A Model of Online Misinformation[*]

Daron Acemoglu[†] Asuman Ozdaglar[‡] James Siderius[§]

August 19, 2022

Abstract

We present a model of online content sharing where agents sequentially observe an article and decide whether to share it with others. This content may or may not contain *misinformation*. Agents gain utility from positive social media interactions but do not want to be called out for propagating misinformation. We characterize the (Bayesian-Nash) equilibria of this social media game and show sharing exhibits strategic complementarity. Our first main result establishes that the impact of homophily on content virality is non-monotone: homophily reduces the broader circulation of an article, but it creates echo chambers that impose less discipline on the sharing of low-reliability content. This insight underpins our second main result, which demonstrates that social media platforms interested in maximizing engagement tend to design their algorithms to create more homophilic communication patterns ("filter bubbles"). We show that platform incentives to amplify misinformation are particularly pronounced for low-reliability content likely to contain misinformation and when there is greater polarization and more divisive content. Finally, we discuss various regulatory solutions to such platform-manufactured misinformation.

Keywords: echo chambers, fake news, filter bubbles, homophily, misinformation, networks, social media

JEL classification: D83, D85, P16

[*] This paper builds on and replaces our earlier working paper on the same topic, entitled "Misinformation: Strategic Sharing, Homophily, and Endogenous Echo Chambers". We are grateful to Hunt Allcott, Jackie Dewart, Matthew Gentzkow, Ali Makhdoumi, Azarakhsh Malekian, Mohamed Mostagir, Ro'ee Levy, Francesca Parise, Davy Perlman, Alexander Wolitzky, and numerous theory lunch seminar participants, as well as NSF and the Sloan Foundation.

[†] Massachusetts Institute of Technology, NBER, and CEPR, daron@mit.edu

[‡] Massachusetts Institute of Technology, asuman@mit.edu

[§] Massachusetts Institute of Technology, siderius@mit.edu

1 Introduction

Social media has become a major 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 by 2019, over 70% of Americans reported receiving at least *some* of their news from social media (Levy (2021)). At the same time, there is growing concern about misinformation in social media, including made-up news stories such as those claiming that there were no mass shootings under Donald Trump or that Hillary Clinton approved ISIS weapon sales.[1] Some recent evidence also suggests that misinformation on social media has impacted critical decisions such as vaccinations against COVID-19 (see Pennycook et al. (2018); Pennycook et al. (2020b)).

Although there is yet no consensus on what promotes the spread of falsehoods and misleading content on social media, two sets of factors have been emphasized. The first is the presence of echo chambers, which arise when individuals communicate and share content with like-minded users (Sunstein (2018), Lazer et al. (2018)). Törnberg (2018) and Vicario et al. (2016) show that echo chambers reinforce existing political viewpoints and tend to propagate misinformation. Social media, much more than traditional media, allows individual users to choose who and what they listen to, and thus echo chambers may be an unavoidable side effect. However, there is also evidence that echo chambers are a result of the "filter bubbles" that platform algorithms create (see Levy (2021) on Facebook). The second factor conjectured to have fueled misinformation is the general political polarization in many countries, and especially the United States,[2] and there is some preliminary evidence suggesting that polarization has indeed contributed to selective exposure to questionable content on social media as well (Guess et al. (2018)). Despite the importance and salience of these issues, we do not currently have a framework to understand how online interactions impact the spread of misinformation and what factors shape incentives for sharing low-reliability content.

In this paper, we develop a parsimonious model of online sharing behavior in the presence of misinformation, and as a first step, we focus on the behavior of fully Bayesian agents.[3] Our model is inhabited by a set of N agents. Each agent has a prior about the state of the world ("ideological bias"), and is connected to the rest of the users via a network, which is given by agents' friends and acquaintances, and is also shaped by the algorithms of the social media platform. A news article, defined by an underlying type (truthful or containing misinformation), a message (right-wing or left-wing), and a level of reliability (which determines the likelihood of misinformation), is then seeded at one of the agents. The message and the level of reliability of the article are common knowledge, while whether it is truthful or contains misinformation is unobserved, and agents form beliefs about this

[1] See https://www.snopes.com/fact-check/mass-shootings-under-trump/ and https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html, respectively.

[2] While there has been some debate about whether polarization has been mainly among 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)).

