NBER WORKING PAPER SERIES

MISINFORMATION: STRATEGIC SHARING, HOMOPHILY, AND ENDOGENOUS ECHO CHAMBERS

Daron Acemoglu
Asuman Ozdaglar
James Siderius

Working Paper 28884
http://www.nber.org/papers/w28884

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2021

We are grateful to Hunt Allcott, Matthew Gentzkow, Ro'ee Levy, Davy Perlman, Alexander Wolitzky, MIT theory lunch participants, and MIT LIDS tea talk participants for comments. We also gratefully acknowledge financial support from the NSF, the Sloan Foundation, and MIT Quest. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Misinformation: Strategic Sharing, Homophily, and Endogenous Echo Chambers
Daron Acemoglu, Asuman Ozdaglar, and James Siderius
NBER Working Paper No. 28884
June 2021
JEL No. D83,D85,P16

ABSTRACT

We present a model of online content sharing where agents sequentially observe an article and must decide whether to share it with others. The article may contain misinformation, but at a cost, agents can fact-check it to determine whether its content is entirely accurate. While agents derive value from future shares, they simultaneously fear getting caught sharing misinformation. With little homophily in the "sharing network", misinformation is often quickly identified and brought to an end. However, when homophily is strong, so that agents anticipate that only those with similar beliefs will view the article, misinformation spreads more rapidly because of echo chambers. We show that a social media platform that wishes to maximize content engagement will propagate extreme articles amongst the most extremist users, while not showing these articles to ideologically opposed users. This creates an endogenous echo chamber—filter bubble—that makes misinformation spread virally. We use this framework to understand how regulation can encourage more fact-checking by online users and mitigate the consequences of filter bubbles.

Daron Acemoglu
Department of Economics, E52-446
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
daron@mit.edu

Asuman Ozdaglar
Department of Electrical Engineering
and Computer Science
Massachusetts Institute of Technology
77 Massachusetts Ave, E40-130
Cambridge, MA 02139
asuman@mit.edu

James Siderius
Department of Electrical Engineering
and Computer Science
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139
USA
siderius@mit.edu

1 Introduction

Social media has become an increasingly important source of information for many Americans. Leading up to the 2016 US presidential election, around 14% of Americans indicated social media as their primary source of news (Allcott and Gentzkow (2017)), and over 70% of Americans reported receiving at least *some* of their news from social media (Levy (2020)). Many of these news stories made wildly false claims, for example, arguing that there were no mass shootings under Donald Trump[1] or Hillary Clinton approved ISIS weapon sales.[2] These false stories spread virally on social media sites, with "falsehood diffusing significantly farther, faster, deeper, and more broadly than the truth in all categories of information," according to Vosoughi et al. (2018).

Two main forces, echo chambers and political polarization, may have exacerbated the spread of misinformation. Facebook, where the vast majority of news consumption on social media takes place, employs algorithms that create "filter bubbles," which limit the exposure of users to counter-attitudinal news outlets and instead promote "politically congenial" content (Lazer et al. (2018), Levy (2020)). These dynamics have been shown to reinforce political viewpoints and exacerbate cascades of misinformation (see, for instance, Törnberg (2018) and Vicario et al. (2016)). At the same time, from 1996 to 2016, political polarization has increased significantly according to a variety of measures.[3] This polarization may have fueled more selective exposure to news consumption, as well as growing exposure to misinformation, particularly for individuals with more extreme views (Guess et al. (2018)). Engaging in selective exposure is further facilitated when information can be easily consumed and shared on social media sites like Facebook by scrolling through a news feed with nearly limitless content (Bakshy et al. (2015)).

In this paper, we develop a parsimonious model of online behavior in the presence of misinformation, homophily, and political polarization. As a first step in this endeavor, we abstract from the various behavioral biases of individuals when they engage with online information (see, for instance, Buchanan (2020) and Pennycook and Rand (2018)), and focus on the behavior of fully Bayesian agents, which we show generates rich strategic interactions and dynamics.

Agents in our model choose to interact with an article by either ignoring it, inspecting it for misinformation, or simply sharing it (the importance of the inspection mechanism in practice is demonstrated in Guriev et al. (2020)). Each agent has a different ideological leaning, captured by his or her (henceforth, simply her) priors, and every article has a particular political viewpoint, which may agree or disagree with an agent's prior beliefs. The article may also contain misinformation. We suppose that agents are situated in a "sharing network"—reflecting their social network as well as the algorithms used by the platform—which determines who they share articles with. We first take this network as given and then endogenize it. Agents derive value from getting additional shares or

[1] https://www.snopes.com/fact-check/mass-shootings-under-trump/

[2] https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html

[3] While there has been some debate about whether polarization has been isolated mainly to politicians (see Fiorina et al. (2008) and Prior (2013)), there is considerable evidence that polarization has also risen among the general public (see Pew Research Center (2014) and Abramowitz (2010)).

"retweets" on the stories they share, but are afraid to get called out for spreading misinformation. This creates an incentive to share stories that (i) are unlikely to contain misinformation, and/or (ii) are likely to be shared, and not scrutinized, by others in one's social circle. We also assume that articles at first spread slowly (are shared with a few people), and after a while, they become viral and are shared with a larger subset of those in the agent's relevant social circle.

We characterize the (sequential) equilibria of the aforementioned game. Before the article has gone viral, and is being shared slowly, agents who disagree with it (i.e., those who fall below a certain ideological belief cutoff) simply ignore it, while those who agree with it always share it. This is consistent with the selective exposure effect identified in Guess et al. (2018). Once an article enters the viral phase, agents who disagree with the article inspect it to see if it contains misinformation, whereas agents who agree with it share without inspection.

Our game features neither strategic substitutes nor strategic complements. When other agents share more, an agent might worry that nobody is inspecting and becomes more likely to inspect more and thus share less, which pushes towards strategic substitutes. On the other hand, when others are expected to inspect more in the future, the agent may reason that she is unlikely to be found out to have propagated misinformation and may share more, thus pushing towards strategic complements. Despite these rich strategic interactions, we prove that an equilibrium always exists. We also provide conditions under which the most viral—the "all-share"—equilibrium, one where all agents in the population share the article, exists and is unique.

We present two sets of main results. The first is a series of comparative static results showing when misinformation is more likely to spread virally. Most importantly, we show that when there is little homophily in the sharing network, and agents are likely to be exposed to cross-cutting content (including counter-attitudinal articles), extreme messages with misinformation are unlikely to survive for very long. Perhaps paradoxically, we find that in this baseline environment with limited homophily, political polarization and/or a small increase in homophily actually *help* reduce the extent of misinformation. Both of these effects occur because agents will encounter strongly counter-attitudinal news, which they will investigate and expose if they find it to contain misinformation. A small increase in homophily (or more political polarization) encourages further inspection from agents who disagree with the article, because they have become more likely to share it with others, who like themselves, disagree with it. Their behavior in turn further disciplines agents who agree with the story to be cautious about sharing without inspection.

However, these conclusions are reversed when the network has strong homophily. In this case, agents who agree and disagree with the article interact rarely. Consequently, an increase in homophily creates an echo chamber where agents become less concerned about others inspecting their article and are more likely to spread misinformation. Notably, this viral behavior is exacerbated when the article has an extreme message or polarization is strong. Because the article is unlikely to fall in the hands of any individual who opposes it, extremism fuels misinformation amongst those who agree with it.

Our second set of main results identify conditions under which platforms will design algorithms for content sharing in a way that both creates ideological homophily in the sharing network and propagates misinformation. In particular, we assume that the platform wishes to maximize engagement but is indifferent to whether this engagement is with misinformation or truthful content. We demonstrate that when all sources of news are reasonably moderate, the platform prefers to inspect the veracity of the articles itself and then tags them as "verified". It also does not create filter bubbles. This strategy is beneficial for the platform because tagged accurate information triggers a share cascade that will not be interrupted by an exposure of misinformation. In contrast, if extremist content is available, the platform's incentives are diametrically different: it prefers not to inspect the veracity of such extreme content, designs its algorithm to create strong filter bubbles, and propagates the extremist content. This preference is rooted in the fact that like-minded agents will share extremist content among themselves when they understand they are in an echo chamber, which leads to high engagement. We also establish that the platform's incentive to create filter bubbles and share extremist messages becomes even stronger when there is severe polarization. Overall, these results imply that a particularly pernicious form of misinformation—contained in the most extremist messages and exacerbated with user polarization—will spread virally.

Lastly, we briefly consider the problem of regulation of social media—say by a planner who wants to limit misinformation on the platform. We consider three different policies: news quality revelation, censorship, and network regulation (e.g., an ideological segregation standard for the network induced by platform algorithms). First, we show that a policy that reveals an article's provenance (and the source's quality) can decrease misinformation by allowing users to separate their scrutiny based on the likelihood of misinformation.. However, we also find that such policies can sometimes backfire and actually increase the spread of misinformation, because they discourage agents from inspecting articles that are not tagged as originating from unreliable sources. Second, we show a limited form of censorship may be effective in reducing misinformation, and even the threat of such censorship may motivate platforms not to create filter bubbles and propagate extremist messages. Lastly, we consider regulation that targets the sharing network and filter bubbles directly, by setting a requirement that reduces the amount of acceptable homophily present in the network.

Related Literature. Our paper builds on a large body of work on misinformation and fake news, both empirical and theoretical. In addition to the literature mentioned previously, several other papers on misinformation are related to our findings.

Much previous work has focused on the susceptibility of boundedly-rational agents to engage with misinformation. In Acemoglu et al. (2010) and Acemoglu et al. (2013), the existence of persuasive agents can impede information aggregation and enable misinformed beliefs to survive, and sometimes even become dominant, in the population. In Mostagir et al. (2019) and Mostagir and Siderius (2021), a strategic principal who wants to persuade agents of an incorrect belief can often distort the learning process by leveraging social connections and echo chambers to propagate misinformation. Similarly, models of misinformation “contagion”—without Bayesian agents or strategic decisions—have been studied in Budak et al. (2011), Nguyen et al. (2012), and Törnberg (2018). Our contribution is different because, in our model, the spread of misinformation is rooted in this strategic interactions of Bayesian agents and profit-maximizing platforms.

There is a growing literature on information design by platforms, most often building on the concept of Bayesian persuasion (Kamenica and Gentzkow (2011) and Kamenica (2019)). Candogan and Drakopoulos (2017) study how a platform with private knowledge of content’s accuracy should optimally signal to rational users whether to engage with it, while Chen and Papanastasiou (2019) and Keppo et al. (2019) consider more manipulative actions by platforms, including strategic seeding of information or “cheap talk” signals about quality. Also related are works on reputation and media bias. Motivated by the 2016 presidential election, Allcott and Gentzkow (2017) study the incentives of certain outlets to present misleading news, while Gentzkow and Shapiro (2006), Hsu et al. (2020) and Allon et al. (2019) explore other strategic reasons for media bias. Among other things, our paper differs from these works because our focus is on social media interactions among users with heterogeneous prior beliefs in potentially homophilic networks. As a result, none of these papers generate viral spread of misinformation driven by user sharing behavior or endogenous echo chambers.

The most closely related work to ours is the important paper by Papanastasiou (2018), who studies a model where agents hold heterogenous ideological beliefs and digest and potentially inspect a news article sequentially. Our work is different in three important dimensions. First, consistent with much of the empirical literature on social networks, our model incorporates the strategic complementarity of online sharing behavior.[4] This strategic complementarity is at the root of all of our main results, and arises in our model because agents care about whether others will inspect and find them to be propagating misinformation. In contrast, Papanastasiou (2018) there is no peer feedback on sharing decisions and hence the game he studies is one of strategic substitutes (making his analysis more straightforward as a result). Second, and relatedly, echo chambers play no role in Papanastasiou (2018) (and in fact, an agent is completely indifferent to any share cascades that might result from her initial sharing decision). Our reading is that the vast majority of empirical work in this area finds socializing and status seeking are among the top reasons individuals share news on social media (see Lee et al. (2011)), and as a result, viral misinformation is typically associated with social reinforcement among peers in one’s network (see, for example, Törnberg (2018), Centola (2010), and Centola and Macy (2007)). Third, the platform analysis in Papanastasiou (2018) assumes that the platform wants to reduce the spread of misinformation, so the results focus on when self-policing is more effective than platform inspections. As a result, our results on how engagement-maximizing platforms will propagate misinformation by creating filter bubbles are, to the best of our knowledge, completely new.

The rest of the paper is organized as follows. We introduce our basic model, with exogenous

[4] For example, Eckles et al. (2016) find evidence that feedback or “encouragement” from peers about Facebook posts contributes significantly to future behavior and posting. In our model, as opposed to an inherent dislike for spreading misinformation, agents simply fear peer “shaming” from being identified as someone who spreads misinformation. See also Taylor and Eckles (2018) and Aral and Dhillon (2018).

sharing network, in the next section. Section 3 characterizes the (sequential) equilibria of this model and provides some basic comparative static results. Section 4 studies the effects of homophily by focusing on a special class of sharing networks that correspond to a set of the "islands" of like-minded individuals who are less closely linked to those in other islands. Section 5 endogenizes the sharing network as a result of the algorithmic choices of a platform that aims to maximize engagement. Section 6 discusses a range of regulations aimed at containing misinformation. Section 7 concludes, while all proofs are provided in the Appendix.

2 Model

There is an underlying state of the world $\theta \in \{L, R\}$, for example, corresponding to whether the left-wing or the right-wing candidate is more qualified for political office. Agents have heterogeneous prior (ideological) beliefs about θ, and agent i's prior that $\theta = R$ is denoted by b_i and has distribution $H_i(\cdot)$. For concreteness, throughout the paper we will take the state to be $\theta = L$.

We assume there are N agents in the population, who share a news article according to a *sharing network* defined by matrix $\mathbf{P}$ of link probabilities:

$$\mathbf{P} \equiv \begin{pmatrix} 0 & p_{12} & \cdots & p_{1N} \\ p_{21} & 0 & \cdots & p_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1} & p_{N2} & \cdots & 0 \end{pmatrix}$$

where p_{ij} is the probability that agent i has a link to agent j. We denote by $\mathcal{N}_i$ as the set of agents attached to agent i with an outgoing link (i.e., agent i's "neighborhood"). The sharing network reflects both an individual's social circle and the algorithms the platform uses for sharing articles. We take it as given throughout this and the next two sections, and endogenize it in Section 5.

We next describe how news articles are generated and how they spread through the population. We then present the payoffs of the agents. Finally, we introduce the solution concept we use to analyze the equilibria of this game.

2.1 Misinformation and News Generation

Each news story has a three-dimensional type (s, ν, m). The type $s \in \mathbb{R}$ indicates the *provenance* (or source) of the news. The type ν is the article's *veracity*, which can either be $\mathcal{T}$, to indicate the article is truthful, or $\mathcal{M}$, to indicate the article contains *misinformation*. The type $m \in \mathbb{R}$ is the *message*, which corresponds to the article's viewpoint. Without loss of generality, we assume that lower-valued messages advocate for left-wing ideas, whereas higher messages advocate for right-wing ideas, as seen in Figure 1. The message $m^* = 0$ is a moderate message that splits the message space into left and right-wing news.[5]

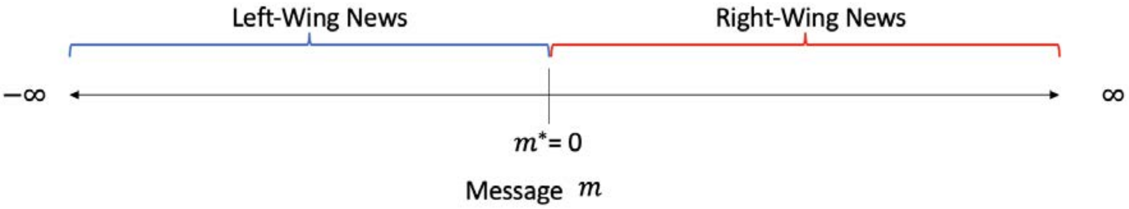

Figure 1. Interpretation of Article's Message.

Misinformation vs Fake News. Our focus in this paper is on articles that contain misinformation, interpreted as items that include misleading information or arguments, with the explicit or predictable purpose of influencing (a subset of) the public. Articles containing "fake news" (demonstrably false information) make up a small subset of those containing misinformation. For example, a news item that favorably describes a report denying climate change, without putting this in the context of the criticisms that this report has received from scores of experts or hundreds of other reports reaching the opposite conclusion, would be one containing misinformation according to this definition. Recent work finds that this type of misinformation is much more common than explicit fake news (e.g., Allen et al. (2020), Guess et al. (2019), Grinberg et al. (2019)). For example, Allen et al. (2020) report that less than 0.2% of political content on social media is overtly false.

Generation Process. Each article's (three-dimensional) type (s, ν, m) is drawn according to the following i.i.d process:

(i) The provenance of the news s is drawn from a Lebesgue-integrable density function $g(\cdot)$.

(ii) The veracity of the article is $\nu = \mathcal{T}$ with probability q_s and $\nu = \mathcal{M}$ with probability $1 - q_s$.[6]

  (a) If the news story is truthful (i.e., $\nu = \mathcal{T}$), then the message is generated according to a smooth distribution $F(\cdot|\theta)$ which depends on the underlying state θ and has density $f(\cdot|\theta)$ bounded above and bounded below away from 0. We assume truthful messages are informative about the state, so the likelihood ratio $f(m|\theta = R)/f(m|\theta = L)$ is (strictly) increasing in m. This implies, for example, that strong right-wing messages are a stronger indication of $\theta = R$ than weak right-wing messages or left-wing messages. We refer to this property of f as the *monotone likelihood ratio property* (MLRP).

  (b) If the news story contains misinformation (i.e., $\nu = \mathcal{M}$), we assume that it is *anti-correlated* with the state, meaning that the message is distributed according to $f(m|\theta = \neg\theta)$, where $\neg\theta$

[5] This is without loss of generality: for general m^*, one can simply make the change of variables $\tilde{m} = m - m^*$ to realign the moderate message to $\tilde{m}^* = 0$.

[6] Note that while agents hold different prior beliefs about the state θ, they hold the same (prior) belief about the veracity of the news, given by the (objective) probability of truthfulness, q_s.

is the opposite of state θ (i.e, $\neg\theta = R$ when $\theta = L$ and vice-versa). This assumption is adopted for convenience. In general, our definition implies that misinformation is potentially, but not entirely, misleading, and this can be modeled as a different correlation structure between the underlying state and messages under misinformation than the truth. We show in Appendix B.6 that our results generalize to this case, though the relevant thresholds are more complicated, and thus we opt for the pure anti-correlation formulation here for expositional simplicity.

In the baseline model, we assume that (s, ν) is not known to the agents—in particular, the platform does not provide any information about the veracity of the news article. Therefore, the *ex-ante* probability an article is truthful is $q \equiv \int_{-\infty}^{\infty} q_s \, g(s) \, ds$ and the agent's belief π_i about ν is updated according to Bayes' rule, conditional on her prior belief b_i.

2.2 Lifetime of an Article

Time is discrete $t = 1, 2, \ldots$. Without loss of generality, we assume there is just a single article in circulation. At $t = 1$, this article is given to some seed agent i^* chosen uniformly at random. When receiving the article, an agent plays one of three actions: share the article, inspect the article, or kill the article. We denote the action of agent i as a_i and the share, inspect, and kill actions as $\mathcal{S}, \mathcal{I}$, and $\mathcal{K}$, respectively. Agents see the message m of the article before taking their action a_i. The share action means the agent chooses to pass on the article to others without fact-checking it first. The inspect action means the agent fact-checks the article, which perfectly reveals whether the article is truthful or contains misinformation, and then shares it with others if it is truthful. The kill action means the agent just ignores the article.

Every article goes through three phases in its "lifetime" which we call the *initial phase*, the *viral phase*, and the *obsolescent phase*. Formally, the article, given to seed agent i^* at $t = 0$, always starts in the initial phase. The article then transitions to the viral phase at random, which is determined by a Poisson clock. Once in the viral phase, a new Poisson clock starts, which determines when the article then transitions to the obsolescent phase. The details of what occurs during each of the phases is provided next.

Initial Phase. When agent i shares the article in the initial phase, the article is passed on to exactly one neighbor of agent i, determined uniformly at random from the agents in i's neighborhood. If agent i kills the article, the article dies and skips immediately to the obsolescent phase. Conditional on the article not being killed up until some point in time, there is a random chance the article goes "viral" and begins attracting more attention, determined by a Poisson clock with parameter λ_1. Once this clock ticks (and given all agents have shared until now), the article transitions to the viral phase.

Viral Phase. In the viral phase, when agent i shares the article, it is passed to $\gamma > 1$ neighbors of agent i instead of just 1 (again, chosen uniformly at random from agent i's neighborhood). While the article is viral, there is a random chance it becomes "yesterday's news" and agents lose interest and stop sharing. This also occurs randomly: a new Poisson clock starts as soon as the viral phase begins, with parameter λ_2. When this clock ticks, the article immediately transitions into the obsolescent phase.

Obsolescent Phase. After the article becomes obsolete, we assume the article is inspected by an outside source. For instance, after the article has been around for awhile, a third-party fact checking organization, such as Politifact, inspects.[7] Thus, once the article is no longer being shared on social media, society discovers the ground truth about whether the article did in fact contain misinformation.

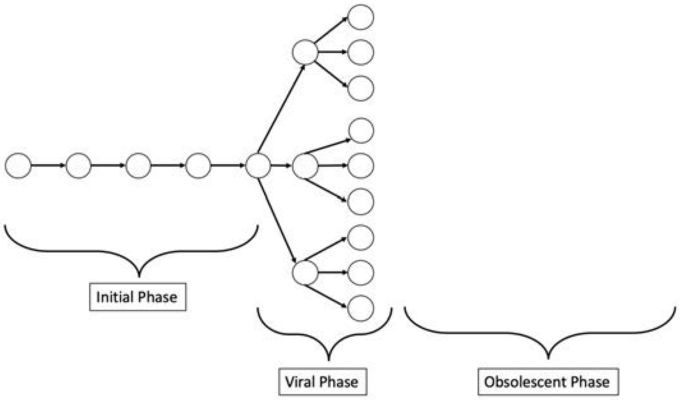

Figure 2. The lifetime of an article.

A pictorial version of the process is shown in Figure 2. Intuitively, this three-phase structure is motivated by the typical dynamics of articles over social media, for example as documented in the work of Vosoughi et al. (2018). The average lifetime of an article for both real (green) and fake (red) stories is shown in Figure 3, courtesy of Vosoughi et al. (2018). For articles containing misinformation, while it took nearly 1k minutes to attract 1k unique users, 40k users were obtained in about 10k minutes, but only in rare circumstances were more users obtained after that.[8] The viral phase is also reminiscent of the "hot" designation that content receives after being shared consistently.[9]

2.3 Strategic Sharing and Exposure to Misinformation

After observing the message m of the article, any agent i who receives the article at some period $t \geq 1$ chooses its action $a_i \in \{\mathcal{S}, \mathcal{I}, \mathcal{K}\}$. If she chooses to share ($a_i = \mathcal{S}$), then no fact-check is done and the

[7] This final inspection can occur for a number of reasons, which need not be related to another news source investigating the story. For instance, the virality of the Pizzagate conspiracy came to a halt when someone invaded the pizza parlor and revealed the story to have misinformation.

