ is not a scalar-multiple of $\mathbf{v}^*$, and so we obtain the strong inequality:

$$||\mathbf{W}^k\mathbf{v}||_2 < ||\mathbf{QW}^k\mathbf{v}||_2 \leq ||\mathbf{v}||_2$$

Moreover, we obtain the weak inequality on the eigenvalues of $\mathbf{W}$:

$$||\mathbf{Wv}||_2 \leq ||\mathbf{QWv}||_2 \leq ||\mathbf{v}||_2$$

The weak inequality shows the eigenvalues of $\mathbf{W}$ lie (weakly) within the unit circle. Since the eigenvalues of $\mathbf{W}^k$ are k-powers of the eigenvalues of $\mathbf{W}$, we see by the strong inequality that no eigenvalue can lie precisely on the unit circle. □

Lemma 3. *Under Assumption 1, the beliefs of the agents, $\boldsymbol{\pi}_t$, converge almost surely to some $\boldsymbol{\pi}_\infty$.*

Proof. Fix $\delta > 0$. Recall that $\boldsymbol{\pi}_t = \boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) + \mathbf{W}\boldsymbol{\pi}_{t-1}$ for the DeGroot agents and we can treat stubborn agents as DeGroots with $\theta_i = 1$ and $\gamma_i = 0$. Notice by induction one can show that $\mathbf{0} \leq \boldsymbol{\pi}_t \leq \mathbf{1}$ (it is a belief): because $\boldsymbol{\pi}_0 = q\mathbf{1}$ and every belief update is a convex combination of $\text{BU}(\mathbf{h}_t)$, which lies between $\mathbf{0}$ and $\mathbf{1}$, and neighboring beliefs in the period $t-1$, which by the inductive hypothesis lie between 0 and 1, $\boldsymbol{\pi}_t$ must lie between $\mathbf{0}$ and $\mathbf{1}$. Moreover, by Lemma 1, $\text{BU}(\mathbf{h}_t)$ converges to a constant vector almost surely. Now let us write:

$$\begin{aligned} ||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 &= ||\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})\boldsymbol{\pi}_{t-1}||_2 \\ &= ||\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})(\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_{t-1}) + \mathbf{W}\boldsymbol{\pi}_{t-2})||_2 \\ &= ||\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})(\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_{t-1}) + \mathbf{W}(\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_{t-2}) + \mathbf{W}\boldsymbol{\pi}_{t-3}))||_2 \\ &= ||\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})(\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_{t-1})) - (\mathbf{I} - \mathbf{W})\mathbf{W}(\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_{t-2})) - (\mathbf{I} - \mathbf{W})\mathbf{W}^2\boldsymbol{\pi}_{t-3}||_2 \end{aligned}$$

Repeating this, we see that for any $t \geq M - 2$:

$$||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 = \left|\left|\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}) \sum_{k=0}^{M} (\mathbf{W}^k \text{BU}(\mathbf{h}_{t-k-1})) - (\mathbf{I} - \mathbf{W})\mathbf{W}^{M+1}\boldsymbol{\pi}_{t-M-2}\right|\right|_2$$

Because $\mathbf{W}$ has a spectral radius which is strictly less than 1 by Lemma 2, we know that $\lim_{k\to\infty} \mathbf{W}^k = \mathbf{0}$. Moreover, since $\boldsymbol{\pi}_t$ is bounded between $\mathbf{0}$ and $\mathbf{1}$, we know there exists some M^* such that $||(\mathbf{I} -$

$\mathbf{W})\mathbf{W}^{M^*+1}\boldsymbol{\pi}_{t-M^*-2}||_2 \leq \frac{\delta}{3}$ and $||\boldsymbol{\theta} \odot \mathbf{W}^{M^*+1}\mathbf{1}||_2 \leq \frac{\delta}{3}$. Thus, for this value of M^* and any $t \geq M^* - 2$:

$$||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 \leq \left|\left|\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}) \sum_{k=0}^{M^*} \mathbf{W}^k \text{BU}(\mathbf{h}_{t-k-1})\right|\right|_2 + \frac{\delta}{3}$$

Because $\text{BU}(\mathbf{h}_t)$ converges to a constant almost surely by Lemma 1, we know there exists T^* almost surely such that for all $t > T^*$, $||\boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W})(\text{BU}(\mathbf{h}_t) - \text{BU}(\mathbf{h}_{t-k-1}))||_2 < \frac{\delta}{3(M^*+1)}$ for all $0 \leq k \leq M^*$ by the Cauchy criterion of convergence. Thus, for all $t > \max\{M^* - 2, T^*\}$:

$$\begin{aligned}
||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 &\leq \left|\left|\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}) \sum_{k=0}^{M} (\mathbf{W}^k \text{BU}(\mathbf{h}_t))\right|\right|_2 + \frac{2\delta}{3} \\
&< \left|\left|\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}^{M^*+1}) \text{BU}(\mathbf{h}_t)\right|\right|_2 + \frac{2\delta}{3} \\
&\leq \left|\left|\boldsymbol{\theta} \odot \mathbf{W}^{M^*+1}\mathbf{1}\right|\right|_2 + \frac{2\delta}{3}
\end{aligned}$$

(Note that because the spectral radius of $\mathbf{W}$ is less than 1 by Lemma 2, $\sum_{k=0}^{M} \mathbf{W}^k = (\mathbf{I} - \mathbf{W})^{-1}(\mathbf{I} - \mathbf{W}^{M+1})$.) Recall we chose M^* such that the first term in the last expression does not exceed $\delta/3$. Thus, $||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 < \delta$ for all $t > \max\{M^* - 2, T^*\}$, which completes the proof. □