[3] Myopic reactions and biased behavior appear to play some role, for example, via the "confirmation bias" in social media behavior (see, e.g., Buchanan (2020) and Pennycook and Rand (2019)), but we believe that the Bayesian benchmark we construct already generates a number of empirically-relevant and rich results. We view incorporating realistic and relevant behavioral biases as a next step in this research agenda.

component.

Given these beliefs, the agent in question decides whether to ignore, dislike, or share the news article. If it is shared, the article moves from the agent sequentially to her connections on social media, who are then faced with the same choices. If the article is ignored or disliked, it does not get past the agent. We assume that agents receive utility when their shared content is re-shared and incur a cost when it is disliked. The former aspect captures the role of positive engagement in social media, while the latter represents the reputation loss from being called out for sharing content containing misinformation. Agents additionally receive utility from disliking (or calling out) items that they believe contain misinformation.

We characterize the Bayesian-Nash equilibria of this sequential game and prove that these equilibria always exist and are in cutoff strategies. In particular, our payoff structure implies that an individual will share any item that she believes is truthful with a high probability and will dislike articles that she believes to contain falsehoods. Items with intermediate beliefs will be ignored. Beliefs about the truthfulness of articles are formed on the basis of the article's reliability and message, and agents' prior beliefs/ideology. Moreover, we establish that ours is a game of strategic complements: when others are more likely to share an item, each agent also becomes more likely to do so. As a result, we show that the set of equilibria forms a lattice, with well-defined most-sharing and least-sharing equilibria. All else equal, low-reliability articles are shared less, while articles that are "sensational" (either because they have provocative content or have broad appeal for other reasons) are shared more.

We present two main results. First, we study the implications of the (social media) network structure. We establish non-monotone comparative statics with respect to the degree of homophily (which determines how likely agents are to be connected to others who are ideologically similar to them). Low levels of homophily ensure that agents are likely to be exposed to cross-cutting content, including "counter-attitudinal articles" that advocate views opposed to theirs. This in turn ensures that misinformation is unlikely to survive for very long. Perhaps paradoxically, for high-reliability articles, an increase in homophily reduces content virality. This is because greater homophily makes it less likely that an article escapes a given community, reducing its circulation throughout the network.

More interestingly, when relevant news items have low reliability, homophily increases virality. This is because, countering the circulation effect, high homophily also creates a perverse incentive effect: knowing that shared articles will be seen by like-minded individuals, agents become more likely to share questionable content. Strategic complementarities amplify this effect, because when others are expected to share, the benefits from sharing are greater and being called out for spreading misinformation becomes less likely. It is particularly telling that homophily leads to the viral spread of low-reliability content, which are the ones more likely to contain misinformation.

We also show that political polarization and politically divisive articles are more likely to spread virally when they are low-reliability and the level of homophily is already high, generating an echo chamber-like social media environment. Strategic complementarities tend to amplify these pernicious effects of political polarization and divisive content as well.

Our second main result turns to social media platforms' algorithm design choices. We assume that platforms maximize engagement (in order to increase revenues from advertisements). Under this assumption, we establish a striking result: when the relevant articles have low reliability, social media platforms design algorithms that increase homophily and create filter bubbles, propagating misinformation. Intuitively, high-reliability content tends to spread anyway because most users recognize it as such and share it, and low homophily contributes to its spread by increasing its circulation throughout the network. In contrast, low-reliability content will be ignored or disliked by agents who disagree with its message and believe it to contain misinformation. Engagement with low-reliability content can be boosted if the platform ensures that it remains among users ideologically aligned with its message, who would be willing to share it with like-minded others without fear of being called out by users with different ideologies. Hence, creating filter bubbles becomes an attractive strategy for engagement-maximizing platforms. It is particularly troublesome that such filter bubbles are created precisely when the relevant content is low-reliability and likely to contain misinformation.

If platform algorithms are propagating misinformation, can public policy counter and discourage this type of behavior?[4] The answer is yes, but with some caveats. In the last part of the paper, we discuss four different types of regulatory policies, and in each case, we show how they may reduce misinformation but also point out the possibility that, if they are not designed well, they can backfire and exacerbate the problem.