[8] According to https://buzzsumo.com/blog/the-most-shared-facebook-content-posts-videos/, the most shared content on Facebook in 2017 received over 22 million shares. Facebook had over 2 billion active users in 2017 though (https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/), so only a very small fraction of users interact with even the most viral posts.

[9] For example, sites like Reddit algorithmically identify such "hot" content and place it at the top of related subreddits (https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9) and Facebook is reported to use various strategies to attract users to this type of content (https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/).

article progresses as described in Section 2.2 (i.e., passes to one agent in $\mathcal{N}_i$ in the initial phase and passes to γ agents in $\mathcal{N}_i$ in the viral phase). If she chooses to inspect ($a_i = \mathcal{I}$), then the agent first fact-checks the article, which requires effort, and pays a cost K; if the article is truthful, she shares it,[10] and if it contains misinformation, the article is revealed as containing misinformation and payoffs are realized. We describe these next.

Each agent discounts the downstream consequences (i.e., friends of friends' decisions and so on) at the rate $\beta \geq 0$. When $\beta > 0$, payoffs include the implications of subsequent interactions on social media (e.g., retweeting a retweet). But nothing in our analysis precludes the possibility that β is close to or equal to zero.

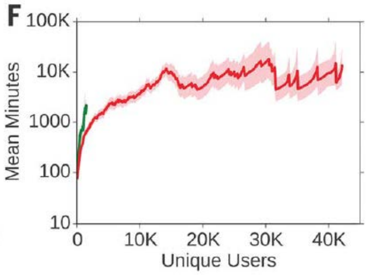

Figure 3. Average unique users as a function of time.

We refer to the downstream network of agents receiving the article, rooted at agent i, following i's original share as i's *sharing subtree*. Let t_i be the time at which agent i receives the article, τ_1 the (random) time at which the first clock ticks and the article transitions to the viral phase, and τ_2 the (random) time at which the second clock ticks *after* the first clock, and the article transitions to the obsolescent phase. All sharing ceases in agent i's sharing subtree after either: (i) the article is inspected and found to contain misinformation by some agent in i's sharing subtree, or (ii) the article is inspected by the outside source (regardless of its veracity). We denote by T_i^* this (random) time for agent i. Naturally, an agent's payoff depends on which of the three actions ($\mathcal{S}, \mathcal{I}$, or $\mathcal{K}$) she takes:

Kill ($a_i = \mathcal{K}$). When an agent kills the article, she simply *ignores* it, and it is not shared with others.[11] The payoff from killing the article is normalized at 0.

Share ($a_i = \mathcal{S}$). The payoff from sharing the article is proportional to the (discounted) number of agents who share that article in the future. The utility from sharing is $S_i \equiv \kappa \cdot \sum_{\tau=1}^{\infty} \beta^{\tau-1} S_{i,\tau}$, where $S_{i,\tau}$ is the number of (indirect) shares (in her sharing subtree) resulting from agent i's share τ periods later, κ captures the marginal social utility from an additional share, and β is the discount factor. Formally,

[10] Therefore, the agent will always share an article she knows to be truthful with probability 1. This is an implication of the payoffs we present next, but it is convenient incorporate it into the description of the environment.

[11] If we removed the kill/ignore option and forced agents to either share or inspect the news article, this would simplify the analysis but would also lead to a highly counterfactual prediction that is not present in our model: truthful articles would spread farther than articles containing misinformation, whereas the evidence seems to be the opposite; e.g., Vosoughi et al. (2018).

the value of the shares at time τ is:

$$S_{i,\tau} = \begin{cases} 0, \text{ if } t_i + \tau \geq T_i^* \\ 1, \text{ if } t_i + \tau < \min\{\tau_1, T_i^*\} \\ \gamma S_{i,\tau-1}, \text{ if } \tau_1 \leq t_i + \tau < \min\{\tau_1 + \tau_2, T_i^*\} \end{cases}$$

In other words, there are three cases: (i) once sharing comes to an end, there are no more payoffs, (ii) in the initial phase, conditional on the article's survival, each period leads to one additional share, and (iii) in the viral phase, conditional on the article's survival, the sharing agent receives γ times the shares she received in the previous period (due to the branching of the share process).[12]

There is also a possible punishment from sharing misinformation. In particular, an agent i who chose $\mathcal{S}$, in addition to her sharing payoff, faces a punishment $C\beta^{T_i^*-t_i-1}$ when the article is inspected at time $T_i^* > t_i$ by an agent in her sharing subtree (with no sharing payoff or punishment if $t_i = T_i^*$), and found out to contain misinformation. In Appendix B.8, we derive this cost from a reputational mechanism, but it is more convenient to impose it as an assumption here. This type of reputational cost has been documented in Altay et al. (2020), who find in a Facebook field study that users who had passed on misinformation in the past experienced bad reputation in their future shares.

Inspect ($a_i = \mathcal{I}$). If an agent inspects, she automatically pays an inspection cost $K > 0$, regardless of the outcome of the inspection (see Section 5.2 for heterogeneous inspection costs). If the article contains misinformation, the punishments described in the sharing section are realized, and the article dies. If the article is viral, the agent receives a positive payoff $\delta > 0$ from exposing misinformation in an article, but there is no such payoff in the initial phase.[13]

If the article is truthful, the agent shares it and receives the same payoff as the agent who elects $a_i = \mathcal{S}$ (of course, after paying the cost K). Consequently, the payoff from inspecting is given by:

$$\begin{cases} -K, & \text{if } \nu = \mathcal{M} \text{ and } \mathcal{I} \text{ is played in the initial phase} \\ \delta - K, & \text{if } \nu = \mathcal{M} \text{ and } \mathcal{I} \text{ is played in the viral phase} \\ \kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1} S_{i,\tau} - K, & \text{if } \nu = \mathcal{T} \text{ and } \mathcal{I} \text{ is played (regardless of phase)} \end{cases}$$

2.4 Payoff Assumptions

In this section, we introduce additional assumptions about payoff parameters. We define two objects, $v_{initial}$ and v_{viral}, that will play an important role in our analysis. More specifically, these are the

[12] In practice, agents may receive utility (or higher utility) from others liking their posts and shares rather than the number of further shares, but the implications of this modification would be similar in our setting and would only further encourage sharing in the presence of homophily (see Section 4).

[13] The feature that there are weaker incentives to expose misinformation in the initial phase is also supported by the data. For example, moderators on Reddit typically concentrate the majority of their time on "hot" articles and not content that does not get shared often. There are also "community awards" that Reddit moderators can use to entice Reddit users to engage in responsible behavior (e.g., fact-checking) of top posts.

expected payoffs from a share cascade in the initial phase and viral phase, respectively, when it is common knowledge that the focal article is truthful (and all agents share):

(i) *Expected initial phase share cascade,* $v_{initial}$: Recall that for agents in the initial phase, the article gets shared with exactly one other agent, until the first Poisson clock ticks, and then the article enters the viral phase. The probability the clock does not tick after each share is $e^{-\lambda_1}$. If the clock ticks at random time τ_1, the share cascade expected in the initial phase (before the clock ticks and transitions to the viral phase) is given by:

$$\kappa \sum_{\tau_1=1}^{\infty} \beta^{\tau_1-1}\tau_1(1-e^{-\lambda_1})e^{-\lambda_1(\tau_1-1)} = \frac{\kappa(1-e^{-\lambda_1})}{(1-\beta e^{-\lambda_1})^2} \tag{1}$$

Next, we consider the share cascades that occur *after* the article enters the viral phase (which it is guaranteed to under the assumption all agents share). This corresponds to the expected number of nodes in a branching process with parameter γ, discounted after each level by $\beta e^{-\lambda_2}$, where $e^{-\lambda_2}$ captures the likelihood of the viral phase continuing after each period. Thus, the expected (initial) share cascade from the viral phase is:

$$\kappa \sum_{\tau_1=1}^{\infty}\sum_{\tau_2=1}^{\infty} \beta^{\tau_1+\tau_2-1}\left(\sum_{\tau'=1}^{\tau_2}\gamma^{\tau'}\right) e^{-\lambda_1(\tau_1-1)}(1-e^{-\lambda_1})e^{-\lambda_2(\tau_2-1)}(1-e^{-\lambda_2})$$

$$= \frac{\kappa\beta(1-e^{-\lambda_1})(1-e^{-\lambda_2})}{(1-\beta e^{-\lambda_2})(1-\beta e^{-\lambda_1})(1-\beta\gamma e^{-\lambda_2})} \tag{2}$$

with $v_{initial}$ being the sum of both of the above expressions.

(ii) *Expected viral phase share cascade,* v_{viral}: Similar to the above analysis, the expected share cascade is the expected number of nodes in a branching process with parameter γ discounted according to $\beta e^{-\lambda_2}$ after each level. In particular, the expected viral phase share cascade is given by:

$$v_{viral} = \kappa \sum_{\tau_2=1}^{\infty} \beta^{\tau_2-1}\left(\sum_{\tau'=1}^{\tau_2}\gamma^{\tau'}\right) e^{-\lambda_2(\tau_2-1)}(1-e^{-\lambda_2}) = \frac{\kappa\gamma(1-e^{-\lambda_2})}{(1-\beta e^{-\lambda_2})(1-\beta\gamma e^{-\lambda_2})}$$

Because the sharing payoff is relatively small in the initial phase, given that agents are only sharing with one neighbor, we assume $v_{initial} < K$. This ensures that while the article is gaining traction, agents will either choose to share or simply ignore (kill) the content, but will not invest resources to verify the article. In other words, agents decide between the two most extreme actions; those that agree with the article share it and those who do not simply ignore it, and no agents investigate the veracity of the article. Our assumption is consistent with Bakshy et al. (2015), who show that many users on Facebook either engage with content substantially or not at all—in the terminology of the current paper, they either share or kill the article.

On the other hand, in the viral phase, sharing is much more lucrative because of the anticipation of large share cascades, which will be the case when $v_{viral} > K$. These two inequalities can hold simultaneously only when there is a substantial gap between $v_{initial}$ and v_{viral}. Figure 4 depicts a sample simulation of $v_{initial}$ and v_{viral}, with the line at K separating the two, as we have assumed.

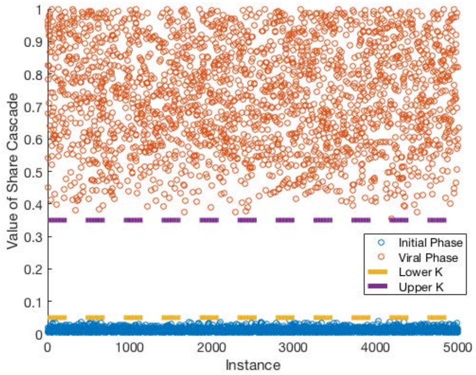

Figure 4. Blue = $v_{initial}$; Orange = v_{viral}; Yellow = Minimal K separation; Purple = Maximal K separation.

Agents additionally derive utility from debunking well-known—in our context viral—articles that contain misinformation. However, if $\delta < K$, even an agent who knows the article to contain misinformation *for a fact* would not inspect it. For this reason, we also impose $\delta > K$. In summary, we impose:

Assumption 1. Parameter values satisfy $v_{initial} < K < \min\{v_{viral}, \delta\}$.

Assumption 1 guarantees that agents in the initial phase decide between sharing and killing, and in the viral phase between sharing and inspecting. Thus, the sharing process can end only in the initial phase if an article is killed by an agent who strongly disagrees with it. This might constitute a *type-I error*, because the article may or may not actually contain misinformation. It is also possible that an agent shares an article with misinformation, which constitutes a *type-II error*. Because there will be no killing without inspection in the viral phase, there can only be type-II errors at this stage.

2.5 Information Structure and Solution Concept

We assume that agents are not aware of calendar time, the prior sharing process, or any social network connections beyond their own.[14] However, because agents do know their own social connections (e.g., their "friends"), they will have *some* information about who shared the article with her.

When receiving an article from another agent j, agent i has access to the share behavior of agent j, and thus can observe how many *other* agents also received the article from agent j. Such information can be deduced easily on social media sites like Facebook by seeing how many "likes" or "shares" your friend received on her post. This is depicted pictorially in Figure 5.

[14] This assumption is consistent with the findings in Breza et al. (2018), who demonstrate that, in practice, agents rarely know much about social interactions beyond their direct neighborhoods. It is also consistent with the fact that most users do not seem to have a detailed understanding of how the newsfeed works (see Pew Research Center (2014)).

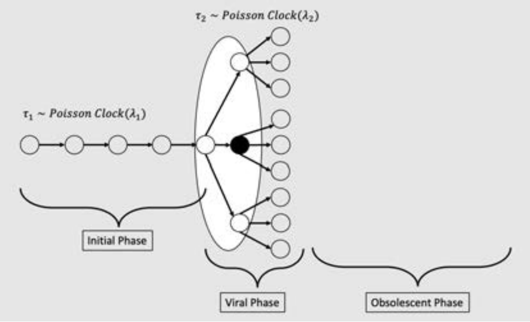
$\tau_2 \sim$ Poisson Clock(λ_2)
$\tau_1 \sim$ Poisson Clock(λ_1)
Initial Phase
Viral Phase
Obsolescent Phase

Figure 5. The dark agent observes only what is in white shading (nothing in the shadows). She sees the agent who shared it with her, and how many agents she shared it with, but nothing else.

For our solution concept, we concentrate on (pure-strategy) sequential equilibria of this game. In particular, the information set I for agent i specifies the phase and the fact that the article is still alive by the time agent i receives it. For each information set, agents hold a probability distribution (belief assessment Π) over calendar time and past actions of agents determined by Bayes' rule. Let $\sigma = (\sigma_{initial}, \sigma_{viral})$ be strategy profiles in the initial and viral phases, respectively. We say that strategy profile σ is *sequentially rational* if and only if at each information set I, the agent who moves at I chooses the action (kill, inspect, or share) that maximizes her expected utility given belief assessment Π and given that all agents play according to σ in the continuation game. Then a sequential equilibrium is a pair (σ, Π) such that there exists a sequence (σ^d, Π^d) where (i) $\lim_{d\to\infty}(\sigma^d, \Pi^d) \to (\sigma, \Pi)$, (ii) $\sigma^d_{initial}$ is totally mixed between killing and sharing and σ^d_{viral} is totally mixed between inspecting and sharing, (iii) Π^d is derived from σ^d using Bayes' rule, and (iv) (σ, Π) is sequentially rational. We often will discuss the equilibrium strategy profile σ without explicitly providing Π.

3 Equilibria in General Networks

In this section, we characterize the structure of equilibria for any stochastic network structure $\mathbf{P}$. To do so, we fix the message m of the article received by the seed agent at $t = 1$. By the MLRP, $f(m|\theta = R) = f(m|\theta = L)$ for a unique m, which we set to 0, which is without loss of generality (see footnote 5), dividing the message space of the article into left-wing and right-wing news. We thus refer to the article as *left-wing* when $m < 0$ and *right-wing* when $m > 0$.

3.1 Cutoff Form

When an agent i believes the article is truthful with some *ex-ante* probability $\tilde{q}$,[15] she updates her *ex-post* belief, π_i, of the article's truthfulness depending on the message m according to Bayes' rule:

$$\pi_i = \frac{(f(m|\theta = R)b_i + f(m|\theta = L)(1 - b_i))\tilde{q}}{(f(m|\theta = R)b_i + f(m|\theta = L)(1 - b_i))\tilde{q} + (f(m|\theta = L)b_i + f(m|\theta = R)(1 - b_i))(1 - \tilde{q})}$$

[15] In general, $\tilde{q} \geq q$ because agents anticipate that the article may have been inspected in the past.

Note that π_i is increasing in b_i when the message is right-wing ($m > 0$) but decreasing in b_i when the message is left-wing ($m < 0$). This is intuitive, an agent is more likely to believe an article to be truthful when its message agrees with her prior. Observe also that if an agent's belief of truthfulness increases, then sharing becomes more attractive relative to inspecting or killing.

The following definition leverages this idea:

Definition 1. We say that agent i employs a *cutoff strategy* if there exist b_i^*, b_i^{**} such that:

1. If $m > 0$ (right-wing news) and agent i receives the article in the initial phase, she kills if $b_i < b_i^*$ and shares if $b_i > b_i^*$, whereas if agent i receives the article in the viral phase, she inspects if $b_i < b_i^{**}$ and shares if $b_i > b_i^{**}$.

2. If $m < 0$ (left-wing news) and agent i receives the article in the initial phase, she kills if $b_i > b_i^*$ and shares if $b_i < b_i^*$, whereas if agent i receives the article in the viral phase, she inspects if $b_i > b_i^{**}$ and shares if $b_i < b_i^{**}$.

We say that an (equilibrium) action profile is in cutoff strategies if all agents i employ a cutoff strategy.

Theorem 1. *There exists a cutoff-strategy equilibrium and all equilibria are in cutoff strategies.*

Proof: See Appendix A.

The intuition for Theorem 1 is straightforward. All equilibria are in cutoff strategies because the (posterior) belief of agent i, π_i, that the article is truthful is monotone in her ideology b_i, and the expected utility of sharing is monotone in π_i. Thus, a cutoff (in the ideology or prior space) determines when an agent's best response is to share rather than kill (in the initial phase) or share rather than inspect (in the viral phase). Existence then follows directly as a consequence of Brouwer's fixed point theorem.

Theorem 1 implies that, for any given equilibrium, we know almost surely the equilibrium action of agent i if we know her prior b_i. Note because agent i's prior is drawn randomly from H_i, this induces a probability distribution over the actions of agent i in equilibrium. In particular, when $m > 0$ we know the probability that the agent shares and kills in the initial phase is $1 - H_i(b_i^*)$ and $H_i(b_i^*)$, respectively. Similarly, the agent shares and inspects in the viral phase with probabilities $1 - H_i(b_i^{**})$ and $H_i(b_i^{**})$. These probabilities are flipped when $m < 0$.

Because of the symmetry of the analysis between left-wing and right-wing news, and to avoid any confusion with alternating back and forth, in what follows we fix, without any loss of generality, $m > 0$.[16]

[16] In particular, the same analysis for $m < 0$ holds if one considers $\tilde{f}(m) = f(-m)$ (i.e., a reflection of f across the line $m = 0$), message $\tilde{m} = -m > 0$, and "flips" all the priors of the agents, $\tilde{b}_i = 1 - b_i$.

3.2 All-Share Equilibrium

The most "viral" equilibrium is when all agents share regardless of their prior, i.e., $(b_1^*, b_1^{**}, \dots, b_N^*, b_N^{**}) = \mathbf{0}$ (of course, such an equilibrium may not exist). We call this the *all-share* equilibrium.

Our first observation is that given an all-share equilibrium in the viral phase, there is also an all-share equilibrium for agents in the initial phase:

Lemma 1. *If there is an equilibrium with all-share in the viral phase, then there is an all-share equilibrium (in both phases).*

The proof of this lemma, like all others unless indicated otherwise, is in the (online) Appendix B. (We provide the proofs of Theorems 1, 2, 3 and 7 in Appendix A.)

Lemma 1 proves that to establish existence of an all-share equilibrium in both phases, it is sufficient to check whether an all-share equilibrium can be sustained in just the viral phase. Thus, all-share equilibrium existence only depends on whether sharing is a best response for viral phase agents, *given* that all other viral phase agents are also sharing.

We now give a sketch of the argument for Lemma 1. First, the cutoffs employed by initial-phase agents have no bearing on the strategic decisions of viral-phase agents, because conditional on receiving the article, viral-phase agents know all initial-phase agents must have shared. Thus, if there is an equilibrium with all-share in the viral phase, then any equilibrium cutoffs supported in the initial phase are an equilibrium for both phases. Hence, to establish Lemma 1, it is enough to show that the payoff of playing $a_i = \mathcal{S}$ in the initial phase is positive (and thus is a better response that $a_i = \mathcal{K}$ which gives zero payoff), conditional on all other agents playing $\mathcal{S}$.

Consider some agent i acting in the initial phase when all other agents are sharing. Recall that agent i gets the benefit $v_{initial}$ from share cascades in an all-share equilibrium. We can express $v_{initial} = v^1_{initial} + v^2_{initial}$, where $v^1_{initial}$ is given by equation (1), while $v^2_{initial}$ is given by equation (2) and is equal to the expected share cascade from the viral phase, discounted by β^{τ_1}, and hence $v^2_{initial} = \mathbb{E}[\beta^{\tau_1} v_{viral}]$. Since all-share is an equilibrium of the viral phase, we know the payoff from sharing exceeds that of inspecting which, in turn, exceeds that of killing (Assumption 1), so $v_{viral} \geq \mathbb{E}[\beta^{\tau_2 - 1} C \cdot \mathbf{1}_{\nu=\mathcal{M}}]$—that is, the share cascade payoff exceeds the expected punishment from sharing misinformation. In comparison, the expected punishment for an initial phase agent is $\mathbb{E}[\beta^{\tau_1+\tau_2-1} C \cdot \mathbf{1}_{\nu=\mathcal{M}}] = \mathbb{E}[\beta^{\tau_1}(\beta^{\tau_2-1} C \cdot \mathbf{1}_{\nu=\mathcal{M}})]$, which is the same expected punishment for the viral-phase agents, but discounted by β^{τ_1}. Thus, it must be the case that $v^2_{initial}$ exceeds the expected punishment for the initial phase agents. Since $v^1_{initial} > 0$, $v_{initial} = v^1_{initial} + v^2_{initial}$ must exceed the expected punishment, making sharing in the initial phase a best response.

We remark that Lemma 1 does not extend to uniqueness: a unique all-share equilibrium in the viral phase does not prove it is unique in the initial phase. This is because of the strategic complementarities in sharing in the initial phase. When fewer initial-phase agents share, the payoff to sharing in the initial phase decreases because of lower share cascades. Killing may then be supported in an initial-phase equilibrium even if in the viral phase all agents are sharing.

Existence. Lemma 1 implies that to establish whether an all-share equilibrium exists, it is sufficient to consider whether an all-share equilibrium exists in the viral phase. Let $\underline{b} = \inf\{\, b \,:\, \sum_{i=1}^{N} H_i(b) > 0 \,\}$ (i.e., the lowest prior support across all N agents) and $\mathcal{L}(m) = f(m|\theta = R)/f(m|\theta = L) > 1$ denote the likelihood ratio for message m (which, by assumption, satisfies $m > 0$). We define three mathematical objects (one of which was introduced in Section 2.4):

$$1 - \underline{\pi} = \left(1 + \frac{q}{1-q} \cdot \frac{\mathcal{L}(m)\underline{b} + (1-\underline{b})}{\underline{b} + \mathcal{L}(m)(1-\underline{b})}\right)^{-1}$$

$$v_{viral} = \frac{\kappa\gamma(1-e^{-\lambda_2})}{(1-\beta e^{-\lambda_2})(1-\beta\gamma e^{-\lambda_2})}$$

$$\mathcal{C} = \delta + \frac{C(1-e^{-\lambda_2})}{1-\beta e^{-\lambda_2}}$$

We obtain the following existence result:

Theorem 2. *An all-share equilibrium exists if and only if:*

$$(1-\underline{\pi})(\mathcal{C} - v_{viral}) \leq K$$

Proof: See Appendix A.