Proof of Theorem 1. By Lemma 1 and Lemma 3 we know that both $\text{BU}(\mathbf{h}_t)$ and $\boldsymbol{\pi}_t$ converge almost surely to $\text{BU}(\mathbf{h}_\infty)$ and $\boldsymbol{\pi}_\infty$, respecitlvely. Thus, $\boldsymbol{\pi}_\infty$ must solve the fixed-point problem:

$$\boldsymbol{\pi}_\infty = \boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_\infty) + \mathbf{W}\boldsymbol{\pi}_\infty$$

If not, then the difference between the left-hand side and right-hand side is always some positive amount η, and so every iteration of belief updating changes the belief by at least η, contradicting convergence. By Lemma 2, we know that all eigenvalues of $\mathbf{W}$ lie within the unit circle, so $\mathbf{I} - \mathbf{W}$ is invertible, and thus we can solve this fixed-point problem explicitly:

$$\boldsymbol{\pi}_\infty = (\mathbf{I} - \mathbf{W})^{-1}(\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_\infty))$$

which proves the claim of Proposition 1. □

Proof of Proposition 1. Whenever $p_i > 1/2$, by Lemma 1, the personal Bayesian update component (BU) of the DeGroot update converges almost surely to belief 1 on the true state, so $\text{BU}_i(h_{i,\infty}(\mathbf{0}))(R|S) \overset{a.s.}{\to} 0$. On the other hand, when $p_i = 1/2$ we have $\theta_i = 0$ by Assumption 1, so $\text{BU}_i(\mathbf{h}_\infty(\mathbf{0}))(y'|y) \odot \boldsymbol{\theta} \overset{a.s.}{\to} 0$. This implies that $\text{BU}(\mathbf{h}_\infty(\mathbf{0}))(y'|y) \odot \boldsymbol{\theta} \overset{a.s.}{\to} \mathbf{0}$ trivially. By Proposition 1, we see that for R:

$$\begin{aligned}
\boldsymbol{\pi}_t(R) &\overset{a.s.}{\to} (\mathbf{I} - \mathbf{W})^{-1}(\text{BU}(\mathbf{h}_\infty(\mathbf{0}))(R) \odot \boldsymbol{\theta}) \\
&= (\mathbf{I} - \mathbf{W})^{-1}\mathbf{0} \\
&= \mathbf{0}
\end{aligned}$$

Thus, $\boldsymbol{\pi}_t(S) \overset{a.s.}{\to} \mathbf{1}$, and agents learn the true state almost surely. □

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

By Lemma 3, 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 $\mathcal{B}$ which consists of all the values of b where some agent i has a limit belief $\lim_{t\to\infty} \pi_{i,t} \overset{a.s.}{\to} (1-b)/2$, for some network action $\mathbf{x}$. Note there is only one such b value per agent, given by $1 - 2\pi_{i,\infty}$. Thus, provided there are finitely many agents and finitely many principal influence actions, the set $\mathcal{B}$ is finite, so has measure zero, implying that $(-1,1)\backslash\mathcal{B}$ has full measure. Moreover, every agent either picks the correct terminal action or the incorrect terminal action, almost surely, for all $b \in (-1,1)\backslash\mathcal{B}$.

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

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

by Lemma 3. Since the set of $\mathcal{B}$ contains no b's with an agent holding $\pi_{i,\infty} = (1-b)/2$, we can pick T^* large enough and ζ 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 $k_1 - \varepsilon||\mathbf{x}||_1$ with probability $1-\kappa$ (which we deem the "likely payoff") and some other payoff with probability κ, where k_1 is the number of manipulated agents under (pure) strategy $\mathbf{x}$.

Now suppose two network strategies $\mathbf{x}_1, \mathbf{x}_2$ have a different number of manipulated agents, k_1 and k_2, respectively. If $\mathbf{x}_1$ and $\mathbf{x}_2$ give the same likely payoff, this implies that $k_1 - \varepsilon||\mathbf{x}_1||_1 = k_2 - \varepsilon||\mathbf{x}_2||_1$, which implies that:

$$\varepsilon = \frac{k_1 - k_2}{||\mathbf{x}_1||_1 - ||\mathbf{x}_2||_1}$$

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 expected payoff contribution of any "unlikely" payoff is bounded above by $n \cdot \kappa$, and $\kappa \to 0$). Thus, for the set of irrational ε and $b \in (-1,1)\backslash\mathcal{B}$, the principal plays the strategy over network actions which induces the "likely" outcome of that network action with probability at least $1-\kappa$. In such a strategy, the number of manipulated agents then must be the same, and the all of the principal's optimal strategies are manipulation-invariant. □

Proof of Proposition 2. We prove by induction that $\mathbf{W}^\ell_{ij}$ represents the sum of weighted walks of length ℓ between i and j, not passing through a stubborn agent. The base case of $\ell = 0$ is clear because every agent has a walk of length 0 to themselves of weight 1, and none others. Note that:

$$\begin{aligned}
\mathbf{W}^{\ell+1}_{ij} = [\mathbf{W}\mathbf{W}^\ell]_{ij} &= \sum_{k=1}^n w_{ik}\mathbf{W}^\ell_{kj} \\
&= \sum_{k=1}^n w_{ik} \cdot [\text{weight of walks of length } \ell \text{ between } k \text{ and } j] \\
&= [\text{weight of walks of length } \ell+1 \text{ between } i \text{ and } j]
\end{aligned}$$

Therefore, the total weight of walks between i and j (avoiding stubborn agents) is given by $\sum_{W\in\mathcal{W}_{ij}} w_W = \sum_{\ell=0}^\infty \mathbf{W}^\ell = (\mathbf{I}-\mathbf{W})^{-1}$ since the spectral radius of $\mathbf{W}$ is strictly less than 1, by Lemma 2. Finally, note

that by Proposition 1:

$$\begin{aligned}\pi_{i,\infty}(\mathbf{x}^*) &= (\mathbf{I}-\mathbf{W})_i^{-1}(\boldsymbol{\gamma}(\mathbf{x}^*)\odot\boldsymbol{\theta}) \\ &= \sum_{j=1}^{n}(\mathbf{I}-\mathbf{W})_{ij}^{-1}\gamma_j(\mathbf{x}^*)\theta_j \\ &= \sum_{j=1}^{n}\gamma_j(\mathbf{x}^*)\theta_j\left(\sum_{W\in\mathcal{W}_{ij}} w_W\right) \\ &= \mathcal{D}_i(\boldsymbol{\gamma})\end{aligned}$$