First, we look at potential censorship of articles identified by a regulator as likely containing misinformation. While censorship can help reduce the viral spread of misinformation, it also generates an "implied truth" effect (Pennycook et al. (2020a)) that contributes to the viral spread of questionable content that escapes censorship. Second, we discuss regulations that force platforms to reveal the provenance of articles, making it easier for users to identify falsehoods (e.g., claims originating from less reputable sources, such as InfoWars). Though generally useful and sometimes more powerful than censorship, provenance regulation can also backfire. This is for a related reason: this policy also creates an implied truth effect because individuals rely on other users' verification of the content before them. Third, we discuss "performance targets", where the regulator places limits on the amount of misinformation that circulates on the platform. Such targets tend to better align platform and regulator preferences, but unless they appropriately monitor and penalize platforms for violations, strict targets can exacerbate the spread of misinformation. Lastly, we show how direct regulation of platform algorithms can reduce misinformation, but also point out that the non-monotone effects of homophily imply that such regulations need to be finely calibrated.

Related Literature. Our paper builds on a large body of work on models of misinformation. In addition to the literature mentioned previously, several other papers in this literature 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. (2021) and Mostagir and Siderius (2021), a strategic principal who wants to persuade agents of an incorrect belief can distort the learning

[4] As of August 2021, federal law protects social media platforms from being held responsible for content posted by its users (see Section 230 of the Communications Decency Act of 1996, discussed by `https://hbr.org/2021/08/its-time-to-update-section-230`).

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 relative to this literature is the possibility that misinformation spreads because of the strategic interactions of Bayesian agents and is exacerbated by profit-maximizing platform algorithms.

There is a growing literature on information design by platforms, building for the most part on the concept of Bayesian persuasion (Kamenica and Gentzkow (2011) and Kamenica (2019)). Candogan and Drakopoulos (2020) 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 (2021) 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. (2021) explore other strategic reasons for media bias. Our paper contributes to this literature by highlighting the role of ideological leaning, strategic interactions, ideological homophily, and platform algorithms.

The most closely related work to ours is Papanastasiou (2020) who studies a model where agents hold heterogenous ideological beliefs and digest (and potentially share) a news article sequentially. Our work is different in three important dimensions. First, Papanastasiou (2020) focuses on costly inspection, which makes sharing decisions strategic substitutes, while our model generates strategic complementarities, because individuals care about further shares of the content they share.[5] All of our results and formal analysis turn on strategic complementarities. Second, and relatedly, echo chambers play no role in Papanastasiou (2020).[6] Third, our analysis of engagement-maximization by the platform and its implications for the spread of low-reliability content has no counterpart in Papanastasiou (2020) or any other work in this area we are aware of.[7]

The rest of the paper is organized as follows. The next section introduces our basic environment and describes the information structure and payoffs. Section 3 characterizes the (Bayesian-Nash) 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 "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 the 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 Appendix A.

[5] Our reading of the evidence is that strategic complementarities are more relevant for social media behavior than strategic substitutabilities. For example, Eckles et al. (2016) find evidence that feedback or "encouragement" from peers about Facebook posts have contributed significantly to future behavior and posting. See also Taylor and Eckles (2018) and Aral and Dhillon (2018).

[6] As already noted, echo chambers appear central to the spread of misinformation in practice. See, for example, Lee et al. (2011), Törnberg (2018), Centola (2010), and Centola and Macy (2007)

[7] Papanastasiou (2020) also discusses platform incentives, but assumes that the platform is interested in limiting misinformation. Our reading of the evidence in this instance, too, favors our interpretation, where platforms such as Facebook are (or at the very least used to be before regulatory pressure mounted) fairly indifferent to the presence of misinformation but strongly prioritize engagement maximization.

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 with an *ex ante* distribution $H_i(\cdot)$, which may or may not be the same across agents.

Sharing Network. We assume there are N agents in the population, who share a news item according to a *sharing network* defined by a matrix $\mathbf{P}$ of link probabilities, with p_{ij} denoting the probability that agent i has a link to agent j. We define agent i's neighborhood $\mathcal{N}_i$ as the set of agents attached to her with an outgoing link, and denote her degree or the size of her neighborhood by $|\mathcal{N}_i|$. The sharing network reflects both an individual's social circle and the algorithms the platform uses for promoting shared content. The news item in question could be a news article or a post by one of the users, and throughout we refer to it as an "article".