We provide some context for the three objects used in Theorem 2:

(i) *Worst-case posterior,* $1 - \underline{\pi}$: If agent i knows that all other agents are sharing, then when she receives the article, she knows with probability 1 that no one has inspected it yet. Thus, before reading the message, the agent assigns likelihood q that the article is truthful. After reading message m, she then develops a posterior $1 - \pi_i$ about the likelihood the article contains misinformation. Because $1 - \pi_i$ is monotonically decreasing in b_i when the news is right-wing, we know the highest posterior belief the article contains misinformation is the posterior held by the most left-wing agent, who has prior $\underline{b}$. This is given by the expression $1 - \underline{\pi}$.

(ii) *Expected share cascade,* v_{viral}: This is the discounted total share value of an agent who shares in the viral phase and expects all other future agents to share. The share cascade ends only when the article transitions into the obsolescent phase as explained in Section 2.4.

(iii) *Expected punishment,* $\mathcal{C}$: Note that when all agents share, the only punishment that can occur is when the article transitions into the obsolescent phase and is inspected by an outside source. This is determined by the random clock which ticks at time τ_2 after t_i with probability $(1 - e^{-\lambda_2})e^{-\lambda_2(\tau_2 - 1)}$. To compute the expected punishment from inspection by the outside source, it is enough to calculate the expected discount factor $\mathbb{E}[\beta^{\tau_2 - 1}]$ applied to the (undiscounted) punishment C. We let $\mathcal{C}$ denote this expected discounted punishment plus δ (because the agent simultaneously sacrifices δ when sharing an article containing misinformation). Explicitly:

$$\mathcal{C} = \delta + C \sum_{\tau_2=1}^{\infty} (1 - e^{-\lambda_2}) \beta^{\tau_2 - 1} e^{-\lambda_2(\tau_2 - 1)} = \delta + \frac{C(1 - e^{-\lambda_2})}{1 - \beta e^{-\lambda_2}}$$

Therefore, $\mathcal{C} - v_{viral}$ is the "net punishment" from a share action, which is realized only if the article contains misinformation, and must be compared against the cost of inspection K. To see that the inequality in Theorem 2 is a necessary condition, suppose that $(1 - \underline{\pi})(\mathcal{C} - v_{viral}) > K$. Then the agent with prior $\underline{b}$ prefers to inspect than share when all other agents are sharing; this is a contradiction that all-share is an equilibrium. To see that is sufficient, observe that for every agent i we have that $(1 - \pi_i)(\mathcal{C} - v_{viral}) \leq (1 - \underline{\pi})(\mathcal{C} - v_{viral}) \leq K$ by the definition of $\underline{\pi}$. That is, if the agent with the lowest belief $\underline{b}$ finds it profitable to share, then all agents (with any prior b_i) find it profitable to share. So all-share is in fact an equilibrium. Finally, we note:

Corollary 1. *The existence of an all-share equilibrium does not depend on the stochastic network structure* $\mathbf{P}$.

As should be clear from Theorem 2, none of the necessary and sufficient conditions for existence of an all-share equilibrium depend on the underlying network structure. In an all-share equilibrium, share cascades do not depend on the identities (or ideologies) of the agents that receive the article before or after, and inspection is conducted only by an outside source determined by a random clock. Thus, in this most viral equilibrium, the shape of the sharing network plays little role in the spread of misinformation; we will contrast this with results presented in Sections 4 and 5.

Uniqueness. We next present our uniqueness result for all-share equilibria:

Theorem 3. *The all-share equilibrium is* unique *if both of the following conditions hold:*

$$(1 - \underline{\pi})C \leq \underline{\pi}(1 - e^{-\lambda_1})v_{viral}$$
$$(1 - \underline{\pi})(C + \delta) \leq K$$

Proof: See Appendix A.

To prove the all-share profile is the unique equilibrium profile, we establish lower bound payoffs for sharing (independent of the strategies of other agents) and argue that these must dominate the payoffs from inspecting or killing. This proves that sharing is a dominant strategy for all agents, and thus all-share must be the unique equilibrium.[17]

In the initial phase, the worst sharing payoff attainable for some agent i would occur if all initial-phase agents after her kill and all viral-phase agents after her inspect (the strategies of the agents preceding her do not affect her payoff). The initial-phase agent gets shares *only* when the first clock

[17] Note these conditions are only sufficient (not necessary) for uniqueness and that unlike with existence conditions in Theorem 2, tight uniqueness conditions may depend on the network structure.

immediately ticks and transitions into the viral phase in the following period, which occurs with probability $1 - e^{-\lambda_1}$. It then goes into the hands of a viral-phase agent who inspects it: with probability π_i it is truthful, and then gets shared until it becomes obsolete (which has expected share cascade utility v_{viral}), and with probability $1 - \pi_i$, it is revealed as containing misinformation, and agent i gets no share cascade (and is instead punished). Therefore, the worst-case payoff from sharing is $\pi_i(1 - e^{-\lambda_1})v_{viral} \geq \underline{\pi}(1 - e^{-\lambda_1})v_{viral}$ less the expected punishment. The punishment occurs if either: (i) the Poisson clock does not tick, another initial-phase agent receives it, she kills it, and then it is revealed by the outside source to be misinformation, or (ii) the Poisson clock does tick, a viral-phase agent receives it, she inspects, and it is revealed to contain misinformation then. Both yield the same undiscounted punishment C. Thus, the expected worst-case punishment is $(1 - \pi_i)C \leq (1 - \underline{\pi})C$. Since the payoff from killing is 0, every initial-phase agent in the population finds the worst-case share payoff to be positive if and only if the first condition in Theorem 3 holds.

In the viral phase, the worst sharing payoff attainable for some agent i is if all viral-phase agents who acted before her chose to share but all viral-phase agents acting after her chose to inspect. This leads to the minimal posterior belief the article is truthful (no information is gained from receiving the article) and the maximal potential for punishment. If the article is truthful, the agent receives v_{viral} regardless of whether she inspects; thus, all that needs to be compared are the expected punishment and the inspection cost. If the article contains misinformation, the agent forgoes δ (the exposing utility) and also faces punishment C by not inspecting. For all agents in the population, the punishment of sharing without inspection is upper bounded by $(1 - \underline{\pi})(C + \delta)$. Agent i can avoid this possibility by inspecting, which costs K. However, if the inspection cost exceeds the worst-case punishment, agent i would necessarily prefer to "risk it" and simply share without inspection. This is the second condition in Theorem 3.

3.3 Comparative Statics

Next, we conduct comparative statics on the existence of an all-share equilibrium in general networks. Throughout we will say, with a slight abuse of terminology, that an all-share equilibrium becomes *more likely* following a shift in (some) parameters if an all-share equilibrium is never "destroyed" by making this shift.

Basic comparative statics. We begin with some simple comparative statics on the primitives of our model:

Proposition 1. *All-share equilibria become more likely when κ and γ increase and when C and λ_2 decrease. Moreover, there exists γ^* such that if $\gamma > \gamma^*$, then all-share equilibria become more likely when β increases, whereas if $\gamma < \gamma^*$, all-share equilibria become more likely when β decreases.*

We provide some intuition for this result, along with how it connects to the past empirical literature on social media:

(i) *Social utility of share cascades*: All-share equilibria persist when κ (the marginal utility of each share) increases. As κ increases, the appeal of sharing a story, even if it is likely to contain misinformation, becomes too large to bother with inspection. Duffy et al. (2020) document this effect: online users often share a story that is "too good not to share" even when realizing it is "too good to be true."

One can interpret β as the value an agent assigns to the actions of other agents much further downstream (i.e., how much "friends of friends" matter). This includes both the value of additional shares *and* the expected punishment from sharing misinformation. When this loss C decreases the all-share equilibrium is more easily supported. However, note that shares grow at an exponential rate whereas the punishment C remains a constant. When these shares grow rapidly, the social utility from higher-order shares outweighs any potential loss from being exposed by higher-order friends. In this case, caring significantly about downstream effects tends to promote recklessness over caution.

In contrast to previous social media sites, such as MySpace and Friendster, modern social media, like Facebook and Twitter, are primarily designed to spread information and content among users (Cuthbertson et al. (2015), Edosomwan et al. (2011)). While the former focused on virtual hangouts (MySpace's slogan was "a place for friends"), according to McWilliams (2009) the latter were built to diffuse content quickly and profusely throughout one's social circle (Facebook's promise is "to help you connect and share with the people in your life"). The effusive propagation of content to friends and friends of friends, etc. amplifies the social utility from sharing content and promotes less fact-checking and more engagement. The value of "deep" share cascades is determined by the parameter β.

(ii) *Depth and size of share cascades*: All-share equilibria persist when γ (the rate at which the article spreads) increases and λ_2 (the speed of the clock transitioning to the obsolescent phase) decreases. One can interpret γ as a proxy for the *size* of the share cascade and λ_2^{-1} as a proxy for its *depth*. Figure 6, provided by Vosoughi et al. (2018), shows evidence that the spread of misinformation is more powerful and deeper than truthful articles, indicating that the accuracy of the information is less relevant for the existence of all-share equilibria when γ and λ_2^{-1} are larger.

Message space. We obtain the following comparative statics with respect to m (where, recall, we fix $m > 0$):

Proposition 2. *If $\underline{b} > 1/2$, then a more extreme message ($m' > m$) makes all-share equilibria more likely. Conversely, if $\underline{b} < 1/2$, then a more extreme message ($m' > m$) makes all-share equilibria less likely.*

In environments where $\underline{b} > 1/2$, the entire population holds right-wing beliefs, so (mis)information spreads more easily when the message advocates more strongly for ideas that agree more with the entire population. Guess et al. (2018) find evidence of this: users with strong ideological preferences in

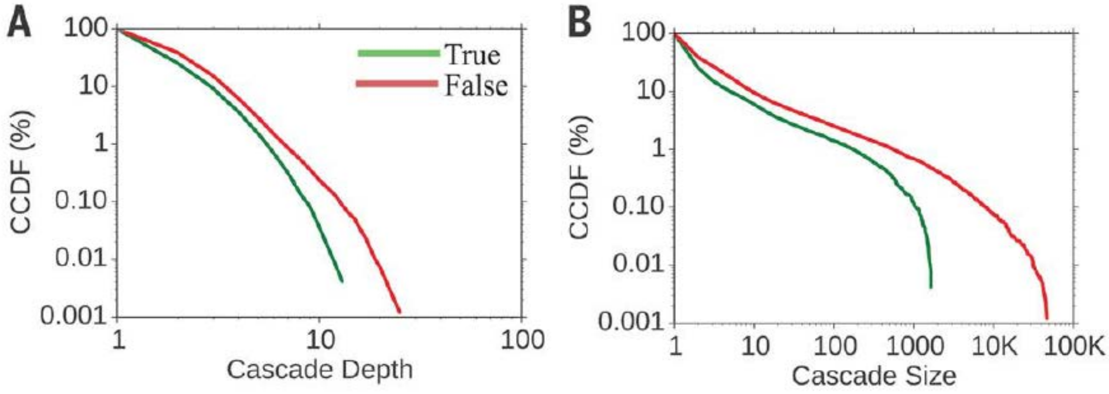

Figure 6. Depth and Size of Share Cascades from Vosoughi et al. (2018).

the 2016 presidential election consumed extreme politically-congruent content more than any other viewpoint (including moderate, but still politically-congruent, material). In environments where $\underline{b} < 1/2$, at least some of the agents hold left-wing priors, and these agents tend to believe stronger right-wing messages are *less likely* to be accurate (as they disagree *more*). Thus, in a population with at least some diversity of beliefs, moderate messages tend to support all-share more often. This result comes as a direct consequence of comparative statics on the worst-case posterior, $1 - \underline{\pi}$.

Next, we consider what it means for the distribution f of truthful messages to become *more informative*:

Definition 2. We say that $\tilde{f}$ is more informative than f if (i) $\tilde{f}(m|\theta = R)/f(m|\theta = R)$ is monotonically increasing in m and (ii) $\tilde{f}(m|\theta = L)/f(m|\theta = L)$ is monotonically decreasing in m.

A more informative message distribution $\tilde{f}$ more clearly differentiates between the R and L states for a given message m, because the likelihood ratio $\tilde{f}(m|\theta = R)/\tilde{f}(m|\theta = L)$ increases more quickly for $m > 0$ than for $f(m|\theta = R)/f(m|\theta = L)$. We observe the following comparative static with respect to informativeness:

Proposition 3. *If $\underline{b} > 1/2$, all-share equilibria are more likely if $\tilde{f}$ is more informative than f. Conversely, if $\underline{b} < 1/2$, all-share equilibria become more likely if $\tilde{f}$ is less informative than f.*

Just as with Proposition 2, the comparative statics depend on the distribution of prior beliefs within the population. As the informativeness of f increases, right-wing agents are encouraged to spread right-wing articles and left-wing agents become more skeptical of them. When the informativeness decreases, right-wing agents are less confident in the message but the left-wing agents are simultaneously less skeptical that it may contain misinformation.

One can interpret the informativeness of the message as how "correlated" or "orthogonal" the message is to state θ (in our case, the political content of the message). For instance, articles which are apolitical in nature, such as "Los Angeles Dodgers Sign 12-year, $365-million Contract Extension with Mookie Betts," have little informativeness relative to more political ones, such as "Obama Signs

Executive Order Banning The Pledge of Allegiance in Schools Nationwide" (Fourney et al. (2017)). Similar intuition as Proposition 2 shows that informativeness/political relevance fuels blind sharing with a right-wing demographic, but admits healthy skepticism with a more diverse set of prior beliefs.

Distribution of Priors. Finally, we consider the effects of perturbing the distribution of priors, H, held by the population:

Proposition 4. *When $\underline{b}$ increases, all-share equilibria become more likely.*

The intuition for Proposition 4 is clear: as the lowest prior in the population increases, the worst-case posterior that the article contains misinformation, $1-\underline{\pi}$, decreases, so right-wing misinformation spreads more easily. This result has consequences for more traditional shifts in the distribution of priors seen in the literature:

Definition 3. We say H_2 is more *right-leaning* if H_2 first-order stochastically dominates H_1. We say H_2 is *more polarized* than H_1 if it satisfies the monotone-single crossing property: $H_2^{-1}(\alpha) - H_1^{-1}(\alpha)$ is a nondecreasing function in α which crosses (zero) at $\alpha^* = 1/2$ with $H_1(1/2) = H_2(1/2) = 1/2$.

Definition 3 encompasses two notions of distributional shifts in priors: one which moves the entire belief distribution toward one state, and another which "stretches" the belief distribution around the moderate candidate (i.e., $b = 1/2$) while preserving an equal distribution of left-wing and right-wing agents (i.e., $H(1/2) = H(1/2) = 1/2$).

When H_2 is more right-leaning or less polarized, then $\underline{b}$ increases: in the former case, the entire distribution shifts to the right (including the lower support), whereas in the latter, the most extreme left-wing agents become less extreme, which increases the lower support closer to $1/2$ as well. Both have the effect of making agents more sympathetic toward right-wing news, even if it contains misinformation.

Increasing $\underline{b}$ can be associated with the rise of unified-demographic social media sites. For instance, Parler, a social media site intended for "right-wing politicians and influencers," arose out of the 2020 presidential election as a social media site for sharing right-wing ideas and has already exceeded 10 million users. Independent third-party fact-checkers have identified the site as a forum for sharing rampant misinformation that goes unchecked: "Along with its success comes the reality that extremist movements like QAnon and the Boogalooers have thrived in the platform's unregulated chaos" (according to PBS (2020)).

4 Island Networks and the Implications of Homophily

We now specialize the network $\mathbf{P}$ in order to focus on one of our main results—the effects of homophily on misinformation. Specifically, we adopt a lower dimensional model of connections, known as *island networks* (or the stochastic block model), where homophily can be parameterized simply.

In island networks, agents are partitioned into k blocks of size $N_1, N_2, \ldots, N_k$, called *islands* each with some constant (but not necessarily equal) share of the population N. Each agent i has a type $\ell_i \in \{1, \ldots, k\}$ corresponding to which block (or "island") she is in. Link probabilities are given by the standard stochastic-block model:

$$p_{ij} = \begin{cases} p_s, \text{ if } \ell_i = \ell_j \\ p_d, \text{ if } \ell_i \neq \ell_j \end{cases}$$

where $p_s \geq p_d$. We refer to this as the *homophily structure.* As a special case, we have $p_s = p_d$, which reduces to just a single-island model. On the other end of the spectrum is the *segregated islands* model, whereby $p_s > 0$ but $p_d = 0$, which corresponds to the network with maximal homophily, where agents interact only with others from their own "community"/island.

We make the following assumption about how priors are distributed on each of the islands:

Assumption 2. The prior distribution for agents on the same island is the same, H_ℓ, and the distributions across the islands can be ordered by a first-order stochastic dominance relation: $H_1 \succeq_{FOSD} H_2 \succeq_{FOSD} \cdots \succeq_{FOSD} H_k$.

Assumption 2 suggests a form of homophily where agents of similar ideological beliefs tend to interact with each other on social media more often than those with opposing beliefs. This is consistent with Bakshy et al. (2015), who show that "friend networks" on Facebook are ideologically segregated, with the median share of friends from the opposing ideology around only 20%. Other evidence from a Twitter field experiment shows that the same ideological homophily may exist on other social media sites. In particular, the connections formed between users on Twitter (in the form of "following") were strongly influenced by whether the users shared a common partisanship (Mosleh et al. (2021)).

By Theorem 1, the equilibrium set consists of cutoff-strategy equilibria of the form $(b_1^*, b_1^{**}, \ldots, b_N^*, b_N^{**})$. In fact, because of the symmetry of the island model, we obtain:

Lemma 2. *In the multiple island model, as* $\min_\ell\{N_\ell\} \to \infty$, *all equilibria are in symmetric island-dependent cutoff strategies. In other words, in every equilibrium, there exists* $\{(b_\ell^*, b_\ell^{**})\}_{\ell=1}^k$ *such that* $b_i^* = b_{\ell_i}^*$ *and* $b_i^{**} = b_{\ell_i}^{**}$ *for all agents* i.

This lemma shows that as the population becomes large, equilibrium cutoffs are no longer agent-dependent, but only island-dependent. This is true in *every* equilibrium, not just some equilibrium. Intuitively, an agent's cutoff strategy depends on the anticipated actions of others acting before and after her, given her network position. Thus, if agent i and agent j's strategies differ, each anticipates different actions before or after her, which may in turn reinforce those different strategies in equilibrium. But, when the population is large, the network positions of any two agents on the same island are identical and the influence that one's strategy has on the "average" strategy of the rest of the population is minimal, and hence their best responses must be identical.

4.1 Single-Island Model

First, we start with the uniform connection (or single-island) model, where any two agents in the population have the same probability of being linked as any other pair of agents. In particular, we assume that $p_{ij} = p \in (0,1)$ for all agents i, j. Similarly, we assume the distribution of priors is the same for all agents, i.e., $H_i = H$ for some distribution H. This is pictured in Figure 7.[18]

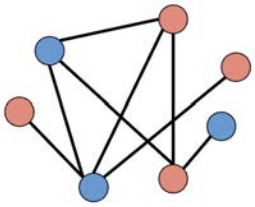

Figure 7. Sample network from uniform-connection model: agents are equally-likely to receive and share content from others of any ideology (blue = left; red = right).

A special case of Lemma 2 shows that in a single-island model, all agents employ the same cutoffs in every equilibrium. Thus, the equilibrium set can be characterized entirely by the set of (b^*, b^{**}) supported in equilibrium.

Equilibrium Structure. In our model, sharing decisions of different agents are neither strategic substitutes nor strategic complements. When others share more, an agent might worry that nobody else has inspected an article and may be tempted to inspect more and thus share less (creating an instance of strategic substitutes), or expecting others in the future not to inspect, she may be encouraged to share as well (an instance of strategic complements). We next prove that despite these mixed incentives, the equilibrium in the single-island case has a lattice structure.

Theorem 4. *The equilibrium set of cutoffs* (b^*, b^{**}) *form a lattice structure according to the natural order.*

Theorem 4 therefore establishes that there is always a minimum and maximum sharing equilibrium (we refer to these as the *extremal equilibria*). In particular, the maximum-sharing equilibrium has fewer agents killing, fewer agents inspecting, and more agents sharing (in both phases), than in any other equilibrium.

The intuition for the lattice structure in Theorem 4 is as follows. First, notice that the cutoff in the initial phase b^* does not affect the strategic decisions of agents acting in the viral phase. When agents change their cutoffs in the initial phase, they affect only the probability that the article will reach the viral phase, but not the posterior probability the article is truthful upon reaching the viral phase (by definition, no agents have inspected up until this point). Because an agent receiving the article in the viral phase knows the article did indeed reach the viral phase, the initial phase cutoff is immaterial.

[18] Observe that an equivalent formulation of the single-island model is to have k islands with prior distributions $H_1, H_2, \ldots, H_k$, but where $p_s = p_d$ (i.e., links between islands are the same as across islands). In this case, as $N \to \infty$ and connections between agents occur uniformly at random, one can replace this model with a single-island, where $H = \sum_{\ell=1}^{k} \frac{N_\ell}{N} H_\ell$.

This implies that independent of b^*, there exists a set of possible b^{**} that form an equilibrium in the viral phase.

Second, for a fixed b^{**}, initial-phase agents always play a game of strategic complementarity with each other (unlike viral-phase agents, where both strategic forces exist). Moreover, notice that b^*, the cutoff for killing, is always *increasing* in b^{**}. This is because, as more viral-phase agents inspect, misinformation becomes more likely to be detected, discouraging share cascades and increasing the threat of punishment. This economic force ensures that the set of equilibrium cutoffs is a lattice, as pictured in Figure 8.

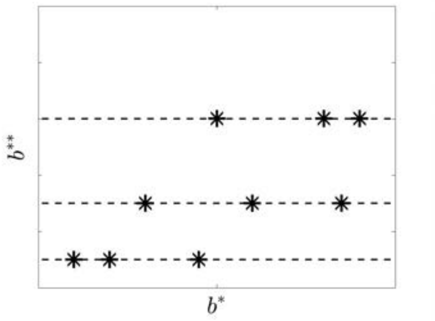

Figure 8. Example of lattice structure of (b^*, b^{**}) equilibrium cutoffs in the single-island model.

