Inequality in Social Learning[*]

Abstract

We study the spread of misinformation in a social network characterized by unequal access to learning resources. Agents use social learning combined with their own signals to uncover an unknown state of the world, and a principal tries to distort this learning process in order to influence their beliefs. A subset of agents throughout the network is endowed with knowledge of the true state. This gives rise to a natural definition of inequality: privileged communities with a large proportion of knowledgeable agents experience a positive externality from their presence and therefore are more resistant to the interference of the principal, whereas marginalized communities who do not have access to these individuals are comparatively disadvantaged. This access is determined by the homophily structure of the network – a highly integrated society has unimpeded access to these knowledgeable agents regardless of which community they reside in, whereas agents in a more segregated society face more restricted access to these resources. We show that the role that inequality plays in the spread of misinformation is highly complex. For instance, communities who hoard resources and deny them to the larger population can end up exposing themselves to *more* misinformation. On the other hand, while more inequality generally leads to worse outcomes, the prevalence of misinformation in society is non-monotone in the level of inequality. This implies that policies that reduce inequality without completely eradicating it can sometimes leave society more prone to misinformation.

1 Introduction

The central question of the social learning literature is whether a society or a group of people will manage to learn an unknown state of the world.[1] We study this question in a social network characterized by unequal access to learning resources: some agents are already endowed with knowledge of the true state, and the rest of the agents update their beliefs about this state based on their own information (that they obtain from news sources) as well as from communicating with their friends. A strategic actor – a *principal* – tampers with the learning process in order to spread misinformation and get agents to mislearn the state of the world. This induces a form of inequality in the network: If some agents are endowed with knowledge of the true state or with a higher ability to learn, then being connected to these agents provides an advantage. Similarly,

[*] Previously titled "Homophily is (not) Bad for Learning"

[1] For example, whether temperatures on earth are increasing over time, as stated in Golub and Jackson (2012).

agents who do not have access to these individuals are comparatively disadvantaged and their learning may be hampered as a result.

This access or lack thereof is determined by the homophily structure of the network. Homophily –the tendency of people to associate with those who are similar to themselves– provides a realistic representation of real-world networks (see Marsden (1987); McPherson et al. (2001)), and in our framework serves as the underlying mechanism through which inequality propagates in society. Privileged communities with easy access to knowledgeable agents are more likely to be protected against misinformation, but if homophily is too high and communities become more insular, then vulnerable communities with few or no knowledgeable agents will have little access to the more affluent communities and their resources. How does this inequality affect the spread of misinformation in the individual communities and society as a whole?

We show that the answer to this question is complex and depends on a myriad of factors. These factors interact to give rise to a range of outcomes, some of which are intuitive – like an increase in inequality leading to the underprivileged communities being worse off– while others are less so. For example, sometimes an increase in inequality can make the *privileged* communities worse off, or sometimes intermediate levels of inequality can make the entire society worse off while under different conditions the same levels of inequality can protect the whole society from misinformation. Understanding the drivers behind this range of outcomes is important because of its potential consequences to social and economic policy: in addition to its effects on learning inequality, homophily is also connected to segregation in society (Currarini et al. (2009)), which leads to well-documented economic inequality.[2] This makes it natural to suggest policies that alleviate the negative effects of homophily and inequality. Our paper shows that these policies should be carefully planned –by taking into account the factors that we identify– in order for them to accomplish their desired goals and not result in unintended consequences.

Throughout the paper we assume that some agents are "knowledgable" – they are endowed with knowledge of the true state, while other agents try to uncover that state through social learning. We do not assume the network structure is deterministic; instead, we assume that the network is generated randomly from a distribution and exploit the connection between random graphs and homophily to examine the effects of the latter on the spread of misinformation. This allows us to develop insights that are useful for resource allocation and policy interventions. By understanding how the inequality structure of society interacts with the incentives of a strategic principal, a social planner can respond to the spread of misinformation by shaping

[2] Wage inequality and differences in labor market participation across different groups are documented in Card and Krueger (1992); Chandra (2000), and Heckman et al. (2000). Calvo-Armengol and Jackson (2004) and Calvó-Armengol and Jackson (2007) show that these inequalities can be explained through network models of homophily.

that inequality structure through endowing some agents with knowledge of the correct state (for example through education or increasing awareness) or by adjusting the homophily levels in society, for example through facilitating communication between communities.

Finally, we remark that the prevailing assumption throughout the social learning literature is that the news that agents obtain is organic: it may or may not be accurate, but it is provided by sources that have no stake in what the agents' beliefs are. In reality, such sources are rarely neutral, and may have an interest in shaping these beliefs in order to direct agents towards taking certain actions. The interference of Cambridge Analytica in the 2016 US presidential election is one example of such belief manipulation. Similarly, recent outbreaks of measles in Eastern Europe and parts of the US have been linked to Russian interference and propaganda whose goal is to convince people that vaccines are harmful in order to make them opt against vaccinating themselves and their children.[3] Our paper provides a robustness check for some of the results in the standard learning models. For example, while homophily is unambiguously identified as a barrier to learning in previous literature (see Golub and Jackson (2012)), its role in this strategic setup is much more nuanced. Successful learning under this strategic interference model therefore implies successful learning in the models in the literature, but the opposite implication is not necessarily correct.

Contribution and Overview of Results. In this paper, we study the effects of inequality on the spread of misinformation, with homophily being the underlying mechanism through which inequality propagates. We show in Theorem 1 and Appendix A that the study of the random networks arising from our homophily models can be reduced to studying the expected network. Given this simplification, we make the following contributions:

Conceptual: We start by giving a definition of learning inequality as a measure of how communities differ in their access to these knowledgeable agents. This is determined by i) the distribution of these agents over different communities, and ii) the level of homophily in society. We then show that the effect of inequality on the spread of misinformation is not monotone, and is shaped by several factors. In particular, the principal's technology for sending signals to agents can be costly, and that cost is a large determinant for how inequality affects the spread of misinformation. Theorem 3 shows that when that cost is negligible, so that the principal can target whomever he wants, then intermediate levels of inequality are always (weakly) worst for society, even more so than extreme inequality. Theorem 5 shows that this conclusion can be reversed when the principal's cost of sending misinformation is no longer negligible, i.e. intermediate

[3] https://www.newsweek.com/russian-trolls-promoted-anti-vaccination-propaganda-measles-outbreak-1332016

levels of inequality can sometimes be best for society. Such a reversal might occur because the incentives for the principal to send misinformation are highly complex when her signaling technology is relatively expensive.

We also show that another factor that determines how inequality shapes the spread of misinformation is the relative population sizes of the different communities in society. Theorem 2 shows that when communities have the same size, an increase in inequality has an expected effect: privileged communities are better off and marginalized communities (with little access to knowledgeable agents) are worse off and more prone to manipulation. However, Theorem 4 shows that if the privileged communities are smaller in size compared to the rest of the underprivileged population –as is typically the case– then an increase in inequality not only hurts the large population, but also hurts the privileged communities themselves.

Lastly, we introduce a novel inequality model through a network structure we call *strong homophily*, which complements the weak homophily model commonly studied in the literature (and throughout most of our paper). Strong homophily provides a more appropriate framework for analyzing societies characterized by a hierarchical structure. Propositions 1 and 2 contrast the effects of weak and strong homophily on the spread of misinformation, and show how the latter always requires substantially more equality to protect society.

Policy Implications: The findings from our model can be used as input to policy makers trying to curb manipulation. By understanding the inequality structure of society, resources can best be directed towards interventions that will protect this society from misinformation. We consider two natural interventions: first, educating specific agents in the network who will then serve as the knowledgeable agents to protect their (and potentially other) communities. Second, increasing connectivity across communities in order to make the network less segregated and provide more access to these knowledgeable agents.

Our policy recommendations are parametrized by the budget that the planner has and the cost of the signaling technology of the principal. Corollary 1 and propositions 3 and 4 show that when the planner's budget is large enough, the principal's signaling costs are low, and communities are of similar size, then it is best to make society as equitable as possible. Corollary 2 follows Theorem 4 to show that if privileged communities are small in size compared to the rest of the population, then any policy that subjects the population at large to inequality is sub-optimal, and that it is in the best interest of privileged communities to champion policies that allocate resources to this large community, otherwise everyone in the population is worse off.

The above policies advocate for equitable distribution of resources, but due to the complexity of the phenomena being studied, exceptions can sometimes occur. As we mentioned above, when the principal's signaling costs are not cheap, some inequality in the network may be better for society that no inequality. Similarly, when signaling costs are cheap, intermediate levels of inequality are worst for society, and hence a budget-constrained planner who cannot drastically reduce extreme inequality and instead opts to just reduce it can unwittingly push society into the susceptible regime. Our results in this case advocate for a minimum budget allocation that allows the planner to reduce inequality enough to bypass the problematic intermediate inequality region. More generally, our model provides a basic framework to think about these decisions in terms of the primitives of the problem represented by the inequality structure, the planner's budget, and the principal's signaling costs.

Related Literature The model we develop builds upon the work of Chandrasekhar et al. (2019) and Mostagir et al. (2019). The first paper shows, experimentally, that agents update their beliefs in ways that are consistent with Bayesian or DeGroot updating,[4] and suggests that the proportion of agent types in society may be driven by contextual factors (e.g. level of education). The second paper takes this finding and builds a theoretical model where a principal tries to manipulate a society with mixed learning types, and shows that Bayesian agents always learn the true state eventually and then can help spread this knowledge to other agents.[5] For this reason, we refer to agents who are knowledgeable about the state as **Bayesian**, which can correspond to either an agent's sophistication and reasoning abilities *or* her deep knowledge of the issue at hand.

The seminal paper of Golub and Jackson (2012) shows the negative effects that homophily has on a society that learns in a DeGroot fashion. Instead of learning rates, our focus is on understanding the role of homophily in whether a network is impervious to manipulation when the learning horizon is long, and when information is potentially provided by a strategic source. Lobel and Sadler (2015) show how the role of homophily in a sequential learning model depends on the density of the network. Homophily in that paper is used to describe alignment of preferences over the agents' decision problem,[6] whereas homophily in our model captures similarities along dimensions (race, age, profession, income, etc.) that can be orthogonal to whatever state the agents are trying to learn.

Bayesian agents in our model are stubborn agents who know the truth. Opinion dynamics with stubborn agents have been studied in Acemoglu et al. (2013) and Yildiz et al. (2013) among

[4] Bayesian agents integrate the opinions of their friends using Bayes' law, while less-sophisticated DeGroot agents take the opinions of their friends at face value and incorporate them into their own beliefs through a weighted average.

[5] This is Theorem 8 in Mostagir et al. (2019).

[6] For example, an agent who is deciding on a restaurant weighs the opinion of her friend differently if she and her friend prefer the same type of food.

others. The recent work of Sadler (2019) extends Yildiz et al. (2013) to random graphs. What differentiates our paper from this literature is the presence of a strategic principal, which gives rise to completely different learning dynamics and implications.

There is recent work on fake news and manipulation. In Candogan and Drakopoulos (2017) and Papanastasiou (2020), there is no strategic news provider; fake news already exists in the system and the focus is on how it can be identified and controlled. Keppo et al. (2019) pay less attention to the social network aspect and focus instead on how how Bayesian agents can be manipulated through selective information dissemination. As mentioned, the manipulation problem as presented in our paper was introduced in Mostagir et al. (2019), where agents interact over a fixed and known network topology. Our paper embeds this model in a random network structure in order to study the role of inequality in the spread of misinformation. In addition, our paper provides a prescriptive component to evaluate which policies may be effective in stopping the spread of misinformation as a function of inequality in society. As mentioned earlier, these policies also speak to issues of community integration and resource allocation.

Our paper assumes only knowledge of the random process from which the network is generated. There is recent literature that tries to recover the network structure from relational data, e.g. Alidaee et al. (2020); Ata et al. (2018) consider a seller who does not know the network structure but, in the presence of externalities, estimates it from transaction data. Other recent work, e.g. Auerbach (2019), tests whether a network was generated from an inhomogeneous random graph model (which includes the class of stochastic block matrices commonly used to model homophily). These methods can be used to estimate the structure of homophily in society and applied as input to our model.

Finally, our paper is also related to diffusion and seeding in random networks, as exemplified by the recent work of Manshadi et al. (2018), Akbarpour et al. (2018), and Sadler (2019). These papers consider the classic problem of which agents to select in order to spread information throughout the network. The primary difference with our work is that we consider an adversarial, strategic principal who is trying to spread his own influence in the network, and our goal is to identify conditions and policies under which we can stop this principal from spreading misinformation, with a specific emphasis on the role of inequality and the social structure of the network in propagating such information.

Organization As mentioned, our paper builds on the model of Mostagir et al. (2019), so we present a summary of that model and the necessary definitions and results in Section 2.1. Section 2.2 introduces the random network formation process and provides a result under which the study of random networks can be reduced to the study of the average network. Section 3 demonstrates some of the main technical results in the paper through a few examples. The formal results follow in Sections 4 and 5, while Section 6 introduces strong homophily and compares it to the weak homophily model. Finally, we discuss possible interventions to prevent the spread of misinformation in Section 7 and conclude the paper in Section 8.

2 Model

We build a framework that embeds the model of Mostagir et al. (2019) into a random network formation model. We introduce the random network model later in this section, but to make the paper self-contained, we first present a high-level summary of the primitives and results from Mostagir et al. (2019) that are releveant for our setup.[7]

2.1 Deterministic Networks: Reduced-Form Model and Solution

We consider a social network with n agents trying to learn a binary state of the world $y \in \{S, R\}$ over time. Time is discrete and agents learn over a finite horizon, $t \in \{1, \ldots, T\}$.[8] At time $t = 0$, the underlying state $y \in \{S, R\}$ is drawn, with $\mathbb{P}(y = S) = q \in (0, 1)$. Agents try to learn the state of the world in order to take an action at time T, with the goal of taking the action that matches the true state, i.e. take action S if the state is S or action R if the state is R.[9] The principal is interested in agents taking action R regardless of what the state of the world is. These payoffs are represented in Table 1, where the two numbers in each cell represent the payoffs to the principal and to the agent, respectively, for the state of nature and agent action combination corresponding to that cell. From the table, we can see that an agent would take action R if her belief that the true state is R is at least equal to $\frac{1-b}{2}$. Equivalently, she would take action S if her belief that the true state is S is at least equal to $\frac{1+b}{2}$. Without loss, we assume throughout that the true state is $y = S$.[10]

News Agents receive news over time in the form of signals, where $s_{i,t} \in \{S, R\}$ is the signal that

[7] For completeness, we include the full technical details in Appendix A.1. These details (e.g. arrival rates of news, explicit form of DeGroot updating, etc.) are not pertinent for our goals but we include them in the appendix for the interested reader. We also detail other formal concepts of the baseline model in Appendix A that are pertinent for the results presented in this paper.

[8] Implicit in this is that agents receive news at the same time. In Appendix A we provide a less parsimonious, but equivalent, model where agents digest news at different rates, among other generalizations, that do not affect the conclusions of the model.

[9] For example, as in Mostagir et al. (2019), they want to learn whether a particular vaccine is safe (state of the world $y = S$) so that they vaccinate (take action S) or whether a vaccine is risky (state of the world $y = R$) so that they would choose to avoid vaccination (take action R).

[10] If $y = R$, the results are trivial because by Proposition 1 in Mostagir et al. (2019), all agents will learn $y = R$ even without interference from the principal.

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

Table 1. Terminal Payoffs. The parameter b is in $[-1, 1]$.

agent i receives at time t. News is either organic or strategic. Organic news is informative and is generated from a process that is correlated with the state of the world, and we assume that the probability that $s_{i,t}$ correctly represents that state is strictly greater than $1/2$. However, a principal may choose specific agents in the network and jam their news process by periodically sending them message $\hat{y} \in \{S, R\}$ that corresponds to the state that he would like them to believe (the principal' targeting strategy is detailed below in the section titled **Manipulation**). Importantly, agents who are targeted by the principal do not know that they are targeted and cannot tell whether a signal they are receiving is organic or strategic. This is akin to agents scrolling through their news feed and seeing both organic and strategic stories without knowing which stories are which.[11]

Learning Agents are connected in a social network and each agent i has a neighborhood of agents denoted by $N(i)$. Society consists of two types of agents: DeGroot and Bayesian agents. Bayesian agents are aware of the correct state y at $t = 0$, and thus serve as stubborn agents with correct beliefs. Equivalently, Bayesian agents update their beliefs about the state in a fully Bayesian way (hence the term, Bayesian) from observing the news and the beliefs of agents in their neighborhood. This equivalence is a consequence of Theorem 8 in Mostagir et al. (2019).[12]

DeGroot agents follow a similar model to Jadbabaie et al. (2012), with two components that go into their learning and belief update process. The first component comes from the DeGroot agent's own personal experience, which is a Bayesian update on the true state *taking the information received (personally) at face value.* In particular, given a vector of signals $\mathbf{s}_{i,t}$ up until time t, we denote this personal experience by $\text{BU}(\mathbf{s}_{i,t})$. The second component comes from a linear aggregation of the opinions of friends. These components are detailed in Appendix A.1. In general, for periods of new information, we assume DeGroot agent i updates her beliefs $\pi_{i,t+1}$

[11] For the sake of the reduced-form model, it is sufficient to think of untargeted agents as reading inorganic (as opposed to organic) news with probability 0, but targeted agents reading strategic news with some exogenous probability between 0 and 1. Again, see Appendix A.1 and Appendix A.2 for more details on the model of Mostagir et al. (2019).

[12] We provide technical conditions in Appendix A.2 that guarantee that this result extends to the random network domain.

at time $t+1$ as follows:

$$\pi_{i,t+1} = \theta \cdot \mathrm{BU}(\mathbf{s}_{i,t}) + \frac{1-\theta}{|N(i)|}\sum_{i=1}^{n} \pi_{j,t}$$

where θ is the weight that agent i places on her own Bayesian update. Note the belief at $t+1$ is a convex combination of one's belief from reading the news and the beliefs of her neighbors at time t. We implicitly assume in this formulation that every agent weights her neighbors equally.

Manipulation At $t = 0$, before the learning process begins, the principal picks an influence strategy $x_i \in \{0,1\}$ for each agent i in the network, where $x_i = 1$ indicates that agent i is targeted by the principal (and will therefore occasionally receive strategic news signals $\hat{y}$). The principal may play any influence strategy $\mathbf{x} \equiv \{x_i\}_{i=1}^{n}$ over the network, and incurs an upfront investment cost $\varepsilon > 0$ for each agent with $x_i = 1$, thus the utility of the principal is the number of agents taking action R less total investment cost.

Agent i is manipulated if she would figure out the correct state in the absence of interference from the principal (i.e. when $\mathbf{x} = \mathbf{0}$), but would mislearn the state when the principal has a profitable network strategy $\mathbf{x} \neq \mathbf{0}$, as $T \to \infty$. Notice that the agent does not have to be directly targeted by the principal to be manipulated. Likewise, an agent may not be manipulated even if she is targeted.

A network is impervious to manipulation if no agents can be manipulated, i.e. if there is no profitable strategy for the principal that results in any agent mislearning the true state. If a network is not impervious, we say it is susceptible.