Misinformation and News Generation. Each article has a three-dimensional type (r, m, ν). Here, $r \in [0, 1]$ indicates the *reliability* of the news, and $m \in \{L, R\}$ is the *message*, which corresponds to the article's viewpoint, for example, whether it argues for a left-wing or right-wing idea. Finally, 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*.[8]

We assume that, at the beginning of the game, the type vector (r, ν, m) is drawn according to the following i.i.d. process:

(i) The article's reliability $r \in [0, 1]$ is drawn from a continuous distribution F with density f.

(ii) The veracity of the article is $\nu = \mathcal{T}$ (contains truthful content) with probability $\phi(r)$ or is $\nu = \mathcal{M}$ (contains misinformation) with probability $1 - \phi(r)$. We assume that ϕ is increasing and differentiable in r, and satisfies $\phi(0) = 0$ and $\phi(1) = 1$, so that the least reliable article always contains misinformation, and as the degree of reliability increases, the likelihood of misinformation monotonically declines and reaches zero.

(iii) If $\nu = \mathcal{T}$ (the article is truthful), then its message is generated as $m = \theta$ with probability $p > 1/2$. Conversely, if $\nu = \mathcal{M}$ (the article contains misinformation), then its message is generated as $m = \theta$ with probability $q \leq 1/2$ and is weakly anti-correlated with the truth.

While m and r are common knowledge (for example, the message m is directly observed and reliability depends on certain commonly-observed characteristics such as source and headline), the third dimension, ν, is unknown to all agents. We assume that agents update their beliefs about ν using Bayes' rule given beliefs about the underlying state θ and the observables (r, m) of the article.

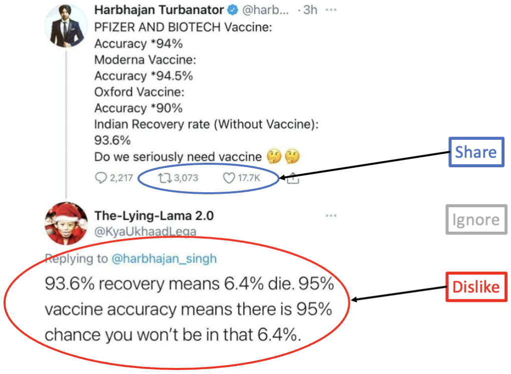
[8] Our focus in this paper is on misinformation, interpreted as items containing misleading information or arguments that can influence (a subset of) the public. Articles containing misinformation are in practice much more numerous than those that can be classified as "fake news", which explicitly propagate demonstrably false information (e.g., Egelhofer and Lecheler (2019), Allen et al. (2020), Guess et al. (2019), Grinberg et al. (2019)). For example, according to this definition a news item that favorably describes a report denying climate change, without putting this in the context of hundreds of other reports reaching the opposite conclusion or mentioning the criticisms that it has received from experts, contains misinformation.

Social Media Behavior. Time is discrete $t = 1, 2, \ldots$. Upon receipt of the article, an agent i can take one of three actions $a_i \in \{\mathcal{S}, \mathcal{I}, \mathcal{D}\}$, as described below:

(i) Share ($\mathcal{S}$): The agent decides to *share* the article and passes it onto others after her.

(ii) Ignore ($\mathcal{I}$): The agent decides to *ignore* the article and does not engage with it.

(iii) Dislike ($\mathcal{D}$): The agent decides to *dislike* the article, which means expressing disagreement or contempt for the content contained in it.



Harbhajan Turbanator @harb... · 3h ···
PFIZER AND BIOTECH Vaccine:
Accuracy *94%
Moderna Vaccine:
Accuracy *94.5%
Oxford Vaccine:
Accuracy *90%
Indian Recovery rate (Without Vaccine):
93.6%
Do we seriously need vaccine 🤔🤔
2,217 3,073 17.7K

The-Lying-Lama 2.0 ···
@KyaUkhaadLega
Replying to @harbhajan_singh
93.6% recovery means 6.4% die. 95% vaccine accuracy means there is 95% chance you won't be in that 6.4%.

Figure 1. Sample Tweet.

The three possible actions are depicted in Figure 1 using Twitter as a sample social media platform. A given user sees the article and decides how to respond to it. She can (i) share it ($\mathcal{S}$), which actively puts it on other social media news feeds; (ii) ignore it ($\mathcal{I}$), where the user simply scrolls past the article; or (iii) actively dislike it ($\mathcal{D}$), expressing derision for the content.