4.2 Multi-Island Model

In contrast to the single-island case, the set of equilibria is in general not a lattice in the multi-island model and a "maximal sharing equilibrium" may not exist. This motivates our definition of virality, which we now provide:

Definition 4. Consider the seeding of some agent i^* with the article at $t = 1$. Let $T_1(i^*), T_2(i^*)$ be the (random) times at which the article's lifetime ends under equilibria σ_1, σ_2, respectively, conditional on agent i^* being seeded. Call $\mathbf{S}_{T(i^*)}$ the total amount of sharing that occurs before stopping time $T(i^*)$.

(i) We say that σ_1 exhibits more *content engagement* than σ_2 if $\max_{i^*} \mathbb{E}[\mathbf{S}_{T_1(i^*)}] \geq \max_{i^*} \mathbb{E}[\mathbf{S}_{T_2(i^*)}]$ unconditional on the article's veracity.

(ii) We say that σ_1 exhibits more *viral misinformation* than σ_2 if $\max_{i^*} \mathbb{E}[\mathbf{S}_{T_1(i^*)}] \geq \max_{i^*} \mathbb{E}[\mathbf{S}_{T_2(i^*)}]$ conditional on the article containing misinformation.

Virality and engagement measure how deep an article can spread when the initial seed is chosen in a way that is most conducive to the article's spread. The difference between the two notions is based on whether the article contains misinformation. Platforms like Facebook may care about content engagement, while policymakers care about viral misinformation. In the single-island model, because there is a well-defined maximal sharing equilibrium, the equilibrium with the most content engagement and the most viral misinformation coincide. But this is not necessarily the case with multiple islands. In the rest of this section, we focus on the equilibrium with most viral misinformation, or for short, on the *most viral equilibrium*.

We next define the other key notion of this section, homophily:

Definition 5. We say (p_s, p_d) has *more homophily* than (p'_s, p'_d) if $p_s \geq p'_s$ and $p_d \leq p'_d$ (with at least one strict).

In other words, an increase in homophily corresponds to either a decrease in the across-island link probability or an increase in the within-island link probability, or both. The net effect is an overall increase in the likelihood that an agent receives the article from someone on her own island, as well as an increase in the likelihood that an agent will pass the article to other agents on her own island. Equivalently, every agent is more likely to interact with agents of congruent ideological beliefs, both before and after receiving the article. We refer to these as *echo chambers*, because content is passed between agents who agree and echo each others' decisions, which worsen as the homophily in the network increases.

As mentioned in works like Bakshy et al. (2015), echo chambers occur as a natural phenomenon when connected social media users tend to have congruent ideological beliefs. Our focus is on how these echo chambers affect the spread of content and misinformation, and for the purposes of Section 4, we take the social network $\mathbf{P}$ (determined by (k, p_s, p_d)) as given. In Section 5, we consider how the designer of the platform can shape these parameters to maximize their personal objectives.

4.3 Comparative Statics: Homophily

We investigate how changes in the homophily parameters affect the virality of misinformation in the most viral equilibrium. The effects of local changes in homophily will depend on whether strategic substitutes or strategic complements forces dominate. Recall that when strategies are *strategic substitutes*, agents want to choose opposing actions from each other; in other words, more sharing in the population creates incentives to share *less*. On the other hand, strategies are *strategic complements* when agents tend to choose similar actions, or put differently, when more sharing entices others to share *more*.

We first observe that, for a fixed strategy of the agents in the viral phase, sharing decisions in the initial phase are strategic complements. This is because sharing by other initial-phase agents increases the likelihood of share cascades and reduces the imminence of punishment from inspection by outside sources.

Strategic interactions in the viral phase are more subtle, however. Agents play a game of strategic complements with *future* agents but a game of strategic substitutes with *past* agents. When more agents inspect, the fear of future punishment from sharing is greater; this makes sharing less attractive because of future inspections, i.e., strategic complements. On the other hand, more inspections indicate to agents that an article in circulation is more likely to be accurate because past agents are likely to have inspected it already, and found the article to be truthful (i.e., the hard work of inspecting has already been done). This makes sharing less attractive when most other agents are sharing (and not inspecting), i.e., strategic substitutes. In general, this makes it ambiguous whether strategic complements or substitutes will "win out" for viral phase agents.

We next define what it means for complements to dominate ("net" complements) or substitutes to dominate ("net" substitutes). Let I_ℓ denote the fraction of agents inspecting connected to an agent on island ℓ in the viral phase:

$$I_\ell \equiv \frac{p_s i_\ell N_\ell + p_d \sum_{\ell' \neq \ell} i_{\ell'} N_{\ell'}}{p_s N_\ell + p_d \sum_{\ell' \neq \ell} N_{\ell'}}$$

Similarly, let $\Delta_\ell(b)$ be the difference in the inspecting and sharing payoff for an agent with prior b in the viral phase. There are *net strategic complements* at prior b for island ℓ if $\frac{\partial \Delta_\ell}{\partial I_\ell}(b) > 0$ whereas there are *net strategic substitutes* if $\frac{\partial \Delta_\ell}{\partial I_\ell}(b) < 0$.

Note a sufficient condition for the complements effect to dominate is that λ_2 or γ is sufficiently large (i.e., the article spreads quickly in the viral phase *or* the clock for transition to obsolescence is fast, or both). Because we think of the viral phase as relatively short-lived (fast clock, so λ_2 is large), but with deep and extensive share cascades (large γ), much of our attention in the rest of the paper will be when net strategic complements exist for viral-phase agents.[19]

With this in mind, we obtain the following result about how homophily affects viral misinformation:

Theorem 5. *Suppose there exists some island ℓ that is all-share (i.e., $b_\ell^* = b_\ell^{**} = 0$) and has net strategic complements. Then, there exist $1 < \underline{p} < \bar{p} < \infty$ such that*

(a) If $p_s/p_d < \underline{p}$ and we have two equal-sized islands, a (marginal) increase in homophily decreases the virality of misinformation;

(b) If $p_s/p_d > \bar{p}$, an increase in homophily increases the virality of misinformation.

Theorem 5 establishes that an increase in homophily has a non-monotonic effect on the spread of misinformation: with little initial homophily, more homophily is beneficial.[20] In contrast, when we start with a significant degree of homophily, increasing it further leads to more viral spread of misinformation. We next provide the intuition for these results.

In both cases, an increase in homophily means that agents are more likely to receive the article from their own island, but also more likely to pass it to their own island. The former effect pushes in the direction of strategic substitutes, while the latter is a source of strategic complements. In part (a), we start in a situation in which all islands are highly connected to each other and, by hypothesis, one of the islands has an all-share equilibrium and strategic complement effects dominate. Now consider an increase in homophily and suppose that the article in question has a right-wing message and the

[19] Additionally, in the simulations reported in Appendix B.9, we find the case of strategic complements to be much more common than strategic substitutes.

[20] This result is stated for two equally-sized islands for technical reasons, because behavior in islands with intermediate ideology turns out to be more complex.

right-wing island is at all-share. All agents now expect that the article will be shared within their island, and this motivates greater sharing in the right-wing island. In the left-wing island, on the other hand, expecting that the article will be shared more with other left-wingers, there will be more inspection. Such inspection disciplines behavior both on the left-wing island and also on the right-wing island, which in this case is well-connected with the left-wing island.

In contrast, in part (b), we start with sufficiently high homophily, so a further increase in homophily makes it more likely that agents find themselves in an echo chamber. This implies that it is now less likely that an article shared within a right-wing island will reach one of the left-wing islands and be inspected there. This removes the disciplining effect that comes from left-wing islands and encourages further sharing and exacerbates the virality of misinformation.

Another implication of Theorem 5(b) is worth noting: when there are islands with sufficiently left-wing and right-wing ideologies, then a very extreme message—a message m with a sufficiently high or a sufficiently low likelihood ratio $\mathcal{L}(m)$—will necessarily lead to all-sharing in the most extreme communities. Thus, it will be the most extreme messages that become most viral.

The results in Theorem 5 are closely linked to the empirical literature on the spread of (mis)information in the presence of echo chambers. Törnberg (2018) and Vicario et al. (2016), among others, demonstrate that homophily in sharing behavior propagates ideologically-congruent ideas, with little incentive to question the source or veracity of this information. Quattrociocchi et al. (2016) present evidence of these echo chambers on Facebook: conspiracy theories and poor science spread rapidly within polarized groups because of confirmation bias and "favored narratives." Levy (2020) shows that social media sites, like Facebook, have recently exploited echo chambers in their algorithms to create "filter bubbles." In particular, Facebook's algorithms tend to artificially create echo chambers in order to maximize engagement with content, at the cost of possibly spreading misinformation. We investigate this result relative to Theorem 5(a) in more detail in Section 5 when we discuss the incentives of a social media platform in shaping recommendation algorithms.

4.4 Comparative Statics: Extremism and Polarization

We now present related comparative statics with respect to the message m and the distribution of priors H in the population with uniform connections. Our main focus is on how *extremism* and *polarization* affect the equilibrium cutoffs of the most viral equilibrium as a function of the homophily structure.

Extremism in the message space is a measure of how strongly a message advocates for one state or the other; increasing m when $m > 0$ or decreasing m when $m < 0$ increases the extremism of the message. Similarly, an increase in polarization, as defined in Definition 3, spreads ideologies on both the left and right sides of the moderate belief $b = 1/2$ toward the more extreme ends of the spectrum, $b = 0$ (extreme left) and $b = 1$ (extreme right).

For simplicity, for the next result, we focus on the case of just two islands, a left-wing and

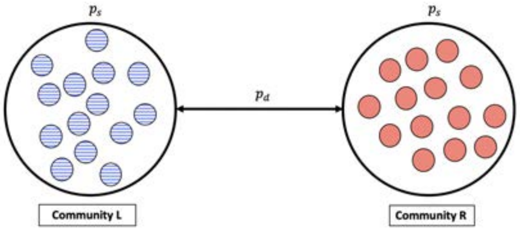

Figure 9. Two-Island Model.

a right-wing one with prior distributions H_L and H_R, respectively, as pictured in Figure 9. Our main comparative result with respect to extremism and polarization follows a similar contrast as in Theorem 5:

Theorem 6. *Let H_R and H_L have disjoint support, i.e., H_R has support on $[\underline{b}_R, \bar{b}_R]$ and H_L has support on $[\underline{b}_L, \bar{b}_L]$, with $\bar{b}_L < 1/2 < \underline{b}_R$. Moreover, suppose there are net strategic complements and the right-wing island is at all-share at the most viral equilibrium. Then:*

(a) In the single-island model (no homophily), an increase [decrease] in the extremism of the message m or in the polarization of H (weakly) decreases [increases] the virality of misinformation;

(b) In the segregated-island model (extreme homophily), an increase [decrease] in the extremism of the message m or in the polarization of H (weakly) increases [decreases] the virality of misinformation.

Theorem 6 shows that the environments where homophily ignites the spread of misinformation through echo chambers are exactly the environments where the forces of extremism and polarization are most dangerous. Similar forces as Theorem 5 drive these comparative statics. An increase in extremism and polarization lead to more discipline from the left-wing community (who disagree with the right-wing ideas even more strongly) but less scrutiny from the right-wing community. When homophily is low (Theorem 6(a)), additional discipline in one community helps control misinformation. However, once homophily is high (Theorem 6(b)), echo chambers become stronger because agents echo even more similar beliefs and actions, leading to more virality on islands where the content begins and remains for a long time.

Theorem 6(a) connects closely to the second major finding of Levy (2020), which finds that engagement with counter-attitudinal news can reduce strong attitudes about politically-congruent, extremist content. We offer a related, but slightly different insight. In social networks with little homophily, where news cuts across ideological lines, agents are exposed more to counter-attitudinal news. Simultaneously, because agents must then share this content with others of opposing ideology, strong feelings about propagating extremist (politically-congruent) content are diminished, because the likelihood of future inspection is higher. Therefore, similar to Theorem 5, messages with more extremist content or higher polarization simultaneously generate forces towards greater sharing and greater disciplining because of inspections of agents with different ideologies. However, when there is sufficiently high homophily, the latter force becomes moot and more extremist messages or greater polarization exacerbate the virality of misinformation.

5 Platform Design and Filter Bubbles

We now turn to another one of our main results: how platform behavior affects misinformation. We assume that the social media platform is interested in maximizing the engagement of users with an article. However, the platform does directly value the welfare of the agents: engagement with misinformation is equally valuable to the platform as engagement with truthful content.

Let there be k communities with disjoint prior distributions $\underline{b}_1 < \bar{b}_1 < \underline{b}_2 < \cdots < \underline{b}_k < \bar{b}_k$.[21] We also assume that the communities are *ideologically symmetric* in the sense that $\underline{b}_\ell = 1 - \bar{b}_{k-\ell+1}$ and $\bar{b}_\ell = 1 - \underline{b}_{k-\ell+1}$ holds for all ℓ. There is also at least one fully left-wing community (i.e., $\bar{b}_1 < 1/2$) and one fully right-wing community (i.e., $\underline{b}_k > 1/2$). For simplicity, we suppose the platform has an initial (political) opinion of $b_{platform}$. This is all common knowledge.

We introduce a time period $t = 0$, before the article is introduced into the population. The platform receives a collection of n articles, drawn i.i.d. from the distribution detailed in Section 2. After viewing all n articles, the platform picks an article $z \in \{1, \ldots, n\}$ and chooses whether to inspect it for truth at cost $K_P > 0$. If after an inspection the platform finds an article to be truthful, it "tags" it as verified, for example, including endorsement from a trusted third-party independent fact-checker. If it discovers it to contain misinformation, it kills this news article and selects another one.[22] We denote by z^* as the article selected by the platform (whether "tagged" as verified or not).

In addition to picking an article to recommend to users of the platform, the platform chooses how content is shared across users. That is, the platform not only picks the seed agent at $t = 1$ to whom it recommends the article z^*, it also chooses the sharing network by selecting the matrix of link probabilities $\mathbf{P}$. The platform's choice of $\mathbf{P}$ can be interpreted as its "algorithm" to determine how users are exposed to content from others in their social circle. This algorithm optimizes over which content is shared between users (of certain communities) and which content is not.[23]

The platform chooses its strategy at $t = 0$ to maximize engagement minus inspection costs. This is motivated by the fact that social media sites, like Facebook, primarily rely on advertising revenue, which becomes more valuable as users increase their activity on the site. For example, 85% of

[21] As in Section 4, we assume these communities are large, so that there are enough users with any distinct ideology that the platform need not worry about exhausting that entire community while the content is still going viral. For example, in the case of two communities, it will (almost) never occur that all right-wing agents will have seen a piece of right-wing content, incentivizing the platform to then push the content to left-wing agents.

[22] This behavior is incentive compatible for the platform, since tagging a truthful article increases its engagement, while maintaining an article with misinformation is inconsistent with paying the cost of inspection in the first place.

[23] For example, the platform starts with a sufficiently connected social network, but can then "hide" or "show" content across other users' shares as it sees fit.

The platform can also easily introduce homophily into the sharing behavior by increasing the probability that an agent shares only amongst her own community, thereby increasing the likelihood that agents only see news that is ideologically congruent. In other words, the platform's algorithm can identify which agents belong to which (ideological) communities, and then can decide how different news articles "cross-over" to different communities during the sharing process.

Facebook's total revenue in 2011 was from advertising, and from 2017-2019, around 98% was.[24] For simplicity (though also realistically), we will assume the inspection cost K_P is small relative to the profits from engagement.

5.1 Extreme Articles and Filter Bubbles

Recall that $\mathcal{L}(m)$ is the likelihood ratio of the article's message: $\mathcal{L}(m) = f(m|\theta = R)/f(m|\theta = L)$. With respect to the articles' messages, we define:

Definition 6. We say message m is *more extreme* than message m' if $\max\{1/\mathcal{L}(m), \mathcal{L}(m)\} \geq \max\{1/\mathcal{L}(m'), \mathcal{L}(m')\}$.

The extremity of the message does not depend on whether it is left-wing or right-wing, just how extreme it is. For instance, the most moderate message is at 0 with $\mathcal{L}(0) = 1/\mathcal{L}(0) = 1$, and increasing the message when $m > 0$ increases $\mathcal{L}(m) > 1$, whereas decreasing the message when $m < 0$ increases $1/\mathcal{L}(m) > 1$. We obtain a simple characterization for the platform's design as a function of the extremity of the message:

Theorem 7. *There exists η such that:*

(a) If $\max_m \max\{1/\mathcal{L}(m), \mathcal{L}(m)\} < \eta$, there exists a sequence $\{z_1, z_2, \ldots, z_n\}$ such that the platform chooses articles in this sequential order until one can be verified, "tags" it as truthful, and then adopts a uniform-connection model;

(b) If $\max_m \max\{1/\mathcal{L}(m), \mathcal{L}(m)\} > \eta$, the platform chooses the most extreme article, does not inspect it, and adopts the segregated islands connection model, i.e., $\mathbf{P}$ has maximal homophily.

Proof: See Appendix A.

Theorem 7 is one of our main results. Part (a) shows that when all articles have relatively moderate content, then the platform behaves responsibly—inspecting the articles for the veracity of their content—and also refrains from introducing algorithmic homophily. In contrast, part (b) demonstrates that when there are articles with sufficiently extremist content, then the platform selects exactly these articles and creates an extreme filter bubble—by designing its algorithms to achieve a sharing network with the greatest homophily. The result is a form of double whammy: it is precisely when articles have extremist content that the platform creates (endogenous) echo chambers, or filter bubbles, and they spread virally within these like-minded communities. Put differently, when there are extremist messages, neither the platform nor the users themselves inspect, and these news items with extremist content spread virally.[25]

[24] See Andrews (2012) and https://www.nasdaq.com/articles/what-facebooks-revenue-breakdown-2019-03-28-0

[25] When the platform can share multiple articles, which may have different degrees of extremism, it can create different sharing networks, with greater homophily for those with more extremist messages. Facebook's algorithms induce different sharing networks depending on features of the message, such as whether it contains videos or political content. See https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/.

It is also worth noting that this theorem builds on but, in many ways, significantly strengthens Theorem 5. In Theorem 5, echo chambers are not always bad for misinformation. In contrast, in part (b) of Theorem 7, homophily is always bad for viral misinformation, precisely because the platform creates homophily when there are news items with extremist content. Also in contradistinction to Theorem 5, the results in Theorem 7 do not require that strategic complements dominate locally.

In addition, Theorem 7 shows that when a platform can shape the network topology through its recommendation algorithm, the echo chamber effect that arises from Theorem 5(b) is exactly what fuels misinformation (whereas in Theorem 5(a), echo chambers were harmless). This corroborates and extends much of the previous literature mentioned earlier on the interaction between recommender systems, echo chambers, and misinformation.

Remark – In Theorem 7, we assume the platform can select any network $\mathbf{P}$ it desires through its recommendation algorithm. This is without loss of generality. If we assume the social network originally begins as an arbitrary island network in Section 4, and the platform can hide and emphasize content across different links, the same result is established.

5.2 Comparative Statics on η

The threshold η in Theorem 7 captures the conditions under which the platform is incentivized to spread misinformation and induce homophily in its sharing algorithm. When η is smaller, these incentives are stronger. Next, we perform some comparative statics on η to understand the circumstances that lend itself most to the platform's adoption of a filter bubble:

Proposition 5. *If there is (a) an increase in polarization; (b) the addition of another community; or (c) an increase in K, then η is non-increasing.*

The three parts of Proposition 5, taken together, show which conditions encourage the platform to adopt "filter bubble" algorithms. Part (a) mimics the conclusion of Theorem 6(b): as polarization increases, articles with even slight ideological bias are consumed by extremist communities with intentionally-constructed echo chambers. Filter bubbles are most advantageous to the platform in these settings. Communities with extreme beliefs (following polarization) feel more strongly about news in general, rarely second-guess politically-congruent news, but often doubt counter-attitudinal news. Increases in political polarization from social media consumption and the lack of much counter-attitudinal news allows the platform to more easily induce effective echo chambers (see Levy (2020)).

Part (b) shows that when new, ideologically-distinct communities form, the platform has an opportunity to target a whole new group of users with politically-congenial material. In part (c), when users have higher inspection costs, they are less likely to investigate content seen on the platform. Similar to the situation in the presence of extremist messages in Theorem 7, when users are less likely to inspect, the algorithm itself becomes less likely to inspect and more likely to create filter bubbles. In fact, part (c) shows the platform has disincentives to make content easy to verify (e.g., will not provide context or direct links to news stories), because inspection limits the extent to which filter bubbles can be effective.

Remark – In practice, certain demographic groups, such as users over 65 years old, appear more likely to accept and spread misinformation, perhaps because of poor media interpretation skills (see Grinberg et al. (2019) and Guess et al. (2019)). This can be introduced in our model as heterogenous inspection costs, K_i, across agents. Our results from Theorem 7 and Proposition 5 extend to this setting: the platform would now have incentives to create filter bubbles amongst communities that both have high K_i and strongly agree with the content. Moreover, in this case, more diversity of K_i would increase the virality of misinformation by providing the platform an opportunity to locate demographic groups that are least likely to fact-check.

Proposition 5 also provides a possible (albeit of course speculative) interpretation for why accelerating political polarization (part (a)) and identity politics (part (b)) in the last 10 years may have come with more aggressive filter bubble algorithms from social media sites (Apprich et al. (2018)). To take the most prominent example in this domain, it is generally agreed that Facebook's algorithms are designed to maximize commercial success, since Facebook's primary method of monetization is through ad revenue.[26] Thus, this platform's "commercial success" is closely linked to user engagement. As the recent documentary *The Social Dilemma* puts it: "The way to think about it is as 2.5 billion Truman Shows. Each person has their own reality with their own facts. Over time you have the false sense that everyone agrees with you because everyone in your news feed sounds just like you." Tellingly in this context, while Facebook cracked down on misinformation prior to the 2020 election in part due to political pressure, Facebook's algorithms have resumed promotion of misinformation in November and December of 2020: "...the measures [Facebook] could take to limit harmful content on the platform might also limit its growth: In experiments Facebook conducted last month, posts users regarded in surveys as 'bad for the world' tended to have a greater reach - and algorithmic changes that reduced the visibility of those posts also reduced users' engagement with the platform...".[27]

6 Regulation

Our analysis so far raises the natural question of what types of regulations might counter the viral spread of misinformation and platform choices leading to excessive ideological homophily. We now briefly discuss three distinct types of regulations that have been discussed in this context: (1) regulations that force platforms to reveal articles' *provenance*; (2) *censorship* (on the basis of extremist

[26] We presume that shares also serve as a good proxy for "time on platform" engagement and ad revenue for social media sites (e.g., time spent clicking on links, reading content, etc).

[27] See *Vanity Fair*:
https://www.vanityfair.com/news/2020/12/with-the-election-over-facebook-gets-back-to-spreading-misinformation and also https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

messages or misinformation); (3) *network regulations*, restricting the extent of ideological homophily or segregation introduced by platform algorithms intended to maximize engagement.

6.1 Article Provenance

For simplicity, we focus on the case in which the sharing network $\mathbf{P}$ is the single-island network. This is without much loss of generality, since, from Theorem 7, the platform will either choose uniform connections throughout the entire network or a completely segregated sharing network, in which case each island has uniform connections.