DeGroot Centrality Determining whether an agent is manipulated is equivalent to computing her limit belief of the incorrect state and checking whether the belief is higher than the cutoff obtained from the payoff table. The concept of *DeGroot Centrality* (DC) in Mostagir et al. (2019) combines the ideas of Katz-Bonacich and PageRank centrality and can be computed using the technique of weighted walks (Appendix A.4 provides a comprehensive primer on how to compute DC). This centrality measure captures how much influence the principal's strategy has on a given agent's belief, and is precisely equal to that agent's belief of the incorrect state. With a slight abuse of notation, we will often refer to π_i as the agent's belief of the correct state (i.e., $\pi_i(S)$), which is equal to 1 minus agent i's DC.

2.2 Random Networks

We embed the above setup into a class of random networks known as *stochastic block* networks, which are based on inhomogeneous Erdos-Renyi graphs. Stochastic block networks were introduced in Holland et al. (1983) and are the focus of the study of homophily in Golub and Jackson

(2012). In these networks, agents interact in well-connected communities, with few links between communities.

Weak and Strong Homophily We differentiate between two cases of interest. In *weakly assortative* networks, agents are more likely to be linked to agents within their community, but when they reach out to agents outside that community, they are equally likely to connect to agents in any other community. In that sense, homophily has a flat hierarchy. This model turns out to be rather accurate in describing friendship and communication patterns, as documented in, e.g. Marsden (1987).

We also introduce *strongly assortative* networks, which are communities ordered by similarity, with agents in neighboring communities more likely to be linked than agents in communities that are farther apart. This model captures the more hierarchical structure that is sometimes observed in society. While this is a natural homophily model, we are not aware of any literature that studies it compared to the much stronger focus on weakly-assortative networks. In this model, each community ℓ has a vector of qualities, $\Lambda_\ell \in \mathbb{R}^L$. Qualities can capture different variables like education, profession, income, etc. Communities are sorted according to their similarity, with the distance metric between communities ℓ and ℓ' given by $d(\ell, \ell') = ||\Lambda_\ell - \Lambda_{\ell'}||_2$. For simplicity, we assume that $L = 1$ (the quality vector is one-dimensional) and thus communities are (strongly) ordered by their Λ_ℓ on a *line topology*.

Formally, both types of random networks consist of k communities (or islands). In both models, for any two agents on the same island, there is a link probability p_s. In the weakly assortative model, there is also a link probability $p_d < p_s$ for any two agents on different islands. However, in the strongly assortative model, agents do not form links with agents on islands outside of their neighboring islands. Agents in community ℓ are linked to agents in community $\ell+1$ or $\ell-1$ with probability p_d, whereas agents in "farther" communities are linked with probability 0,[13] with the exception of island 1 and island k, which are linked to island 2 and island $k-1$ only, respectively.

Inequality A society consists of communities connected together through a homophily structure. The existence of (Bayesian) agents who know the correct state and propagate truthful information gives rise to a degree of inequality across communities. The distribution of these agents over communities along with a weak (or strong) homophily structure induces a form of weak (or strong) inequality. We use the term *privileged* to refer to a community with a higher proportion of Bayesians compared to a *marginalized* community that has a smaller proportion

[13] We can equivalently assume that these link probabilities are positive but decay sufficiently quickly, such as on the order of $\exp(-||\Lambda_\ell - \Lambda_{\ell'}||_2)$. For simplicity of exposition and illustration of the effects of our strong assortative property, we simply set the link probabilities to 0.

of Bayesians. Access to Bayesian agents is closely tied to the homophily structure, which determines how many (direct) connections DeGroot agents in different communities have to these Bayesians. Thus, changes in the homophily of a network also introduce changes to inequality. This working definition of inequality is useful for understanding the examples in Section 3; the formal definition is deferred to Section 4 before we derive our main theorems. We compare the differences between weak and strong inequality on misinformation in Section 6.

2.3 Reduction to Deterministic Networks

Analyzing the random networks generated according to the weak or strong homophily models is difficult. Using the proof technique in the technical note attached to this submission, the following result establishes that as long as the population is large enough, it is sufficient to analyze the deterministic analogues of these random networks:

Theorem 1. *For almost all b,*[14] *as $n \to \infty$, the probability that a random network drawn from the weak or strong inequality models has the same number of manipulated agents as the expected network converges to 1.*[15]

Theorem 1 offers a technical simplification: instead of analyzing the random networks drawn according to the weak and strong homophily models, we can treat these networks as deterministic objects where the weights are chosen proportionally to the probability of link formation. Moreover, because all of our results are invariant to a doubling of the population as long as the proportion of Bayesians and DeGroots remain constant on every island, every example can be extended to the case of $n \to \infty$ and Theorem 1 can be applied.

3 Demonstration of Main Ideas

This section presents three examples that demonstrate the complex role of (weak) inequality in learning and manipulation and serves as an overview of the technical results in the paper. The examples below show that increasing inequality can: i) have divergent effects on different communities, hurting one community and making another better off, or ii) it can hurt the whole society, or iii) it can protect the whole society. These variety of outcomes depend on a myriad of factors like relative community affluence, relative community sizes, and the cost of the manipulation technology.

[14] "Almost all" is meant in the measure theoretic sense; the only exceptions lie on a set of measure 0 in $(-1, 1)$.

[15] The formal definitions of realized and expected networks can be found in the attached technical note.

Section 3.1 provides an example where the role of inequality is consistent with the casual conclusion that increasing homophily (and therefore, increasing inequality) makes some agents in the population more susceptible to the principal's influence. In particular, privileged communities become even more resilient while the additional inequality hurts marginalized communities and leaves them more exposed to belief manipulation. In the next two examples, however, we show that more subtle effects may exist to reverse this conclusion.

In Section 3.2, we show that when communities have different sizes (and thus, unequal influence over other communities), it is possible for increased inequality to make *all* communities worse off, because this inequality directly hurts an influential core of the social network through which a substantial amount of information flows. Therefore, in contrast to the example in Section 3.1, an increase in inequality can can end up hurting even the privileged communities themselves.

Finally, Section 3.3 shows that different levels of inequality can change how the principal chooses his optimal influence strategy: as increasing inequality isolates key privileged communities, the principal may find no profitable strategy that influences even those who are most marginalized, thereby protecting the entire population.

Below, we go through the details of each of these examples.

3.1 Inequality Hurts the Most Marginalized

We consider two communities of equal size and explore the degree of manipulation under different homophily structures. Two of the agents on the first island are Bayesian, compared to only one of the agents on the other island. Thus, the former island is the *privileged* community and the latter island is the *marginalized* community. This setup is pictured in Figure 1. In this example, we vary homophily by setting $p_d = 0.2$ and increasing $p_s > p_d$ (note that as p_s increases, homophily increases). We assume that $\varepsilon = 0$ so that it is costless for the principal to send misinformation (and therefore will send to everyone).

Figure 1 shows the beliefs of the agents in both communities. Recall that manipulation occurs when an agent's belief falls below a certain threshold π^*. We see that as p_s increases, the beliefs of the privileged community move closer to the truth (higher belief) whereas the beliefs of the marginalized community move farther from the truth (lower belief). In addition, there exists a corresponding homophily threshold $\bar{p}_s$ (approximately 0.4 in this example) whereby when $p_s < \bar{p}_s$, there is no manipulation, but when $p_s > \bar{p}_s$, the marginalized community becomes manipulated. An increase in inequality in this example leads to more manipulation in society. Thus, an increase in inequality makes society worse off as the marginalized community becomes

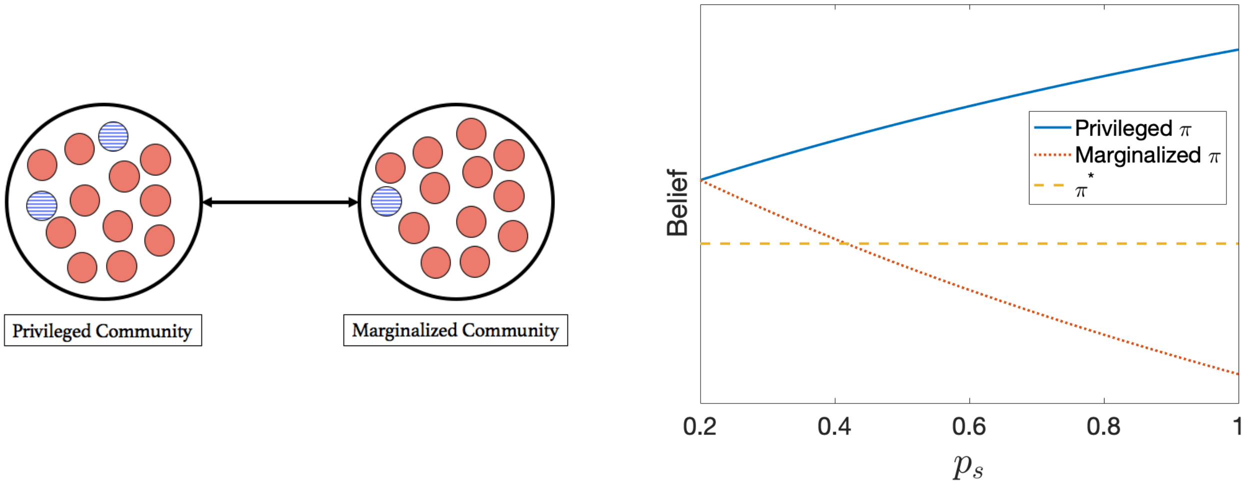


Figure 1. On the left is the setup of Section 3.1, and on the right are the beliefs of the two communities. As homophily increases (i.e. as p_s increases), the beliefs of the privileged community move towards the truth while the marginalized community's beliefs fall below the belief threshold given by the dashed line, leading to the agents in that community taking the incorrect action.

susceptible to misinformation.

3.2 Inequality Hurts Everyone

We now consider three islands in a weak inequality model: a small privileged community with an eighth of the population, a small marginalized community with an eighth of the population, and a large community with the remaining three quarters of the population. Assume that there are three Bayesian agents in the privileged community, one Bayesian agent in the large community, and no Bayesian agents in the marginalized community. This setup is depicted in Figure 2.

Similar to the previous example, we set $\varepsilon = 0$, fix $p_d = 0.2$, and vary p_s to change the amount of homophily in society. The beliefs of the three different communities are shown in Figure 2 as a function of p_s. Given the threshold line π^* in the plot, we see that as homophily increases, the beliefs of *all* agents in the population move farther away from the truth. This is true even for the *privileged* community, despite the fact that agents in this community are forming more direct connections with Bayesian agents who spread truthful information. Such a phenomenon occurs because the size disparity between communities leads to all of them deriving most of their beliefs from the information spreading in the large community, so when inequality hurts this community, it propagates to those who, on the surface, should be benefiting from it (as in the previous example). Once homophily hits $p_s = 0.3$, two communities are manipulated while the privileged community is still immune. As homophily increases further to $p_s = 0.5$, everyone

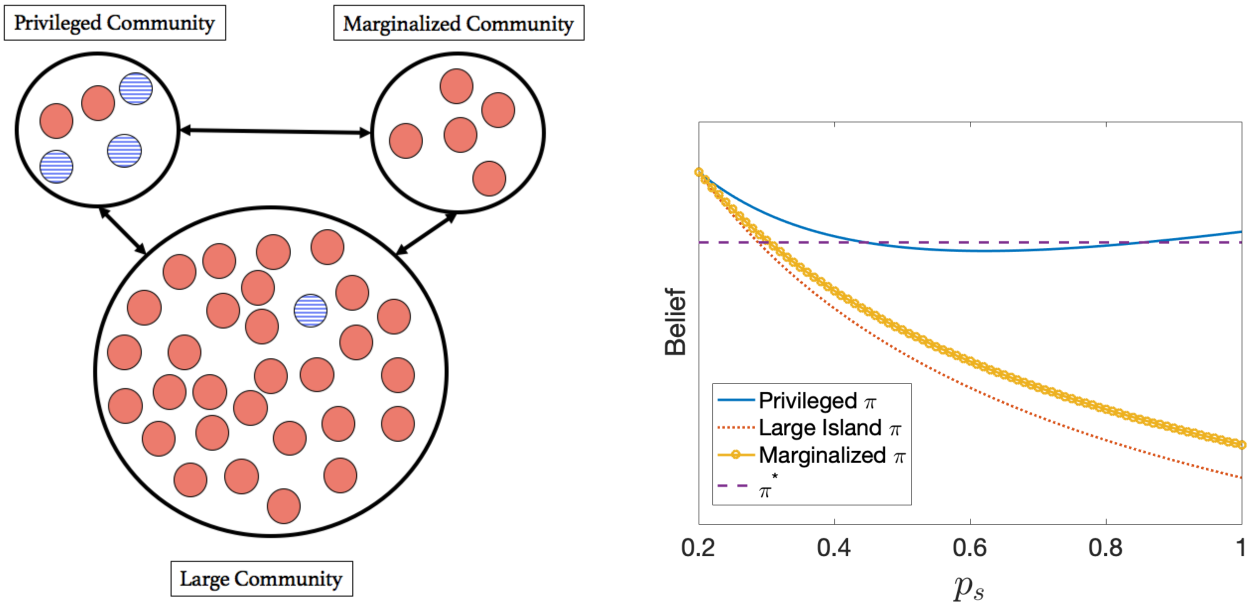


Figure 2. On the left is the setup of Section 3.1, and on the right are the beliefs of the two communities. As homophily increases (i.e. as p_s increases), the beliefs of all communities move away from the truth and fall below the belief threshold, so that all agents take the incorrect action. Further increase in homophily restores some of the beliefs in the privileged community, but does not bring it back to first-best levels.

in the network is manipulated.

After a point, more increase in homophily begins to restore the beliefs of privileged community, but not to the extent of returning these beliefs to first-best levels. This is a consequence of the large community suffering with false beliefs and dragging down the beliefs of those in the privileged community. Thus, once homophily reaches an extreme level, the misinformation rampant on the large island ceases to spread beyond its own community. In that sense, extreme homophily is better than intermediate homophily because at least one of the islands is insulated from the misinformation in the large community.

3.3 Inequality Protects Society

We now give an example to show how the spread of misinformation can be shaped by the interplay between the principal's strategy and the inequality structure of society. Consider three communities of the same size. The privileged community has a 3% Bayesian population, the "average" community has a 1% Bayesian population, and the marginalized community has no Bayesians. Unlike the previous examples, we assume that $\varepsilon \in (4/5, 1)$, so that it is *not free* for the principal to expend resources in manipulating the beliefs of the agents. This setup is depicted

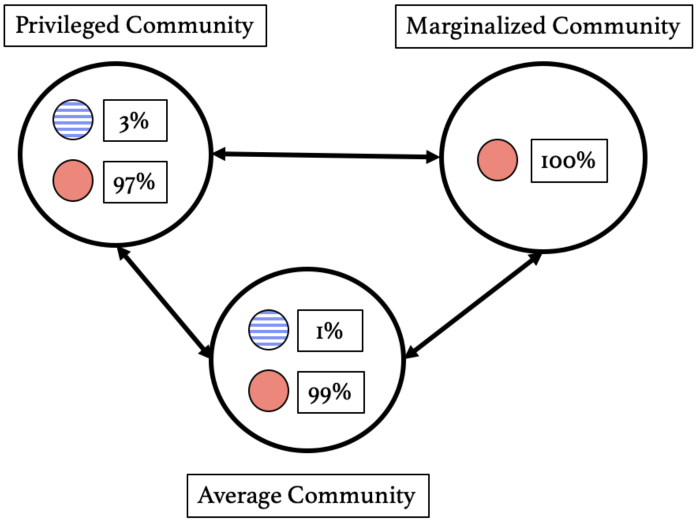


Figure 3. Example of Section 3.3

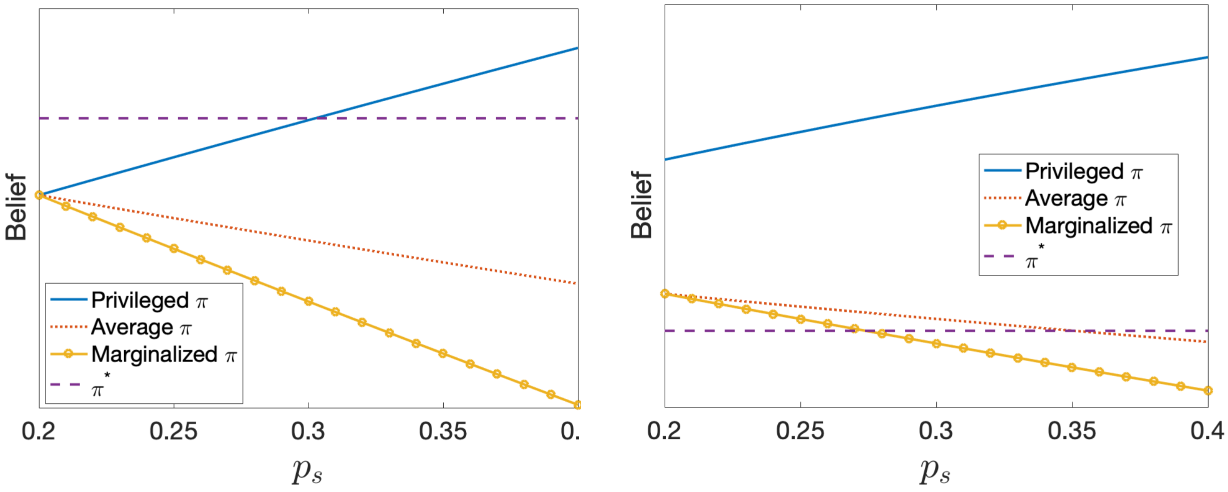
in Figure 3.

Figure 4a shows the beliefs (of the correct state) when the principal sends signals to *everyone* in the population. Suppose there is no homophily, so that $p_s = p_d = 0.2$, and, as always, there is a belief cutoff π^* for taking the correct action. Then under this strategy, all agents are manipulated, and the cost of sending signals is $\varepsilon < 1$, so this is indeed profitable and the network is susceptible.[16]

As homophily increases, the beliefs of the privileged community move closer to the truth and eventually pass the cutoff, thereby insulating them from the strategy where the principal exerts maximal influence over the entire population. For instance, when $p_s > 0.3$ in Figure 4a, the privileged community takes the correct action even though the other two communities do not. However, since $\varepsilon > 4/5 > 2/3$ (the principal is manipulating $2/3$ of the population), a strategy that targets all agents in the population is no longer profitable.

Instead, we investigate whether the principal has a profitable strategy where he incurs less cost but still manipulates the average and marginalized communities. It is relatively easy to show that the principal always prefers to decrease his influence on the privileged community, as he cannot manipulate this community anyway, and, because each community has more connections within their own island, exerting influence *within* the average and marginalized com-

[16] Note that it is not immediate that every agent will be manipulated in equilibrium, just that the network is susceptible (see Mostagir et al. (2019), Corollary 2). However, it can be shown through a more sophisticated argument that if the principal targets at most 5/6 of the population (regardless of the distribution across islands), then he manipulates no one. Thus, the optimal strategy for the principal is to target sufficiently many agents to guarantee that all islands are manipulated.



(a) Beliefs when principal targets all agents

(b) Beliefs when principal sends fewer signals

Figure 4. Beliefs of the communities in Section 3.3. The left plot shows beliefs when the principal targets everyone. The right plot shows beliefs when the principal sends fewer signals to the privileged island.

munities distorts beliefs the most within these communities. For a profitable strategy then, the principal must abandon sending misinformation to $1 - \frac{2}{3\varepsilon}$ proportion of the population; in particular, given $\varepsilon > 4/5$, he must abandon sending misinformation to 1/6 of the population, or 1/2 of the privileged community. The beliefs of the agents under this strategy are pictured in Figure 4b.