As this holds for every i, we have $\mathcal{D}(\boldsymbol{\gamma}) = (\mathbf{I}-\mathbf{W})^{-1}(\boldsymbol{\gamma}\odot\boldsymbol{\theta})$. □

D.2 Section 4

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

$$\boldsymbol{\pi}_\infty(R) = (\mathbf{I}-\mathbf{W}_{\theta'}^{-1})(\boldsymbol{\gamma}\odot\boldsymbol{\theta}')$$

We first prove that the asymptotic bound for DeGroot beliefs is continuous in θ' around $\theta' = 0$. Clearly the network preservation of $\mathbf{W}_{\theta'}$ is continuous in θ', so it is sufficient to prove that as $\theta' \to 0$, $\mathbf{I}-\mathbf{W}_{\theta'}$ is non-singular. To see this, note the eigenvalues of $\mathbf{W}_{\theta'}$ are uniformly bounded away from the unit circle as $\theta' \to 0$ (and thus $\mathbf{I}-\mathbf{W}_{\theta'}$ is non-singular as $\theta' \to 0$), so one can apply the same reasoning as Lemma 2, noting that the existence of at least one stubborn agent guarantees $\mathbf{W}$ is still substochastic. Thus, provided $(\mathbf{I}-\mathbf{W}_{\theta'})^{-1}$ is a continuous operation at $\theta' = 0$, we can substitute $\theta' = 0$ and apply DeGroot centrality with influence vector $\boldsymbol{\gamma} \leq \mathbf{1}$, showing that all DeGroot centralities tend to 0, so beliefs of the correct state tend toward 1. This yields the claim in (a).

Because $\lim_{\theta'\to 1}\mathbf{W}_{\theta'} = \mathbf{0}$, it is obvious that beliefs are continuous at $\theta' = 1$. Moreover, when $\theta' = 1$, any 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 DeGroots as $\mathbf{1}_D$, 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 $\mathbf{1}_D$ with the only difference being that he does not target the DeGroot agent not adjacent to any stubborn agents; call this strategy $\mathbf{x}_{\text{spec}}$. 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 $\mathbf{1}_D$. 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 the strategy $\mathbf{x}_2$ which violates independence with θ_2 by targeting all agents except agent i^* who is manipulated. 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, given that all other agents receive $\gamma_j = 1$ but agent i has $\gamma_i = 0$. 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 $\varepsilon^* > 1 - 1/n$ such that $\theta' \in (\theta^*, \overline{\theta})$ is susceptible (because $\mathbf{x}_{\text{spec}}$ dominates $\mathbf{0}$) but where $\mathbf{x}_D$ 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 3. 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, as $\theta_1 < \overline{\theta}$. 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 Theorem 3(a). □

Proof of Proposition 4. Let there be M manipulated agents under optimal strategy $\mathbf{x}$ with influence cost ε, so the principal has a payoff of $M - \varepsilon||\mathbf{x}||_1$. After we increase ε to ε', suppose the principal manipulates more agents; it necessarily must be the case that $||\mathbf{x}'|| > ||\mathbf{x}||$, otherwise $\mathbf{x}$ is strictly preferred to $\mathbf{x}'$ for any influence cost, so cannot be optimal with ε'. But then, of course:

$$\begin{aligned}
M' - \varepsilon||\mathbf{x}'||_1 &= M' - \varepsilon'||\mathbf{x}'||_1 + (\varepsilon' - \varepsilon)||\mathbf{x}'||_1 \\
&\geq M - \varepsilon'||\mathbf{x}||_1 + (\varepsilon' - \varepsilon)||\mathbf{x}'||_1 \\
&= M - \varepsilon||\mathbf{x}'||_1 + \varepsilon'(||\mathbf{x}'||_1 - ||\mathbf{x}||_1) \\
&\geq M - \varepsilon||\mathbf{x}'||_1 + \varepsilon(||\mathbf{x}'||_1 - ||\mathbf{x}||_1) \\
&\geq M - \varepsilon||\mathbf{x}||_1
\end{aligned}$$

which contradicts the optimality of $\mathbf{x}$ when the influence cost is ε.

Note that the DeGroot centrality of the agents under the same strategy $\mathbf{x}$ does not depend on b, but the cutoff necessary to take the incorrect action is $(1-b)/2$, so is decreasing in b. Thus, the number of manipulated agents (for a fixed network strategy), k, is non-decreasing in b. Therefore, if there exists some strategy $\mathbf{x}$ where $M - \varepsilon||\mathbf{x}||_1 > 0$, then when b increases to b', we know $M' \geq M$, so the same strategy $\mathbf{x}$ yields $M' - \varepsilon||\mathbf{x}||_1 > 0$. By Corollary 2, the network with $b' > b$ is susceptible. □

Proof of Example 1. First, consider the belief cutoff $\pi_{\text{cutoff}}(R) = 0.35$ and where the principal targets agents 1 and 3. Then beliefs of the incorrect state are given by:

$$\boldsymbol{\pi}(R) = \begin{pmatrix} 1 & -1/2 & 0 & 0 \\ -1/3 & 1 & -1/3 & 0 \\ 0 & -1/3 & 1 & -1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1/2 \\ 0 \\ 1/3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.692 \\ 0.385 \\ 0.462 \\ 0 \end{pmatrix},$$

and all three DeGroot agents are manipulated, yielding a payoff of $3-2\varepsilon>0$. Targeting all three agents will also lead to these three agents being manipulated, but increases the cost with no additional benefit. Clearly, if the principal targets no one, then all beliefs of R will be 0, which yields no profit. Thus, the only potential for a better strategy would be if the principal can manipulate two or more agents by sending signals to only one:

1. *Send to agent 1 only*: Only agent 1 is manipulated.

$$\boldsymbol{\pi}(R)=\begin{pmatrix}1&-1/2&0&0\\-1/3&1&-1/3&0\\0&-1/3&1&-1/3\\0&0&0&1\end{pmatrix}^{-1}\begin{pmatrix}1/2\\0\\0\\0\end{pmatrix}=\begin{pmatrix}0.615\\0.231\\0.077\\0\end{pmatrix}$$

2. *Send to agent 2 only*: Only agent 2 is manipulated.

$$\boldsymbol{\pi}(R)=\begin{pmatrix}1&-1/2&0&0\\-1/3&1&-1/3&0\\0&-1/3&1&-1/3\\0&0&0&1\end{pmatrix}^{-1}\begin{pmatrix}0\\1/3\\0\\0\end{pmatrix}=\begin{pmatrix}0.231\\0.462\\0.154\\0\end{pmatrix}$$

3. *Send to agent 3 only*: Only agent 3 is manipulated.

$$\boldsymbol{\pi}(R)=\begin{pmatrix}1&-1/2&0&0\\-1/3&1&-1/3&0\\0&-1/3&1&-1/3\\0&0&0&1\end{pmatrix}^{-1}\begin{pmatrix}0\\0\\1/3\\0\end{pmatrix}=\begin{pmatrix}0.077\\0.154\\0.385\\0\end{pmatrix}$$

Thus, the optimal strategy with $\pi_{\text{cutoff}}=0.35$ is to target agents 1 and 3 and all agents are manipulated.

Now let us consider the case of $\pi_{\text{cutoff}}=0.2$ (manipulation is easier). Once again consider the three cases from before. In the cases when only agent 1 is targeted, both agents 1 and 2 are manipulated; when only agent 2 is targeted, both agents 1 and 2 again end up manipulated; when agent 3 is targeted, only agent 3 is manipulated. Thus there is a strategy that obtains a payoff of $2-\varepsilon>3-2\varepsilon>3-3\varepsilon$. Therefore, no strategy that targets more agents (even if all three agents are manipulated!) beats the strategy of targeting just agent 1 or agent 2. And as before, targeting no one leads to manipulation and gives a payoff of 0. Thus, the optimal strategy with $\pi_{\text{cutoff}}=0.2$ is to target just one agent and manipulate only two. □

D.3 Section 5

Proof of Theorem 4. This follows immediately from Proposition 11, as noted in Appendix A, by taking $I_1=\{1,\dots,n\}$. □

Proof of Proposition 5. Suppose we sprinkle m stubborn agents such that $\lceil n/m\rceil$ is the farthest distance between any two "adjacent" stubborn agents along the ring. Then for all DeGroots i, letting

$j^*(i)$ be the nearest stubborn agent (looking backward):

$$\begin{aligned}\mathcal{D}_i(\mathbf{1}_D) &= 1 - \prod_{\ell=j^*(i)+1}^{i} \left(1 - \frac{1}{n+1}\right) \\ &\leq 1 - \left(\frac{n}{n+1}\right)^{\lceil n/m \rceil} \\ &\leq 1 - e^{-2/m}\end{aligned}$$

For any b and $\boldsymbol{\gamma}$, we have that $\mathcal{D}_i(\boldsymbol{\gamma}) \leq \mathcal{D}_i(\mathbf{1}_D) \leq (1-b)/2$ if $m \geq \frac{2}{\log\left(\frac{2}{1+b}\right)}$, which as we see does not depend on n. Thus setting the constant $m^* = \frac{2}{\log\left(\frac{2}{1+b}\right)}$ obtains the claim. □

Proof of Proposition 6. Because θ is constant in n, when the principal plays $\boldsymbol{\gamma} = \mathbf{1}_D$, the DeGroot centrality of all agents depends only on their distance from the nearest stubborn agent $j^*(i)$, and not the population size n:

$$\begin{aligned}\mathcal{D}_i(\mathbf{1}_D) &= 1 - \prod_{\ell=j^*(i)}^{i} \frac{1}{2} \\ &= 1 - \frac{1}{2^{d(i,j^*(i))}}\end{aligned}$$

where $d(i, j^*(i))$ is the distance between agent i and (stubborn) agent $j^*(i)$. Thus, every DeGroot agent is manipulated if and only if she is (at least) a distance d^* away from her previous stubborn agent. Because a DeGroot agent is manipulated only if $\mathcal{D}_i(\mathbf{1}_D) > (1-b)/2$, we see that $d^* = 1 + \lceil \log_2\left(\frac{1}{1+b}\right) \rceil$.

Clearly, by setting $c = 1$, the network is impervious with $c \cdot n$ stubborn agents because all the agents are stubborn. On the other hand, when $c = 0$ and $\varepsilon < 1$, the principal makes positive utility by targeting every agent in the population and manipulating (almost) everyone, so by Corollary 2, the network is susceptible. Now consider the infimum of all c such that the network with n agents remains impervious with some configuration of $\lfloor c \cdot n \rfloor$ stubborn agents. Call this value c^*, and by the previous two observations, we know that it exists and $c^* \in (0, 1)$.

We first show that $\lfloor c^* \cdot n \rfloor$ stubborn agents makes the network impervious. To do this, we establish that $c^* \cdot n$ is integral. If $c^* \cdot n$ is not integral, then the network with n agents is still impervious with $\lfloor c^* \cdot n \rfloor < c^* \cdot n$ agents, so is impervious with $c^{**} \cdot n$ agents where $c^{**} < c^*$, contradicting the definition of c^*. Thus, $c^* \cdot n$ is integral. Then, it is easy to see $c = c^* + \epsilon$ for small ϵ attains the same manipulation as with c^*, so the network is impervious with $c^* \cdot n$ stubborn agents.