Suppose news is either *reputable* or *sketchy*. In the latter case the article is truthful with *ex-ante* probability $q_s \geq 1/2$, while in the former case, it is truthful with *ex-ante* probability of $q_r > q_s \geq 1/2$. In addition, $G(\cdot)$ follows a binomial distribution, with the probability of reputable news being ϕ and the probability of sketchy news being $1 - \phi$. The next proposition explores the implications of a regulatory policy that requires the platform to reveal the provenance (and hence quality) of the article's source.

Proposition 6. *Let H be a multinomial distribution. Then, there exists $\bar{\phi} < 1$ such that:*

(a) If $\phi > \bar{\phi}$, a policy that reveals the source of the news reduces the virality of misinformation in the most sharing equilibrium;

(b) If $\phi < 1 - \bar{\phi}$, a policy that reveals the source of the news increases the virality of misinformation in the most sharing equilibrium.

This proposition implies that if most news comes from a reputable source, a policy that requires the platform to disclose whether the source is sketchy will cut down on misinformation. The flip side of this, however, is that when most of the news is sketchy, reviewing the provenance of articles will backfire: revealing that a source is reputable hurts the spread of misinformation. The intuition for the latter result is related to the "implied truth" effect found in Pennycook et al. (2020): when sketchy news is flagged by the platform with sufficient frequency, then users tend to believe articles without the tag are more likely to be truthful relative to the case where none of the content is tagged. In our model, this is the inference Bayesian agents draw when most of the news is indeed sketchy.

The next example shows that the paradoxical case in part (b) of this proposition does not require extreme parameter values. Specifically, even if news is equally likely to come from a reputable or sketchy source, a policy of provenance revelation could be counterproductive.

Example 1. Suppose the initial phase is long (λ_1 is small) and so all-share is supported in some equilibrium (the most viral equilibrium) of the initial phase. This allows us to concentrate on the spread of misinformation in the viral phase as a consequence of provenance revelation policies.

When provenance is revealed, equilibria separates into two cases, where $q_r = 0.9$ and $q_s = 0.5$; when it is not revealed, the probability the article is truthful is $q = 0.7$. The (unique) equilibrium in all three cases is shown in Figure 10, which plots the expected payoff from inspecting less sharing, given that

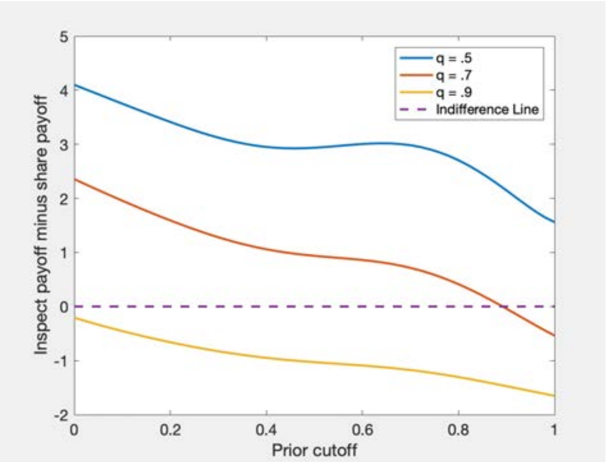

Figure 10. Equilibria for Example 1.

other agents in the population employ (prior) cutoff strategy b^{**}. When the payoff line lies above the indifference cutoff at zero, all-inspect is unique; when the payoff line lies below the indifference cutoff, all-share is unique. Otherwise, the intersection of the payoff with the indifference cutoff determines the equilibrium cutoff strategy, as implied by Theorem 1.

Figure 10 shows that when there is no revelation, because there is sufficiently high probability of the article coming from a sketchy source, over 90% of the agents inspect the article before sharing, just to be safe (when beliefs are uniformly distributed).

In contrast, if the article's provenance is revealed, inspections drop from 90% to 50% on average (recall, half of the articles are sketchy, and equilibria separate into all-share and all-inspect). This leads to a 5% likelihood that an article with misinformation is not inspected as opposed to only a 3% likelihood when the source is kept hidden. ∎

6.2 Censorship

In this subsection, we discuss the censorship policy of a planner wishing to limit misinformation. We capture the strategic interactions between the planner and the platform via a simultaneous-move game. We now describe two versions of this problem, one in which the planner stochastically observes whether an article contains misinformation, and another in which she can only identify how extreme the message of an article is. Since this is the only difference between these two versions, we provide a unified description as follows:

(i) *Planner*: The planner observes the article with some probability $\epsilon > 0$ (capturing the fact that she will have access only to a small fraction of billions of articles circulating on a typical platform). If the planner does not observe the article, there is no censoring. If the planner observes the article, she decides whether to censor it (in which case it ceases to circulate). In the first version, she receives a signal about the accuracy of the article. We suppose that there are type I errors but no type II errors (truthful articles are never misidentified, but misinformation is identified with some probability less than one). In the second version, she observes a signal about the extremism (as before, defined as $\max\{\mathcal{L}(m), 1/\mathcal{L}(m)\}$). In this version, too, there are no type II errors, and if the extremism of the article is greater than η (where η is the same as in Theorem 7), the planner identifies it with some probability less than one.

(ii) *Platform*: The platform chooses its sharing network/recommendation algorithm, $\mathbf{P}$, and decides whether to verify content as discussed in Section 5.

Corollary 2. *In both versions, there is a unique Nash equilibrium and in this equilibrium, the platform inspects all articles, adopts the uniform-connection model for* $\mathbf{P}$*, and the planner does not censor any content.*

This result follows as a straightforward corollary of Theorem 7, but its message is interesting. Censoring some of the content with the most obvious misinformation or the most extremist messages transmutes the architecture of news sharing significantly because it alters the platform's incentives. Once the platform knows that it cannot propagate these articles (which were beneficial for maximizing engagement), it opts for a sharing network with minimal homophily and it behaves responsibly itself, inspecting articles for the veracity of their content. Notably, in the case of censoring content based on extremism, the planner does not need to believe that an article with $\{\mathcal{L}(m), 1/\mathcal{L}(m)\} > \eta$ contains misinformation (or even that it is more likely to contain misinformation). Rather, this censorship policy is effective because, despite its limited reach, it modifies the platform's optimal strategy.

Remark – Note that in the equilibrium, no content is actually censored. Hence, just the threat of censorship is sufficient to remove the platform's incentives to adopt filter bubbles and spread unverified content. This corollary thus contains relevant lessons for the broader literature on censorship of misinformation on the effectiveness of the threat of (limited) censorship (see, for example, Marantz (2019), Cerf (2016), Risch and Krestel (2018)).

6.3 Network Regulations

Finally, we consider limits on ideological homophily. Such limits are similar to "ideological segregation standards", discussed in Sunstein (2018), which aim to encourage moderation by restricting the extent to which content is curated specifically to the ideology or interests of the group of users. For simplicity, we will consider the island networks of Section 4 (as Theorem 7 shows this is the class of networks from which the platform optimally chooses). In terms of our model, the regulation corresponds to a bound on the feasible homophily ratio p_s/p_d that platforms' algorithms will induce. We say that a network regulation is *binding* if it forces the platform to change its choice of $\mathbf{P}$, and we focus on the case of filter bubbles from Theorem 7(b), where there is a possibility of binding regulations.

Corollary 3. *Suppose that* $\max_m \max\{1/\mathcal{L}(m), \mathcal{L}(m)\} > \eta$ *and there is no all-share equilibrium. Then, there exists* $p^* < \infty$ *such that a regulation* $p_s/p_d \leq p^*$ *is binding for the platform and minimizes* *misinformation by inducing the platform to inspect all articles and adopt the uniform-connection model for* $\mathbf{P}$.

Note that in general $p^* > 1$ so that the planner should not necessarily eliminate all homophily (recall Theorem 5(a)). In fact, computing the exact value of p^* is far from straightforward because of the rich strategic interactions discussed in Section 4.

7 Conclusion

This paper has developed a simple model of the spread of misinformation over social media platforms. A group of Bayesian agents with heterogeneous priors receive and share news items (articles) according to a stochastic sharing network, determined by the social media platform. Articles may be truthful and thus informative about an underlying (political) state, or may contain misinformation, making them anti-correlated when the underlying state. Upon receiving an article, an agent can decide to share it with others, kill it, or inspect it at a cost to find out its veracity. Misinformation spreads when agents decide to share an article without inspecting it.

Though simple and parsimonious, the model encapsulates several rich strategic interactions. Agents receive utility from sharing an article, but suffer a (reputational) punishment if they are found out to have shared misinformation. When an agent expects others to have inspected the article, she is less likely to inspect it because misinformation is likely to have been eliminated before reaching her. This creates a force towards strategic substitutes in the game. In contrast, when an agent expects the article to face high likelihood of inspection in the future, she is more likely to inspect it herself because she does not want to be discovered to have shared misinformation. This generates a force towards strategic complements.

The ideological congruence between an agent and those in her sharing network, which we capture with the notion of homophily, matters for sharing decisions. Because individuals are more likely to inspect articles that disagree with their prior beliefs, an agent will be more cautious in sharing an article that disagrees with the views of those in her sharing network.

We provide several comparative static results for (sequential) equilibria of this dynamic game. Some of those are very intuitive, though still useful for interpreting a range of results in the emerging empirical literature on social media and misinformation. For example, we find that when a community holds exclusively right-wing (left-wing) beliefs, right-wing (left-wing) misinformation is more likely to spread. More interestingly, greater polarization tends to increase misinformation.

Of particular interest are the comparative statics with respect to homophily—measuring whether agents tend to share articles with others with similar or dissimilar beliefs. We show that when there is very limited homophily to start with, a small increase in homophily reduces the spread of misinformation. The intuition for this result can be seen from an example with two islands, one with right-wing views and the other with left-wing views, and with a news item with a right-wing slant.

With an increase in homophily, the left-wing island starts sharing the article more within itself. But since left-wing agents know that other left-wingers tend to inspect a right-wing article, they themselves become more likely to inspect because of the discipline other left-wingers impose on them and their incentives than spills over to right-wingers who become more cautious in sharing right-wing articles. This disciplining role becomes immaterial, however, when we have a high degree of homophily. Now left-wing and right-wing agents are unlikely to interact, and a further increase in homophily reassures right-wingers that the article they share will not be inspected by left-wingers. In the resulting echo chambers, misinformation spreads rapidly. (We also obtain similar non-monotonic comparative static results with respect to certain types of polarization increases).

Importantly, our framework enables a tractable study of platform incentives in designing algorithms that determine who shares with whom. To do this, we assume that the platform aims to maximize user engagement (which is a good approximation to the objectives of major social media platforms such as Facebook or Twitter), and allow it to choose the stochastic sharing network. We also allow the platform to inspect articles for veracity itself. Our main result is a striking one. When all available articles have moderate messages, the platform chooses a sharing network with minimal homophily and inspects articles to find out whether they contain misinformation. It tags truthful articles, which are then shared among users, who can then further share them, safe in the knowledge that they contain no misinformation. In striking contrast to this case, however, when there are articles with extreme messages, the platform chooses a network with maximal homophily, selects the most extreme article and never inspects it. The article spreads rapidly in the "filter bubble" the platform's algorithms have created—because now ideologically like-minded individuals know that they are unlikely to be caught sharing misinformation in their extreme echo chambers.

We also study regulations aimed at minimizing the spread of misinformation. Providing the provenance of a news item (in particular, whether it comes from a reputable or a sketchy source) can be useful, because this additional information may encourage more inspection for sketchy articles. However, we also show that this type of regulation may backfire, because tagging some of the articles from sketchy sources may make agents less likely to inspect those that are untagged—thus creating a Bayesian version of the "false sense of security". We also show that a limited type of censorship may be effective by preventing articles with the most extremist messages from circulating. This type of regulation is effective precisely because it is such extremist articles that are most likely to be the source of misinformation and, as we have seen, they also encourage platforms to create filter bubbles that maximize homophily. Interestingly, our analysis shows that even the (stochastic) threat of censorship may be enough to discipline platforms.

Our framework was purposefully chosen to be simple and several generalizations would be interesting to consider in future work. Most importantly, our assumption that agents are Bayesian rational should be viewed as a useful, albeit not fully realistic, benchmark. In our setting, it brought out certain new strategic forces—highlighting how inspection decisions can be strategic complements or substitutes, and how the degree of homophily changes agents' strategic behavior. Nevertheless, misinformation may be particularly important when agents have cognitive limitations and are only boundedly rational. Incorporating such considerations is one of the most important directions for future research. Interesting questions that emerge in this case relate to whether the platform, in addition to designing algorithms that create filter bubbles, may also choose strategies that exploit the cognitive limitations of users.

Other theoretical generalizations that might be interesting to consider include extensions to repeated interactions with incomplete information, which would enable agents to also update their beliefs about the ideological position of others in their sharing network. More challenging but arguably more interesting would be to endogenize the reputational concerns by assuming that others update about whether an individual is an extremist or has an unreliable type. In this case, the existing reputational capital of an agent will determine how likely she is to risk sharing misinformation. We can also use this extended setup with repeated interactions to study how agents update their initial political views. When there is limited misinformation, agents will gradually learn the true state. In contrast, when there is a significant probability of misinformation, agents will be uncertain about how to interpret articles that disagree with their priors and this may put a limit to learning (see Acemoglu et al. (2016)).

Despite its simplicity, our model makes several new empirical predictions, most notably related to the non-monotonic effects of homophily and polarization and to platform incentives and algorithmic decisions. Investigating these predictions empirically as well as generating new stylized facts about patterns of this information cascades on social media, is another important and exciting area for future research.

A Selected Proofs

Proof of Theorem 1. We fix the message $m > 0$ and note that a symmetric analysis applies when $m < 0$ using the reflection principle of Footnote 16. Let σ_i denote the strategy profile of agent i, which maps priors on $[0,1]$ to a distribution over actions $\{\mathcal{S}, \mathcal{I}, \mathcal{K}\}$ and $\boldsymbol{\sigma}$ as the joint strategy profile (not necessarily in cutoff strategies). Then there exists some $\rho_j(\boldsymbol{\sigma})$ which is the probability the article has been inspected upon agent j's receipt of the article, which depends on $\boldsymbol{\sigma}$. Conditional on receiving the article, the agent's posterior belief $\pi_j(m, \boldsymbol{\sigma})$ of the article being truthful is given by the system:

$$\tilde{q} = \frac{q}{(1-q)(1-\rho_j(\boldsymbol{\sigma})) + q}$$
$$\pi_j(m, \boldsymbol{\sigma}) = \frac{(f(m|\theta = R)b_j + f(m|\theta = L)(1-b_j))\tilde{q}}{(f(m|\theta = R)b_j + f(m|\theta = L)(1-b_j))\tilde{q} + (f(m|\theta = L)b_j + f(m|\theta = R)(1-b_j))(1-\tilde{q})}$$

(For simplicity, we omit the dependence of $\tilde{q}$ on $\boldsymbol{\sigma}$.) Similarly, let $\phi_{j,s}(\boldsymbol{\sigma})$ be the probability the article is inspected within j's sharing subtree for the first time at s (after the agent j shares it), which also depends on $\boldsymbol{\sigma}$. Let us consider the initial phase and viral phase separately, where we know based on the assumptions in Section 2.4, killing dominates inspecting in the former whereas inspecting dominates killing in the latter. Note that the share cascade $S_{i,\tau}$ depends on $\boldsymbol{\sigma}$ and the veracity of the article ν. However, it is easy to see that for a fixed $\boldsymbol{\sigma}$, $S_{i,\tau}(\boldsymbol{\sigma}, \mathcal{T}) \geq S_{i,\tau}(\boldsymbol{\sigma}, \mathcal{M})$ state-wise. Thus, the initial-phase agent solves:

$$a_j \in \arg \max_{a'_j \in \{\mathcal{K}, \mathcal{S}\}} u_j(a'_j) \equiv \begin{cases} 0, & \text{if } a'_j = \mathcal{K} \\ \kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1}(\pi_j(m, \boldsymbol{\sigma}) S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{T}) + (1 - \pi_j(m, \boldsymbol{\sigma})) S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{M})) & \\ -C(1 - \pi_j(m, \boldsymbol{\sigma})) \sum_{s=1}^{\infty} \beta^{s-1} \phi_{j,s}(\boldsymbol{\sigma}), & \text{if } a'_j = \mathcal{S} \end{cases} \tag{3}$$

where we observe the payoff from $a'_j = \mathcal{S}$ is monotonically increasing in $\pi_j(m, \boldsymbol{\sigma})$. For a viral-phase agent, she chooses the action:

$$a_j \in \arg \max_{a'_j \in \{\mathcal{I}, \mathcal{S}\}} u_j(a'_j) \equiv \begin{cases} \kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1} \pi_j(m, \boldsymbol{\sigma}) S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{T}) + \delta(1 - \pi_j(m, \boldsymbol{\sigma})) - K, & \text{if } a'_j = \mathcal{I} \\ \kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1}(\pi_j(m, \boldsymbol{\sigma}) S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{T}) + (1 - \pi_j(m, \boldsymbol{\sigma})) S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{M})) & \\ -C(1 - \pi_j(m, \boldsymbol{\sigma})) \sum_{s=1}^{\infty} \beta^{s-1} \phi_{j,s}(\boldsymbol{\sigma}), & \text{if } a'_j = \mathcal{S} \end{cases} \tag{4}$$

Note that $\mathcal{S}$ is preferred to $\mathcal{I}$ if and only if:

$$(1 - \pi_j(m, \boldsymbol{\sigma})) \left(C \sum_{s=1}^{\infty} \beta^{s-1} \phi_{j,s}(\boldsymbol{\sigma}) + \delta - \kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1} S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{M}) \right) \leq K$$

Either $\kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1} S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{M}) > C \sum_{s=1}^{\infty} \beta^{s-1} \phi_{j,s}(\boldsymbol{\sigma}) + \delta$, in which case the inequality holds for all values of π_j (i.e., agent j shares regardless of its prior), or is decreasing in $\pi_j(m, \boldsymbol{\sigma})$ (whereas the right-hand side is constant). Therefore, either the viral-phase agent j plays share unconditional on her prior (which is a special case of the cutoff form) or once again the payoff from $a'_j = \mathcal{S}$ is monotonically increasing in $\pi_j(m, \boldsymbol{\sigma})$. Note, rewriting $\pi_j(m, \boldsymbol{\sigma})$:

$$\pi_j(m, \boldsymbol{\sigma}) = \frac{(f(m|\theta = R)/f(m|\theta = L) b_j + (1 - b_j))\tilde{q}}{(f(m|\theta = R)/f(m|\theta = L) b_j + (1 - b_j))\tilde{q} + (b_j + f(m|\theta = R)/f(m|\theta = L)(1 - b_j))(1 - \tilde{q})} \tag{5}$$

which we see is increasing in b_j for $m > 0$ by the MLRP property (i.e., that $f(m|\theta = R)/f(m|\theta = L) > 1$). Thus the payoff for $\mathcal{S}$ over either $\mathcal{K}$ (initial phase) or $\mathcal{I}$ (viral phase) is monotonically increasing in $\pi_j(m, \boldsymbol{\sigma})$. Thus, there exist some cutoffs b_j^*, b_j^{**} for each agent j that determine whether she chooses $\mathcal{S}$ instead of $\mathcal{K}$ or $\mathcal{I}$ taking $\boldsymbol{\sigma}$ as given.

Finally, to show existence of a cutoff equilibrium, we construct the map $\psi : [0, 1]^{2N} \to [0, 1]^{2N}$ which maps a vector of cutoffs $\boldsymbol{\sigma} = (b_1^*, b_1^{**}, \ldots, b_N^*, b_N^{**})$ to another vector of cutoffs that are best responses to the first set of cutoffs. (By the argument above, the best response to any strategy profile $\boldsymbol{\sigma}$ takes the cutoff form, so this map exists.) Continuity of ψ follows from the continuity of $S_{j,\tau}(\boldsymbol{\sigma},\cdot)$, $\rho_j(\boldsymbol{\sigma})$, $\phi_s(\boldsymbol{\sigma})$, and $\pi_j(m,\boldsymbol{\sigma})$ in $\boldsymbol{\sigma}$. Thus, ψ is a continuous map from a compact set to itself, which by Brouwer's fixed point theorem implies that ψ has at least one fixed point. This fixed point, is by definition, an equilibrium. ∎

Proof of Theorem 2. Note that $\tilde{q} = q$ in Equation (5) because in an all-share equilibrium, no agent inspects and so the posterior belief the article is truthful is exactly equal to her prior. Agent j updates her belief for a message $m > 0$ as follows:

$$\begin{aligned}
1-\pi_j(m,\boldsymbol{\sigma}) &= \frac{(b_j + f(m|\theta=R)/f(m|\theta=L)(1-b_j))(1-q)}{(f(m|\theta=R)/f(m|\theta=L)b_j + (1-b_j))q + (b_j + f(m|\theta=R)/f(m|\theta=L)(1-b_j))(1-q)} \\
&= \frac{(b_j+\mathcal{L}(m)(1-b_j))(1-q)}{(\mathcal{L}(m)b_j+(1-b_j))q+(b_j+\mathcal{L}(m)(1-b_j))(1-q)} \\
&= \left(1+\frac{(\mathcal{L}(m)b_j+(1-b_j))q}{(b_j+\mathcal{L}(m)(1-b_j))(1-q)}\right)^{-1} \\
&= \left(1+\frac{q}{1-q}\cdot\frac{\mathcal{L}(m)b_j+(1-b_j)}{b_j+\mathcal{L}(m)(1-b_j)}\right)^{-1}
\end{aligned}$$

(a) *Sufficiency*: To see if an all-share equilibrium can be supported, we need to check that all agents in the viral phase have a best response to share given that all other agents are also sharing (and then leverage Lemma 1). An agent receives share utility v_{viral} when she plays $\mathcal{S}$, but faces an expected punishment of $C(1-\pi_i)\sum_{\tau_2=1}^{\infty}\beta^{\tau_2-1}(1-e^{-\lambda_2})e^{-\lambda_2(\tau_2-1)} = (1-\pi_i)(\mathcal{C}-\delta)$. Thus, her payoff from $\mathcal{S}$ is $v_{viral}-(1-\pi_i)(\mathcal{C}-\delta)$. When this agent plays $\mathcal{I}$, she shares and receives v_{viral} only if the inspection reveals that $\nu=\mathcal{T}$. If $\nu=\mathcal{M}$, the she receives δ for exposing misinformation. Thus, playing $\mathcal{I}$ gives payoff $\pi_i v_{viral}+(1-\pi_i)\delta-K$. Comparing these payoffs, we see that $\mathcal{S}$ is a best response for agent i if $(1-\pi_i)(\mathcal{C}-v_{viral})\le K$. Finally, we note that $(1-\pi_i)$ is maximized at prior belief $\underline{b}$. Thus, if this inequality holds for $\underline{b}$, it holds for all values of b that exist in the population, and sharing is a best response for all agents.