Notice though that under this strategy, when $p_s = 0.3$, none of the agents in the average community are manipulated either, while the agents in the marginalized community are. This in turn implies that this strategy is unprofitable, as the principal expends $\frac{5\varepsilon n}{6} > \frac{2}{3}n$ but only receives a benefit of $\frac{1}{3}n$. Instead, the principal must abandon sending misinformation to $1 - \frac{1}{3\varepsilon} > 7/12$ proportion of the population. Once again, it can be shown that the principal is better off targeting the marginalized community directly, before sending misinformation to the other communities. The principal's maximal influence, while still being possibly profitable, comes from sending everyone on the marginalized community misinformation and then either: (i) not sending signals to the privileged community but sending signals to $1/12$ of the average community, or (ii) not sending signals to the average community but sending signals to $1/12$ of the privileged community. The beliefs under both these strategies are pictured in Figure 5.[17]

As can be seen, under either of these strategies there is not enough misinformation sent to distort even the marginalized island's beliefs. The network is *impervious* to manipulation be-

[17] Any convex combination of targeting agents in both the average and privileged communities lead to the same conclusion.

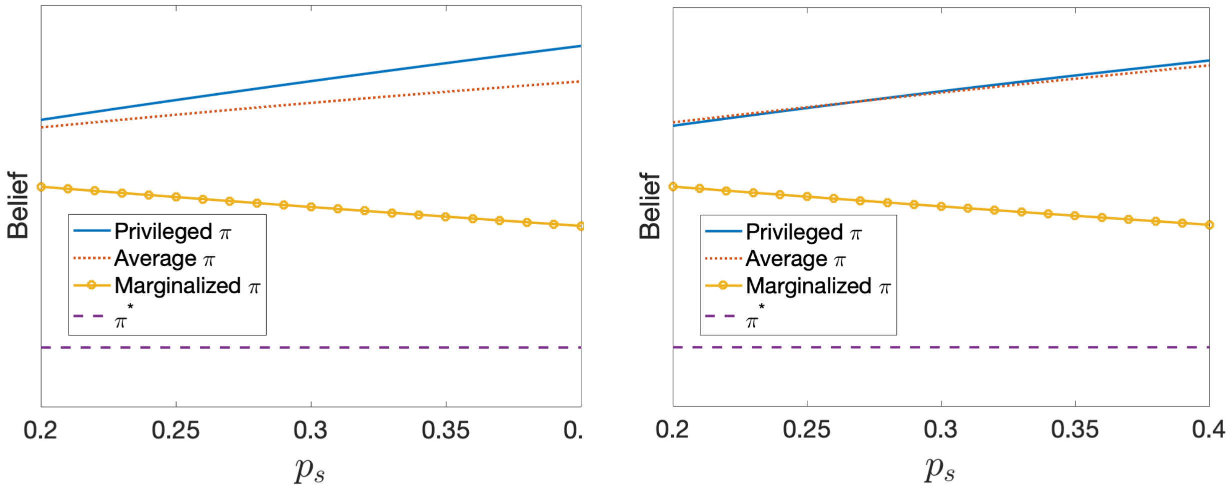


Figure 5. Beliefs of the communities in Section 3.3 when the principal attempts to manipulate only the marginalized community. On the left (right) are the beliefs when the principal targets a fraction of the average (privileged) community in addition to the marginalized community. Neither strategy is profitable as no one is manipulated.

cause of this domino effect: as one community becomes more insulated, its beliefs move closer to the truth and spill over to the next community and the process repeats until all communities are protected and the principal had no profitable strategy. As such, the presence of *some* inequality can end up protecting everyone by discouraging strategic manipulation of beliefs. We call this phenomenon *protection contagion* and explore it further in Section 5.

4 Inequality and the Spread of Misinformation

In this section, we focus our attention on the weak inequality model when the cost of sending misinformation, ε, is close to 0. In such instances, the principal's optimal strategy is trivial: he sends misinformation to everyone in the network. This decouples the belief dynamics from the principal's strategy and allows us to study these dynamics in isolation. The case where $\varepsilon \gg 0$, and when the principal's optimal strategy is non-trivial, is the focus of Section 5.

4.1 Inequality and Network Homophily

In the weak inequality model, a society $(p_s, p_d, \mathbf{m})$ is specified by three objects:[18]

1. p_s: the communication within islands (i.e., the within-island link probability).

[18] We use "communities" and "islands" interchangeably throughout.

2. p_d: the communication across islands (i.e., the across-island link probability).

3. $\mathbf{m} \equiv (m_1, \ldots, m_k)$: the vector of Bayesian counts for each island $\ell \in \{1, \ldots, k\}$.

We refer to the pair (p_s, p_d) as the *homophily* of the network; these parameters completely determine the social network structure. We always assume $p_s \geq p_d$. Next we define what it means for one society to exhibit less inequality than another:

Definition 1. (Inequality) We say that society $(p_s, p_d, \mathbf{m})$ exhibits *less inequality* than society $(p'_s, p'_d, \mathbf{m}')$ if:

(a) There is more communication across islands; namely, $p_d \geq p'_d$;

(b) There is less communication within islands; namely, $p_s \leq p'_s$;

(c) The distribution of Bayesian agents across groups is more "equally distributed"; formally, $\mathbf{m}'/\mathrm{s}$ is a majorization[19] of $\mathbf{m}/\mathrm{s}$.

with at least one condition strict.

If Society A has less inequality than Society B, this suggests two features. First, any agent in Society A is more likely to talk to agents outside her own island, relative to the higher inequality Society B. This is a direct consequence of less homophily. Second, less inequality also implies less inequality in terms of direct connections to Bayesian agents: any two agents in Society A are more likely to have a similar number of (weighted) connections to Bayesians as compared to Society B. The most equitable distribution of Bayesians occurs when they are the same constant fraction of the population on every island.

Inequality provides a partial (as opposed to total) ordering on societies. This occurs for the following reasons. First, if we simultaneously increase homophily and more evenly distribute Bayesian agents, then we create more even access to resources but also restrict how communities share these resources, resulting in an ambiguous inequality comparison. For example, assuming equal island sizes, a society described by $(p_s, p_d, m_1, m_2) = (0.5, 0.5, 3, 0)$ is no more or less equitable than a society described by $(p'_s, p'_d, m'_1, m'_2) = (0.8, 0.2, 2, 1)$. Second, majorization itself defines only a partial order, so it is possible that two Bayesian distributions $\mathbf{m}$ and $\mathbf{m}'$ are not comparable. For instance, assuming equal island sizes, $(m_1, m_2, m_3) = (1, 1, 4)$ is no more or less equitable than $(m'_1, m'_2, m'_3) = (0, 3, 3)$. For this reason, we use the following definition in order to compare inequality structures:

[19] A majorization $\mathbf{x}'$ of $\mathbf{x}$ satisfies (i) $\sum_{\ell=1}^{k} x_\ell = \sum_{\ell=1}^{k} x'_\ell$ and (ii) $\sum_{\ell=1}^{\ell^*} x_\ell \geq \sum_{\ell=1}^{\ell^*} x'_\ell$ for all $\ell^* \in \{1, \ldots, k\}$, where the components of $\mathbf{x}$ and $\mathbf{x}'$ are sorted in ascending order (see Marshall et al. (2011)). An equivalent condition is whether one can transform $\mathbf{m}'$ into $\mathbf{m}$ via a sequence of "Robin Hood" operations: one can recover $\mathbf{m}$ from $\mathbf{m}'$ via a sequence of transferring Bayesians from islands with more Bayesians to those islands with fewer (see Arnold (1987)).

Definition 2. We say a society has the *most inequality* if there exists no other society with (strictly) more inequality. We say a society has the *least inequality* if there exists no other society with (strictly) less inequality. We say a society has *intermediate inequality* if it is neither a society with the most or least inequality.

Note that a society with the most inequality necessarily has homophily structure $(p_s, p_d) = (1, 0)$ and a society with the least inequality necessarily has $(p_s, p_d) = (0.5, 0.5)$. However, all the analysis that follows is continuous in p_s and p_d, and so holds for (non-empty) open intervals around these homophily parameters as well. Finally, we remind the reader that we use the term *marginalized* to refer to a community that has a smaller proportion of Bayesians compared to a *privileged* community (which has a higher proportion of Bayesians).

4.2 Misinformation in Equal-Sized Communities

We start with a basic setup in which there is a constant number of islands, k, arranged according to the weak inequality model. Each island has an equal share of the population $s_1 = s_2 = \cdots = s_k = 1/k$. In the same vein as Section 3.1, we show that the intuition of "increased inequality is bad for learning" is accurate in the special case where a) islands have equal sizes (as in Golub and Jackson (2012)) and b) the only criterion is whether society as a whole is impervious (i.e. no agent is manipulated) or not, rather than the number of agents manipulated.

Theorem 2. *If Society $(p_s, p_d, \mathbf{m})$ is susceptible to manipulation and has less inequality than Society $(p'_s, p'_d, \mathbf{m}')$, then Society $(p'_s, p'_d, \mathbf{m}')$ is also susceptible to manipulation.*

In other words, Theorem 2 states that there is an inequality threshold[20] whereby increasing inequality eventually flips the network from impervious to susceptible. This result corroborates the evidence that inequality hurts learning; in particular, inequality always negatively affects learning in the most marginalized communities. However, Theorem 2 does not claim that total manipulation —the *number* of manipulated agents— is monotone in the degree of inequality. In particular, once the network becomes susceptible, it may be possible that increasing inequality leads to a reduction in the extent of manipulation, though it does not return the network to its first-best state of imperviousness. This property holds generally:

Theorem 3. *For any society with $k \geq 3$ islands of equal size and m total Bayesians:*

[20] Because there is only a partial ordering of societies, this threshold holds two of three inequality parameters constant while changing the third one. For example, if the Bayesian distribution is the parameter being changed, then one can apply the threshold for any partially ordered sequence of distributions.

(i) For a given b, m, if there is an impervious network for some inequality structure, the network with the least inequality is impervious;

(ii) For all b, m, there always exists a network with intermediate inequality that has (weakly) more manipulation than some network with more inequality.

(iii) There exist values for b, m such that the network of (ii) with intermediate inequality has strictly *more manipulation than some network with more inequality.*

Theorem 3 states that an "intermediate" amount of inequality is worse than an extreme amount of inequality, which in turn is worse than no inequality at all. While removing all inequality improves learning, simply reducing inequality in an extremely homophilous society can actually lead to worse learning and manipulation outcomes.

Underlying the previous result is the fact that social connections have both positive and negative externalities. On one hand, they serve as a transmission mechanism for spreading the (correct) beliefs of Bayesian agents. However, they also allow the principal to spread misinformation in a more effective way, by using social forces to manipulate other agents as well. When homophily is strong, the principal cannot use one community to influence another. These missing connections can prevent the principal from manipulating certain communities, who had previously derived their beliefs from more marginalized communities when homophily was not too extreme. On the other hand, when homophily is quite weak, access to Bayesians is relatively similar across islands, which allows them to communicate truth most effectively. It is the intermediate homophily case that often acts as a perfect breeding ground for manipulating beliefs.

This result provides a sleek connection to models of contagion in financial networks (see Acemoglu et al. (2015), Babus (2016), Kanak (2017), for example). Similar to the degree of homophily in our setting, in these models, connections both serve to reduce and exacerbate the propagation of negative forces. On one hand, when a bank's linked institutions are in distress, the bank finds itself less well-capitalized and more likely to default. However, when a bank faces an idiosyncratic or temporary problem, it can rely on neighboring (safe) institutions to protect it from insolvency. Hence, the stability of a financial network can be subtle, and the effect of increased interconnectivity is typically ambiguous, just as with social learning in the presence of homophily and inequality.

4.3 Misinformation with Different Community Sizes

We now consider the case when communities are not the same size. We begin with the following definition:

Definition 3. We say that an island ℓ is *least privileged* if (i) its belief of the correct state is the least of any island (i.e., $\pi_\ell \leq \pi_{\ell'}$ for all ℓ') and (ii) it has the least Bayesian percentage of the population (i.e., $m_\ell / s_\ell \leq m_{\ell'} / s_{\ell'}$ for all ℓ').

Observe that condition (i) is also equivalent to island ℓ having the largest DeGroot centrality. Note that with islands of the same size, condition (i) holds if and only if condition (ii) holds for island ℓ, so is redundant. However, with islands of different sizes, because influence is asymmetrical across islands, neither condition implies the other.

As we saw in Section 3.2, when communities have different population sizes, the results of the previous section need not hold. When there is a large community, it is possible that additional inequality can hurt *the entire society*. Because most communities draw their beliefs from the belief of the "masses," the effect of inequality on the masses determines how society as a whole is affected by inequality:

Theorem 4. *Suppose there are k islands of unequal sizes. Assume the largest island, island 1, is the least privileged. For almost all b (see Footnote 14), there exists size threshold $\bar{s}$ such that if $s_1 > \bar{s}$, the number of manipulated islands is monotonically increasing in inequality, provided that island 1 remains the least privileged.*

Theorem 4 states that if we are to decrease inequality with a large least privileged island, manipulation can only decrease. Put more simply, if the masses are the least privileged, then decreasing inequality helps everyone, including very privileged communities. This is because these communities still form a sizable number of connection with the large island, just by virtue of the size disparity, and hence draw a large part of their beliefs from there. Indeed, as inequality decreases, Bayesian agents in privileged communities can have their voices amplified through talking to agents in the large community, who then spread these beliefs over the network (including back to DeGroot agents in the privileged communities). The flip side of this is that if the masses are the least privileged, increasing inequality helps no one: in fact, moving resources from the masses to the privileged communities ends up making both the masses and the privileged communities worse off. Theorem 4 thus establishes that inequality benefits society as a whole (in a Pareto sense) if it benefits the large community that wields heavy influence. Likewise, even the privileged islands should want to move their resources to reduce inequality.

Note the assumption that island 1 is the least privileged (and remains so after decreasing inequality) cannot be dispensed with. If island 1 is simply underprivileged, non-monotone comparative statics might exist following an increase in inequality. The intuition is as follows. While this inequality hurts island 1's access to more privileged communities' resources, it also exposes island 1 less to communities which are less affluent than itself and may have more misinformed beliefs. This effect does not exist, of course, when island 1 starts off as the least well-off community.

5 Strategic Influence and Inequality

The results in Section 4 are obtained under the assumption that the cost of sending misinformation is negligible, and therefore the principal targets every agent in the population. This $\varepsilon = 0$ case enabled us to measure how misinformation propagates as a function of the inequality structure in society, without introducing strategic considerations on the part of the principal.

We now relax this by assuming $\varepsilon \gg 0$, and for simplicity also assume that all communities are the same size. The latter assumption allows us to isolate the effects of the principal's strategy from the population size effects identified in Sections 3.2 and 4.3. The $\varepsilon \gg 0$ assumption requires a strategic choice by the principal of who to target, and is of critical importance in understanding the spread of misinformation in the presence of a strategic actor.

We start with an example where ε varies from small to large in a society with two communities, and observe that *some* inequality can protect the entire society by initially protecting the privileged community. Theorem 5 extends this to an arbitrary number of communities by showing that such a network where intermediate inequality is best for society always exists when the principal faces non-negligible signaling costs. This contrasts with Theorem 3, where intermediate inequality is not only never optimal, but is always (weakly) worst for society when signaling costs are low. Finally, we conclude with some numerical experiments that show how manipulation changes as a function of simultaneously varying the investment cost and the inequality structure under the principal's optimal strategy.

5.1 Illustrative Example

Suppose there are two islands of equal size. We assume that 5% of the population is Bayesian, and we have a fixed homophily structure of $(p_s, p_d) = (0.5, 0.2)$. Moreover, suppose that $b = 0$, so that agents choose the action corresponding to the state they believe is more likely. We explore how inequality, in the form of the distribution of Bayesian agents across the islands, affects the principal's strategy as we vary the cost ε:

1. *Extreme inequality*: Suppose Bayesians constitute 10% of the population of the first island and the second island has no Bayesians. Let κ_ℓ denote the proportion of agents targeted by the principal on island ℓ. Then the fraction (of the total population) of agents manipulated,

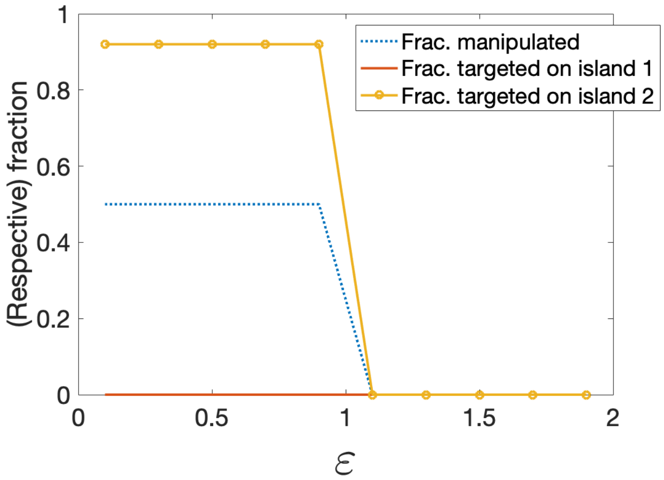


Figure 6. The principal's optimal strategy for various values of ε under extreme inequality. Depending on the curve, (Respective) fraction refers to either the fraction of the entire population manipulated, or the fraction of a given island that is targeted. For example, when $\epsilon = 0.5$, the principal targets no one on the first island and almost everyone on the second island and ends up manipulating half the population.

along with the proportions of island 1 and island 2 targeted by the principal (i.e., κ_1 and κ_2) are given in Figure 6. With extreme inequality, the principal "gives up" on the island with more Bayesians, and sends no misinformation to any agents on this island ($\kappa_1 = 0$). On the other hand, he sends misinformation to almost all of the second island and manipulates everyone on that island, until the cost of sending signals exceeds a threshold $\bar{\varepsilon} > 1$.

2. *Intermediate inequality*: Now suppose the first island has 7.5% Bayesians and the second island has 2.5% Bayesians. The principal's strategy and resulting manipulation are shown in Figure 7. The principal sends misinformation to everyone on the second island, but importantly, this alone is not enough to manipulate the agents on that island: he also has to send misinformation to the first island in order to be able to manipulate the second island. However, this strategy targets more agents on the whole than when there is extreme inequality, and thus is more expensive. After a point $\bar{\varepsilon} < 1$, the principal has no profitable strategy. This network is therefore more resilient than one with extreme inequality, since the cost range that allows the principal to (profitably) spread misinformation is smaller.