At time $t = 1$, we assume that some initial seed agent i^* first engages with the article. If the article is shared by agent i, it is passed to all $j \in \mathcal{N}_i$. In contrast, following ignore or dislike, the article does not propagate past agent i.

Payoffs. Let us define shares after i as $S_i = |\{j \in \mathcal{N}_i : a_j = \mathcal{S}\}|$ and dislikes after i as $D_i = |\{j \in \mathcal{N}_i : a_j = \mathcal{D}\}|$. Agent i's utility can then be written as

$$U_i = \begin{cases} 0, & \text{if } a_i = \mathcal{I} \\ \tilde{u}\mathbf{1}_{\nu=\mathcal{M}} - \tilde{c}, & \text{if } a_i = \mathcal{D} \\ u\mathbf{1}_{\nu=\mathcal{T}} - c\mathbf{1}_{\nu=\mathcal{M}} + \kappa S_i - dD_i, & \text{if } a_i = \mathcal{S} \end{cases} \tag{1}$$

where $\mathbf{1}$ is the indicator function (equal to 1 if true and 0 otherwise). Here, $\tilde{u}, \tilde{c}, u, c, \kappa$ and d are strictly positive parameters, which we discuss below.

(i) We normalize payoffs following ignore, $\mathcal{I}$, to $U_i = 0$.

(ii) Payoffs from dislike, $\mathcal{D}$, depend on whether the article contains misinformation. We assume, in particular, that disliking has a cost of $\tilde{c} > 0$, regardless of whether the article is truthful (because of, say, the effort required to actively call out misinformation). In addition, disliking an article containing misinformation has a benefit of $\tilde{u} > \tilde{c}$, because individuals like calling out misleading articles. This formulation implies that disliking is never preferred to ignoring for an article that is truthful with probability 1, and is always preferred to ignoring for an article that contains misinformation with probability 1.

(iii) Following a decision to share, $\mathcal{S}$, an agent receives utility from two sources. First, agents receive utility from sharing truthful content, but incur a cost from sharing misinformation. This explains the first component of utility following $\mathcal{S}$, $U_i^{(1)} = u\mathbf{1}_{\nu=\mathcal{T}} - c\mathbf{1}_{\nu=\mathcal{M}}$. Second, agents enjoy positive feedback from their peers (such as likes, or in our setting re-shares), but are negatively affected by dislikes. This is captured by the second component of utility $U_i^{(2)} = \kappa S_i - dD_i$.[9] In this formulation, the parameter κ captures the importance of "popularity" for the agent's sharing decision, while d represents the extent to which she cares about negative reactions. In Appendix B we provide a simple microfoundation for disutility from negative reactions based on reputational concerns.

Information Structure and Solution Concept. Agents are not aware of, and have uniform prior over, when the article was first introduced onto social media, the prior sharing process, and the structure of the social network (though the link matrix $\mathbf{P}$ is common knowledge).[10] Moreover, while any agent i knows the distribution $\{H_i\}_{i=1}^N$ of beliefs in the population, she does not know any agent j's belief (ideology) b_j. We focus on Bayesian-Nash equilibria, and refer to these as "equilibria" for short.

To eliminate trivial and unrealistic equilibria, we assume that the sensationalism of an article is upper bounded by $\bar{\kappa} = (c\tilde{c} - u(\tilde{u} - \tilde{c}))/(\tilde{u}N)$. This assumption guarantees that there is never an equilibrium where *every* agent *always* shares *all* articles. It also eliminates equilibria where agents may share and dislike, but never ignore.

Discussion—The basic assumptions introduced above are consistent with salient patterns of behavior and information structure in social media. First, as documented in studies such as Pennycook et al. (2021), users want to share content they believe to be truthful and not contain misinformation. Second, while users derive value from peer encouragement and re-shares on social media (Eckles et al. (2016)), they also suffer reputational costs when they get called out for sharing misinformation (see, for example, evidence from Facebook in Altay et al. (2020)). Finally, social media users often engage in criticisms of available content and inform others about misinformation (see, for example, Kim et al. (2020) for evidence in the context of 2018 midterm elections).

[9] Equivalently, the terms κS_i and dD_i could be replaced by arbitrary functions $\varphi_S(\kappa, S_i)$ and $\varphi_D(d, D_i)$ that satisfy $\varphi_S(0,\cdot) = \varphi_S(\cdot, 0) = \varphi_D(0,\cdot) = \varphi_D(\cdot, 0) = 0$ and have (weakly) increasing differences. This generalization captures a broad range of peer feedbacks based on sharing different types of content, beyond the additive structure we adopted for notational simplicity in the text.