(b) *Necessity*: Suppose that $(1-\underline{\pi})(\mathcal{C}-v_{viral})>K$. Then there exists some agent in the population (an agent near prior $\underline{b}$) who prefers to play $\mathcal{I}$ instead of $\mathcal{S}$ when all other agents are sharing. Because this is a profitable deviation, an all-share equilibrium is not supported.

∎

Proof of Theorem 3. We prove this by establishing conditions under which sharing is a dominant strategy in both phases, thus making it the unique equilibrium. Consider some strategy profile $\boldsymbol{\sigma}$ played by the rest of the agents in the population and let $\boldsymbol{\sigma}^*$ be the strategy profile where all initial-phase agents kill and viral-phase agents inspect. Clearly $u_i(\mathcal{S}|\boldsymbol{\sigma}) \ge u_i(\mathcal{S}|\boldsymbol{\sigma}^*)$ for all initial-phase agents i, and thus if $\mathcal{S}$ is a best-response under $\boldsymbol{\sigma}^*$ it is a dominant strategy. If the article contains misinformation, then at time t_i+1, the article is inspected and the punishment C is realized, which occurs with probability $(1-\pi_i)$. On the other hand, if the article is truthful, with probability $1-e^{-\lambda_1}$ is transitions to the viral phase, where the share cascade v_{viral} is realized for the agent. Thus, sharing is a dominant strategy for initial-phase agent i if $(1-\pi_i)C \leq \pi_i(1-e^{-\lambda_1})v_{viral}$. Because the left-hand side is maximized when $b_i = \underline{b}$ and right-hand side is minimized also when $b_i = \underline{b}$, $\mathcal{S}$ is a dominant strategy for all agents in the population if the inequality holds when $\pi_i = \underline{\pi}$.

For the viral phase, note that $\tilde{q} \geq q$ for all strategy profiles $\boldsymbol{\sigma}$, and $u_i(\mathcal{S})$ is always increasing in q, so it is sufficient to prove $\mathcal{S}$ is dominant when agent i holds ex-ante beliefs q about the truthfulness of the article. The least share utility and greatest potential for social punishment occurs for the strategy profile $\boldsymbol{\sigma}^*$ where all agents inspect in the viral phase. If agent i in the viral phase plays $\mathcal{S}$, she receives v_{viral} if the article is truthful (probability π_i) and otherwise is punished by C in the next period. If agent i plays $\mathcal{I}$, she receives v_{viral} if the article is truthful (probability π_i) and otherwise gets δ (probability $1-\pi_i$), but always pays K to inspect. Combining these facts shows that $\mathcal{S}$ is a best response under $\boldsymbol{\sigma}^*$ if $(1-\pi_i)(C+\delta) \leq K$. Once again, the left-hand side is maximized when $b_i = \underline{b}$, so $\mathcal{S}$ is a dominant strategy for all agents in the population if the inequality holds when $\pi_i = \underline{\pi}$. ∎

Proof of Theorem 7. As the platform wishes to maximize engagement and the cost of inspection K_P is small, the platform will inspect the article unless it can choose a network $\mathbf{P}$ where the article will be shared (and not inspected or killed) amongst all agents who could receive it, with probability 1. For a fixed left-wing article (i.e., $m < 0$), π_j (the belief the article is truthful) is monotonically decreasing in agent j's prior b_j. Similarly, for a fixed right-wing article (i.e., $m > 0$), π_j is monotonically increasing in agent j's prior b_j. From the same reasoning as Theorem 1, the payoff from $\mathcal{S}$ relative to $\mathcal{I}$ and $\mathcal{K}$ (in the initial and viral phases, respectively) is increasing in the agent's belief that the content is truthful, π_j. Note that communities are assumed to be large and ideologically distinct, and given $\lambda_1, \lambda_2 < \infty$, the article will reach only a finite subset of agents within these communities, if the platform were to share it with them. Thus, if the platform does not verify the article, it will always choose $\mathbf{P}$ such that there are no links across communities, and the article will either be shared with (and remain within) community 1 or community k.

Recall the article will either be credibly verified by the platform and tagged, which would lead to an all-share equilibrium, or the platform will pick $\mathbf{P}$ and the initial seed i^* as to have an all-share equilibrium amongst the agents connected to i^*. As we saw in Proposition 3, the belief π_j for any agent j in community k is increasing in $\mathcal{L}(m)$ (because $\underline{b}_k > 1/2$). Similarly, the belief π_j for an agent j in community 1 is increasing in $1/\mathcal{L}(m)$ (because $\bar{b}_1 < 1/2$). By ideological symmetry, we observe that π_j for agent j having prior $\bar{b}_1$ or $\underline{b}_k$ with message m and likelihood ratio $\mathcal{L}(m)$ or $1/\mathcal{L}(m)$, respectively, is identical. Moreover, because community $\underline{b}_k > 1/2$, π_j for $j \in \ell_k$ is monotonically increasing in m (Proposition 4). Similarly, because $\bar{b}_1 < 1/2$, π_j for $j \in \ell_1$ is monotonically decreasing in m. Thus, there exist cutoffs for the platform, m_L^* and m_R^*, where it knows that if $m < m_L^*$ or $m > m_R^*$, sharing the article with community 1 or community k (respectively) leads to all-share within this community. Because of ideological symmetry, the cutoff is determined by $\mathcal{L}(m)$ and $1/\mathcal{L}(m)$, with the former being the cutoff for right-wing messages (shared with community k) and the latter being a cutoff for left-wing messages (shared with community 1).

Finally, note that if no article exists such that community 1 or community k has an all-share equilibrium, the platform verifies the content. If it fails to do this, $\mathbb{E}[\mathbf{S}_{T(i^*)}]$ is not maximized because with positive probability the article both contains misinformation and it is inspected by a user of the platform. The platform will verify the articles sequentially based on its own belief the articles are truthful, $b_{platform}$ (e.g., if $b_{platform} > 1/2$, it will verify the messages in order from greatest to least). Once the article is tagged as verified, the share cascade is determined by how many agents the article reaches before the obsolescent phase (as there will be no inspections or kills). Thus, the platform chooses $\mathbf{P}$ such that every agent has a complete neighborhood as to maximize the likelihood the article is passed to other agents who have not already received it. In particular, it is optimal to simply pick $\mathbf{P}$ to be the uniform-connection model with sufficiently high link probability. ∎

References

Abramowitz, Alan I. (2010), *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. Yale University Press.

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz (2016), "Fragility of asymptotic agreement under bayesian learning." *Theoretical Economics*, 11, 187–225.

Acemoglu, Daron, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar (2013), "Opinion Fluctuations and Disagreement in Social Networks." *Mathematics of Operations Research*, 38, 1–27. Publisher: INFORMS.

Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi (2010), "Spread of (mis)information in social networks." *Games and Economic Behavior*, 70, 194–227.

Allcott, Hunt and Matthew Gentzkow (2017), "Social media and fake news in the 2016 election." *Journal of economic perspectives*, 31, 211–36.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts (2020), "Evaluating the fake news problem at the scale of the information ecosystem." *Science Advances*, 6, eaay3539. Publisher: American Association for the Advancement of Science Section: Research Article.

Allon, Gad, Kimon Drakopoulos, and Vahideh Manshadi (2019), "Information Inundation on Platforms and Implications." In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, 555–556, Association for Computing Machinery, New York, NY, USA.

Altay, Sacha, Anne-Sophie Hacquin, and Hugo Mercier (2020), "Why do so few people share fake news? It hurts their reputation." *New Media & Society*, 1461444820969893. Publisher: SAGE Publications.

Andrews, Lori (2012), "Facebook is using you." *The New York Times*, 4.

Apprich, Clemens, Florian Cramer, Wendy Hui Kyong Chun, and Hito Steyerl (2018), *Pattern Discrimination*. meson press. Accepted: 2020-05-04T14:50:47Z.

Aral, Sinan and Paramveer S. Dhillon (2018), "Social influence maximization under empirical influence models." *Nature Human Behaviour*, 2, 375–382. Number: 6 Publisher: Nature Publishing Group.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic (2015), "Exposure to ideologically diverse news and opinion on Facebook." *Science*, 348, 1130–1132. Publisher: American Association for the Advancement of Science Section: Report.

Breza, Emily, Arun G. Chandrasekhar, and Alireza Tahbaz-Salehi (2018), "Seeing the forest for the trees? An investigation of network knowledge." *arXiv:1802.08194 [physics, stat]*. ArXiv: 1802.08194.

Buchanan, Tom (2020), “Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation.” *PLOS ONE*, 15, e0239666. Publisher: Public Library of Science.

Budak, Ceren, Divyakant Agrawal, and Amr El Abbadi (2011), “Limiting the spread of misinformation in social networks.” In *Proceedings of the 20th international conference on World wide web*, WWW ’11, 665–674, Association for Computing Machinery, New York, NY, USA.

Candogan, Ozan and Kimon Drakopoulos (2017), “Optimal Signaling of Content Accuracy: Engagement vs. Misinformation.” SSRN Scholarly Paper ID 3051275, Social Science Research Network, Rochester, NY.

Centola, Damon (2010), “The spread of behavior in an online social network experiment.” *science*, 329, 1194–1197. Publisher: American Association for the Advancement of Science.

Centola, Damon and Michael Macy (2007), “Complex Contagions and the Weakness of Long Ties.” *American Journal of Sociology*, 113, 702–734. Publisher: The University of Chicago Press.

Cerf, Vinton G. (2016), “Information and misinformation on the internet.” *Communications of the ACM*, 60, 9–9.

Chen, Li and Yiangos Papanastasiou (2019), “Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation.” SSRN Scholarly Paper ID 3456139, Social Science Research Network, Rochester, NY.

Cuthbertson, Richard, Peder Inge Furseth, and Stephen J. Ezell (2015), “Facebook and MySpace: The Importance of Social Networks.” In *Innovating in a Service-Driven Economy: Insights, Application, and Practice* (Richard Cuthbertson, Peder Inge Furseth, and Stephen J. Ezell, eds.), 145–158, Palgrave Macmillan UK, London.

Duffy, Andrew, Edson Tandoc, and Rich Ling (2020), “Too good to be true, too good not to share: the social utility of fake news.” *Information, Communication & Society*, 23, 1965–1979. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2019.1623904.

Eckles, Dean, René F. Kizilcec, and Eytan Bakshy (2016), “Estimating peer effects in networks with peer encouragement designs.” *Proceedings of the National Academy of Sciences*, 113, 7316–7322. Publisher: National Academy of Sciences Section: Colloquium Paper.

Edosomwan, Simeon, Sitalaskshmi Kalangot Prakasan, Doriane Kouame, Jonelle Watson, and Tom Seymour (2011), “The history of social media and its impact on business.” *Journal of Applied Management and entrepreneurship*, 16, 79–91.

Fiorina, Morris P., Samuel A. Abrams, and Jeremy C. Pope (2008), “Polarization in the American Public: Misconceptions and Misreadings.” *The Journal of Politics*, 70, 556–560. Publisher: The University of Chicago Press.

Fourney, Adam, Miklos Z. Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz (2017), “Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election.” In *CIKM*, volume 17, 6–10.

Gentzkow, Matthew and Jesse M. Shapiro (2006), “Media Bias and Reputation.” *Journal of Political Economy*, 114, 280–316. Publisher: The University of Chicago Press.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer (2019), “Fake news on Twitter during the 2016 U.S. presidential election.” *Science*, 363, 374–378. Publisher: American Association for the Advancement of Science Section: Research Article.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker (2019), “Less than you think: Prevalence and predictors of fake news dissemination on Facebook.” *Science Advances*, 5, eaau4586. Publisher: American Association for the Advancement of Science Section: Research Article.

Guess, Andrew, Brendan Nyhan, and Jason Reifler (2018), “Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign.” *European Research Council*, 9, 4.

Guriev, Sergei, Emeric Henry, and Ekaterina Zhuravskaya (2020), “Checking and Sharing Alt-Facts.” SSRN Scholarly Paper ID 3603969, Social Science Research Network, Rochester, NY.

Hsu, Chin-Chia, Amir Ajorlou, and Ali Jadbabaie (2020), “News Sharing, Persuasion, and Spread of Misinformation on Social Networks.” SSRN Scholarly Paper ID 3391585, Social Science Research Network, Rochester, NY.

Kamenica, Emir (2019), “Bayesian Persuasion and Information Design.” *Annual Review of Economics*, 11, 249–272. _eprint: https://doi.org/10.1146/annurev-economics-080218-025739.

Kamenica, Emir and Matthew Gentzkow (2011), “Bayesian Persuasion.” *American Economic Review*, 101, 2590–2615.

Keppo, Jussi, Michael Jong Kim, and Xinyuan Zhang (2019), “Learning Manipulation Through Information Dissemination.” *Available at SSRN 3465030*.

Lazer, David MJ, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, and David Rothschild (2018), “The science of fake news.” *Science*, 359, 1094–1096. Publisher: American Association for the Advancement of Science.

Lee, Chei Sian, Long Ma, and Dion Hoe-Lian Goh (2011), “Why Do People Share News in Social Media?” In *Active Media Technology* (Ning Zhong, Vic Callaghan, Ali A. Ghorbani, and Bin Hu, eds.), Lecture Notes in Computer Science, 129–140, Springer, Berlin, Heidelberg.

Levy, Roee (2020), “Social Media, News Consumption, and Polarization: Evidence from a Field Experiment.” SSRN Scholarly Paper ID 3653388, Social Science Research Network, Rochester, NY.

Marantz, Andrew (2019), *Antisocial: Online Extremists, Techno-Utopians, and the Hijacking of the American Conversation*. Penguin. Google-Books-ID: wG1QwgEACAAJ.

McWilliams, Jenna (2009), “How Facebook beats MySpace.” Section: Opinion.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G. Rand (2021), “Shared partisanship dramatically increases social tie formation in a twitter field experiment.” *Proceedings of the National Academy of Sciences*, 118.

Mostagir, Mohamed, Asuman E. Ozdaglar, and James Siderius (2019), “When is Society Susceptible to Manipulation?” SSRN Scholarly Paper ID 3474643, Social Science Research Network, Rochester, NY.

Mostagir, Mohamed and James Siderius (2021), “Inequality in Social Learning.” *Working Paper*.

Nguyen, Nam P., Guanhua Yan, My T. Thai, and Stephan Eidenbenz (2012), “Containment of misinformation spread in online social networks.” In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci ’12, 213–222, Association for Computing Machinery, New York, NY, USA.

Papanastasiou, Yiangos (2018), “Fake News Propagation and Detection: A Sequential Model.” SSRN Scholarly Paper ID 3028354, Social Science Research Network, Rochester, NY.

PBS (2020), “Right-wing users flock to Parler as social media giants rein in misinformation.” Section: Nation.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David G. Rand (2020), “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings.” *Management Science*, 66, 4944–4957. Publisher: INFORMS.

Pennycook, Gordon and David G. Rand (2018), “Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning Than by Motivated Reasoning.” SSRN Scholarly Paper ID 3165567, Social Science Research Network, Rochester, NY.

Pew Research Center (2014), “Political Polarization in the American Public.”

Prior, Markus (2013), “Media and Political Polarization.” *Annual Review of Political Science*, 16, 101–127. _eprint: https://doi.org/10.1146/annurev-polisci-100711-135242.

Quattrociocchi, Walter, Antonio Scala, and Cass R. Sunstein (2016), “Echo Chambers on Facebook.” SSRN Scholarly Paper ID 2795110, Social Science Research Network, Rochester, NY.

Risch, Julian and Ralf Krestel (2018), “Delete or not delete? Semi-automatic comment moderation for the newsroom.” In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 166–176.

Sunstein, Cass R. (2018), *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press. Google-Books-ID: oVBLDwAAQBAJ.

Taylor, Sean J. and Dean Eckles (2018), "Randomized Experiments to Detect and Estimate Social Influence in Networks." In *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks* (Sune Lehmann and Yong-Yeol Ahn, eds.), Computational Social Sciences, 289–322, Springer International Publishing, Cham.

Törnberg, Petter (2018), "Echo chambers and viral misinformation: Modeling fake news as complex contagion." *PLOS ONE*, 13, e0203958. Publisher: Public Library of Science.

Vicario, Michela Del, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi (2016), "The spreading of misinformation online." *Proceedings of the National Academy of Sciences*, 113, 554–559. Publisher: National Academy of Sciences Section: Physical Sciences.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018), "The spread of true and false news online." *Science*, 359, 1146–1151. Publisher: American Association for the Advancement of Science Section: Report.

B Online Appendix (not for publication)

B.1 Auxiliary Lemmas

Lemma B.1. *In every (sequential) equilibrium with cutoffs $(b_1^*, b_1^{**}, \ldots, b_N^*, b_N^{**})$, the viral-phase cutoffs $(b_1^{**}, \ldots, b_N^{**})$ are best responses to any set of initial-phase cutoffs $(\tilde{b}_1^*, \ldots, \tilde{b}_N^*)$.*

Proof of Lemma B.1. In a sequential equilibrium, every viral-phase assigns possibly vanishing (but always non-zero) probability to the information set I that the article is not killed during the initial phase. The cutoffs played during the initial phase, conditional on viral-phase agent i's receipt of the article, have no bearing of her belief about the article's veracity nor the payoffs she will receive following her actions. Because this information node is reached with positive probability, her best response cannot depend on the actions of agents in the initial phase. ∎

Lemma B.2. *For a single-island model, the most-sharing equilibrium cutoff in the viral phase takes one of the following three forms: (i) all-share, (ii) all-inspect, or (iii) $b^{**} \in (0,1)$ such that if $w(b)$ denotes $\Delta(b)$ (the difference in inspecting and sharing payoff) when b itself is the viral phase cutoff, then dw/db at cutoff b^{**} is negative.*

Proof of Lemma B.2. Consider the curve $w(b)$ mapping prior cutoffs to the difference between inspect and share payoffs. All intersections of the form $w(b^{**}) = 0$ can be supported in some equilibrium for the viral phase cutoff b^{**}, by Lemma B.1 and Theorem 1. Moreover, if $w(0) < 0$, then all-share is an equilibrium (in both phases by Lemma 1), which is most viral. Thus, it is sufficient to consider $w(0) > 0$. If $w(b) > 0$ for all b, then all-inspect is the unique strategy profile for the viral phase played in every equilibrium (via Lemma B.1), so must be the most viral one. Otherwise, $w(b)$ crosses 0 for the first-time from $w(0) > 0$, which must be from above. Thus, the slope of the tangent line at this first intersection (which is the most viral equilibrium) must negative. ∎

Lemma B.3. *An increase in polarization of beliefs can always be constructed via the following process: take every belief b_i and either (i) add some $\epsilon_i > 0$ to b_i if $b_i > 1/2$, or (ii) subtract some $\epsilon_i > 0$ to b_i if $b_i < 1/2$.*

Proof of Lemma B.3. For part (i), note that $H_1(b_i^1) = \alpha > 1/2$, so by single-crossing at $H_1^{-1}(1/2) = H_2^{-1}(1/2)$, we know that $H_2^{-1}(\alpha) - H_1^{-1}(\alpha) > 0$. Thus, for some $b_i^2 > b_i^1$, we have $H_2^{-1}(\alpha) = b_i^2$, or in other words, $H_2(b_i^2) = \alpha$. Setting $\epsilon_i = b_i^2 - b_i^1 > 0$ in this fashion for all $b_i > 1/2$ accomplishes claim (i). For part (ii), note that $H_1(b_i^1) = \alpha < 1/2$, so by single-crossing at $H_1^{-1}(1/2) = H_2^{-1}(1/2)$, we know that $H_2^{-1}(\alpha) - H_1^{-1}(\alpha) < 0$. Thus, for some $b_i^2 < b_i^1$, we have $H_2^{-1}(\alpha_1) = b_i^2$, or in other words, $H_2(b_i^2) = \alpha$. Setting $\epsilon_i = b_i^1 - b_i^2 > 0$ in this fashion for all $b_i < 1/2$ accomplishes claim (ii). ∎

B.2 Omitted Proofs from Section 3

Proof of Lemma 1. By Lemma B.1, we know that it is sufficient to show that when all agents play $a_i = \mathcal{S}$, it is a best response for agent j acting in the initial phase to play $a_j = \mathcal{S}$, regardless of her prior b_j. By

Assumption 1, we know that killing dominates inspecting for the initial-phase agent, so it is enough to prove the payoff of sharing is positive, *conditional* on it being a best response for agent i in the viral phase to play $\mathcal{S}$ when all other agents in the viral phase play $\mathcal{S}$ as well.

Note that agent i gets share utility $v_{initial}$ given by:

$$\begin{aligned}
v^1_{initial} &= \frac{\kappa(1-e^{-\lambda_1})}{(1-\beta e^{-\lambda_1})^2} > 0 \\
v^2_{initial} &= \kappa \sum_{\tau_1=1}^{\infty}\sum_{\tau_2=1}^{\infty} \beta^{\tau_1+\tau_2-1}\left(\sum_{\tau'=1}^{\tau_2}\gamma^{\tau'}\right) e^{-\lambda_1(\tau_1-1)}(1-e^{-\lambda_1})e^{-\lambda_2(\tau_2-1)}(1-e^{-\lambda_2}) \\
v_{initial} &= v^1_{initial} + v^2_{initial}
\end{aligned}$$

and gets (expected) social punishment:

$$\mathcal{C}_{initial} \equiv C(1-\pi_i)\sum_{\tau_1=1}^{\infty}\sum_{\tau_2=1}^{\infty}\beta^{\tau_1+\tau_2-1}e^{-\lambda_1(\tau_1-1)}(1-e^{-\lambda_1})e^{-\lambda_2(\tau_2-1)}(1-e^{-\lambda_2})$$

Whereas agent j acting in the viral phase receives share utility:

$$v_{viral} = \kappa\sum_{\tau_2=1}^{\infty}\beta^{\tau_2-1}\left(\sum_{\tau'=1}^{\tau_2}\gamma^{\tau'}\right)e^{-\lambda_2(\tau_2-1)}(1-e^{-\lambda_2})$$

and gets (expected) social punishment:

$$\mathcal{C}_{viral} \equiv C(1-\pi_i)\sum_{\tau_2=1}^{\infty}\beta^{\tau_2-1}e^{-\lambda_2(\tau_2-1)}(1-e^{-\lambda_2})$$

Because sharing for agent i is a better response than inspecting, which in turn dominates killing by Assumption 1, we know that $v_{viral} - \mathcal{C}_{viral} > 0$. However, observe that:

$$\begin{aligned}
v^2_{initial} &= \sum_{\tau_1=1}^{\infty}\beta^{\tau_1}e^{-\lambda_1(\tau_1-1)}(1-e^{-\lambda_1})v_{viral} \\
\mathcal{C}_{initial} &= \sum_{\tau_1=1}^{\infty}\beta^{\tau_1}e^{-\lambda_1(\tau_1-1)}(1-e^{-\lambda_1})\mathcal{C}_{viral}
\end{aligned}$$

Therefore,

$$v^2_{initial} - \mathcal{C}_{initial} = \sum_{\tau_1=1}^{\infty}\beta^{\tau_1}e^{-\lambda_1(\tau_1-1)}(1-e^{-\lambda_1})(v_{viral}-\mathcal{C}_{viral}) > 0$$

by assumption. Since $v^1_{initial} > 0$, we have that $v_{initial} - \mathcal{C}_{initial} > 0$, which implies that sharing is a best response for agent i in the initial phase. ∎

Proof of Corollary 1. This is a direct consequence of the conditions in Theorem 2, which depend on

$q, m, \underline{b}, \gamma, \kappa, \lambda_2, \beta, \delta,$ and C, but not on the network structure $\mathbf{P}$.