3. *No inequality*: Finally, suppose both islands have 5% Bayesians. The plot of the principal's strategy and resulting manipulation are shown in Figure 8. Similar to the case of extreme inequality, there is a threshold $\bar{\varepsilon} > 1$ such that the principal has no profitable strategy above that threshold, and so this inequality structure is again less resilient than the case of

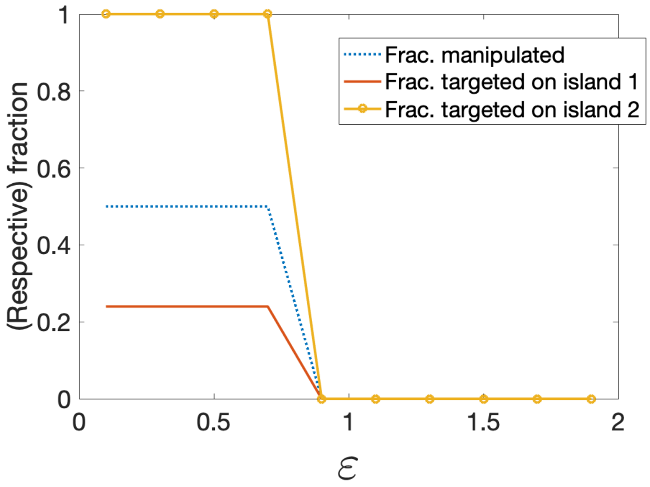


Figure 7. The principal's optimal strategy for various values of ε under *some* inequality. Depending on the curve, (Respective) fraction refers to either the fraction of the entire population manipulated, or the fraction of a given island that is targeted.

intermediate inequality. However, unlike the case of extreme inequality, the entire population is manipulated before this threshold is met. Thus, for $\varepsilon < 1$, the absence of inequality leads to maximal manipulation relative to the other inequality structures.

The computation of the principal's optimal strategy when there are two islands with the same population, as in the example above, can be easily generalized via the algorithm presented in Appendix B.3.

5.2 Protection Contagion: The Case for *Some* Inequality

Recall from Section 3.3 that when there was no inequality, the principal had a profitable strategy to target and manipulate everyone. With some inequality, however, the principal was unable to manipulate one of the more privileged communities, which in turn made his strategy too expensive. To maintain a profitable strategy, the principal had to reduce his direct influence on that community in order to save costs, while still trying to retain the same extent of (indirect) overall influence on the other communities. However, this reduction made the principal unable to manipulate the next privileged community, which similarly led to him reducing his direct influence in that community, and so on. We refer to this cascade effect as *protection contagion.*

This effect is not an artifact of Section 3.3, or the example presented in the previous section. In fact, when the principal has intermediate costs for sending misinformation, protection contagion can sometimes lead to a complete unraveling of his influence when there is some inequality in the network. This is summarized in the next result.

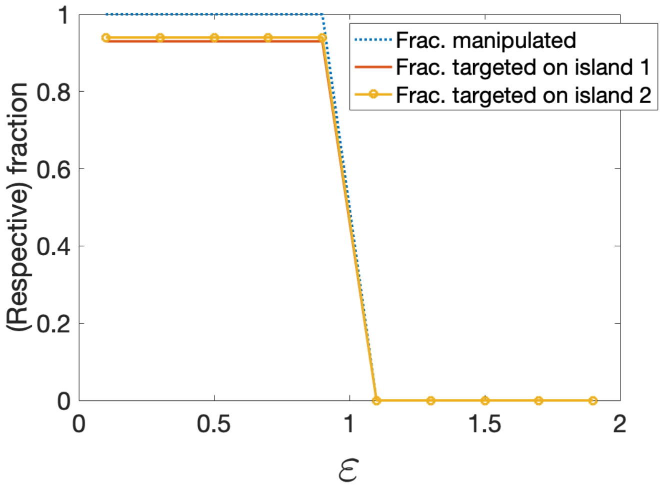


Figure 8. The principal's optimal strategy for various values of ε under no inequality. Depending on the curve, (Respective) fraction refers to either the fraction of the entire population manipulated, or the fraction of a given island that is targeted.

Theorem 5. *Suppose there are $k \geq 2$ islands of equal size. There exists $b^* < b^{**}$, $\varepsilon^* < \varepsilon^{**}$, such that if $b \in (b^*, b^{**})$ and $\varepsilon \in (\varepsilon^*, \varepsilon^{**})$, there exists a network with intermediate inequality that is impervious, despite every network with the most inequality being susceptible, and every network with the least inequality admitting strictly more manipulation than networks with the most inequality.*

Theorem 5 describes a range where intermediate inequality is best for protecting society from the spread of misinformation. Suppose we order the communities based on their privilege, i.e. the proportion of Bayesian agents in the population, and protect the most privileged community from manipulation. This protection forces the principal to decrease his effort in this community to try and maintain a profitable strategy. By doing so, the beliefs in that community move closer towards the truth, and because there is still *some* communication across communities, this provides a positive externality to the rest of the network. The principal then is unable to manipulate the next privileged community, and so stops targeting that community as well, leading to a recursive process that repeats for all communities, and the principal cannot target anyone while retaining a positive payoff. However, if inequality becomes extreme, this contagion effect fails to take place: the positive spillovers from protecting one community are minimal in the face of gross inequality. Extreme homophily leads to little communication across communities and so protecting one community still leaves the rest exposed to misinformation.

Note the connection between Theorem 5, when $\varepsilon \gg 0$, and Theorem 3(a), when $\varepsilon \approx 0$. Theorem 3(a) states that if some inequality model is impervious, then the least inequality attains imperviousness. This is not the case for $\varepsilon \gg 0$; in particular, Theorem 5 states that it may be

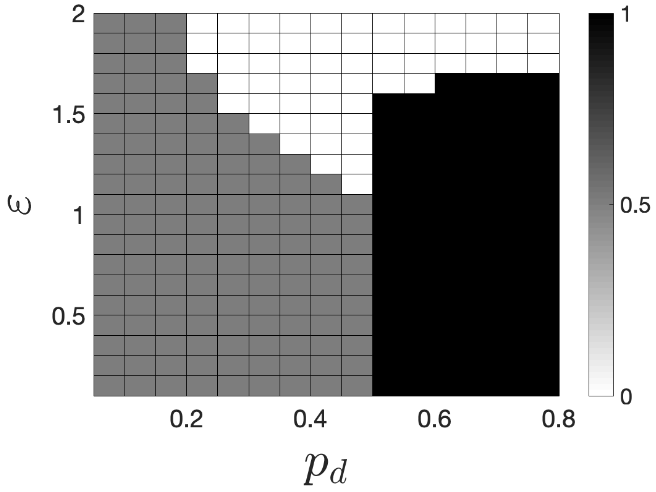


Figure 9. Heat map showing fraction of the population manipulated as a function of p_d (cross-island connectivity) and the signaling cost ε. Note that increasing p_d implies decreasing homophily/inequality. Light blocks indicate no manipulation, while gray (dark) blocks indicate half (all) the population is manipulated.

possible for an intermediate inequality model to be the only model that attains imperviousness.

5.3 Numerical Simulations

We provide results from two numerical simulations that illustrate the non-monotonic behavior from the previous section on the broader parameter space. Recall that we can increase the level of inequality by increasing homophily or by having a more uneven Bayesian distribution between islands of equal size. We simulate both of these scenarios. In the first simulation, we vary the extent of homophily through varying p_d (while holding p_s fixed). The second simulation varies the distribution of Bayesians across the islands. In both cases, we simultaneously vary the cost ε that the principal faces.

Homophily. We fix $p_s = 0.8$ and take $b = 0$, so that an agent takes an action based on the state she believes is most likely. There is a total population of 1000 agents split equally across two islands; one island has 80 Bayesians and the other has the remaining 20.

Figure 9 shows the results of this simulation. In the range of $\varepsilon \in (1.1, 1.7)$, we notice the non-monotonicity described in Theorem 5 as we increase p_d (i.e., decrease homophily/inequality). For small values of p_d (large homophily), half the agents are manipulated. As we decrease homophily through increasing p_d, we transition to a region where the network is impervious. Finally, as homophily decreases further, we end up in a region where *all* agents in the network are manipulated. This is the same effect seen in the example of Section 5.1.

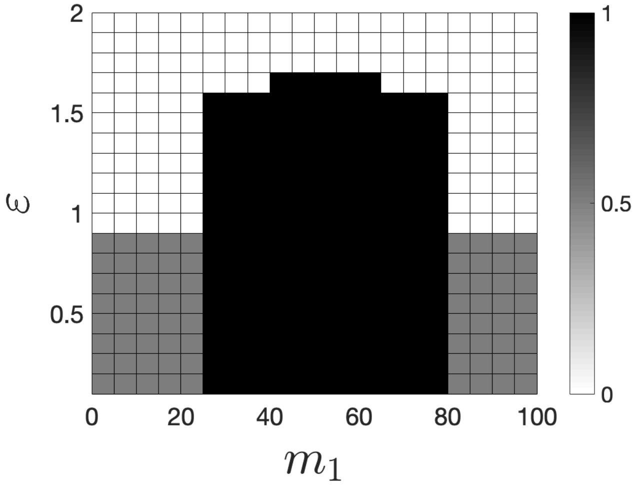


Figure 10. Heat map showing fraction of the population manipulated as a function of the number of agents on the first island, m_1, and the principal signaling cost ε. Light blocks indicate no manipulation, while gray (dark) blocks indicate half (all) the population is manipulated.

Bayesian distribution. We fix $(p_s, p_d) = (0.5, 0.2)$ and take $b = 0$. There is a total population of 1000 agents, split over two islands of equal size, and we vary the number of Bayesians, m_1, on the first island from 0 to 100 (with the other island containing the remainder, $m_2 = 100 - m_1$) .

The results are shown in Figure 10. Inequality between islands is most severe when $m_1 = 0$ or $m_1 = 100$, with the least inequality at $m_1 = 50$. In the range $\varepsilon \in (0.9, 1.7)$, we see that the network is impervious provided there is sufficient inequality in the distribution of Bayesians; otherwise, all agents are manipulated. This inequality protects one island from manipulation and, through protection contagion, prevents the principal from having any profitable strategy.

6 Weak vs Strong Inequality

Up until now, we have focused almost entirely on the weakly assortative network model, where agents associate more with those in their own group, but do not differentiate their social interactions amongst "other" groups. In this section, we compare how this model of inequality differs from the strongly assortative model, where agents further differentiate their social interactions by only affiliating with groups who have characteristics that are *close* –in the sense defined in Section 2.2– to those of their own group. We consider how this type of strong inequality affects society at large relative to the weaker notion of inequality we have studied. Our main finding, which we make precise over the next two sections, is that "good" strong inequality models are still much worse, in terms of manipulation, than "bad" weak inequality models. Thus, the exis-

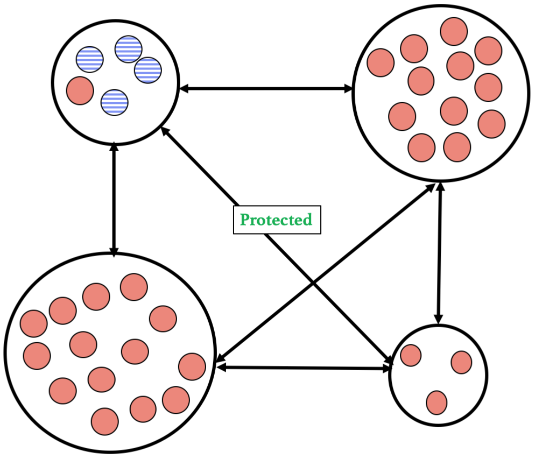


Figure 11. An illustration of Proposition 1: under weak homophily, a linear number of Bayesian agents is enough to prevent manipulation anywhere in the network. This could include stacking them all on one island.

tence of hierarchical homophily structures is much more detrimental to the spread of misinformation.

6.1 Weak Inequality

Section 4 and Section 5 documented the effects of weak inequality on manipulation. To make the comparison between strong and weak inequality most transparent, we consider *worst-case* inequality for weak inequality i.e., the inequality structure that makes the principal most easily able to manipulate. Toward this end, the following result establishes a condition on the *total number of Bayesians* in the weak homophily model needed for imperviousness, *independent* of their placement across communities:

Proposition 1. *For any (p_s, p_d), there exists $\bar{\theta}$ such that if $\theta < \bar{\theta}$, there exists a constant $c < 1$ such that if there are $m = cn$ Bayesian agents anywhere, then* any *weak inequality model (regardless of the number of communities k) is impervious. Moreover, c is increasing in p_s and decreasing in p_d.*

In other words, there exists a threshold c whereby if a proportion c of the population is Bayesian, the principal will be unable to manipulate anyone, regardless of the depth of weak inequality present. This includes the most extreme inequality configuration where all the Bayesians are on one island, and the rest of the $k-1$ islands are all DeGroot (for any k). A visual depiction of Proposition 1 is given in Figure 11. The assumption that θ is not too large ensures that agents use social learning as a primary means of learning; clearly when θ is too close to 1, the presence of Bayesian agents is irrelevant because agents place too much weight on their own (manipulated) news.

Proposition 1 also sheds some light on whether homophily helps or hurts the worst-case Bayesian lower bound. Because m is increasing in p_s and decreasing in p_d, we see the number of Bayesians needed to apply Proposition 1 increases as we increase inequality through the network homophily structure. This result reinforces the general idea that increasing inequality makes it more challenging for society to avoid manipulation, despite the exceptions presented earlier. The intuition is clear: as homophily becomes more severe, configurations like that of Figure 11 do little to help communities with few to no Bayesians.

6.2 Strong Inequality

For illustration, we assume that the first community has m Bayesian agents and all other communities consist of DeGroot agents. At the end of this section, we discuss the robustness of the result to other configurations. There are k communities which may or may not be the same size and we fix (p_s, p_d). In the strong inequality model we obtain a much different result from Proposition 1:

Proposition 2. *For any $\theta > 0$ and $c < 1$, there exist $\bar{k}$ (independent of k) and $\varepsilon > 0$ where all communities except $\bar{k}$ are manipulated, even with cn Bayesian agents.*

Proposition 2 shows the stark difference between weak and strong inequality. First, with weak inequality, we can always find a proportion c such that cn Bayesians will make the network impervious, even with rampant (weak) inequality. On the other hand, we can never find such a proportion c in the strong inequality model: no constant fraction guarantees society is safe from manipulation because the influence of Bayesian agents is too diluted under strong homophily, as seen in Figure 12. Second, the strong inequality network is not only susceptible, but manipulation is actually *ubiquitous* in society. Note that $\bar{k}$ does not depend on k, so when there are several communities, only a vanishing fraction of them will not be manipulated. Except for a very small set of communities who happen to have close ties to Bayesian agents, almost all communities will be negatively impacted by the existence of strong inequality.

The intuition for the result is as follows. Notice that with strong inequality, as in Figure 12, agents receiving misinformation communicate their beliefs both forwards and backwards, which leads to more propagation of misinformed beliefs. This creates a strong *echo chamber effect*, where the influence from misinformation, as reflected in the agents' beliefs, gets inflated because they fail to recognize their own influence on their own neighboring islands' beliefs. For agents who are not in communities extremely close to the Bayesian community, this echo cham-

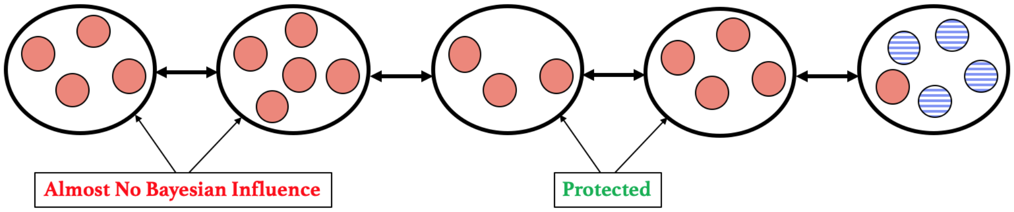


Figure 12. An illustration of Proposition 2: even with many Bayesian agents, strong homophily allows the principal to manipulate plenty of agents in the network. In the figure above, communities which are not "close enough" to the Bayesians will not be very influenced by their beliefs, so will be manipulated.

ber is strong enough to completely mask any influence the Bayesians might have in spreading accurate beliefs. Contrast this with weak inequality in Figure 11, where every community has some direct interaction with the Bayesians, even if there are no Bayesians in that community itself. This not only provides a direct positive influence on everyone's beliefs, but also prevents these echo chambers from wielding too much power, precisely because other communities are also directly interacting with Bayesians.

Robustness. Finally, we consider how robust these strong inequality results are to the initial setup. Suppose instead of stacking all of the Bayesians on the first island, we instead redistribute them in a way that dampens these echo chamber effects. Would this mitigate the effects of Proposition 2? An affirmative answer to this question requires this redistribution to be significant. From Proposition 2, it is easy to see that any island with a Bayesian cannot protect more than a constant number of communities on either side of it. Therefore, the number of Bayesians would need to be dispersed very evenly across all communities to have any hope of preventing manipulation. For example, simply moving the Bayesians to a more central community or distributing them across a couple of islands throughout would have no significant effect, and the conclusion of Proposition 2 remains intact. Thus, while agents can be protected in the strong homophily model, the requirements on the Bayesian distribution are much stricter: nearly every island has to have some Bayesian agents of its own, which requires drastically less inequality.

Second, Mostagir et al. (2019) show that higher *density*, while not a perfect measure, is often related to lower manipulation (for instance, see Theorem 4 in Mostagir et al. (2019)). It is clear that the average degree with strong inequality will be lower than that of weak inequality, so a natural question is to wonder whether this difference in density is what drives the difference in manipulation we observe between the two models. For concreteness, assume we have k communities of the same size in both the strong (with p_s, p_d) and weak inequality (with p'_s, p'_d)

models. In the strong inequality model, we take $p_s = \alpha p'_s$ and $p_d = \alpha p'_d$, where $\alpha = \frac{p_s+(k-1)p_d}{p_s+2p_d}$.[21] This equalizes the average degree (i.e., connections) of the strong and weak inequality models, but has no effect on any of the beliefs of the agents (or on their DeGroot centralities).[22] Therefore, we see that the differences in density alone cannot explain the differences seen across the two models.

7 Optimal Interventions

We now discuss the role that a social planner has in combating misinformation. We consider two possible interventions: Bayesian interventions and homophily interventions. In the former, we assume the planner may improve the sophistication type of a subset of agents, perhaps through targeted education. In the latter, the planner may decrease the extent of homophily through efforts to integrate communities (i.e. by increasing p_d). The social planner wants to enact a policy that protects as many agents as possible from manipulation.

We say a policy is *optimal* if it minimizes the number of manipulated agents. Similarly, we say some policy X *dominates* another policy Y if all agents' beliefs of the correct state are higher under X than under Y. While an optimal policy is never dominated, there may be non-optimal policies that are not dominated, and thus lie at the Pareto frontier of effective interventions.

We can write the beliefs of the agents, $\boldsymbol{\pi}$, as:

$$\boldsymbol{\pi}(p_s, p_d, \mathbf{m}, \mathbf{x}) = \left(\frac{\mathbf{I}}{1-\theta} - \boldsymbol{B}(p_s, p_d, \mathbf{m}, \mathbf{x})\right)^{-1} \boldsymbol{a}$$

where $\boldsymbol{B}$ is a function of (i) the homophily structure (p_s, p_d), (ii) the distribution $\mathbf{m}$ of Bayesians across islands $\mathbf{m}$, and (iii) the principal's strategy $\mathbf{x}$. Recall that the belief (of the correct state) threshold is given by $\frac{1+b}{2}$ and the principal wants to maximize the number of agents whose beliefs fall below this threshold, less the total cost of manipulation, so as before, the principal solves:

$$\mathbf{x}^*(p_s, p_d, \mathbf{m}) = \arg\max_{\mathbf{x}} \sum_{i=1}^{n} \left(\mathbb{1}_{\boldsymbol{\pi}_i(p_s,p_d,\mathbf{m},\mathbf{x})<(1+b)/2} - \varepsilon x_i\right)$$

Then the planner solves the min-max optimization problem:

$$\min \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{\pi}_i(p_s,p_d,\mathbf{m},\mathbf{x}^*(p_s,p_d,\mathbf{m}))<(1+b)/2}$$

[21] For this, we have to naturally assume p'_s and p'_d are not too large so that $p_d < p_s < 1$ and this is possible. Otherwise, there is no way to equalize the average degrees of the two models.