[10] Hence, agents do not know the exact interactions and sharing patterns outside their neighborhood, which is consistent with the evidence in Breza et al. (2018). That being said, because the equilibrium sharing process is Markovian, this assumption can be relaxed by replacing $\mathbf{P}$ with the adjacency matrix.

3 Equilibria in General Networks

In this section, we characterize the structure of equilibria for any sharing network structure $\mathbf{P}$ and provide various comparative statics. Without loss of generality (and ease of exposition), we fix the article's message as $m = R$ for the remainder of the paper.[11]

3.1 Cutoff Strategies and Strategic Complementarities

When agent i receives an article with reliability r and message $m = R$, she updates her *(ex post)* belief, π_i, that the article is truthful according to Bayes' rule:

$$\pi_i = \frac{(pb_i + (1-p)(1-b_i))\phi(r)}{(qb_i + (1-q)(1-b_i))(1-\phi(r)) + (pb_i + (1-p)(1-b_i))\phi(r)}. \tag{2}$$

Clearly, π_i is increasing in b_i since an agent is more likely to believe in an article's veracity when its message agrees with her prior. Moreover, π_i is increasing in r, as the agent updates more on the basis of more reliable articles.

We can also see that the payoff to sharing ($\mathcal{S}$) increases in π_i, since the first component of utility, $U_i^{(1)}$, is increasing in π_i (as the individual would like to share truthful articles), while $U_i^{(2)}$ is independent of π_i. With a similar reasoning, the payoff to disliking ($\mathcal{D}$) is decreasing in π_i, whereas the payoff to ignoring ($\mathcal{I}$) is independent of π_i. This monotone behavior of payoffs will lead to best-response decision rules for agents that take the form of cutoff strategies, as we explain next.

We say that agent i employs a *cutoff strategy* if there exists $b_i^*(r)$ and $b_i^{**}(r)$ such that agent i chooses $\mathcal{S}$ when $b_i > b_i^{**}(r)$, chooses $\mathcal{I}$ when $b_i^*(r) < b_i < b_i^{**}(r)$, and chooses $\mathcal{D}$ when $b_i < b_i^*(r)$. Cutoff strategies in our context imply that agents who strongly agree with an article tend to share it, agents who strongly disagree with it tend to choose dislike, and those with intermediate beliefs typically ignore the article.

We will see in the next theorem that all equilibria are in cutoff strategies. This means, in particular, that an equilibrium can be summarized by cutoff vectors $(\mathbf{b}^*, \mathbf{b}^{**}) = (b_1^*, b_1^{**}, \dots, b_N^*, b_N^{**})$. Furthermore, these cutoffs $b_i^*(r)$ and $b_i^{**}(r)$ will both be decreasing in r, so that as reliability increases, an article becomes more likely to be shared and less likely to be disliked.

We can also note that our social media game exhibits *strategic complementarities*. To see this, observe that when others share more—meaning that b_i^{**} (weakly) decreases for all i—the second component of utility, $U_j^{(2)}$, increases for each agent j, and this raises the overall utility of sharing and encourages more sharing. Similarly, when others reduce their likelihood of disliking, meaning that now b_i^* (weakly) decreases for all i, this reduces the likely cost of sharing misinformation by mistake, also raising $U_j^{(2)}$. Strategic complementarities capture an important dimension of social media interactions—utility feedback from others' behavior tends to encourage agents to cohere with those behaviors.

Equilibrium Structure. The next theorem shows that an equilibrium always exists and is in cutoff

[11] To see that this is without loss of generality, observe that the analysis applies identically with an $m = L$ message but with complementary priors $b_i' = 1 - b_i$.

strategies. At the same time, strategic complementarities ensure that there is a well-defined structure to the set of equilibria. To make this more concrete, we say that an equilibrium $(\mathbf{b}^*, \mathbf{b}^{**})$ has *uniformly more sharing* than other $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ if $\mathbf{b}^* \preceq \hat{\mathbf{b}}^*$and $\mathbf{b}^{**} \preceq \hat{\mathbf{b}}^{**}$ (where $\preceq$ is the component-wise order). This, in particular, means that all the thresholds for each agent is (weakly) lower in the former equilibrium (recall that lower thresholds mean more sharing).