Proof of Proposition 1. We use the conditions of Theorem 2, which are tight. Only v_{viral} depends on κ and γ, which is an increasing function of both. This decreases the left-hand side of the expression in Theorem 2 while holding the right-hand side constant, making all-share equilibria more likely. Only $\mathcal{C}$ depends on C, which is a decreasing function in C, and decreases the left-hand side of the expression in Theorem 2 while holding the right-hand side constant. Note that $\mathcal{C}$ is increasing in λ_2, whereas v_{viral} is decreasing in λ_2, so the left-hand side of the expression in Theorem 2 is increasing in λ_2. Finally, v_{viral} and $\mathcal{C}$ are both functions of β with:

$$\frac{\partial v_{viral}}{\partial \beta} = \frac{\gamma\kappa(1-e^{-\lambda_2})e^{-\lambda_2}(1+\gamma(1-2\beta e^{-\lambda_2}))}{(1-\beta e^{-\lambda_2})^2(1-\beta\gamma e^{-\lambda_2})^2}$$
$$\frac{\partial \mathcal{C}}{\partial \beta} = \frac{C(1-e^{-\lambda_2})e^{-\lambda_2}}{(1-\beta e^{-\lambda_2})^2}$$

Since $\gamma \geq 2$, we know that $1 - 2\beta e^{-\lambda_2} > 0$ as we require $\beta\gamma e^{-\lambda_2} < 1$, otherwise the game is trivial because viral-phase sharing gives an infinite payoff. Thus, both v_{viral} and $\mathcal{C}$ are increasing functions in β. Finally observe that $\partial\mathcal{C}/\partial\beta$ is not a function of γ, but $\partial v_{viral}/\partial\beta$ is an increasing function in γ, so there exists a threshold γ^* such that $\partial\mathcal{C}/\partial\beta < \partial v_{viral}/\partial\beta$ for all $\gamma > \gamma^*$. $\blacksquare$

Proof of Proposition 2. As in the proof of Proposition 1, we use the conditions of Theorem 2. Only $(1-\underline{\pi})$ is a function of the message m. Notice:

$$\frac{\partial(1-\underline{\pi})}{\partial m} = \frac{(1-2\underline{b})(1-q)q}{(\underline{b}(1-2q)(1-\mathcal{L}(m)) - q\mathcal{L}(m) + q + \mathcal{L}(m))^2}\frac{\partial\mathcal{L}(m)}{\partial m}$$

By the MLRP, we know that $\partial\mathcal{L}(m)/\partial m > 0$. Thus, $1-\underline{\pi}$ is decreasing in m when $\underline{b} > 1/2$ and increasing in m when $\underline{b} < 1/2$. In the former case, an increase in m decreases the left-hand side of the expression in Theorem 2, making all-share equilibria more likely. In the latter case, an increase in m increases the left-hand side of the expression in Theorem 2, making all-share equilibria less likely. $\blacksquare$

Proof of Proposition 3. First, we show that if $\tilde{f}$ is more information than f, then $\tilde{\mathcal{L}}(m) > \mathcal{L}(m)$ for all $m > 0$. Note that:

$$\frac{\tilde{\mathcal{L}}(m)}{\mathcal{L}(m)} = \frac{\tilde{f}(m|\theta=R)/\tilde{f}(m|\theta=L)}{f(m|\theta=R)/f(m|\theta=L)} = \frac{\tilde{f}(m|\theta=R)}{f(m|\theta=R)} \cdot \frac{f(m|\theta=L)}{\tilde{f}(m|\theta=L)}$$

which is an increasing function in m and since $\tilde{\mathcal{L}}(0) = \mathcal{L}(0)$, we see that $\tilde{\mathcal{L}}(m) > \mathcal{L}(m)$ for all $m > 0$. Next, we use the conditions in Theorem 2, and note that $\mathcal{L}(m)$ only affects $1-\underline{\pi}$. Differentiating with respect to $\mathcal{L}(m)$:

$$\frac{\partial(1-\underline{\pi})}{\partial\mathcal{L}} = \frac{(1-2\underline{b})(1-q)q}{(\underline{b}(2q-1)(\mathcal{L}(m)-1) - q\mathcal{L}(m) + \mathcal{L}(m) + q)^2}$$

which is negative when $\underline{b} > 1/2$ and positive when $\underline{b} < 1/2$. In the former case, an increase in informativeness with $m > 0$ increases $\mathcal{L}(m)$, which decreases $(1 - \underline{\pi})$, and makes all-share equilibria more likely. In the latter case, an increase in informativeness with $m > 0$ increases $\mathcal{L}(m)$, which increase $(1 - \underline{\pi})$, and makes all-share equilibria less likely. ∎

Proof of Proposition 4. As in the proof of Proposition 1, we use the conditions of Theorem 2. Only $(1-\underline{\pi})$ is a function of the lower support prior $\underline{b}$. Notice:

$$\frac{\partial(1-\underline{\pi})}{\partial\underline{b}} = \frac{(1-\mathcal{L}(m))(1+\mathcal{L}(m))(1-q)q}{(\mathcal{L}(m)q(1-2\underline{b}) + \mathcal{L}(m)\underline{b} - \mathcal{L}(m) + 2q\underline{b} + q + \underline{b})^2}$$

Because $m > 0$, we know that $\mathcal{L}(m) > 1$, so the above expression is negative, and $1 - \underline{\pi}$ is decreasing in $\underline{b}$. This makes all-share equilibria more likely because it decreases the left-hand side of the expression in Theorem 2. ∎

B.3 Omitted Proofs from Section 4

Proof of Lemma 2. Suppose there are two agents i and j on some island ℓ with the same prior belief b. We show that both agents must have the same best response to *any* strategy profile $\boldsymbol{\sigma}$, and thus must employ the same cutoff strategy, as $N_\ell \to \infty$.

Note that because $\min_\ell N_\ell \to \infty$ and (p_s, p_d) are fixed, the degree of every agent grows unboundedly. Thus, given (γ, λ_2) are constant, the probability that agent i or agent j receives the article at any point in the sharing process vanishes as $N \to \infty$. Moreover, agent i (resp. agent j) believes the probability that agent j (resp. agent i) will receive the article after her in the sharing process has vanishing probability (as $N \to \infty$). Let $\hat{\sigma}_i$ be an arbitrary strategy profile for agent i and let $[\boldsymbol{\sigma}_{-i}, \hat{\sigma}_i]$ be a strategy profile that replaces agent i's strategy with strategy $\hat{\sigma}_i$. Because agent i and agent j have symmetric network positions, we know that the following limits coincide for any choice of $\hat{\sigma}_i, \hat{\sigma}_j$:

$$\lim_{N_\ell\to\infty} S_{i,\tau}([\boldsymbol{\sigma}_{-j}, \hat{\sigma}_j], \cdot) = \lim_{N_\ell\to\infty} S_{i,\tau}([\boldsymbol{\sigma}_{-i}, \hat{\sigma}_i], \cdot)$$
$$\lim_{N_\ell\to\infty} \phi_{i,s}([\boldsymbol{\sigma}_{-j}, \hat{\sigma}_j]) = \lim_{N_\ell\to\infty} \phi_{j,s}([\boldsymbol{\sigma}_{-i}, \hat{\sigma}_i])$$

Note that it is impossible that agent j sees agent i shared the article with her and simultaneously that agent i sees agent j shared it with her; without loss of generality, suppose i does not receive it from j. Then $\rho_i([\boldsymbol{\sigma}_{-j}])$ does not depend on j and therefore $\lim_{N_\ell\to\infty} \rho_i([\boldsymbol{\sigma}_{-j}, \hat{\sigma}_j])$ must converge to the same limit regardless of $\hat{\sigma}_j$, and in particular converges to $\rho_i(\boldsymbol{\sigma})$. This moreover implies that $\lim_{N_\ell\to\infty} \pi_i(m, [\boldsymbol{\sigma}_{-i}, \hat{\sigma}_i]) = \lim_{N_\ell\to\infty} \pi_i(m, \boldsymbol{\sigma})$, so agent i plays according to $\boldsymbol{\sigma}$.

In turn, this implies that once j receives the article (even if from i), she holds the identical beliefs $\lim_{N_\ell\to\infty} \pi_j(m, \boldsymbol{\sigma}) = \lim_{N_\ell\to\infty} \pi_i(m, \boldsymbol{\sigma})$ because agents i and j hold the same prior beliefs and have the same $\lim \rho_i(\boldsymbol{\sigma}), \lim \rho_j(\boldsymbol{\sigma})$. These show that, in particular, all limits coincide for the true strategy profile

$\boldsymbol{\sigma}$. Thus the payoffs shown in Equations(3) and (4) are identical for agents i and j, which means agents i and j must have identical best responses. ∎

Proof of Theorem 4. By Lemma 2, all equilibria on a single island take the form of two cutoff strategies (one for the initial phase and one for the viral phase), (b^*, b^{**}). We define a *partial equilibrium* of the initial (viral) phase as a cutoff b^* (b^{**}) that could be supported in an equilibrium taking the cutoff of the other phase b^{**} (b^*) as given. The set of equilibrium cutoffs are the partial equilibria that intersect.

Let $B^{**}(b^*)$ be the partial equilibria of the viral phase for a given b^*. By Lemma B.1, initial-phase cutoffs do not affect the viral-phase cutoffs that can be supported in equilibrium. Thus, $B^{**}(b^*)$ has a trivial dependence on b^* and we can call this set of cutoffs B^{**}. Note that $\rho_j(\boldsymbol{\sigma}) = 0$ for the initial phase because inspecting is a dominated strategy, so $\pi_j(m, \boldsymbol{\sigma})$ has no dependence on $\boldsymbol{\sigma}$, and we can just write $\pi_j(m)$. Moreover, the payoff from sharing in the initial phase is then given by:

$$\kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1}(\pi_j(m) S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{T}) + (1 - \pi_j(m)) S_{j,\tau}(\boldsymbol{\sigma}, \mathcal{F})) - C(1 - \pi_j(m)) \sum_{s=1}^{\infty} \beta^{s-1} \phi_{j,s}(\boldsymbol{\sigma})$$

which is supermodular in the proportion of the population who is sharing in both the initial and viral phases (i.e., $1 - H(b^*)$ and $1 - H(b^{**})$). Thus, the partial equilibrium cutoffs $B^*(b^{**})$ in the initial phase form a lattice according to Tarski's theorem and by Topkis's theorem, have monotone comparative statics in $1 - H(b^{**})$ (which is monotone in b^{**}). Thus, the set of (full) equilibrium cutoffs is given by the lattice $\{(b^*, b^{**}) \,|\, b^{**} \in B^{**}, b^* \in B^*(B^{**})\}$. ∎

Proof of Theorem 5. We prove part (a) first. Without loss of generality, let island 1 be the one at all-share. Then:

$$I_1 = \frac{i_2 N_2}{(p_s/p_d) N_1 + N_2}$$
$$I_2 = \frac{(p_s/p_d) i_2 N_2}{N_1 + (p_s/p_d) N_2}$$

and following an increase in p_s or decrease in p_d, I_1 decreases but I_2 increases. When p_s/p_d is sufficiently close to 1, we have the single-island model of Section 4.1, and both islands have nearly the same cutoffs b_1^*, b_2^* in the initial-phase and b_1^{**}, b_2^{**} in the viral phase by Lemma 2. Thus, for any $\epsilon > 0$, we can take p_s/p_d sufficiently close to 1 to bound the killing probability and inspection probability on island 2 to at most ϵ, before the increase in homophily.

Similarly, both islands have almost the same difference in payoffs from inspecting and sharing, $\Delta_1 \approx \Delta_2$ and the same exposure to individuals inspecting, $I_1 \approx I_2$, when starting near the single-island model. Thus, given that island 1 has net strategic complements, by continuity, island 2 does as well; hence, both $\partial \Delta_1(b^{**})/\partial I_1 > 0$ and $\partial \Delta_2(b^{**})/\partial I_2 > 0$ hold. A marginal increase in homophily yields:

$$\frac{\partial \Delta_1(b^{**})}{\partial (p_s/p_d)} = \frac{\partial \Delta_1(b^{**})}{\partial I_1} \frac{\partial I_1}{\partial (p_s/p_d)} < 0$$
$$\frac{\partial \Delta_2(b^{**})}{\partial (p_s/p_d)} = \frac{\partial \Delta_2(b^{**})}{\partial I_2} \frac{\partial I_2}{\partial (p_s/p_d)} > 0$$

Because island 1 is at all-share, and the payoff from marginally increasing homophily decreases $\Delta_1(b^{**})$, which implies the payoff from playing $a_i = \mathcal{S}$ is higher relative to $a_i = \mathcal{I}$ after this increase. Hence, all-share remains in the viral phase. Whether all-share remains in the initial phase is indeterminate, but assuming it does only increases the virality of misinformation (we cannot leverage Lemma 1 here because island 2 is only ϵ close to all-share, but is not required). Thus, if we want to show the increase in homophily decreased the virality of misinformation, it is sufficient to assume that there is all-share in both phases on island 1.

Conversely, island 2 (at the most viral equilibrium) can either be at all-share or an equilibrium where $\Delta(b^{**})$ crosses the indifference line $\Delta = 0$ from above (and thus, has a negative tangent line) by Lemma B.2. In the first case, both islands are at all-share (so the most viral equilibrium has all-share for the initial phase too, per Lemma 1), so an increase in homophily does not increase or decrease virality by Corollary 1. In the second case, the curve Δ_2 shifts upward, moving the intersection with the indifference line to some higher b_2^{**}. Because this was an internal equilibrium (by assumption), the fraction of inspections on island 2 increases in the most viral equilibrium (i.e., $\partial i_2/\partial(p_s/p_d) > 0$). By the same argument as in Theorem 4, this necessarily increases the fraction of agents killing in the initial phase as well as on island 2, given that island 1 is at all-share. Once again, for ease and to show a stronger claim, we will assume the fraction of agents killing in the initial phase remains the same for the purposes of studying virality (and engagement) with an article containing misinformation.

To find the most viral equilibrium (for misinformation), we need to consider the expected engagement with an article containing misinformation. Because we have assumed that the strategy profile of initial phase agents has remained unchanged, it is sufficient to simply consider engagement that stems from interactions in the viral phase. Engagement satisfies the following fixed-point equation:

$$\mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 1}] = \gamma + \gamma e^{-\lambda_2} \left(\frac{p_s/p_d}{1+p_s/p_d} \mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 1}] + \frac{1-i_2}{1+p_s/p_d} \mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 2}] \right) \quad (6)$$
$$\mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 2}] = \gamma + \gamma e^{-\lambda_2} \left(\frac{1}{1+p_s/p_d} \mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 1}] + \frac{(p_s/p_d)(1-i_2)}{1+p_s/p_d} \mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 2}] \right) \quad (7)$$

Note that when $i_2 = 0$ (i.e., exactly at $p_s = p_d$), engagement has no dependence on (p_s, p_d):

$$\mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 1}] = \mathbb{E}[\mathbf{S}_{T(i^*)} \,|\, \text{island 2}] = \frac{\gamma}{1-\gamma e^{-\lambda_2}}$$

for all p_s/p_d, which implies that for fixed $i_2 = 0$,

$$\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 1]}{\partial(p_s/p_d)} = \frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 2]}{\partial(p_s/p_d)} = 0 \tag{8}$$

locally at $p_s = p_d$. Differentiating at $p_s = p_d$ and accounting for the fact i_2 is a function of (p_s, p_d), as discussed before, and using Equation (8) we get that:

$$\begin{aligned}\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 1, i_2(p_s)]}{\partial(p_s/p_d)} =& \gamma e^{-\lambda_2}\Big(\frac{1}{(1+(p_s/p_d))^2}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 1] - \frac{1-i_2}{(1+(p_s/p_d))^2}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 2] \\ &- \frac{1}{1+p_s/p_d}\frac{\partial i_2}{\partial(p_s/p_d)}\Big)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 2, i_2(p_s)]}{\partial(p_s/p_d)} =& \gamma e^{-\lambda_2}\Big(-\frac{1}{(1+(p_s/p_d))^2}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 1] + \frac{1-i_2}{(1+(p_s/p_d))^2}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 2] \\ &- \frac{p_s/p_d}{1+p_s/p_d}\frac{\partial i_2}{\partial(p_s/p_d)}\Big)\end{aligned}$$

with the change in total engagement being roughly:

$$\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 1, i_2(p_s)]}{\partial(p_s/p_d)} + \frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } 2, i_2(p_s)]}{\partial(p_s/p_d)} = -\frac{\partial i_2}{\partial(p_s/p_d)} < 0$$

as $\partial i_2/\partial(p_s/p_d) > 0$ by our previous argument. Thus, total engagement with homophily near $p_s/p_d = 1$ decreases and establishes the claim from part (a).

For part (b), we first show that given some island ℓ has *strategic complements* and an *all-share equilibrium*, an increase in homophily preserves the all-share equilibrium on this island. Without loss, we again will assume $\ell = 1$. We have:

$$\begin{aligned} I_1 &= \frac{\sum_{\ell'=2}^{k} i_{\ell'} N_{\ell'}}{(p_s/p_d)N_1 + \sum_{\ell'=2}^{k} N_{\ell'}} \\ \implies \frac{\partial I_1}{\partial(p_s/p_d)} &= -\frac{N_1 \sum_{\ell'=2}^{k} i_{\ell'} N_{\ell'}}{((p_s/p_d)N_1 + \sum_{\ell'=2}^{k} N_{\ell'})^2} < 0\end{aligned}$$

for all p_s/p_d and . Thus, similar to the proof in part (a), we have:

$$\frac{\partial \Delta_1(0)}{\partial(p_s/p_d)} = \frac{\partial \Delta_1(0)}{\partial I_1}\frac{\partial I_1}{\partial p_s} < 0$$

for all p_s/p_d due to strategic complementarity. As a result, the difference in sharing payoff and inspecting becomes larger as compared to before the increase in homophily (for the worst-case prior of $b_i = 0$), so all-share on island 1 is still part of the most viral equilibrium.

Second, we establish that for sufficiently high homophily ratio $\tilde{p}$, all $p_s/p_d > \tilde{p}$ satisfy the property that all-share remains on island 1 in the initial phase as well in the most viral equilibrium. We employ a similar technique to the proof of Lemma 1. Recall that $v_{initial} > v^2_{initial} > \mathcal{C}_{initial}$, so sharing is a best response when others in the initial phase are also sharing. Next, we compute the probability that if the article starts with agent $i \in \ell_1$ in the initial phase that it ever leaves island 1 (and thus could be killed or inspected in the future). A loose union bound gives us:

$$
\begin{aligned}
&\frac{p_d(N-N_1)}{p_s N_1 + p_d(N-N_1)} \sum_{\tau_1=1}^{\infty} \sum_{\tau_2=1}^{\infty} \left(\tau_1 + \left(\sum_{\tau'=1}^{\tau_2} \gamma^{\tau'} \right) \right) e^{-\lambda_1(\tau_1-1)-\lambda_2(\tau_2-1)} (1-e^{-\lambda_1})(1-e^{-\lambda_2}) \\
=&\frac{N-N_1}{(p_s/p_d-1)N_1+N} \left(\frac{\kappa(1-e^{-\lambda_1})}{(1-e^{-\lambda_1})^2} + \frac{\kappa(1-e^{-\lambda_1})(1-e^{-\lambda_2})}{(1-e^{-\lambda_1})(1-e^{-\lambda_2})(1-\gamma e^{-\lambda_2})} \right)
\end{aligned}
$$

which is finite and converges toward 0 as $p_s/p_d \to \infty$. Taking p_s/p_d sufficiently large bounds this probability, call it ζ, below $(v_{initial} - \mathcal{C}_{initial})/(v_{initial} + C)$, so that $(1-\zeta)v_{initial} - \zeta C - \mathcal{C}_{initial} > 0$ and all-share is still supported in the initial phase.

In the trivial case where all of the islands are at all-share, we know by Corollary 1 that following any change in homophily will not destroy the all-share equilibrium, so the virality of misinformation does not improve. Thus, we assume without loss of generality that there exists some island ℓ^* with either kill rate $k_{\ell^*} > 0$ or inspection rate $i_{\ell^*} > 0$.