[22] Technically, this equalizes the average degree for only the islands in the "middle" of the line, but not those on the ends. However, assuming there are a large number of communities, this difference will be negligible.

The combinatorial nature of these problems preclude a general solution. We derive the optimal policies for some special cases and show the nuances of optimal policies via simulation. For a number of cases, we prove the optimal policy attempts to minimize inequality. However, this is not always true: if the planner cannot completely eradicate inequality, then sometimes measures that only slightly reduce it can be counterproductive.

7.1 Bayesian Interventions

We consider the possibility of endowing some agents in the population with Bayesian abilities. We assume that this process is costly and that the planner's budget constraint is of the form $\sum_{i=1}^{n} \mathbb{1}_{\text{type}(i)=B} \leq M$, where the (non-bolded) letter B designates a Bayesian agent and M is an integer. Note that the planner will always use the entire budget in an optimal policy.

Intervention with Large Budget and Cheap Signals. First, we derive the optimal policy when the planner's budget is sufficiently large and the principal's cost of sending signals is nearly free:

Corollary 1. *Suppose that the budget M is large enough so that it is possible for the planner to make the network impervious when ε is small. Then if all islands are the same size, the optimal policy is to minimize inequality.*

This result is in-line with the conclusion of Theorem 2, which argues that when imperviousness is possible, the least inequality is always first-best. As a special case, if M is big enough to make the Bayesian populations equal on every island, then this is the optimal policy.

Note the assumption that M is big enough that imperviousness is attainable for small ε is necessary. First, if imperviousness is attainable only for a given $\varepsilon \gg 0$, then it is possible that a configuration with some inequality may be optimal, as we show next. This is a direct consequence of Theorem 5. Second, if M is small and so imperviousness is impossible, then an optimal policy may involve creating some inequality to protect at least a fraction of the population. For instance, placing Bayesians equally may lead to every island being manipulated, whereas stacking them all on one island would protect at least this island.

Intervention with Costly Signals. We simulate the optimal policy for the planner when the principal's signals are costly. Consider Figure 13 with budget $M = 4$, $n = 100$, two islands of equal size, and homophily structure $(p_s, p_d) = (0.8, 0.2)$. We investigate whether the optimal policy can make the network impervious as a function of ε.

As ε ranges from 0 to 0.5 (small ε region), there is no distribution of Bayesians that admits imperviousness. Corollary 1 allows us to quickly check this by distributing the Bayesians evenly

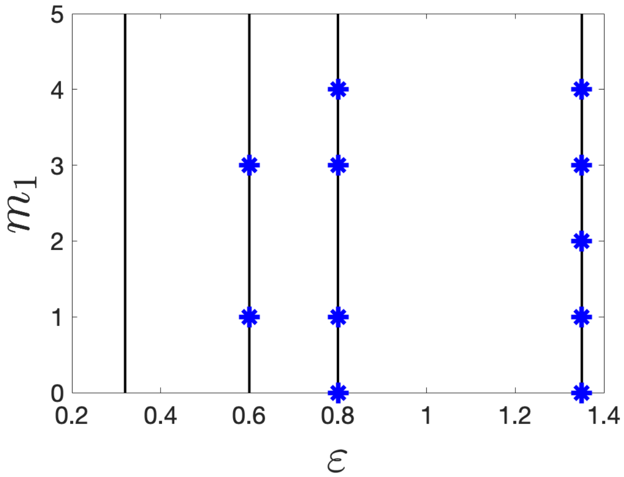


Figure 13. Policies that obtain imperviousness with budget $M = 4$. The number of Bayesians on the first island is m_1 and ε is the signaling cost of the principal. Every highlighted point is a Bayesian placement that leads to imperviousness, e.g. $m_1 = 1$ and $m_2 = 4 - 1 = 3$ when $\varepsilon = 0.6$.

across the two islands and checking if manipulation exists, which it does. As ε becomes slightly higher than 0.5, the most inequitable distribution (4 Bayesians on one island and 0 on the other) or the most equitable distribution $(m_1, m_2) = (2, 2)$ leads to manipulation. On the other hand, an unequal distribution of $(m_1, m_2) = (3, 1)$ or $(m_1, m_2) = (1, 3)$ leads to imperviousness. When ε continues to increase, *only the even distribution* $(m_1, m_2) = (2, 2)$ *makes society susceptible*, i.e., $(0, 4)$ and $(4, 0)$ are also impervious, and splitting the Bayesians equally across both islands is the worst distribution for society. Of course, eventually, all distributions are impervious when the cost becomes too prohibitive for the principal to have a profitable strategy. In summary, for the planner, the most equitable distribution $(m_1, m_2) = (2, 2)$ is weakly dominated by every other distribution over the *entire* cost range of ε.

Therefore, in the case of costly signals for the principal, the planner may want to introduce some inequality in the Bayesian distribution. This is precisely to generate the protection contagion effect documented in Theorem 5.

Intervention with One Large Island. We next consider a setting where there is one large island and many small islands; without loss, let the large island be island 1:

Corollary 2. *When ε is small, there exists $\bar{s}$ such that if $s_1 > \bar{s}$, any policy that makes island 1 the least privileged is dominated by a policy with more Bayesians on this island, provided that such a policy does not already use the entire budget on island 1.*

Corollary 2 complements Theorem 4: when there is a single large island, a policy which subjects the masses (on the large island) to inequality is not only sub-optimal, it actually hurts *all* agents in society. In particular, if we start with a configuration of many small privileged communities and one large under-privileged community, then even the privileged communities are negatively impacted by taking resources (Bayesians) from the under-privileged community. Therefore, the social planner *and* those agents in privileged communities should (rationally) support expending more resources on the least privileged community.

7.2 Homophily Interventions

We now consider a fixed homophily model with parameters (p_s, p_d^o) and Bayesian distribution $\mathbf{m}$. We assume the social planner pays a positive, convex cost $\phi(p_d - p_d^o)$ with $\phi(0) = 0$ to increase (or decrease) connections between islands. As before, we assume the planner has a budget to spend; that is, the planner must satisfy $\phi(p_d - p_d^o) \leq Budget$.

For the remainder of this section, we focus on the case where the principal's signaling cost ε is small. Similar conclusions to the previous section apply when $\varepsilon \gg 0$. Our first result shows that an equally-distributed Bayesian policy eliminates the need for a homophily intervention:

Proposition 3. *If Bayesians are equally distributed (i.e., $m_\ell = Ms_\ell$ for all ℓ), then $p_d = p_d^o$ is an optimal policy.*

When Bayesians are distributed proportional to the islands' populations, the beliefs of all agents in society are the same regardless of the homophily parameters. Therefore, no additional intervention is necessary because access to Bayesian agents is perfectly equitable.

Large Budget. We first focus on the case where the planner's homophily budget is fairly large. Based on our observations in Theorem 2 and Theorem 3 we have:

Proposition 4. *Suppose the budget is greater than $\phi(p_s - p_d^o)$ and large enough to make the network impervious. Then if all islands have the same size, $p_d = p_s$ is the optimal policy when ε is sufficiently small.*

When homophily can be fully corrected, some intervention is desirable. By Theorem 3, we know that if the budget is big enough for the planner to implement $p_d = p_s$, this obtains the first-best outcome. Thus, the optimal policy is always to completely eliminate any homophily that exists in the network.

Small Budget. When the planner's budget is limited, implementing $p_d = p_s$ may not be feasible. In this case, it may not be optimal to simply reduce inequality by minimizing $(p_s - p_d)$; as we saw in Theorem 3, often some inequality can be worse than extreme inequality. In other words,

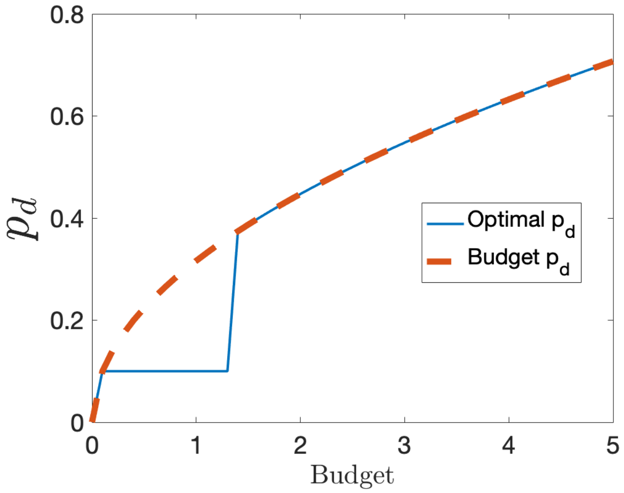


Figure 14. Optimal (Simulated) p_d with limited budget. The dotted curve is the level of homophily achieved by spending the entire budget available, whereas the solid curve is the optimal level of homophily given that budget. A planner should always ask for a minimum budget (around 1.3 in the figure) allocation that makes these two curves coincide.

a planner who simply helps decrease inequality without eradicating it completely can do unintended harm. In fact, unlike with Bayesian interventions, the planner may not want to use the entire budget.

We simulate the optimal p_d with three islands of equal size, $n = 999$, which have 100, 60, and 10 Bayesians, respectively. We assume the cost function $\phi(\alpha) = 10\alpha^2$, $p_d^o = 0$, and $p_s = 0.8$. Note the principal can set $p_d \leq \sqrt{Budget/10} < 0.8$ provided that the budget is at most 5. As in Proposition 3, we assume the cost of the principal's signaling technology is small, i.e., $\varepsilon \approx 0$.

In Figure 14, we show both the optimal homophily (p_d) choice of the planner and the maximum p_d attainable by the budget. We see that when the budget is small, the planner prefers to leave relatively extreme homophily in society as opposed to making the more substantial correction allowed by the budget. Once the budget exceeds a threshold (around 1.3 in the figure), however, the planner uses all of it up to remove as much homophily from the network as possible. Thus, a planner tasked with stopping misinformation through reducing inequality should always ask for a budget allocation that is at least equal to that threshold, in order to guarantee that this reduction will indeed achieve the desired effect and be beneficial to society.

8 Conclusion

This paper analyzes the role of inequality in social learning when the information that agents receive is a mixture of organic news and news originating from a strategic actor. This setup resembles many scenarios where an information provider may have their own agenda and exerts costly effort to influence agents to take certain actions. Inequality in society results from the distribution of knowledgeable agents who know the true state of the world. Privileged communities have a higher proportion of these agents, and the homophily structure of the network determines access to these agents across communities.

We show that the role that inequality plays in the spread of misinformation is shaped by relative community privilege, relative community sizes, and the cost of the principal's signaling technology. This leads to a range of outcomes depending on how these factors interact. For example, when the privileged communities are small in size compared to the population at large, as is often the case, then an increase in inequality not only makes the large population worse off, but it also makes the privileged communities themselves more prone to misinformation, thus it is in the privileged communities' best interest to encourage allocating resources to the larger community. Generally, the spread of misinformation in not monotone in the level of inequality in the network. Even more so, intermediate levels of inequality can be worst for society when the signaling costs of the principal is low, but can be best for society when the principal's costs are high.

In a similar vein, policies that counteract manipulation depend on the principal's signaling costs as well as the social planner's budget. When signaling costs are low, so that the principal can target everyone, then the planner's optimal policies with a large budget involve eradicating inequality through equitable Bayesian placement and removing homophily from the network. When signaling costs are no longer trivial, the planner needs to be more careful, as inequality extremes might be worse for society than intermediate inequality regimes. This is a consequence of a protection contagion phenomenon that precludes the principal from having a profitable manipulation strategy. Generally, the complexity of computing these optimal strategies provides a wealth of interesting algorithmic challenges that can be explored further and constitute a promising area for future work.

Finally, our model provides a basic framework to analyze the phenomena described in the paper in terms of the primitives of the problem represented by the inequality structure, the planner's budget, and the strategic injection of misinformation. Given the salience of these points in modern social learning environments, the model provides a step towards understanding the complex interactions of these factors, and offers guidelines that can help inform policies that aim to reduce inequality and protect society from misinformation.

Appendix

A Technical Conditions and Model Details

Appendix A.1 provides more technical details about the deterministic model, while Appendix A.2 give conditions under which the Bayesian learning results hold in the random network generation model. In Appendix A.3 we show how to adapt the model in Mostagir et al. (2019) to one with different communities with different levels of access to resources. Finally, Appendix A.4 demonstrates the main methods used in the proofs of this paper.

A.1 News Generation and Belief Evolution

The following model details are from Mostagir et al. (2019) and are presented here for contextualization of Section 2.1.

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

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

i.e., the signal is correlated with the underlying truth.

(b) **News from Principal**: In addition to the organic news process, there is a principal who may also generate news of his own. At $t = 0$, the principal picks an influence state $\hat{y} \in \{S, R\}$. The principal then picks an influence strategy $x_i \in \{0, 1\}$ for each agent i in the network. If the principal chooses $x_i = 1$, for any agent i, then he (the principal) generates news according to an independent Poisson process with (possibly strategically chosen) intensity λ_i^* which is received by all agents where $x_i = 1$. We assume the principal commits to sending signals at this intensity, which may not exceed some threshold $\bar{\lambda}$.

(c) **News Observations**: Agents are unable to distinguish news sent by the principal or that organically generated. We denote by $\hat{t}_1^{(i)}, \hat{t}_2^{(i)}, \ldots$ the arrival of *all* news, either from organic sources or from the principal, for agent i. At each time $\hat{t}_\tau^{(i)}$, if the news is organic, the agent gets a signal according to the above distribution, whereas if the news is sent from the principal, she gets a signal of $\hat{y}$.

(d) **DeGroot Update**: DeGroots use a simple learning heuristic to update beliefs about the underlying state from other agents. We assume every DeGroot agent believes signals arrive according to a Poisson process and all signals are independent over time with $\mathbb{P}\left(s_{i,\hat{t}_\tau^{(i)}} = y\right) = p_i$ (i.e., takes the news at face value). DeGroot agents form their opinions about the state both through their own experience (i.e., the signals they receive) and the beliefs of their neighbors. Given history $h_{i,t} = (s_{i,\hat{t}_1^{(i)}}, s_{i,\hat{t}_2^{(i)}}, \ldots, s_{i,\hat{t}_{\tau_i}^{(i)}})$ up until time t with $\tau_i = \max\{\tau : \hat{t}_\tau^{(i)} \leq t\}$, each agent forms a personal belief about the state according to Bayes' rule. Let $z_{i,t}^S$ and $z_{i,t}^R$ denote the number of S and R signals, respectively, that agent i received by time t; then the DeGroot agent has a direct "personal experience":

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

As mentioned in Section 2.1, each DeGroot i then updates her belief for all $k\Delta < t \leq (k+1)\Delta$ according to:

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

for some weights θ_i, α_{ij} with $\theta_i + \sum_{j=1}^{n} \alpha_{ij} = 1$, and Δ is a time period of short length. Note for simplicity in this paper we assume $\theta_i = \theta$ for all DeGroots i and $\alpha_{ij} = \frac{1-\theta}{|N(i)|}$ for all $j \in N(i)$.

(e) **Bayesian Update**: Bayesian agents know the network $\mathbf{G}$ and the signal structures $\{p_i\}_{i=1}^n$. Each Bayesian observes the history of beliefs in her neighborhood $N(i)$ for all time $t' \leq t$. Moreover, the Bayesian is aware the principal may be strategic and has accurate conjectures about the equilibrium (influence) strategy of the principal (as is typical in a Nash equilibrium). At time $t + dt$, the Bayesian agents makes a Bayesian update about the state given her private history of signals and her history of observed neighbor beliefs, forming π_{t+dt}.

(f) **DeGroot Centrality Vector**: To figure out the limit beliefs of the DeGroot agents when the principal targets everyone who is not a Bayesian, denote by $\boldsymbol{\gamma}$ the vector in $\{0,1\}^n$ that designates which agents are targeted by the principal and let $\gamma_i = x_i = 1$ wherever agent i is DeGroot and $\gamma_i = 0$ everywhere else. DeGroot centrality, which is equivalent to the belief in the incorrect state in the limit is then given by $\mathcal{D}(\boldsymbol{\gamma}) = (\mathbf{I} - \mathbf{W})^{-1}\boldsymbol{\gamma} = \sum_{k=0}^{\infty} \mathbf{W}^k\boldsymbol{\gamma}$, where $\mathbf{I}$ is the identity and $\mathbf{W}$ is the adjacency matrix of weights given in Mostagir et al. (2019). DeGroot agent i is manipulated if her belief in the false state is above the cutoff, i.e. if $\mathcal{D}_i(\boldsymbol{\gamma}) > (1-b)/2$. More detailed methods on the computation of DeGroot centrality is given in A.4.

A.2 Reduction of Bayesians to Stubborn Agents

Consider a sequence of growing societies $\mathcal{S}_n$ each with n agents. Agent n is born in society $\mathcal{S}_n$ with sophistication type type$(n) \in \{B, D\}$, signal intensity $\lambda_n \in \mathbb{R}_+$, and signal strength $p_n \in (1/2, 1)$, which never changes. We make the following assumption:

Assumption 1. Consider the vectors $\boldsymbol{\lambda}_n \equiv \{\lambda_i\}_{i=1}^n$, $\mathbf{p}_n \equiv \{p_i\}_{i=1}^n$. Then, $\mathbf{p}_n, \boldsymbol{\lambda}_n$ satisfy Assumption 2 from Mostagir et al. (2019) for all n.

Recall that our average network consists of a set of "average" neighborhoods for the agents. For DeGroot agents, who simply perform linear aggregations of their neighbors's beliefs, it is immediate how to define the "average" neighborhood. But because Bayesians update according to Bayes' rule, and Bayesians know which links are realized in their neighborhood, we cannot simply take the Bayesian's belief update as a (linear) expectation of the updates across these realizations. Therefore, we require additional conditions. For this, we make the additional assumption:

Assumption 2. The realized network is connected almost surely.[23] Moreover, the weights α_{ij}

[23] By "connected" we mean there is a path between any two agents i, j, where the "links" are given by positive DeGroot weights and Bayesian neighborhoods.

between any two agents i, j in the realized network is perturbed by some random ϵ_{ij} from a continuous distribution over finite support $F(\cdot)$, with $\lim_{n\to\infty} \epsilon_{ij} \overset{a.s.}{\to} 0$ for all agents i, j.

The first condition is guaranteed in both the weak and strong inequality models as $n \to \infty$. The second condition, while more stringent, considers small perturbations to the network weights to guarantee genericity. Without it, there are some special cases where a Bayesian agent might have an identification problem when observing the beliefs of her neighbors. This is discussed in more detail in Mostagir et al. (2019). Under both of these assumptions, Bayesian agents may be treated as stubborn agents as the learning horizon $T \to \infty$.