By the definition of c^*, any fewer stubborn agents that $c^* \cdot n$ must make the network susceptible. With a non-sprinkled configuration, consider $\bar{d}_{\text{sprinkled}}$ and $\bar{d}_{\text{non}}$, the maximum distance from a stubborn agent for any DeGroot agent in the sprinkled and non-sprinkled configurations, respectively. By definition of "sprinkled," $\bar{d}_{\text{sprinkled}} < \bar{d}_{\text{non}}$. We claim that some DeGroot agent i on the chain between two stubborn agents which attains a distance of $\bar{d}_{\text{non}}$ must be manipulated. If not, the network is impervious when all agents are at a distance (less than or equal to) $\bar{d}_{\text{non}}$ from the last stubborn agent, as the principal employs identical strategies on identical length chains between two stubborn agents, by symmetry. Note that $\bar{d}_{\text{sprinkled}} = \lceil 1/c \rceil - 1$ and $\bar{d}_{\text{non}} \geq \bar{d}_{\text{sprinkled}} + 1$. Thus, the network is impervious with $c^{**} \cdot n$ stubborn agents, where c^{**} is the largest number such that $c^{**} \cdot n$ is integral and $\lceil 1/c^{**} \rceil = \lceil 1/c^* \rceil + 1$ (which is guaranteed to exist for large n). Clearly $c^{**} < c^*$, a contradiction of the definition of c^*. Thus, the non-sprinkled configuration is susceptible.

To see that $c^* \in \Theta(1)$, note that if $n_1 = kn_2$ for $k \in \mathbb{N}$ and $c \cdot n_1$ is integral, then the network with

$c \cdot n_1$ stubborn agents is impervious to manipulation if and only if the network with $c \cdot n_2$ stubborn agents is impervious. This is immediate from the fact that any path of length z between two stubborn agents with n_1 agents in the population along the ring can be transformed into k paths of length z between two stubborn with n_2 agents along the ring. Again, because of symmetry, the principal must employ identical strategies on all k copies of the z-length path. □

Proof of Theorem 5. We can order the agents by their location on the ring, starting from some arbitrary agent 1. Fix the principal's influence vector $\boldsymbol{\gamma}$. We can write the DeGroot centrality of (DeGroot) agent i in network $\mathbf{G}_\eta$ as:

$$\mathcal{D}_i(\boldsymbol{\gamma}) = \left(\frac{\eta}{n+1} + \frac{1-\eta}{2}\right)\gamma_i + \frac{1-\eta}{2}\mathcal{D}_{i-1}(\boldsymbol{\gamma}) + \sum_{j=1}^{n}\frac{\eta}{n+1}\mathcal{D}_j(\boldsymbol{\gamma})$$

Summing over both sides, we obtain:

$$\begin{aligned}\sum_{i=1}^{n}\mathcal{D}_i(\boldsymbol{\gamma}) &= \left(\frac{\eta}{n+1} + \frac{1-\eta}{2}\right)||\boldsymbol{\gamma}||_1 + \frac{1-\eta}{2}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) + \frac{\eta(n-m)}{n+1}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) \\ &= \frac{(1-\eta)n + (1+\eta)}{2(n+1)}||\boldsymbol{\gamma}||_1 + \frac{\eta(n-1-2m) + (n+1)}{2(n+1)}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma})\end{aligned}$$

This gives us

$$\frac{(n+1) - \eta(n-2m-1)}{2(n+1)}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) = \frac{(1-\eta)n + (1+\eta)}{2(n+1)}||\boldsymbol{\gamma}||_1 \Longrightarrow \sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) = \frac{(1-\eta)n + (1+\eta)}{(n+1) - \eta(n-2m-1)}||\boldsymbol{\gamma}||_1$$

Let us call $\zeta(\boldsymbol{\gamma}) \equiv \frac{(1-\eta)n+(1+\eta)}{(n+1)-\eta(n-2m-1)}||\boldsymbol{\gamma}||_1$. If stubborn agents form a continuous chain or are there are only $o(n)$ many, then there exists a continuous chain in the ring of DeGroots that grows unboundedly in n (without any stubborn agents agents along the chain). Let agent i^* be the first DeGroot on such a chain. If the principal targets all agents along this chain, then:

$$\begin{aligned}\mathcal{D}_{i^*}(\boldsymbol{\gamma}) &= \frac{\eta}{n+1} + \frac{1-\eta}{2} + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma}) \\ \mathcal{D}_i(\boldsymbol{\gamma}) &= \frac{\eta}{n+1} + \frac{1-\eta}{2} + \frac{1-\eta}{2}\mathcal{D}_{i-1}(\boldsymbol{\gamma}) + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma})\end{aligned}$$

Solving the recursion, for an agent at location τ away from i^*, we see that:

$$\begin{aligned}\mathcal{D}_\tau(\boldsymbol{\gamma}) &= \left(\frac{(1-\eta)n + (1+\eta)}{2(n+1)} + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma})\right)\sum_{\tau'=0}^{\tau-1}\left(\frac{1-\eta}{2}\right)^{\tau'} \\ &= \left(\frac{(1-\eta)n + (1+\eta)}{2(n+1)} + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma})\right)\frac{1-((1-\eta)/2)^\tau}{1-(1-\eta)/2} \\ &\overset{\tau\to\infty}{\to} \frac{(1-\eta)n + (1+\eta)}{(n+1)(1+\eta)} + \frac{2\eta}{(n+1)(1+\eta)}\zeta(\boldsymbol{\gamma})\end{aligned}$$