Theorem 1.

(i) There exists a Bayesian-Nash equilibrium;

(ii) All equilibria are in cutoff strategies;

(iii) The set of cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$ *forms a lattice, and thus there exists a least-sharing and most-sharing equilibrium.*

The structure of equilibria characterized in Theorem 1 facilitates our analysis, enabling us to focus on two sets of thresholds, $(\mathbf{b}^*, \mathbf{b}^{**})$, which are themselves monotone in the reliability of the article in question. Because of strategic complementarities, there can be multiple equilibria: when others are choosing to share an article with middling reliability, this further encourages sharing because one's own post will circulate more, increasing the utility from sharing. Conversely, if the same article with middling reliability is not shared by others, the payoff to sharing is reduced, while the cost of being found out to have circulated misinformation remains constant. This then discourages sharing.

Theorem 1 also shows that, despite this multiplicity, there are two focal equilibria on which we can concentrate: the equilibrium with the smallest vector of cutoffs (*most-sharing equilibrium*) and the equilibrium with the largest vector of cutoffs (*least-sharing equilibrium*).

Finally, the theorem's characterization provides an explicit measure of the amount of sharing. Recall that agent i's prior belief b_i is drawn *ex ante* from the distribution H_i. Hence, in an equilibrium with cutoff b_i^{**} for agent i, the *ex ante* likelihood that this agent will share is $1 - H_i(b_i^{**})$. Therefore, the most sharing equilibrium, which has the smallest equilibrium b_i^{**} for all i, has the highest likelihood that any agent i will share the article in question.

3.2 Content Virality and Comparative Statics (for Fixed $\mathbf{P}$)

In this subsection we provide comparative statics for the most and least sharing equilibria as we change the parameters of the social media game, but holding the sharing network $\mathbf{P}$ fixed. We discuss comparative statics with respect to the network in the next section. Toward this goal, we define the notion of *content virality*, which captures the expected spread of an article in the sharing network.

Content Virality. Formally, we define content virality as follows. We suppose that an article is seeded at some agent i^* at $t = 1$. We then define $\mathbf{S}_{i^*}$ as the (random) proportion of the population that shares when agent i^* is the seed agent in the most-sharing equilibrium σ. We say σ_1 has more *content virality* than σ_2 if $\max_{i^*} \mathbb{E}_{\sigma_1}[\mathbf{S}_{i^*}] \geq \max_{i^*} \mathbb{E}_{\sigma_2}[\mathbf{S}_{i^*}]$. In words, our notion of content virality compares the spread of an article provided that it starts from the seed that is most favorable to its ultimate circulation. The reason we start from the most favorable seed is that, as we will see in Section 5, social media platforms have an incentive to implement sharing algorithms that place articles in such favorable seeds. For future reference, we also note that content virality is also the same as expected overall engagement with an article, conditional on favorable seeding.

Quantity of Misinformation, Sensationalism, and Reputation. The next proposition shows how the quantity of misinformation, sensationalism, and reputational concerns affect content virality. We define less misinformation as a shift of the function ϕ to some $\phi' \geq \phi$ (pointwise). The parameter κ captures how *sensational* the article is: higher κ implies that agents receive greater value from future shares, because these shares are associated with others paying more attention or perhaps being entertained more by the relevant posts. This greater utility is independent of the content's veracity. We think of κ varying at the level of articles, so that some articles will be more sensational than others. Finally, the parameter d proxies for for the importance of *reputational concerns.* Higher d means that dislikes are more damaging, which corresponds to the agent being more concerned about receiving many dislikes. We think of d as varying at the level of communities (certain communities of users, for example, academics, may have more reputational concerns).

Proposition 1. *Less misinformation, higher sensationalism, and weaker reputational concerns lead to greater content virality.*

These results are intuitive and immediate.[12] Holding constant the reliability of the article, less misinformation reduces the cost associated with sharing, triggering more aggressive sharing by all agents. This prediction is consistent with Pennycook and Rand (2019) who show low-reliability content (e.g., Breitbart or Infowars) is not typically shared by attentive social media users, regardless of partisanship. This proposition also clarifies that viral spread of misinformation is not a mechanical effect in our model: if anything, less reliable articles that are more likely to contain misinformation are less likely to become viral. We will see that other aspects of social media interactions, in particular the topology of the sharing network, are often responsible for viral spread of misinformation.