A.3 Model Adaptation to Stochastic-Block Networks

In the original model of Mostagir et al. (2019), the authors derived results for arbitrary (but deterministic) network structures, where the principal must make (essentially) a binary decision for each agent whether to send her misinformed signals. In this paper, we instead adapt this model for random social networks which embed a notion of inequality in the form of unequal access to educational resources. At the core of the random network process are "islands" (or communities) where agents within an island have similar resources, but can have different resources from agents on different islands. Thus, the principal's optimal strategy instead is more subtle: he can target a fraction of the population on a given island (a rational number between 0 and 1). Because of symmetry, the principal does not care which agents on the island he actually targets, just the total percentage. Recall agents put θ weight on their own personal experience. This implies that, in the standard model, there will be two belief types within a given island, depending on whether the principal targets the agent directly or not.

To avoid this, and to add parsimony to the model, we assume agents will perform the personal-experience Bayesian update using the "average" news sent to the island. In particular, if the principal targets κ_ℓ fraction of the population on an island, the Bayesian update part of the belief update converges to $1 - \kappa_\ell$ as $T \to \infty$. This implies that all agents' beliefs on the same island will be the same, and allows us to study manipulation in the context of how certain communities are affected versus others.

Note this adaptation does not change much from the standard model. For instance, if θ is not too large (and agents rely significantly on social learning), then the differences in beliefs of two agents on the same island will be small in the standard setup. Thus, our results generalize easily to the case of the standard model as well, where agents are treated as individual binary decisions for the principal.

Finally, we note that all results implicitly assume that $n \to \infty$, because only under these conditions does Theorem 1 equate the manipulation in random networks with that of their expected counterparts. Thus, while Bayesian counts are technically discrete objects, because all of our results are closed under multiplication of the Bayesians and DeGroot populations on each island by the same constant, we can think of "Bayesian proportions" on each island and need not worry about whether such proportions divide the population size without remainder.

A.4 DeGroot Centrality: General Methods

Note that given a fixed strategy for the principal, the beliefs of (DeGroot) agents within a given island are the same due to symmetry as $n \to \infty$.[24] Here, we will introduce the general methodology for determining whether a population (or community) whose structure is randomly drawn

[24] See Appendix A.3.

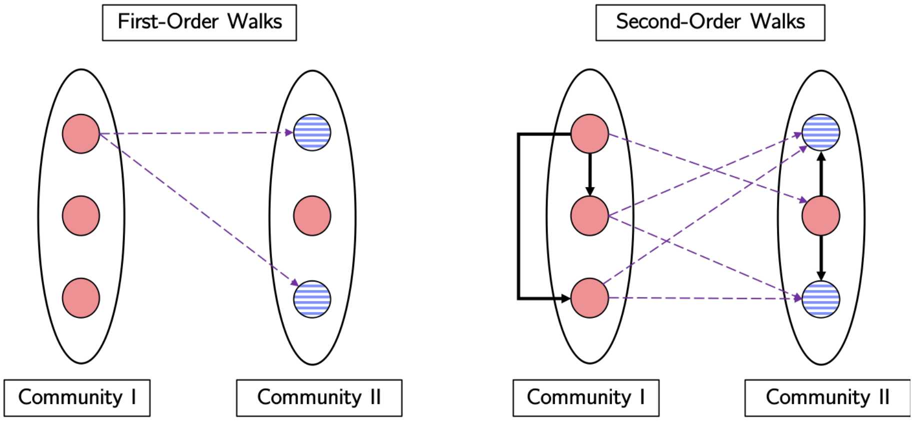


Figure 15. An illustration of computing weighted walks to Bayesian agents, for agent 1. Solid circles are DeGroot agents and shaded circles are Bayesians. Solid lines represent higher weights "within-community links" than dashed lines. Consider the top-left agent, and for each walk, multiply the weights of the links along the walk. The figure on the left shows a first-order walk, i.e. a walk of length 1, which consists of the link directly connecting that agent to a Bayesian agent. The second-order walk displayed on the right consists of walks of length 2, so that there is a link to another DeGroot agent who is linked to a Bayesian agent, and the weight of that walk is the product of the two link weights and so on. Total weighted walks is the sum over all orders (i.e., walk lengths) of walks $1, 2, \ldots$.

from either the strong or weak homophily model is susceptible to manipulation. Let us define some notation:

(i) κ_ℓ is the fraction of island ℓ targeted by the principal (note that who he specifically targets on the island is immaterial);

(ii) $\mathcal{N}_\ell$ is the "neighborhood" of island ℓ; in the weak homophily model it is equal to $\{\ell' | \ell' \neq \ell\}$ whereas in the strong homophily model it is equal to:

$$\begin{cases} \{2\}, \text{ if } \ell = 1 \\ \{\ell-1, \ell+1\}, \text{ if } \ell \in \{2, \ldots, k-1\} \\ \{\ell-1\}, \text{ if } \ell = k \end{cases}$$

One can compute DeGroot centralities by simply counting the weighted walks to misinformed agents, as in Figure 15. However, DeGroot centrality computations are easiest when considering the linear recursive formulation of weighted walks, as in Figure 16.

This calculation is done in two parts. The first part involves computing weighted walks to Bayesian agents from agent i living on some island ℓ, which we denote as w_ℓ^B. The second part involves computing weighted walks to DeGroots who *do not directly* consume misinformation from the principal, which we denote as w_ℓ^D. The belief of the agent on island ℓ is then given by $w_\ell = w_\ell^B + w_\ell^D$.

We can calculate this explicitly as:

$$\frac{w_\ell^B}{1-\theta} = \frac{p_s m_\ell + \sum_{\ell' \in \mathcal{N}_\ell} p_d m_{\ell'}}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} + (1-\theta) \frac{p_s(s_\ell n - m_\ell)}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} w_\ell^B + (1-\theta) \frac{\sum_{\ell' \in \mathcal{N}_\ell} p_d(s_{\ell'} n - m_{\ell'})}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} w_{\ell'}^B$$

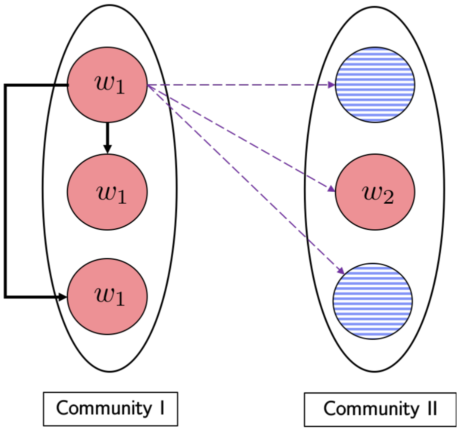


Figure 16. An analytic approach to computing Bayesian walks for the top-left agent (which equals her belief of the true state). Each agent's sum of walks equals a weighted-average of her neighbors' sums of walks.

Similarly, we have:

$$\frac{w_\ell^D}{1-\theta} = \theta\frac{1-\kappa_\ell}{1-\theta} + \theta\frac{p_s(1-\kappa_\ell)(s_\ell n - m_\ell) + \sum_{\ell'\in\mathcal{N}_\ell} p_d(1-\kappa_{\ell'})(s_{\ell'}n - m_{\ell'})}{p_s s_\ell n + \sum_{\ell'\in\mathcal{N}_\ell} p_d s_{\ell'} n}$$
$$+(1-\theta)\frac{p_s(1-\kappa_\ell)(s_\ell n - m_\ell)}{p_s s_\ell n + \sum_{\ell'\in\mathcal{N}_\ell} p_d s_{\ell'} n} w_\ell^D + (1-\theta)\frac{\sum_{\ell'\in\mathcal{N}_\ell} p_d(1-\kappa_{\ell'})(s_{\ell'}n - m_{\ell'})}{p_s s_\ell n + \sum_{\ell'\in\mathcal{N}_\ell} p_d s_{\ell'} n} w_{\ell'}^D$$

In particular, belief w_ℓ^B and w_ℓ^D both admit linear matrix equations with a closed-form solutions:

$$\frac{\mathbf{I}\mathbf{w}^B}{1-\theta} = \boldsymbol{a}^B + \boldsymbol{B}^B\mathbf{w} \Longrightarrow \mathbf{w}^B = \left(\frac{\mathbf{I}}{1-\theta} - \boldsymbol{B}^B\right)^{-1}\boldsymbol{a}^B$$
$$\frac{\mathbf{I}\mathbf{w}^D}{1-\theta} = \boldsymbol{a}^D + \boldsymbol{B}^D\mathbf{w} \Longrightarrow \mathbf{w}^D = \left(\frac{\mathbf{I}}{1-\theta} - \boldsymbol{B}^D\right)^{-1}\boldsymbol{a}^D$$

where the total belief of the correct state is $w_\ell = w_\ell^B + w_\ell^D$, which is the complement of agent i on island ℓ's DeGroot centrality (i.e., $\mathcal{D}_\ell = 1 - w_\ell$).

B Proofs

Preliminaries The following notation is used throughout the proofs. The vector $\gamma \in \{0,1\}^n$ denotes which agents are targeted by the principal, and the DeGroot Centrality vector resulting from this targeting is denoted by $\mathcal{D}(\gamma)$. DeGroot agent on island ℓ is manipulated if $\mathcal{D}_\ell(\gamma) > (1-b)/2$. Equivalently, we write π_ℓ or w_ℓ as the belief of an agent on island ℓ (the latter explicitly referring to walk counting, but is equivalent to π_ℓ).

B.1 Section 2

Proof of Theorem 1. Note that Assumption 1 from (redacted to preserve anonymity – please see Assumption 1 in attached technical note) holds because both the strong and weak inequality models are drawn from an inhomogenous Erdos-Renyi model. Assumption 2 from (attached technical note) is also satisfied because θ is constant and the weak and strong inequality models are connected almost surely. Similarly, the expected degrees condition is satisfied because the expected degrees grow linearly in n in both the weak and strong inequality models. Moreover, the normal society condition holds trivially because θ is the same for all agents. Therefore, we can apply Theorem 1 from (attached technical note) for DeGroot centrality on an arbitrary γ, i.e., $\lim_{n\to\infty}\mathbb{P}\left[||\tilde{\mathcal{D}}^{(n)}(\gamma)-\bar{\mathcal{D}}^{(n)}(\gamma)||_\infty > \epsilon\right]=0$. Thus, as $n\to\infty$:

$$\lim_{n\to\infty}\left(\mathbb{P}\left[\tilde{\mathcal{D}}_i^{(n)} < (1-b)/2 < \bar{\mathcal{D}}_i^{(n)} \text{ for some } i\right]+\mathbb{P}\left[\bar{\mathcal{D}}_i^{(n)} < (1-b)/2 < \tilde{\mathcal{D}}_i^{(n)} \text{ for some } i\right]\right)=0$$

except for countably many b. Thus, for generic b, the number of manipulated agents is the same under both the expected and realized networks in the weak and strong inequality models, as $n\to\infty$. □

B.2 Section 4

Proof of Theorem 2. Notice the network is susceptible to manipulation (with high probability) if there exists an agent j with $\mathcal{D}_j(\mathbf{1}_D) > (1-b)/2$; similarly, the network is impervious to manipulation (with high probability) if for all agents j, $\mathcal{D}_j(\mathbf{1}_D) < (1-b)/2$. We show that increasing homophily leads to an increase in the inequality of DeGroot centralities (i.e., $\mathcal{D}(\mathbf{1}_D)$). We have the system of equations:

$$\begin{aligned}\frac{1}{1-\theta}\mathbf{w} =& \frac{k}{n(p_s+(k-1)p_d)}\begin{pmatrix} p_s m_1 + p_d\sum_{\ell\neq 1} m_\ell \\ p_s m_2 + p_d\sum_{\ell\neq 2} m_\ell \\ \cdots \\ p_s m_k + p_d\sum_{\ell\neq k} m_\ell \end{pmatrix} \\ &+\frac{k(1-\theta)}{n(p_s+(k-1)p_d)}\begin{pmatrix} p_s(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ p_d(n/k-m_1) & p_s(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ \cdots & \cdots & \cdots & \cdots \\ p_d(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_s(n/k-m_k) \end{pmatrix}\mathbf{w}\end{aligned}$$

which is equivalent to:

$$\begin{aligned}\frac{k}{(1-\theta)n(p_s+(k-1)p_d)}\mathbf{w} =& \begin{pmatrix} p_s m_1 + p_d\sum_{\ell\neq 1} m_\ell \\ p_s m_2 + p_d\sum_{\ell\neq 2} m_\ell \\ \cdots \\ p_s m_k + p_d\sum_{\ell\neq k} m_\ell \end{pmatrix} \\ &+(1-\theta)\begin{pmatrix} p_s(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ p_d(n/k-m_1) & p_s(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ \cdots & \cdots & \cdots & \cdots \\ p_d(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_s(n/k-m_k) \end{pmatrix}\mathbf{w}\end{aligned}$$

Without loss of generality suppose that island k has the least Bayesian agents of any island. Consider the map T given by:

$$T : \mathbf{w} \mapsto \begin{pmatrix} p_s m_1 + p_d \sum_{\ell \neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell \neq 2} m_\ell \\ \cdots \\ p_s m_k + p_d \sum_{\ell \neq k} m_\ell \end{pmatrix} + (1-\theta) \begin{pmatrix} p_s(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ p_d(n/k - m_1) & p_s(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ \cdots & \cdots & \cdots & \cdots \\ p_d(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_s(n/k - m_k) \end{pmatrix} \mathbf{w}$$

We claim that T has the property that $(w_\ell \geq w_k) \implies T(w_\ell) \geq T(w_k)$. Suppose that $w_\ell \geq w_k$, then:

$$\begin{aligned} m_\ell + (1-\theta) w_\ell (n/k - m_\ell) &\geq m_k + (1-\theta) w_\ell (n/k - m_k) \\ &\geq m_k + (1-\theta) w_k (n/k - m_k) \end{aligned}$$

which moreover implies that

$$\begin{aligned} & p_s(m_\ell + (1-\theta) w_\ell (n/k - m_\ell)) + p_d(m_k + (1-\theta) w_k (n/k - m_k)) + \sum_{\ell' \neq \ell, k} p_d(m_{\ell'} + (1-\theta) w_{\ell'}(n/k - m_{\ell'})) \\ \geq & p_d(m_\ell + (1-\theta) w_\ell (n/k - m_\ell)) + p_s(m_k + (1-\theta) w_k (n/k - m_k)) + \sum_{\ell' \neq \ell, k} p_d(m_{\ell'} + (1-\theta) w_{\ell'}(n/k - m_{\ell'})) \end{aligned}$$

because $p_s > p_d$. Because p_s, p_d, n are fixed, this suggests the map $\frac{k}{(1-\theta)n(p_s + (k-1)p_d)} \cdot T$ also has this property, so any fixed point of T must have $w_\ell \geq w_k$ by Brouwer's fixed point theorem for all islands ℓ. Since the system is linear and non-singular, there is a unique fixed-point with $w_\ell \geq w_k$. This implies the DeGroot centrality of the island with the least Bayesians is always maximal and determines whether the network is impervious.

For the remainder of this part of the proof, we define a new operator T which maps $\mathbf{w}$, parametrized by p_s, p_d, and $\mathbf{m}$, respectively. We show the following: (i) $T|p_s$ is decreasing in p_s for w_k, (ii) $T|p_d$ is increasing in p_d for w_k, and (iii) $T|\mathbf{m}$ subject to $\sum m_\ell = m$ is increasing with every "Robin Hood" operation that puts more Bayesians on island k for w_k.[25] This result suffices in order to show that there exists a fixed-point of T which obeys the desired properties of Theorem 2 (by Brouwer[26]), and by linearity, this fixed-point is unique.

1. Decreasing in p_s: Let us define T as:

$$\begin{aligned} T|p_s : \mathbf{w} \mapsto & \frac{1}{p_s + p_d} \Bigg[\begin{pmatrix} p_s m_1 + p_d \sum_{\ell \neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell \neq 2} m_\ell \\ \cdots \\ p_s m_k + p_d \sum_{\ell \neq k} m_\ell \end{pmatrix} \\ & + (1-\theta) \begin{pmatrix} p_s(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ p_d(n/k - m_1) & p_s(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ \cdots & \cdots & \cdots & \cdots \\ p_d(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_s(n/k - m_k) \end{pmatrix} \mathbf{w} \Bigg] \end{aligned}$$

[25] Note that other "Robin Hood" operations do not affect w_k, so does not affect the imperviousness of the network.

[26] In particular, let (w_1, w_2) be the old fixed-point and (w_1', w_2') the new fixed point. We illustrate for the case of increasing p_s: all other cases are similar. By increasing p_s, we know that T maps all w_1 larger and all w_2 smaller. Therefore, the convex compact set $[w_1, 1] \times [0, w_2]$ maps into itself, which implies the new fixed-point (w_1', w_2') lies in this set.

Computing directly:

$$\frac{\partial T(w_k|p_s)}{\partial p_s} = p_d \frac{(m_k + (1-\theta)(n/k - m_k)w_k) - \sum_{\ell \neq k}(m_\ell + (1-\theta)(n/k - m_\ell)w_\ell)}{(p_s + p_d)^2} < 0$$

where the inequalities follow from the analysis above.

2. Increasing p_d: Let us define T in the same way as in (1), except parametrized by p_d. Then in exactly the same way:

$$\frac{\partial T(w_k|p_d)}{\partial p_d} = -\frac{p_s}{p_d}\frac{\partial T(w_k|p_s)}{\partial p_s} > 0$$

which is the desired result.

3. Majorization: Assume that we remove a Bayesian from island ℓ^* and add it to island k, with the assumption that $m_k + 1 \leq m_{\ell^*} - 1$. There are two cases: (i) island k still has the fewest Bayesians (and thus the greatest DeGroot centrality), or (ii) some other island had the exact same number of Bayesians as island k. In the latter case, the majorization does not affect whether the network is impervious or susceptible. In the former case, define T as:

$$T|\mathbf{m} : \mathbf{w} \mapsto \Bigg[\begin{pmatrix} p_s m_1 + p_d \sum_{\ell \neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell \neq 2} m_\ell \\ \cdots \\ p_s m_k + p_d \sum_{\ell \neq k} m_\ell \end{pmatrix}$$
$$+ (1-\theta) \begin{pmatrix} p_s(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ p_d(n/k - m_1) & p_s(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ \cdots & \cdots & \cdots & \cdots \\ p_d(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_s(n/k - m_k) \end{pmatrix} \mathbf{w} \Bigg]$$

Computing the directional derivative along the gradient $\mathbf{u} = \mathbf{e}_k - \mathbf{e}_{\ell^*}$ (where $\mathbf{e}_i$ is the vector of all 0's except for a 1 in ith spot):

$$\frac{\partial T(w_k)|\mathbf{m}}{\partial m_k} - \frac{\partial T(w_k)|\mathbf{m}}{\partial m_{\ell^*}} = p_d\,(1 - (1-\theta)w_k) - p_s\,(1 - (1-\theta)w_{\ell^*})$$

Note that because $w_k \leq w_{\ell^*}$, $(1 - (1-\theta)w_k) \geq (1 - (1-\theta)w_{\ell^*})$. Because $p_s > p_d$, the above expression is positive.