It is easy to verify that $\mathcal{D}_\tau(\boldsymbol{\gamma})$ is decreasing in η. When $\eta = 0$, the principal can obtain a payoff that grows unboundedly in n by manipulating $\omega(1)$ agents along this chain of DeGroots, and no strategy that manipulates only $O(1)$ agents does better; therefore, $\omega(1)$ agents are manipulated. This reason-

ing continues to hold as long as $\mathcal{D}_\tau(\boldsymbol{\gamma}) \geq (1-b)/2$, and since $\boldsymbol{\gamma} = \mathbf{1}_D$ is profitable (given $\varepsilon < 1$), the condition $\mathcal{D}_\tau(\mathbf{1}_D) \geq (1-b)/2$ is both necessary and sufficient for imperviousness. Finally, by monotonicity and continuity of $\mathcal{D}_\tau(\boldsymbol{\gamma})$, we are guaranteed there exists η^* such that $\mathcal{D}_\tau(\boldsymbol{\gamma}) > (1-b)/2$ when $\eta < \eta^*$ and $\mathcal{D}_\tau(\boldsymbol{\gamma}) < (1-b)/2$ when $\eta > \eta^*$. □

Proof of Theorem 6. Because the stubborn agents are placed symmetrically and the network is symmetric itself, we know that $\mathcal{D}_i(\boldsymbol{\gamma}) = \mathcal{D}_j(\boldsymbol{\gamma})$ for all DeGroots $i, j \in D$. By definition, there exists ϕkm DeGroot-stubborn connections in the network. Once again, by symmetry, all (DeGroot) agents are adjacent to the same number of stubborn agents, m_*. We can compute m_* by computing the average connections to stubborn agents:

$$m_* = \frac{\phi km}{k(n-m)} = \phi \frac{m}{n-m}$$

By the recursive definition of DeGroot centrality, we see that:

$$\begin{aligned} \mathcal{D}(\mathbf{1}_D) &= \frac{1}{1+k} + \frac{k}{1+k} \cdot \left(1 - \phi \frac{m}{n-m}\right) \mathcal{D}(\mathbf{1}_D) \\ &= \frac{1}{1+k} + \frac{k}{1+k} \frac{n-(1+\phi)m}{n-m} \mathcal{D}(\mathbf{1}_D) \\ \implies \mathcal{D}(\mathbf{1}_D) &= \frac{n-m}{(\phi k - 1)m + n} \end{aligned}$$

Simply rearranging with the observation that an agent is manipulated with $\mathcal{D}(\mathbf{1}_D) \leq (1-b)/2$, we see that if the principal plays $\mathbf{1}_D$, there is no manipulation if and only if $\phi kn/(n-m) \geq (1+b)/(1-b)$. Since $\mathcal{D}(\boldsymbol{\gamma}) \leq \mathcal{D}(\mathbf{1}_D)$ for all $\boldsymbol{\gamma}$, we see the network is impervious when this inequality holds. □

D.4 Section 6

Proof of Proposition 7. Consider the learning dynamics given by $\boldsymbol{\mu}_{t+1} = \tilde{\theta}_i \cdot \boldsymbol{\mu}_0 + \tilde{\mathbf{W}} \boldsymbol{\mu}_t$. By the same reasoning as in Theorem 1, we see that:

$$\boldsymbol{\mu}_t \overset{a.s.}{\to} (\mathbf{I} - \tilde{\mathbf{W}})^{-1} (\boldsymbol{\mu}_0 \odot \tilde{\boldsymbol{\theta}}) \equiv \boldsymbol{\mu}_\infty$$

Thus, as $T \to \infty$, it is sufficient to consider the learning dynamics given by:

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\mu}_\infty \cdot \boldsymbol{\theta} \odot \mathrm{BU}_i(h_{i,t+1}) + (\mathbf{1} - \boldsymbol{\theta} \odot \boldsymbol{\mu}_\infty) \oslash (\mathbf{1} - \boldsymbol{\theta}) \odot \mathbf{W} \boldsymbol{\pi}_t$$

where $\oslash$ is element-wise division. This is equivalent to the original learning dynamics, under a network preservation (see Definition 3) with $\boldsymbol{\theta}' = \boldsymbol{\mu}_\infty \cdot \boldsymbol{\theta}$. Plugging in the expression for $\boldsymbol{\mu}_\infty$, combined with the asymptotic beliefs given in Theorem 1, obtains the result. □

Proof of Proposition 8. Consider the case of concave cost with $X^* \geq \bar{X}$. Let M^*, M^{**} be the number of manipulated agents in the linear and concave cost cases, respectively, and X^*, X^{**} the number of targeted agents in the linear and concave cost cases, respectively. If $M^{**} < M^*$ (so $X^{**} < X^*$), then consider the payoff in the linear cost case from implementing the concave cost strategy:

$$\begin{aligned} M^{**} - \varepsilon X^{**} &= M^* - \varepsilon X^* + (M^{**} - M^*) - \varepsilon (X^{**} - X^*) \\ &\geq M^* - \varepsilon X^* + (M^{**} - M^*) - (C(X^{**}) - C(X^*)) \\ &\geq M^* - \varepsilon X^* \end{aligned}$$

where the first inequality follows from the fact that $C(\cdot)$ is concave and $C(X^*) \geq C(\bar{X})$, and the second follows from the assumptions on M^{**}, M^*. But this contradicts the optimality of (M^*, X^*) in the linear cost case.