Next, consider the engagement fixed-point equation for island 1 that we studied in part (a), but for the more general island model and looking at both phases of the sharing game:

$$
\begin{aligned}
\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; initial}] = 1 &+ e^{-\lambda_1} \Big(\frac{(p_s/p_d)N_1}{(p_s/p_d)(N_1-1)+N} \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; initial}] \\
&+ \sum_{\ell'=2}^{k} \frac{(1-k_{\ell'})N_{\ell'}}{(p_s/p_d-1)N_1+N} \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'\text{; initial}] \Big) \\
&+ \gamma(1-e^{-\lambda_1}) \Big(\frac{(p_s/p_d)N_1}{(p_s/p_d-1)N_1+N} \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] \\
&+ \sum_{\ell'=2}^{k} \frac{(1-i_{\ell'})N_{\ell'}}{(p_s/p_d-1)N_1+N} \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'\text{; viral}] \Big) \\
\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] = \gamma &+ \gamma e^{-\lambda_2} \Big(\frac{(p_s/p_d)N_1}{(p_s/p_d-1)N_1+N} \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] \\
&+ \sum_{\ell'=2}^{k} \frac{(1-i_{\ell'})N_{\ell'}}{(p_s/p_d-1)N_1+N} \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'\text{; viral}] \Big)
\end{aligned}
$$

We prove that $\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]/\partial(p_s/p_d) > 0$ for sufficiently high homophily; the case for the initial phase follows similarly. The derivative (with respect to p_s/p_d) of the term associated with island 1 is:

$$
\gamma e^{-\lambda_2} \left(\frac{N_1(N-N_1)}{((p_s/p_d-1)N_1+N)^2} \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] + \frac{(p_s/p_d)N_1}{(p_s/p_d-1)N_1+N} \frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]}{\partial(p_s/p_d)} \right)
$$

whereas the derivative of the term associated with island ℓ' is:

$$\gamma e^{-\lambda_2}\Big(- \frac{(1-i_{\ell'})N_{\ell'}N_1}{((p_s/p_d-1)N_1+N)^2}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}] - \frac{N_{\ell'}}{(p_s/p_d-1)N_1+N}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]\frac{\partial i_{\ell'}}{\partial(p_s/p_d)}$$
$$+\frac{(1-i_{\ell'})N_{\ell'}}{(p_s/p_d-1)N_1+N}\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]}{\partial(p_s/p_d)}\Big)$$

Rearranging, we have that:

$$\begin{aligned}\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]}{\partial(p_s/p_d)} = & \frac{\gamma e^{-\lambda_2}[(p_s/p_d-1)N_1+N]}{((1-\gamma e^{-\lambda_2})p_s/p_d-1)N_1+N}\Big(\frac{N_1(N-N_1)}{((p_s/p_d-1)N_1+N)^2}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] \\ & -\sum_{\ell'=2}^{k}\Big[\frac{(1-i_{\ell'})N_{\ell'}N_1}{((p_s/p_d-1)N_1+N)^2}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}] \\ & + \frac{N_{\ell'}}{(p_s/p_d-1)N_1+N}\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]\frac{\partial i_{\ell'}}{\partial(p_s/p_d)} \\ & - \frac{(1-i_{\ell'})N_{\ell'}}{(p_s/p_d-1)N_1+N}\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]}{\partial(p_s/p_d)}\Big]\Big)\end{aligned}$$

Multiplying all sides by $((p_s/p_d-1)N_1+N)^2$:

$$\begin{aligned}&((p_s/p_d-1)N_1+N)^2\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]}{\partial(p_s/p_d)} \\ =&\frac{\gamma e^{-\lambda_2}[(p_s/p_d-1)N_1+N]}{((1-\gamma e^{-\lambda_2})p_s/p_d-1)N_1+N}\Big(N_1(N-N_1)\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] \\ -&\sum_{\ell'=2}^{k}\Big[(1-i_{\ell'})N_{\ell'}N_1\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}] \\ +&N_{\ell'}((p_s/p_d-1)N_1+N)\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]\frac{\partial i_{\ell'}}{\partial(p_s/p_d)} \\ -&(1-i_{\ell'})N_{\ell'}((p_s/p_d-1)N_1+N)\frac{\partial \mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]}{\partial(p_s/p_d)}\Big]\Big)\end{aligned}$$

We argue that both the last two terms go to 0 (for all ℓ') as $p_s/p_d \to \infty$:

(i) Penultimate term: Note that as $p_s/p_d \to \infty$, $i_{\ell'} \to i^*_{\ell'}$ for some $i^*_{\ell'}$ because the island model converges to the segregated islands model, which will yield most sharing equilibrium on each individual island (as in Theorem 4). Moreover, in the vicinity of $i^*_{\ell'}$, $i_{\ell'}$ will converge monotonically to $i^*_{\ell'}$ because either strategic complements or substitutes will dominate near $i^*_{\ell'}$ and the effect on $i_{\ell'}$ will be consistent after some $p_s/p_d \geq p^*$. Hence, integrating $\partial i_{\ell'}/\partial(p_s/p_d)$ after p^* reveals that it must converge superlinearly, as linear monotone converge would imply $i^*_{\ell'}$ is unbounded, which it clearly cannot be. Thus, $\lim_{p_s/p_d \to \infty} N_{\ell'}((p_s/p_d-1)N_1+N)\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]\frac{\partial i_{\ell'}}{\partial(p_s/p_d)} = 0$.

(ii) Final term: A symmetric argument to the one above shows that $((p_s/p_d - 1)N_{\ell'} + N)^2\partial\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]/\partial(p_s/p_d)$ is a function of $N_1((p_s/p_d - 1)N_{\ell'} + N)\partial\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]/\partial(p_s/p_d)\Big]$ among the other parameters from the equation above.

Thus, $\partial\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]/\partial(p_s/p_d)$ must converge to 0 at the same rates of convergence in p_s/p_d, which is quadratic.

Therefore, for large enough p_s/p_d, the above is approximately equal to:

$$\begin{aligned}
((p_s/p_d-1)N_1+N)^2\frac{\partial\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]}{\partial(p_s/p_d)} &= \frac{\gamma e^{-\lambda_2}}{1-\gamma e^{-\lambda_2}}\Big(N_1(N-N_1)\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] \\
&\quad - \sum_{\ell'=2}^{k}(1-i_{\ell'})N_{\ell'}N_1\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island } \ell'; \text{viral}]\Big) \\
&\geq \frac{\gamma e^{-\lambda_2}}{1-\gamma e^{-\lambda_2}}\Big(N_1(N-N_1)\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] \\
&\quad - \sum_{\ell'=2}^{k}(1-i_{\ell'})N_{\ell'}N_1\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]\Big)
\end{aligned}$$

By assumption $i_{\ell^*} > 0$ for some island ℓ^*; thus,

$$\begin{aligned}
((p_s/p_d-1)N_1+N)^2\frac{\partial\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]}{\partial(p_s/p_d)} &> \frac{\gamma e^{-\lambda_2}}{1-\gamma e^{-\lambda_2}}\Big(N_1(N-N_1)\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}] \\
&\quad - \sum_{\ell'=2}^{k}N_{\ell'}N_1\mathbb{E}[\mathbf{S}_{T(i^*)} \mid \text{island 1; viral}]\Big) = 0
\end{aligned}$$

∎

Proof of Theorem 6. We prove this theorem for a decrease in message extremism or polarization, and the proof for an increase is analogous.

For part (a), we start with the case of message extremism, then polarization, and then show how these conclusions decrease virality on the single island.

We know that $b^* < 1/2$ and $b^{**} < 1/2$ because the right-wing island is at all-share. Following an decrease in the extremism of the (right-wing) message m, the belief of a left-wing agent π_i that the article is truthful increases, as observed in Proposition 2. As seen in Theorem 1, sharing payoffs net inspecting or killing are monotone in π_i. Thus, in the viral phase, $\Delta(b^{**})$ decreases. If the extremal equilibrium is all-share, it remains all-share. If the extremal equilibrium is all-inspect, inspections can only increase. If the extremal equilibrium crosses the indifference line $\Delta = 0$ at some interior cutoff b^{**} (at negative slope), then b^{**} decreases as Δ shifts down for all cutoffs below $1/2$. These are the only three possibilities by Lemma B.2. Thus, inspections decrease in the most viral equilibrium and applying Topkis's theorem for the initial phase shows that kills must decrease in the initial-phase extremal equilibrium too.

Following a decrease in the polarization, by Lemma B.3, we know that $i = H(b^{**})$ has decreased since $b^{**} < 1/2$ (i.e., inspections have gone down, holding the cutoff b^{**} fixed). By strategic complementarity, note that $\partial\Delta(b^{**})/\partial I > 0$, so once again $\Delta(b^{**})$ decreases, and the same conclusion of the previous paragraph applies to show that b^* and b^{**} both decrease. Finally, this implies that

$H(b^*)$ and $H(b^{**})$ also decrease (because of the reduction in polarization), so kills and inspections also decrease in the most viral equilibrium.

Finally, note that in the single-island model, if the inspection rate and kill rate are both lower in a new equilibrium, virality necessarily increase, because the choice of the seed agent i^* at $t = 0$ is immaterial to engagement with misinformation, as the connections are uniformly at random. Thus, part (a) shows virality of misinformation increases as intended.

For part (b), because we are in the segregated islands model, we can apply the comparative statics of Theorem 2 to the right-wing island. Following a decrease in extremism of the message, all-share becomes less likely on this island by Proposition 2. Similarly, following a decrease in the polarization, all-share becomes less likely on this island by Proposition 4 because all agents on the right-wing island hold beliefs $\underline{b}_R > 1/2$. Starting with an initial seed agent i^* on the right-wing island guarantees all-share, which is the most viral equilibrium for misinformation. Decreasing extremism or polarization might destroy the all-share equilibrium on the right-wing island, which would necessarily destroy the all-share on the left-wing island (if it existed, see the discussion following Definition 3 in the text), and lead to inspections on both islands. Any choice of i^* would lead killing and/or inspection with positive probability. This establishes the claim for part (a) that virality of misinformation indeed decreases.

B.4 Omitted Proofs from Section 5

Proof of Proposition 5. For part (a), we note that by Lemma B.3, an increase in polarization necessarily (weakly) decreases $\bar{b}_1 < 1/2$ and (weakly) increases $\underline{b}_k > 1/2$. By Proposition 4, this necessarily increases the likelihood of all-share equilibria within these communities when $\mathbf{P}$ is chosen to be the segregated-islands model. Because of the monotonicity in messages for all-share equilibria (Proposition 2) when $\bar{b}_1 < 1/2$ and $\underline{b}_k > 1/2$, this implies that η weakly decreases following this increase in polarization.

For part (b), the addition of a new community that neither becomes community 1 or community k does not affect the platform's decision problem, as described in the proof of Theorem 7. However, if the additional community ends up being the most left or right extreme one (i.e., becomes community 1 or community k), then this decreases $\bar{b}_1$ or increases $\underline{b}_k$, and the same comparative static with respect to polarization holds in this case.

Finally, for part (c), the platform only need check whether an all-share equilibrium can be sustained in community 1 or community k when adopting the segregated-islands model of $\mathbf{P}$. This is equivalent to checking the conditions of Theorem 2. The inspection cost K appears only on the right-hand side of these conditions, and by increasing K, all-share equilibria become more likely. Given that $\bar{b}_1 < 1/2$ and $\underline{b}_k > 1/2$, this allows for a greater right-hand side of the inequality in Theorem 2 and still permit all-share. Because $(1 - \underline{\pi})$ is decreasing in the extremity of the message (when the platform targets community 1 or community k), this permits the platform to adopt the algorithm of Theorem 7(b) with a less extreme message, thus decreasing the threshold η. ∎

B.5 Omitted Proofs from Section 6

Proof of Proposition 6. It is sufficient to prove parts (a) and (b) separately for thresholds $\phi_1, \phi_2 \in (0,1)$, respectively, and then take $\bar{\phi} = \max\{\phi_1, 1-\phi_2\}$. Note that q (when provenance is not revealed) satisfies $q = \phi q_r + (1-\phi)q_s$.

(a) In the single-island model, we compare the most-sharing equilibrium before provenance is revealed to the most-sharing equilibria (one for sketchy and one for reputable news) after provenance is revealed. We denote the former cutoff as (b^*, b^{**}) and the latter cutoffs as (b_r^*, b_r^{**}) and (b_s^*, b_s^{**}) for reputable and sketchy news, respectively.

We focus on the most-sharing equilibrium when the news source is revealed as reputable, (b_r^*, b_r^{**}). This falls in one of three categories by Lemma B.2 (and these cutoffs can be analyzed before considering the initial-phase cutoffs, per Lemma B.1).

First, if all-inspect is the most viral with reputable news, it is unique, and then is unique for sketchy news and when provenance is hidden. Moreover, by the arguments in Theorem 4, sharing in the initial-phase extremal equilibria will be monotonically affected by sharing in the viral phase, which is as low as it can be, as well as q, the likelihood the article is truthful. Thus, taking ϕ_1 close to 1, guarantees $b_r^* = b^*$ and $b_r^{**} = b^{**} = 1$, whereas for the sketchy source $b_s^* < b^*$ and $b_s^{**} = 1$. This means viral misinformation is reduced given $q_s < q_r$.

Second, is that all-share is the most viral equilibrium and the most-sharing equilibrium is all-share in both phases by Lemma 1. Any policy here, including revealing provenance, (weakly) reduces the virality of misinformation

Third, is that $w(b)$ intersects the indifference curve at b^{**} with a negatively-sloped tangent line. Once again, taking ϕ_1 close to 1 guarantees that $b_r^* = b^*$ and $b_r^{**} \to b^{**}$ for the most-sharing equilibrium. However, revealing the ex-ante probability the article is truthful is q_s increases $\Delta(b^{**})$, which increases the viral-phase cutoff for sketchy news to b_s^{**}, and thus increases inspections of sketchy news. By the argument in Theorem 4, this also increases killing in the initial phase in the most-sharing equilibrium.

(b) Conversely, we focus on the most-sharing equilibrium when the news source is revealed as sketchy, (b_s^*, b_s^{**}). As before, the most-sharing equilibrium in the viral phase falls in one of three categories by Lemma B.2 (which are unaffected by initial-phase strategies, per Lemma B.1).

First, is that all-inspect is the unique viral-phase equilibrium for sketchy news. Then, when ϕ_2 is sufficiently close to 0, all-inspect is unique when provenance is not revealed. Moreover, by the arguments in Theorem 4, sharing in the initial-phase extremal equilibria will be monotonically affected by sharing in the viral phase, which is as low as it can be. Thus, revealing provenance can only increase the virality of misinformation.

Second, is that the most viral equilibrium is all-share. Similar to the first case in (a), when ϕ_2 is close to 0, revealing provenance retains the all-share equilibrium for both news sources.

Third, is that $w(b)$ intersects the indifference curve at b^{**} with a negatively-sloped tangent line. As before, when ϕ_2 is sufficiently close to 0, $b_s^* = b^*$ and $b_s^{**} = b^{**}$. Revealing q_r for a reputable source decreases $\Delta(b^{**})$ and decreases the cutoff from b^{**} to some $b_r^{**} < b^{**}$. This implies there is more sharing in the initial phase as well when the news source is revealed as reputable. Hence, the revelation of reputable news increases the virality of misinformation for this type of news while not affecting the virality of sketchy news.

∎

Proof of Corollary 2. Following Theorem 7, if the platform does not tag an article and either (i) it *does* contain misinformation or (ii) $\eta > \max\{\mathcal{L}(m), 1/\mathcal{L}(m)\}$, and in either case, the planner censors it with positive probability. This strictly decreases $\mathbb{E}[\mathbf{S}_{T(i^*)}]$ given the platform does not inspect. When K_P is sufficiently small, this implies the platform should instead inspect first, and by Theorem 7, choose the uniform-connection model for $\mathbf{P}$. Because all content is tagged, the planner need not censor any of it.

∎

Proof of Corollary 3. Note that $p^* = 1$ gives the outcome of Theorem 7(a) and $p^* = \infty$ gives the outcome of Theorem 7(b). When $p_s/p_d < \bar{p}$ for $\bar{p}$ sufficiently close to 1, the probability of inspection during the share cascade is bounded away from 0, given there is no all-share equilibrium (which does not depend on $\mathbf{P}$ per Corollary 1). Thus, taking K_P to be sufficiently small, the platform prefers to inspect and tag the article than risk this loss in engagement in the event of the article containing misinformation (which occurs with positive probability). ∎

B.6 Misinformation Structures

In Section 2, we assumed that the message distribution for misinformation (i.e., articles with $\nu = \mathcal{F}$) followed $f(m|\theta = \neg\theta)$, i.e., a mirror image of truthful information. Here, we relax that assumption, and allow for a more general structure of misinformation. Let $\omega(m|\theta)$ be the likelihood (density) of misinformation given state θ, where ω divides the message space at $m^* = 0$ as f does, which is without loss of generality. We assume that misinformation is less strongly correlated with the truth than accurate information (which includes, as special cases, noise or anti-correlation). Thus, while misinformation is less strongly grounded to the truth, it is still possible for misinformation to advocate for the correct state θ. We make the following assumption:

Assumption B.1. Misinformation is less informative than true information in the sense that $f(m|\theta = R)/\omega(m|\theta = R)$ is increasing in m and $f(m|\theta = L)/\omega(m|\theta = L)$ is decreasing in m.

Our anti-correlation assumption in the text is a special case of Assumption B.1 (given the MLRP property of f). Formally, Assumption B.1 requires the ratio of the likelihood of accurate information advocating for the correct state to the likelihood of misinformation advocating for the correct state be monotone in the strength of the message. The following result looks at how the posterior belief of an article's veracity (for some agent i) is affected by her initial belief b_i:

Proposition B.1. *Fix a right-wing message $m > 0$ and let $\tilde{q}$ be the ex-ante probability (conditional on receiving the article) that it is truthful. There exists $\alpha \in (0,1)$ such that all agents with $b_i < \alpha$ have $\pi_i < \tilde{q}$ whereas all agents with $b_i > \alpha$ have $\pi_i > \tilde{q}$.*

In the baseline model (Section 2), left-wing agents believe more extreme right-wing news is more likely to contain misinformation, whereas right-wing agents believe more extreme left-wing news is. In this baseline, $\alpha = 1/2$. However, for more general structures, while α need not be equal to $1/2$, the qualitative properties of the model remain intact. In particular, any results or definitions that explicitly divide the prior space into beliefs less than $1/2$ and greater than $1/2$ can be be generalized using the α of Proposition B.1, whereas all others remain entirely unaffected.

B.7 Proof of Proposition B.1

Under the more general misinformation structure of Appendix B.6, the posterior belief the article is truthful is:

$$\pi_i = \frac{(\mathcal{L}(m)b_i + (1-b_i))\tilde{q}}{(\mathcal{L}(m)b_i + (1-b_i))\tilde{q} + \left(\frac{\omega(m|\theta=R)}{f(m|\theta=L)}b_i + \frac{\omega(m|\theta=L)}{f(m|\theta=L)}(1-b_i)\right)(1-\tilde{q})}$$

Differentiating with respect to b_i:

$$\frac{\partial \pi_i}{\partial b_i} \propto (1-\tilde{q})\tilde{q}\left(\mathcal{L}(m)\frac{\omega(m|\theta=L)}{f(m|\theta=L)} - \frac{\omega(m|\theta=R)}{f(m|\theta=L)}\right)$$

which is positive if and only if:

$$\mathcal{L}(m) > \frac{\omega(m|\theta=R)}{\omega(m|\theta=L)}$$

When $m = 0$, we know that $\mathcal{L}(m) \cdot \frac{\omega(m|\theta=L)}{\omega(m|\theta=R)} = 1$, and by Assumption B.1, we know $\mathcal{L}(m) \cdot \frac{\omega(m|\theta=L)}{\omega(m|\theta=R)}$ is increasing, so $\mathcal{L}(m) \cdot \frac{\omega(m|\theta=L)}{\omega(m|\theta=R)} > 1$ for all $m > 0$, which satisfies the desired inequality. So π_i is monotonically increasing in b_i. Finally, note that $\pi_i = \tilde{q}$ when:

$$b_i = \frac{\frac{\omega(m|\theta=L)}{f(m|\theta=L)} - 1}{\left(\frac{\omega(m|\theta=L)}{f(m|\theta=L)} - 1\right) + \left(\mathcal{L}(m) - \frac{\omega(m|\theta=R)}{f(m|\theta=L)}\right)}$$

Note that when $m > 0$, $\omega(m|\theta = L)/f(m|\theta = L) > 1$ and that $f(m|\theta = R) > \omega(m|\theta = R)$ by Assumption B.1, so both $\frac{\omega(m|\theta=L)}{f(m|\theta=L)} - 1 > 0$ and $\mathcal{L}(m) - \frac{\omega(m|\theta=R)}{f(m|\theta=L)} > 0$. Thus, $b_i = \alpha \in (0,1)$. ∎

B.8 Endogenous Reputation

In the baseline model of Section 2, we assume that agents care about getting caught sharing misinformation, in the form of an exogenous discounted societal punishment C, following an inspection (in her sharing subtree) revealing the article contains misinformation. In this section, we show that this formulation can be microfounded by an endogenous concern to build a reputation of sharing only quality content.

At $t = 0$, every agent is born as either a *careless* user, a *cautious* user, or a *normal* user.[28] The careless user and cautious user are behavioral types that share and kill/inspect every article with probability 1, respectively, whereas a normal user is a fully rational agent of the model in Section 2. The user is born careless with probability $\zeta^{\text{careless}} > 0$ and born cautious with probability $\zeta^{\text{cautious}} > 0$ which is i.i.d. across all agents and immutable. If an article is revealed as containing misinformation (either by another agent in the sharing subtree or by the outside source), the sharers of this misinformation are broadcast to all agents in the population. Consequently, after the article is out of circulation, agents in the population hold posterior beliefs $\zeta_i^{\text{careless}}$ about agent i's likelihood of being the careless type.

We assume each agent i intrinsically values the population's perception $1 - \zeta_i^{\text{careless}}$ (i.e., the public belief that she is not a careless user). This utility can exist for a few reasons. First, the agent may want other agents to believe she is socially responsible and only shares content she believes to likely be true, and does not spread "click-bait" articles with likely misinformation. Second, a bad reputation will also cut down on future share cascade opportunities: with higher $\zeta_i^{\text{careless}}$, agents will inspect more often following agent i's share because of the strategic substitutes effect.[29] In particular, we assume the social punishment takes the form $\tilde{C}(\zeta_i^{\text{careless}} - \zeta^{\text{careless}})$, where $\tilde{C}$ is a scaling factor that determines an agent's aversion to reputation loss.

For agents who share accurate news or kill/inspect, we observe $\zeta_i^{\text{careless}} \leq \zeta^{\text{careless}}$, so as in the baseline model, there is no societal punishment.[30] On the other hand, for those who shared misinformation, $\zeta_i^{\text{careless}} > \zeta^{\text{careless}}$ for all such agents i. Note that because agent i's belief distribution is unknown, all sharers of misinformation have the same $\zeta_i^{\text{careless}} > \zeta^{\text{careless}}$, with the (non-zero) reputation gap determined by averaging the Bayesian posterior of each agents' likelihood of being careless over all the equilibrium cutoffs (per Theorem 1). This translates to a model with some reputation cost C and $\beta = 1$, but where C is endogenously determined as an equilibrium object.

B.9 Numerical Simulations

We now present some simulations to investigate how often all-share equilibria exist and are unique. When all-share exists, we also investigate the likelihood that strategic complements exist at this

[28] Note in equilibria other than all-share, the cautious user is superfluous to the reputation argument (i.e., careless and normal users suffice).

[29] Note the strategic complements effect here is absent because the probability the article *ends up* at agent i is vanishing as $N \to \infty$.

[30] The benefit from revealing misinformation, δ, can also be microfounded in this way if the agent who inspects is revealed: in this event, $\zeta_i^{\text{careless}} < \zeta^{\text{careless}}$, and there is in fact a reputational reward for exposing misinformation.

equilibrium, i.e., at the cutoff $b^{**} = 0$. In these simulations, we draw the model primitives in the following ranges, all independently and uniformly at random: $q \in (1/2, 1)$, $f(m|\theta = L) \in (1/2, 1)$, $f(m|\theta = R) \in (1, 2)$, $C \in (5, 20)$, $\beta \in (0, 1)$, $K \in (0, 5)$, $\delta \in (K, 5)$, $\lambda_1 \in (0, 4)$, $\lambda_2 \in (0, 4)$, $\gamma \in (1, \min(5, 1/\beta))$, and $\kappa \in (0, 1/5)$. We employ rejection sampling to make sure Assumption 1 holds in all simulations.

Property	Sufficient Condition	Truth
Existence	34.3%	34.3%
Uniqueness	7.3%	16.6%
Complements	98.6%	98.6%

Because existence and complements criteria for all-share have closed-form expressions, we have both necessary and sufficient conditions for each, which means the underlying truth and the derived conditions yield the same results. We see that existence and uniqueness of all-share equilibria occur relatively often. However, more importantly, we see that the complements property is *extremely common* in all-share equilibria. Thus, in the what follows, we will often focus on environments where strategic complements dominates for agents in the viral phase.