Lastly, we need to argue that in the case of islands of equal size, increased DeGroot centrality inequality (i.e., the island with larger centrality increases its centrality while the other's centrality decreases) cannot make the network go from susceptible to impervious. To check if the network is impervious, all that needs to be checked is $\max_i \mathcal{D}_i(\mathbf{1}_D) > (1-b)/2$. When inequality of the DeGroot centrality increases, then $\max_i \mathcal{D}_i(\mathbf{1}_D)$ increases, and so the above inequality is more likely to be satisfied when inequality is increased. Therefore, the network can go from impervious to susceptible, but not the other direction. □

Proof of Theorem 3. Part (i) is a direct implication of Theorem 2: it is impossible for an increase in inequality to make the network switch from susceptible to impervious. Thus, if some inequality configuration makes the network impervious, it must necessarily be the case that the network with the least inequality is impervious.

For parts (ii) and (iii), consider the following construction of the inequality structures. If for all choices of $(\mathbf{m}, p_d)$ the DeGroot centralities of all of the islands are monotone in p_s (either monotonically increasing or decreasing) then choose p_s sufficiently close to 1 such that manipulation is the same under this inequality structure and $p_s = 1$ (i.e., extreme homophily). Such a $p_s < 1$ is guaranteed because centralities are continuous in the inequality parameters. Otherwise, there exists $(\mathbf{m}, p_d)$ such that at least one island has non-monotone centrality in p_s. First, note that all centralities are concave in p_s; this can be seen from considering the map in the proof of Theorem 2 for p_s and noting that:

$$\frac{\partial^2 T(w_\ell|p_s)}{\partial p_s^2} = \left(m_\ell + (1-\theta)(n/k - m_\ell)w_\ell - \sum_{\ell \neq k} (m_\ell + (1-\theta)(n/k - m_\ell)w_\ell) \right) \cdot \frac{p_s^2 - p_d^2}{(p_s + p_d)^4} > 0$$

because both terms in the above expression are positive. (Recall that w_ℓ is equal to 1 minus centrality, so convexity of w_ℓ corresponds to concavity of centrality.) Second, note that the centrality curve of an island with more Bayesians always lies above an island with fewer (this was shown in Theorem 2). Suppose some island that exhibits the non-monotonicity of centrality in p_s; if there are multiple, pick the island whose centrality apex occurs at the largest value for p_s (call this is the "special" island); call this value p_s^*. If we choose b so that $(1-b)/2$ lies just below the apex of the centrality curve, then this island will be manipulated for p_s^*, but not at $p_s = 1$ because the centrality curve for the special island is concave (and thus is decreasing after p_s^*). All islands with more Bayesians than the special island are protected when there is the most inequality, and all islands with fewer Bayesians than the special island are manipulated under p_s^*. Similarly, every other island has either: (i) monotonically decreasing centrality, (ii) monotonically increasing centrality, or (iii) non-monotone centrality. In the case of (i) and (iii), by assumption, the island cannot be manipulated at $p_s = 1$ but is at p_s^*. Moreover for islands of type (ii), the centrality must naturally lie above the centrality curve of the special island, so is manipulated at p_s^*. Thus, this intermediate inequality structure has strictly more manipulation than the most inequality.

It just remains to show that for every $k \geq 3$, there exists a choice of m and distribution $\mathbf{m}$ that has at least one non-monotone centrality curve. For this, we provide an explicit example for $k = 3$. The parameters are given by $\theta = 1/20$, $p_d = .2$, $m_1 = 44$, $m_2 = 29$, $m_3 = 0$, $n = 1000$, for three islands. The plot is given in Figure 17.

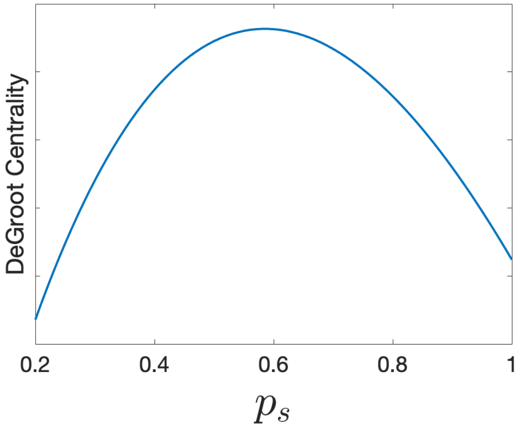


Figure 17. DeGroot centrality for island 2 as a function of p_s.

To generalize to more communities, simply add a community with all DeGroot agents. This will make the centrality of the special island increase for intermediate inequality relative to most inequality because the centrality of the all DeGroot island will exceed that of the special island. Thus, there will still be manipulation with intermediate inequality, but not with the most inequality. □

Proof of Theorem 4. As in the proof of Theorem 2, let us consider each of the inequality cases separately and define corresponding maps T. We show that *all* beliefs decrease following an increase in p_s, decrease in p_d, or a reverse "Robin Hood" operation.

1. $\underline{p_s}$: Let us define T as:

$$T|p_s : \mathbf{w} \mapsto \left[\begin{pmatrix} \frac{p_s m_1 + p_d \sum_{\ell \neq 1} m_\ell}{p_s s_1 + (1-s_1) p_d} \\ \frac{p_s m_2 + p_d \sum_{\ell \neq 2} m_\ell}{p_s s_2 + (1-s_2) p_d} \\ \cdots \\ \frac{p_s m_k + p_d \sum_{\ell \neq k} m_\ell}{p_s s_k + (1-s_k) p_d} \end{pmatrix} + (1-\theta) \begin{pmatrix} \frac{p_s(s_1 n - m_1)}{p_s s_1 + (1-s_1) p_d} & \frac{p_d(s_2 n - m_2)}{p_s s_1 + (1-s_1) p_d} & \cdots & \frac{p_d(s_k n - m_k)}{p_s s_1 + (1-s_1) p_d} \\ \frac{p_d(s_1 n - m_1)}{p_s s_2 + (1-s_2) p_d} & \frac{p_s(s_2 n - m_2)}{p_s s_2 + (1-s_2) p_d} & \cdots & \frac{p_d(s_k n - m_k)}{p_s s_2 + (1-s_2) p_d} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{p_d(s_1 n - m_1)}{p_s s_k + (1-s_k) p_d} & \frac{p_d(s_2 n - m_2)}{p_s s_k + (1-s_k) p_d} & \cdots & \frac{p_s(s_k n - m_k)}{p_s s_k + (1-s_k) p_d} \end{pmatrix} \mathbf{w} \right]$$

Computing directly for island 1:

$$\begin{aligned} \frac{\partial T(w_1|p_s)}{\partial p_s} &= p_d \frac{(1-s_1)\left(m_1 + (1-\theta)(ns_1 - m_1)w_1\right) - s_1 \sum_{\ell \neq 1} (m_\ell + (1-\theta)(ns_\ell - m_\ell)w_\ell)}{(p_s s_1 + p_d(1-s_1))^2} \\ &= \frac{1}{(p_s s_1 + p_d(1-s_1))^2} \left(\frac{m_1}{s_1} + (1-\theta)(n - m_1/s_1)w_1 - \frac{\sum_{\ell \neq 1} m_\ell}{1-s_1} - (1-\theta) \sum_{\ell \neq 1} \frac{ns_\ell - m_\ell}{1-s_1} w_\ell \right) \end{aligned}$$

By assumption, $m_1/s_1 < \sum_{\ell \neq 1} m_\ell/(1-s_1)$ and $w_1 \leq w_\ell$, thus,

$$\begin{aligned} \frac{\partial T(w_1|p_s)}{\partial p_s} &< \frac{w_1(1-\theta)}{(p_s s_1 + p_d(1-s_1))^2} \left(n - m_1/s_1 - \sum_{\ell \neq 1} \frac{ns_\ell - m_\ell}{1-s_1} \right) \\ &= \frac{w_1(1-\theta)}{(p_s s_1 + p_d(1-s_1))^2} \left(n - m_1/s_1 - n + \frac{m - m_1}{1-s_1} \right) \\ &= \frac{w_1(1-\theta)}{(p_s s_1 + p_d(1-s_1))^2} \frac{s_1 m - m_1}{s_1(1-s_1)} < 0 \end{aligned}$$

Thus, the beliefs of the agents on island 1 decrease following an increase in p_s. Then observe that for other islands $\ell \neq 1$:

$$\left(\frac{1}{1-\theta} - (1-\theta) \frac{p_s(s_\ell n - m_\ell)}{n(p_s s_\ell + p_d(1-s_\ell))} \right) w_\ell = \frac{p_s m_\ell + p_d \sum_{\ell' \neq \ell} m_{\ell'}}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta) \sum_{\ell' \neq \ell} \frac{p_d(s_{\ell'} n - m_{\ell'})}{n(p_s s_\ell + p_d(1-s_\ell))} w_{\ell'}$$

when s_1 is sufficiently close to 1, the above simplifies to:

$$\frac{1}{1-\theta} w_\ell = \frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta) \frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))} w_1$$

Both $\frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))}$ and $\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))}$ are decreasing in p_s, and we just showed that w_1 is decreasing in p_s. Thus belief of island ℓ is also decreasing in p_s.

2. $\underline{p_d}$: Recall that

$$\frac{\partial T(w_1|p_d)}{\partial p_d} = -\frac{p_s}{p_d}\frac{\partial T(w_1|p_s)}{\partial p_s} > 0$$

   And the expression for w_ℓ when s_1 is sufficiently close to 1 is:

$$\frac{1}{1-\theta}w_\ell = \frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))}w_1$$

   Both $\frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))}$ and $\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))}$ are increasing in p_d, and we showed prior that w_1 is increasing in p_d. Thus belief of island ℓ is also increasing in p_d.

3. Majorization: We consider a "Robin Hood" operation that adds a Bayesian to the large island. Once again we compute the directional derivative along the gradient $\mathbf{u} = \mathbf{e}_1 - \mathbf{e}_{\ell^*}$ for some island ℓ^*:

$$\frac{\partial T(w_1|\mathbf{m})}{\partial m_1} - \frac{\partial T(w_1|\mathbf{m})}{\partial m_{\ell^*}} = \frac{p_s(1-(1-\theta)w_1) - p_d(1-(1-\theta)w_\ell)}{p_s s_1 + (1-s_1)p_d} > 0$$

   because $w_1 \leq w_\ell$ and $p_s \geq p_d$. Similarly, when s_1 is sufficiently close to 1:

$$\frac{1}{1-\theta}w_\ell = \frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))}w_1$$

   which is increasing in m_1 given that w_1 is increasing in m_1.

   Note that when s_1 is sufficiently large, a reverse "Robin Hood" operation that does not impact island 1 will have little effect on the beliefs of the islands. Thus, provided that no island's centrality lies directly at $(1-b)/2$ (which holds for b on a set of full measure), this operation will have no impact on which agents are manipulated.

□

B.3 Section 5

Algorithm. Let the marginalized island be the island with fewer Bayesians and the privileged island be the island with more Bayesians (if the number of Bayesians is equal, label them arbitrarily). Moreover, let κ_m, κ_p be the proportion of agents targeted on the marginalized and privileged islands, respectively. The principal's optimal strategy can be computed as follows:

(i) Consider the strategy where the principal targets all agents.

   (a) If neither island is manipulated, the principal's optimal strategy is $\mathbf{x} = \mathbf{0}$.

   (b) If both islands are manipulated, then decrease κ_m until the belief on the marginalized island matches that of the privileged island or $\kappa_m = 0$. Then decrease κ_m and κ_p one-for-one,[27] (if $\kappa_m = 0$, just decrease κ_p), until both (identical) beliefs fall below the cutoff. (Note that if $\kappa_p = \kappa_m = 0$, no island will be manipulated, so such κ_p, κ_m always exist.) Record the payoff $1 - \frac{\kappa_p + \kappa_m}{2}\varepsilon$.

[27] Formally, "one-for-one" here means decrease them simultaneously so that the beliefs of both island remain identical as these decrease, which may not necessarily correspond to the same change for κ_m as κ_p to achieve this.

(c) If just the marginalized island is manipulated, decrease κ_p until the marginalized island's belief is below π^*. If $\kappa_p = 0$ still results in the marginalized island's manipulation, begin decreasing κ_m until this island is no longer manipulated. If the payoff of $\frac{1}{2} - \frac{\kappa_p + \kappa_m}{2}\varepsilon > 0$, this is the principal's optimal strategy.

(ii) If case (i)(b) holds, consider the strategy where the principal targets no one and no island is manipulated. The principal then increases κ_m until the marginalized island is manipulated; if the marginalized island is not manipulated at $\kappa_m = 1$, then the principal begins increasing κ_p until the marginalized island is manipulated. (If $\kappa_p = \kappa_m = 1$ and the marginalized island is still not manipulated, then default to case (i)(a).) Record the payoff $\frac{1}{2} - \frac{\kappa_p + \kappa_m}{2}\varepsilon$.

(iii) Compare the recorded payoffs and the payoff of 0. If the largest payoff is 0, the principal's optimal strategy is $\mathbf{x} = \mathbf{0}$. Otherwise, the principal should employ the strategy (either (i)(b) or (ii)) corresponding to the largest payoff.

Proposition 5. *The aforementioned algorithm is correct, i.e., it provides the principal's optimal strategy for any ε.*

Proof of Proposition 5. The principal first decides the cheapest way to manipulate 0, 1, and 2 islands, respectively, and then compares the payoffs and chooses whichever is maximal. The cheapest way to manipulate no islands is $\mathbf{x} = \mathbf{0}$, which is always the recommended outcome of the algorithm for 0 islands.

To manipulate one island, it is always cheaper to manipulate the marginalized island. Note w_m^B does not depend on κ_m or κ_p, and that $\partial w_m^D / \partial \kappa_m > \partial w_m^D / \partial \kappa_p$. Thus, the optimal way to manipulate the marginalized island is to decrease κ_p as much as possible, leaving $\kappa_m = 1$, until the marginalized island is no longer manipulated. If it is still manipulated at $\kappa_p = 0$, then the only more cost effective strategy would be to decrease κ_m until you lose this island. This is precisely the strategy identified in (i)(c) and (ii). In the case of (i)(c), we know this to be the optimal strategy if it beats manipulating no one, because it is impossible to manipulate both islands.

To manipulate two islands, we show that it is always optimal for either: (1) $\kappa_m = 0$ and κ_p is chosen smallest to manipulate both islands, or (2) both islands have the same belief at the optimal strategy (right at the cutoff $(1+b)/2$). Recall that $\partial w_\ell^D / \partial \kappa_\ell > \partial w_\ell^D / \partial \kappa_{\ell'}$ for both islands, where $\ell' \neq \ell$. The privileged island will have beliefs closer to the truth when $\kappa_m = \kappa_p$, it must be the case that $\kappa_p \geq \kappa_m$ in the optimal strategy for 2-island manipulation. Of these strategies, (1) is clearly then optimal provided that $\kappa_m = 0$ does not cause the marginalized island's belief to lie above the cutoff (and thus, not manipulate both islands). Otherwise, if some island ℓ's belief is strictly above the cutoff, then one can decrease κ_ℓ a small amount and increase $\kappa_{\ell'}$ for $\ell' \neq \ell$ by a smaller amount, again, because $\partial w_\ell^D / \partial \kappa_\ell > \partial w_\ell^D / \partial \kappa_{\ell'}$ for both islands ℓ. This is cheaper than the previous strategy, a contradiction. So (2) must be optimal.

Finally, step (iii) checks whether 0-island, 1-island, or 2-island manipulation is the most profitable. □

Proof of Theorem 5. The beliefs of all agents will be identical in the network with the least inequality. Let $m(b)$ be the maximum number of Bayesians such that if the Bayesians are all distributed evenly across the islands, every agent is manipulated if the principal targets every DeGroot, which is a function of b. Moreover, because of symmetry, it is clear that manipulating every agent or manipulating no agent is the principal's optimal strategy, and in particular for $\varepsilon < 1$, manipulating every agent is a profitable strategy. (Note that this does not imply that $\gamma = \mathbf{1}_D$ is optimal, just that manipulating every agent is optimal.) When we move to an inequality configuration with the most inequality, there are two cases: (1) $m(b) \leq n/k - 1$ or (2) $m(b) \geq n/k$. In case (1), we stack all of the Bayesians on a single island and set $p_s = 1$ and $p_d = 0$, noting that there is at least one DeGroot on this island. It is clear that this means the island with all the Bayesians will have a decrease in their DeGroot centrality, which by definition of $m(b)$, will protect this island from manipulation. At the same time, this configuration is still susceptible to manipulation, because the islands with all DeGroots agents will be manipulated given $\varepsilon < 1$. In case (2), we make island 1 contain all Bayesians and one DeGroot, and then distribute the remaining Bayesians equally amongst the rest of the islands. For sufficiently large n, this will always (strictly) decrease the centrality of the one DeGroot agents on the island with concentrated Bayesians, thereby protecting her from manipulation; at the same time, the DeGroots on the other islands will continue to be manipulated, as their centrality does not decrease (and may increase).

Finally, we show there exists a model with intermediate inequality that is impervious for some open interval of $\varepsilon \in (\varepsilon^*, \varepsilon^{**})$ and $b \in (b^*, b^{**})$. Once again there are two cases: (1) $m(b) \leq n/k - 1$ and (2) $m(b) \geq n/k$. In the former case, put all of the Bayesians on island 1 along with one DeGroot, as before. In the latter case, put $n/k - 1$ Bayesians on island 1 along with one DeGroot, and then distribute the rest of the Bayesians evenly amongst the remaining islands. If $0 < p_d < p_s < 1$, then we have a model of intermediate inequality where the beliefs (of the correct state) of the agents on island 1 exceed those on the other islands (which are identical because of symmetry).[28] Because the beliefs of the agents on island 1 exceed that of other islands, we know the principal cannot manipulate the DeGroot on island 1, even if he were to target every DeGroot in the population. Next, we show that for any $\delta < 1$, there exists some p_d and $b > -1$ such that the principal needs to target at least δ proportion of the DeGroots on island 1 *and* at least δ proportion of the DeGroots on the rest of the islands.[29] We know that $\pi_1 > \pi_\ell$ when the principal targets every DeGroot for all $\ell \neq 1$ given that $p_d < p_s$. Thus, we can always choose b such that $(1+b)/2$ is arbitrarily close to π_ℓ but still satisfies $\pi_1 > (1+b)/2 > \pi_\ell$. Given $p_d > 0$, any (substantial) deviation from $\boldsymbol{\gamma} = \mathbf{1}_D$, in the ∞-norm, leads to some island $\ell \neq 1$ not being manipulated. Because the principal should always enact a symmetric strategy with respect to all islands $\ell \neq 1$, we then either have (i) the network is impervious, or (ii) the principal should send signals to at least δ proportion of each island, where δ can be arbitrarily close to 1. In the latter case, the payoff of the principal is no more than $n\left(\frac{k-1}{k}\right) - m(b) - (n - m(b))\delta\varepsilon$, again, for δ arbitrarily close to 1. Thus, there exists $\varepsilon < 1$ such that the principal's payoff from this strategy is negative. Hence, the network is impervious with a model of intermediate inequality. □

B.4 Section 6

Proof of Proposition 1. Suppose there are n agents in the network, and there are m Bayesians. We denote by $w_{\ell \to r}$ the weighted walks from an agent on island ℓ to any Bayesian on island r. For

[28] This was shown in the proof of Theorem 2.