Now consider convex costs with $\bar{X}^* \geq \bar{X}$. Let M^*, M^{**} be the number of manipulated agents in the linear and convex cost cases, respectively, and X^*, X^{**} the number of targeted agents in the linear and convex cost cases, respectively. If $M^{**} > M^*$ (so $X^{**} > X^*$), then consider the payoff in the convex cost case from implementing the linear cost strategy:

$$\begin{aligned} M^* - \varepsilon X^* &= M^{**} - \varepsilon X^{**} + (M^* - M^{**}) - (C(X^*) - C(X^{**})) \\ &\geq M^{**} - \varepsilon X^{**} + (M^* - M^{**}) - \varepsilon(X^* - X^{**}) \\ &\geq M^{**} - \varepsilon X^{**} \end{aligned}$$

This contradicts the optimality of (M^{**}, X^{**}) in the convex cost case. Finally, note that in the convex cost case it is always true that when $X^* < \bar{X}$ that $C(X^*) < \varepsilon X^*$. Therefore, the strategy in the linear cost case necessarily obtains positive payoff for the principal, which means it improves on $\mathbf{x} = \mathbf{0}$, and so by Corollary 2 the network is susceptible. □

Proof of Proposition 9. Let $\mathcal{D}_i^k(\boldsymbol{\gamma})$ and $\mathcal{D}_i(\boldsymbol{\gamma})$ denote the DeGroot centrality in k-cut subnetwork and the original network, respectively, under $\boldsymbol{\gamma}$. Suppose we have a k-cut subnetwork that is impervious to manipulation, so for any network strategy $\mathbf{x}$, we have that $\sum_{i \in D} \mathbf{1}_{\mathcal{D}^k(\boldsymbol{\gamma}(\mathbf{x})) > (1-b)/2} - \varepsilon x_i \leq 0$. Because $\varepsilon_u = 0$, it is sufficient to consider strategies $\mathbf{x}$ with $\gamma_u = 1$, as they dominate the strategies with $\gamma_u = 0$.

Consider the principal applying strategy $\mathbf{x}$ in the original network. First, we show the DeGroot centrality of every agent in the k-cut subnetwork ($\mathcal{D}_i^k$) is at least that in the original network ($\mathcal{D}_i$). In the k-cut subnetwork, since $\gamma_u = 1$, we have that $\mathcal{D}_u^k(\boldsymbol{\gamma}) = 1$ which is clearly an upper bound on all $\mathcal{D}_v(\boldsymbol{\gamma})$ for $v \in \mathcal{K}$ in the original network. Consider the recursive definition of DeGroot centrality, in both the k-cut subnetwork and the original network:

$$\begin{aligned} \mathcal{D}_i(\boldsymbol{\gamma}) &= \theta_i \gamma_i + \sum_{j=1}^{n} \mathcal{D}_j(\boldsymbol{\gamma}) \\ \mathcal{D}_i^k(\boldsymbol{\gamma}) &= \theta_i \gamma_i + \sum_{j=1}^{n-k} \mathcal{D}_j^k(\boldsymbol{\gamma}) + \mathcal{D}_u^k(\boldsymbol{\gamma}) \end{aligned}$$

which admits a unique fixed point. Note the above is an increasing map in $\{\mathcal{D}_j(\boldsymbol{\gamma})\}_{j=1}^n$, and since $\alpha_{iu} = \sum_{j \in \mathcal{K}} \alpha_{ij}$ with $\mathcal{D}_u^k(\boldsymbol{\gamma}) \geq \mathcal{D}_u(\boldsymbol{\gamma})$, the fixed point $\{\mathcal{D}_j(\boldsymbol{\gamma})\}_{j=1}^n$ in relation to the fixed point $\{\mathcal{D}_j^k(\boldsymbol{\gamma})\}_{j=1}^{n-k} \cup \mathcal{D}_u^k(\boldsymbol{\gamma})$ must satisfy $\mathcal{D}_j(\boldsymbol{\gamma}) \leq \mathcal{D}_j^k(\boldsymbol{\gamma})$ for all $j \in \{1, \ldots, n-k\}$. Therefore, for all $\mathbf{x}$:

$$\sum_{i \in D} \mathbf{1}_{\mathcal{D}(\boldsymbol{\gamma}(\mathbf{x})) > (1-b)/2} - \varepsilon x_i \leq \sum_{i \in D} \mathbf{1}_{\mathcal{D}^k(\boldsymbol{\gamma}(\mathbf{x})) > (1-b)/2} - \varepsilon x_i \leq 0$$

which means $\mathbf{x} = \mathbf{0}$ is optimal, and there is no manipulation in the original network. □

Proof of Corollary 1. The local density result of Proposition 11 guarantees that the k-cut subnetwork is impervious to manipulation (with the exception of vertex u) when the log-diameter condition is met. Then applying Proposition 9 shows that the original network is k-impervious. □

Proof of Proposition 10. By Lemma 1, $\text{BU}_i(S|h_{i,t})$ converges to 1 if $z_{i,t}^S - z_{i,t}^R \to \infty$ (or $z_{i,t}^R / z_{i,t}^S \to 0$) and it converges to 0 if $z_{i,t}^R - z_{i,t}^S \to \infty$ (or $z_{i,t}^S / z_{i,t}^R \to 0$). As the DeGroot agents update their beliefs mechanically, any strategy where the principal sends mixed messages (i.e., $\hat{y}_i \neq R$) is dominated by

one where he sends $\hat{y}_i = R$. Note that if the principal sends messages at average intensity λ_i^*, then her signal distribution is given by:

$$\mathbb{P}[s_{i,t} = R|\theta = S] = \frac{\lambda_i^*}{\lambda + \lambda_i^*} + \frac{\lambda}{\lambda + \lambda_i^*}(1 - p_i)$$

which is greater than $1/2$ (so $z_{i,t}^S/z_{i,t}^R \to 0$) when $\lambda_i^* > \lambda(2p_i - 1)$ and less than $1/2$ (so $z_{i,t}^R/z_{i,t}^S \to 0$) when $\lambda_i^* < \lambda(2p_i - 1)$. Note that since $\tilde{\varepsilon}$ is continuous, the difference in average cost between $\lambda_i^* = \lambda(2p_i - 1) - \delta$ and $\lambda_i^* = \lambda(2p_i - 1) + \delta$ shrinks to 0 when $\delta \to 0$, so the principal maximizes her payoff by other choosing an average intensity of $\lambda_i^* = \lambda(2p_i - 1) + \delta$ for vanishing $\delta \to 0$, or $\lambda_i^* = 0$. Moreover, since $\tilde{\varepsilon}$ is convex, the optimal targeting policy that minimizes cost but obtains an average targeting intensity λ_i^* is the constant function $\lambda_i^*(t) = \lambda_i^*$. This obtains exactly an average cost of $\tilde{\varepsilon}(\lambda_i^*)$. □

D.5 Proofs of Appendix Theorems

We now prove any results that were relegated to the appendix.