[29] Note that there is technically only one DeGroot on island 1, so targeting δ proportion of one DeGroot seems nonsensical. However, Appendix A.3 reconciles this: because the analysis is always closed under multiplication of each island by the same proportion of Bayesians and DeGroots, we can always expand the size of the DeGroot population such that δ proportion (assuming $\delta \in \mathcal{Q}$) of $\delta(n/k - m_1)$ is an integer.

$\ell \neq r$, we can write:

$$\begin{aligned}
w_{\ell\to r} &\geq (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + np_d(1-s_\ell))} + (1-\theta)^2\left(\frac{np_s s_\ell w_{\ell\to r} + p_d(ns_r - m_r)w_{r\to r} + np_d\sum_{\tau\neq r,\ell} s_\tau w_{\tau\to r}}{np_s s_\ell + np_d(1-s_\ell)}\right)\\
&\geq (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2\left(\frac{np_s s_\ell \underline{w}_r + p_d(ns_r - m_r)(\underline{w}_r + w_{r\to r} - \underline{w}_r) + np_d(1-s_\ell-s_r)\underline{w}_r}{np_s s_\ell + np_d(1-s_\ell)}\right)\\
&= (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2\left(\underline{w}_r + \frac{np_d s_r(w_{r\to r} - \underline{w}_r) - p_d m_r w_{r\to r}}{np_s s_\ell + np_d(1-s_\ell)}\right)\\
&= (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2\left(\frac{p_s s_\ell + p_d(1-s_\ell-s_r)}{p_s s_\ell + p_d(1-s_\ell)}\underline{w}_r + \frac{p_d(ns_r - m_r)}{n(p_s s_\ell + p_d(1-s_\ell))}w_{r\to r}\right)\\
&\geq (1-\theta)^2\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2\left(\frac{p_s s_\ell + p_d(1-s_\ell-s_r)}{p_s s_\ell + p_d(1-s_\ell)}\underline{w}_r + \frac{p_d(ns_r - m_r)}{n(p_s s_\ell + p_d(1-s_\ell))}w_{r\to r}\right)
\end{aligned}$$

where $\underline{w}_r = \min_{\tau\neq r} w_{\tau\to r}$. This implies that:

$$\underline{w}_r \geq (1-\theta)^2\frac{p_d m_r + p_d(ns_r - m_r)w_{r\to r}}{n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r}$$

Similarly,

$$\begin{aligned}
w_{r\to r} &= (1-\theta)\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2\frac{p_s(ns_r - m_r)w_r^r + p_d\sum_{\tau\neq r} s_\tau w_r^\tau}{np_s s_r + np_d(1-s_r)}\\
&\geq (1-\theta)\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2\left[\underline{w}_r + \frac{np_s s_r(w_{r\to r} - \underline{w}_r) - p_s m_r w_{r\to r}}{np_s s_r + np_d(1-s_r)}\right]\\
&= (1-\theta)\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2\left[\frac{p_d(1-s_r)}{p_s s_r + p_d(1-s_r)}\underline{w}_r + \frac{p_s(ns_r - m_r)}{n(p_s s_r + p_d(1-s_r))}w_{r\to r}\right]\\
&\geq (1-\theta)^2\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2\left[\frac{p_d(1-s_r)}{p_s s_r + p_d(1-s_r)}\underline{w}_r + \frac{p_s(ns_r - m_r)}{n(p_s s_r + p_d(1-s_r))}w_{r\to r}\right]
\end{aligned}$$

which moreover implies that

$$w_{r\to r} \geq (1-\theta)\frac{p_s m_r + (1-\theta)np_d(1-s_r)\underline{w}_r}{\theta p_s ns_r + np_d(1-s_r) + (1-\theta)p_s m_r}$$

Combining these two results we get:

$$\begin{aligned}
&\underline{w}_r \geq (1-\theta)^2\frac{p_d m_r + p_d(ns_r - m_r)(1-\theta)\frac{p_s m_r + (1-\theta)np_d(1-s_r)\underline{w}_r}{\theta p_s ns_r + np_d(1-s_r) + (1-\theta)p_s m_r}}{n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r}\\
\Longrightarrow &\left[n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r\right]\underline{w}_r\\
&\geq (1-\theta)^3\frac{p_d m_r + p_d(ns_r - m_r)\frac{p_s m_r + (1-\theta)np_d(1-s_r)\underline{w}_r}{\theta p_s ns_r + np_d(1-s_r) + (1-\theta)p_s m_r}}{n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r}\\
\Longrightarrow &\left[n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r\right]\underline{w}_r\\
&\geq (1-\theta)^3 p_d m_r + (1-\theta)^3 p_d(ns_r - m_r)\frac{p_s m_r}{\theta p_s ns_r + np_d(1-s_r) + (1-\theta)p_s m_r}\\
&+ (1-\theta)^4 p_d(ns_r - m_r)\frac{np_d(1-s_r)}{\theta p_s ns_r + np_d(1-s_r) + (1-\theta)p_s m_r}\underline{w}_r
\end{aligned}$$

Note that:

$$\left([n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r]\left[\theta p_s n s_r + n p_d(1-s_r) + (1-\theta)p_s m_r\right] - (1-\theta)^4 p_d^2 (n s_r - m_r)(1-s_r)\right)\underline{w}_r$$
$$\geq (1-\theta)^3 p_d m_r\left[\theta p_s n s_r + n p_d(1-s_r) + (1-\theta)p_s m_r\right] + (1-\theta)^3 p_d(n s_r - m_r)$$

Therefore, we can write $\underline{w}_r \geq N(n)/D(n)$, where:

$$N(n) \equiv (1-\theta)^3 p_d m_r\left[\theta p_s n s_r + n p_d(1-s_r) + (1-\theta)p_s m_r\right] + (1-\theta)^3 p_d(n s_r - m_r)$$
$$D(n) \equiv [n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r]\left[\theta p_s n s_r + n p_d(1-s_r) + (1-\theta)p_s m_r\right] - (1-\theta)^4 p_d^2 (n s_r - m_r)(1-s_r)$$

If there are $m = c_r n$ Bayesian agents on island r, as $n \to \infty$, we have that:

$$\underline{w}_r \geq \frac{(1-\theta)^3 p_d c_r}{\theta(p_s \bar{s} + p_d(1-\bar{s}))}$$

where $\bar{s} = \max_{\tau \in \{1,\ldots,k\}} s_\tau$. Thus, the network is impervious as long as $\underline{w}_r > (1+b)/2$. This moreover implies the network is impervious if $c_r \geq \frac{\theta(p_s\bar{s}+p_d(1-\bar{s}))(1+b)}{2(1-\theta)^3 p_d}$ for any island r. By the pigeonhole principle, there must be an island with at least c/k proportion of the population that is Bayesian. Thus taking $c = k\frac{\theta(p_s\bar{s}+p_d(1-\bar{s}))(1+b)}{2(1-\theta)^3 p_d}$ and applying the result for the island r which has the largest proportion of Bayesian agents, we see the network is impervious to manipulation. Moreover $c < 1$ provided that θ is not too large. □

Proof of Proposition 2. For each θ and c we construct a strong inequality model where all but $\bar{k}$ communities are manipulated. Put all cn Bayesians on the first island on the line topology and let all other islands contain only DeGroots and be the same size as each other. We assume that the principal attempts to manipulate the last $k - \bar{k}$ communities along the line. We compute $\mathcal{D}_\ell(\mathbf{1})$ for every island by counting Bayesian walks for every island; we denote these walks by w_ℓ which is equivalent to $1 - \mathcal{D}_\ell(\gamma)$. For island 2, we have the recursion:

$$w_2 = (1-\theta)\frac{p_d s_1}{p_d(s_1+s_3) + p_s s_2} + (1-\theta)^2\frac{p_s s_2 w_2 + p_d s_3 w_3}{p_d(s_1+s_3)+p_s s_2}$$
$$\implies w_2 = (1-\theta)\frac{p_d s_1}{p_d(s_1+s_3) + p_s s_2 - (1-\theta)^2 p_s s_2} + (1-\theta)^2\frac{p_d s_3 w_3}{p_d(s_1+s_3)+p_s s_2 - (1-\theta)^2 p_s s_2}$$

For $\ell \geq 3$:

$$w_\ell = (1-\theta)^2\frac{p_d(w_{\ell-1}s_{\ell-1} + w_{\ell+1}s_{\ell+1}) + p_s w_\ell s_\ell}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell}$$
$$\implies w_\ell = (1-\theta)^2\frac{p_d(w_{\ell-1}s_{\ell-1} + w_{\ell+1}s_{\ell+1})}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell}$$

with $w_1 = 1$ because the island consists of all Bayesian agents. We first show that $\lim_{\ell\to\infty} w_\ell$ must exist. To do this, we show that w_ℓ is monotonically decreasing in ℓ. We know there is a unique fixed point for $\mathbf{w}$, so if we prove that a decreasing sequence of w_ℓ maps to another decreasing sequence of w_ℓ, then by Brouwer's fixed point theorem the unique solution must be a decreasing in ℓ. Note that:

$$\begin{aligned}
&w_\ell \leq (1-\theta)^2 \frac{p_d(w_{\ell-1}s_{\ell-1} + w_\ell s_{\ell+1})}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell} \\
\Longrightarrow &\left(1 - \frac{p_d s_{\ell+1}}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell}\right) w_\ell \leq (1-\theta)^2 \frac{p_d w_{\ell-1} s_{\ell-1}}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell} \\
\Longrightarrow &w_\ell \leq (1-\theta)^2 \frac{p_d w_{\ell-1} s_{\ell-1}}{p_d s_{\ell-1} + p_s s_\ell - (1-\theta)^2 p_s s_\ell} \leq w_{\ell-1}
\end{aligned}$$

where the final inequality follows from the fact that $\beta\frac{\alpha}{\alpha+\delta} < 1$ for $\alpha, \beta, \delta \in (0,1)$. Thus, w_ℓ converges to some w_∞. Note that w_∞ must satisfy the fixed-point equation:

$$w_\infty = (1-\theta)^2 \frac{p_d(s_{\ell-1}+s_{\ell+1})w_\infty}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell}$$

Again, since $\frac{p_d(s_{\ell-1}+s_{\ell+1})}{p_d(s_{\ell-1}+s_{\ell+1})+p_s s_\ell-(1-\theta)^2 p_s s_\ell} < 1$, clearly $w_\infty = 0$.

Now, suppose the principal attempts to manipulate $k - \bar{k}$ communities at the end of the line. By our previous result, we know that for every $\delta > 0$, there exists a sufficiently large $\bar{k}$, such that $w_{\bar{k}+1} < \delta$. Therefore, for any b, the principal can manipulate $k - \bar{k}$ of the islands at the end of the line. This yields a payoff of $\sum_{\ell=\bar{k}+1}^{k} ns_k - n\varepsilon(1-s_1)$, which is positive for some sufficiently small $\varepsilon > 0$. Thus, the network is susceptible to manipulation and all but $\bar{k}$ islands are manipulated. □

B.5 Section 7

Proof of Corollary 1. By assumption, the budget is large enough to make the network impervious. Thus, by Theorem 2, the network is impervious if the current distribution of Bayesians $\mathbf{m}$ is majorized by every other distribuiton (i.e., inequality cannot be reduced by a more equal redistribution of Bayesians). Minimizing inequality with a Bayesian intervention accomplishes this. □

Proof of Corollary 2. Leveraging Theorem 4, we know that decreasing inequality when the big island is the most underprivileged does not introduce more manipulation, and in fact, might reduce it. Assuming the policy does not put all of the Bayesians on island 1, we know such a redistribution is a different feasible policy. In the proof of Theorem 4, this is shown to be true because decreasing inequality increases all agents' beliefs. This is exactly the definition of a dominant policy. □

Proof of Proposition 3. We show that if the Bayesians are assigned proportional Bayesians to island populations, i.e., $m_\ell = M \cdot s_\ell$, then all DeGroot centralities are equal. Then, homophily has no effect (beliefs are the same on every island), so setting $p_d = p_d^o$ is optimal, and always feasible because it costs nothing. Consider the map T:

$$T\,:\,\mathbf{w} \mapsto (1-\theta)\begin{pmatrix} \frac{p_s m_1 + \sum_{\ell\neq 1} p_d m_\ell}{np_s s_1 + np_d(1-s_1)} \\ \cdots \\ \frac{p_s m_k + \sum_{\ell\neq k} p_d m_\ell}{np_s s_k + np_d(1-s_k)} \end{pmatrix} + (1-\theta)^2 \begin{pmatrix} \frac{p_s(ns_1-m_1)}{np_s s_1+np_d(1-s_1)} & \frac{p_d(ns_2-m_2)}{np_s s_1+np_d(1-s_1)} & \cdots & \frac{p_d(ns_k-m_k)}{np_s s_1+np_d(1-s_1)} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{p_d(ns_1-m_1)}{np_s s_k+np_d(1-s_k)} & \frac{p_d(ns_2-m_2)}{np_s s_k+np_d(1-s_k)} & \cdots & \frac{p_s(ns_k-m_k)}{np_s s_k+np_d(1-s_k)} \end{pmatrix} \mathbf{w}$$

Note we will simply plug in $\mathbf{w} = w^*\mathbf{1}$ to T and show it is a fixed point for some constant w^*. For island ℓ:

$$\begin{aligned}
w_\ell &= (1-\theta)\cdot\frac{p_s m_\ell + \sum_{\ell'\neq\ell} p_d m_{\ell'} + (1-\theta)p_s(ns_\ell - m_\ell)w_\ell + (1-\theta)\sum_{\ell'\neq\ell} p_d(ns_{\ell'} - m_{\ell'})w_{\ell'}}{np_s s_\ell + np_d(1-s_\ell)} \\
&= (1-\theta)\cdot\frac{p_s s_\ell M + \sum_{\ell'\neq\ell} p_d s_{\ell'} M + (1-\theta)p_s(ns_\ell - s_\ell M)w^* + (1-\theta)\sum_{\ell'\neq\ell} p_d(ns_{\ell'} - s_{\ell'} M)w^*}{np_s s_\ell + np_d(1-s_\ell)} \\
&= (1-\theta)\cdot\frac{M(p_s s_\ell + p_d(1-s_\ell)) + (1-\theta)w^*(np_s s_\ell + np_d(1-s_\ell) - M(p_s s_\ell + p_d(1-s_\ell)))}{np_s s_\ell + np_d(1-s_\ell)} \\
&= (1-\theta)\cdot\frac{M + (1-\theta)w^*(n-M)}{n}
\end{aligned}$$

The above expression has no dependence on ℓ. Letting $w^* = \frac{M(1-\theta)}{M(1-\theta)+\theta n}$, we see then that $w_\ell = w^*$, which completes the proof. This has no dependence on p_d, so all p_d are optimal, including p_d^o. □

Proof of Proposition 4. The condition that the budget exceeds $\phi(p_s - p_d^o)$ is to guarantee that $p_d = p_s$ is feasible. By Theorem 2, given that all islands are the same size, a network with less inequality cannot transition from impervious to susceptible. Thus, removing all homophily (i.e., reducing inequality the most through the homophily parameters) must make the network impervious given this is possible for some homomphily structure. Thus, setting $p_s = p_d$ makes the network impervious which is obviously an optimal policy. □

References

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

Acemoglu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi (2015), “Systemic Risk and Stability in Financial Networks.” *American Economic Review*, 105, 564–608, URL https://www.aeaweb.org/articles?id=10.1257/aer.20130456.

Akbarpour, Mohammad, Suraj Malladi, and Amin Saberi (2018), “Diffusion, seeding, and the value of network information.” In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 641–641, ACM.

Alidaee, Hossein, Eric Auerbach, and Michael P Leung (2020), “Recovering network structure from aggregated relational data using penalized regression.” *arXiv preprint arXiv:2001.06052*.

Arnold, Barry C. (1987), *Majorization and the Lorenz Order: A Brief Introduction*. Lecture Notes in Statistics, Springer-Verlag, New York, URL https://www.springer.com/gp/book/9780387965925.

Ata, Baris, Alexandre Belloni, and Ozan Candogan (2018), “Latent agents in networks: Estimation and pricing.” *arXiv preprint arXiv:1808.04878*.

Auerbach, Eric (2019), “Measuring differences in stochastic network structure.” *arXiv preprint arXiv:1903.11117*.

Babus, Ana (2016), “The formation of financial networks.” *The RAND Journal of Economics*, 47, 239–272, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1756-2171.12126.

Calvo-Armengol, Antoni and Matthew O Jackson (2004), “The effects of social networks on employment and inequality.” *American economic review*, 94, 426–454.

Calvó-Armengol, Antoni and Matthew O Jackson (2007), “Networks in labor markets: Wage and employment dynamics and inequality.” *Journal of economic theory*, 132, 27–46.

Candogan, Ozan and Kimon Drakopoulos (2017), “Optimal signaling of content accuracy: Engagement vs. misinformation.” *Operations Research, forthcoming*.

Card, David and Alan B Krueger (1992), “School quality and black-white relative earnings: A direct assessment.” *The Quarterly Journal of Economics*, 107, 151–200.

Chandra, Amitabh (2000), “Labor-market dropouts and the racial wage gap: 1940-1990.” *American Economic Review*, 90, 333–338.

Chandrasekhar, Arun G, Horacio Larreguy, and Juan Pablo Xandri (2019), “Testing models of social learning on networks: Evidence from two experiments.” *Econometrica*.

Currarini, Sergio, Matthew O Jackson, and Paolo Pin (2009), “An economic model of friendship: Homophily, minorities, and segregation.” *Econometrica*, 77, 1003–1045.

Golub, Benjamin and Matthew O Jackson (2012), “How homophily affects the speed of learning and best-response dynamics.” *The Quarterly Journal of Economics*, 127, 1287–1338.

Heckman, James J, Thomas M Lyons, and Petra E Todd (2000), “Understanding black-white wage differentials, 1960-1990.” *American Economic Review*, 90, 344–349.

Holland, Paul W, Kathryn Blackmond Laskey, and Samuel Leinhardt (1983), “Stochastic blockmodels: First steps.” *Social networks*, 5, 109–137.

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

Kanak, Zafer (2017), “Rescuing the Financial System: Capabilities, Incentives, and Optimal Interbank Networks.” Technical Report 17-17, NET Institute, URL `https://ideas.repec.org/p/net/wpaper/1717.html`.

Keppo, Jussi, Michael Jong Kim, and Xinyuan Zhang (2019), “Learning manipulation through information dissemination.” *Available at SSRN 3465030*.

Lobel, Ilan and Evan Sadler (2015), “Preferences, homophily, and social learning.” *Operations Research*, 64, 564–584.

Manshadi, Vahideh, Sidhant Misra, and Scott Rodilitz (2018), “Diffusion in random networks: Impact of degree distribution.” *Operations Research, forthcoming*.

Marsden, Peter V (1987), “Core discussion networks of americans.” *American sociological review*, 122–131.

Marshall, Albert W., Ingram Olkin, and Barry C. Arnold (2011), *Inequalities: Theory of Majorization and Its Applications*, 2 edition. Springer Series in Statistics, Springer-Verlag, New York, URL `https://www.springer.com/gp/book/9780387400877`.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook (2001), “Birds of a feather: Homophily in social networks.” *Annual review of sociology*, 27, 415–444.

Mostagir, Mohamed, Asu Ozdaglar, and James Siderius (2019), “When is society susceptible to manipulation?” *Working paper*.

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

Sadler, Evan (2019), “Influence campaigns.” *Available at SSRN 3371835*.

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.