Proof of Proposition 11. Fix an agent i and some I_ℓ such that $i \in I_\ell$. It is sufficient to prove that the sum of weighted walks passing through stubborn agents in I_ℓ is bounded below by a constant $\rho(\delta, m)$ which only depends on δ and the number of stubborn agents m in component I_ℓ, with $\lim_{m\to\infty} \rho(\delta, m) = 1$ for all δ. To see this, note that $\mathcal{D}_i(\boldsymbol{\gamma}) \leq (1-\rho(\delta, m))$ for all principal interventions $\boldsymbol{\gamma}$, 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 given that $(1 + b)/2 \leq 1$ and $\lim_{m\to\infty} \rho(\delta, m) = 1$.

Let w_i^B be the sum of weighted walks remaining in I_ℓ that *end* with a stubborn agent in I_ℓ (and avoid other stubborn agents). Since the log-diameter of the I_ℓ component is less than $\log(|I_\ell| + \delta)$, we know that between any two agents $i, j \in I_\ell$, there exists a walk $W_{ij}^* = i \to u_1 \to u_2 \to \cdots \to u_k \to j$ in I_ℓ such that:

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

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

Let us define an *intermediate walk* to be a walk of weight at least $1/(|I_\ell| + \delta)$ between two DeGroot agents $i, j \in I_\ell$. Additionally, let us say a k-weighted walk from DeGroot i ending at some stubborn agent $j \in I_\ell$ 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 $(|I_\ell| - m - 1)^{k-1}$ k-weighted walks from i to any given stubborn agent j. To see this, note that because i has a weighted walk of weight $1/(|I_\ell| + \delta)$ to every other vertex v in I_ℓ (by the above inequality), the number of k-weighted walks is the number of intermediate vertices between i and j which do not include stubborn agents or i itself. Moreover, we note that by the

previous inequality:

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

Putting the pieces together, we have that:

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

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

Proof of Theorem 7. The play of the DeGroots is pinned-down by their beliefs, which when T is large is high probability close to its limit. By Proposition 2, the DeGroot centrality $\mathcal{D}(\boldsymbol{\gamma})$ is equivalent to the belief $\boldsymbol{\pi}_\infty(R)$. We let z_i denote an agent i who is manipulated at the limit. All stubborn agents are knowledgeable and have $\pi_{i,\infty}(R) = 0$. Similarly, we suppose the principal can "elect" to manipulate agent i only if its DeGroot centrality is above $(1 - b)/2$; in other words:

$$\begin{aligned}
&r_i \leq 1 + \mathcal{D}_i(\boldsymbol{\gamma}) - (1 - b)/2 = \mathcal{D}_i(\boldsymbol{\gamma}) + (1 + b)/2 \\
&r_i \in \{0, 1\}
\end{aligned}$$

Finally, note that the principal gains an additional payoff of 1 for each manipulated agent and pays a cost of ε for each $\gamma_i = 1$ (the principal will never set $x_i = 1$ for a stubborn agent i anyway). Combining these we get the integer program in Theorem 7. Every $\mathbf{x}$ except those that try to target stubborn agents can be represented as $(\mathbf{r}, \boldsymbol{\gamma})$, but such $\mathbf{x}$ are dominated by another network action because of stubborn agents do not change their beliefs. Similarly, each feasible $(\mathbf{r}, \boldsymbol{\gamma})$ 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 $(\mathbf{I} - \mathbf{W})^{-1}$ consists of all nonnegative entries, we know that $\mathcal{D}(\mathbf{1}_\mathcal{K}) < \mathcal{D}(\mathbf{1}_D)$. Under (a), every feasible solution requires that $\mathbf{r} = \mathbf{0}$. 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 $\mathbf{r} = \mathbf{1}_{\mathcal{D}_i(\mathbf{1}_\mathcal{K}) > (1-b)/2}$ is a feasible solution to the IP, and the objective yields $||\mathbf{r}||_1 - \varepsilon||\boldsymbol{\gamma}||_1 > 0$ by the assumption in (b). Thus, the feasible solution $(\boldsymbol{\gamma}, \mathbf{r}) = \mathbf{0}$ does not maximize the IP as it gives an objective of 0, so $\mathbf{0} \notin \boldsymbol{\Gamma}^*$, and thus the network is susceptible. □

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.

Allon, Gad and Dennis Zhang (2017), "Managing service systems in the presence of social networks." *Available at SSRN 2673137*.

Bohren, J Aislinn and Daniel N Hauser (2017), "Bounded rationality and learning: A framework and a robustness result."

Candogan, Ozan and Kimon Drakopoulos (2020), "Optimal signaling of content accuracy: Engagement vs. misinformation." *Operations Research*, 68, 497–515.

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.

Liu, Qingmin (2011), "Information acquisition and reputation dynamics." *The Review of Economic Studies*, 78, 1400–1425.

Mostagir, Mohamed (2010), "Exploiting myopic learning." In *International Workshop on Internet and Network Economics*, 306–318, Springer.

Papanastasiou, Yiangos (2020), "Fake news propagation and detection: A sequential model." *Management Science*.

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*.

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.