The comparative static with respect to sensationalism coheres with the patterns documented in Duffy et al. (2020), suggesting that social media participants often share a story that is "too good not to share", and do so even when they realize it is also "too good to be true". The link between reputational concerns and misinformation is also consistent with the evidence that in settings where reputation matters misinformation is less likely (Altay et al. (2020)), and when such reputational concerns are missing, even calling out individuals sharing misinformation is fairly ineffective (Mosleh et al. (2021a)).

Finally, this proposition provides a possible pathway for low-reliability content to become viral. Vosoughi et al. (2018) argued that misinformation spreads farther, faster, deeper and more broadly than truthful news on social media. This evidence was criticized by Grinberg et al. (2019) who showed that, once the effects of sensational news items is controlled for, misinformation does not spread farther (or faster) than truthful content. Proposition 1 provides a rationalization of these patterns.

[12] In fact, the claim in Proposition 1 can be strengthened to the notion of *uniformly more sharing* where *all* agents in the sharing network share with strictly higher probability, which immediately implies higher content virality. We focus on content of virality, both because it is simpler and also because the comparative static results with respect to homophily in the next section do not always lead to uniformly more or less sharing.

All else equal, misinformation does not spread faster than truthful content as in Grinberg et al. (2019). However, because, as observed in Molina et al. (2021) and Kozyreva et al. (2020), sensational content is often low-reliability, these two propositions together imply that misinformation may be more likely to become viral. Whether this happens or not depends on the boost from sensationalism. When this is limited, low-reliability articles containing misinformation spread less because of concerns of users that others will call them out for sharing this content. But when this sensationalism boost is high, misinformation can become viral. Additionally, the strategic complementarity in sharing decisions implies that sufficiently sensational misinformation can become viral, because once an individual thinks others are going to share this sensational item, she becomes much more likely to share herself, even if she has doubts about its veracity.

4 Island Networks and the Implications of Homophily

In this section, we present comparative static results with respect to the sharing network $\mathbf{P}$. Throughout this section, we take the sharing network as given, and then return to how it is shaped by the algorithms of social media platforms in Section 5.

The focus on comparative statics with respect to the sharing network necessitates two modifications from our analysis so far. First, we restrict attention to *island networks* (or equivalently, the stochastic block model), which are lower-dimensional than general networks we have allowed so far. Namely, in an island network, 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 then given as:

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

where $p_s \geq p_d$. Without loss, we assume each of the islands is weakly connected.

Second, we assume the prior distribution for agents on the same island ℓ is the same, and is denoted by H_ℓ. We also assume that islands are ranked according to their belief distributions. In particular, each island ℓ has distribution H_ℓ with support on $[b^{(\ell)}, b^{(\ell+1)}]$, where $1 \geq b^{(1)} > b^{(2)} > \ldots > b^{(k)} > b^{(k+1)} \geq 0$.[13] This implies that lower-indexed islands have stronger right-wing beliefs.

An important advantage of island networks, in addition to their lower-dimensional representation, is that, combined with this ranking assumption, they enable us to model the degree of *homophily*—the extent to which an individual interacts with others that have common characteristics as herself. Common characteristics for us are those that are relevant for prior beliefs, and therefore, by construction, individuals have more in common with those on the same island as themselves. As a result, homophily will be higher when most links are within islands and links between islands are sparse (high p_s and low p_d).[14]

[13] This assumption is adopted for simplicity. Our results generalize if we instead assume that these distributions are ranked in terms of first-order stochastic dominance: $H_1 \succeq_{FOSD} H_2 \succeq_{FOSD} \cdots \succeq_{FOSD} H_k$. However, this generalization requires considerably more formalism and notation, motivating our focus on disjoint supports.

[14] The homophilic structure and greater congruence of beliefs within islands are consistent with the evidence presented in

More formally, we say that an island network with (p_s, p_d) has *more homophily* than an island network with (p'_s, p'_d) if all agents have the same expected degree under both, but where $p_s > p'_s$ and $p_d < p'_d$.[15] From Theorem 1, we know that the equilibrium is in cutoff strategies of the form $(\mathbf{b}^*, \mathbf{b}^{**}) \equiv (b_1^*, b_1^{**}, \dots, b_N^*, b_N^{**})$. However, because there is symmetry within islands, equilibria now take a simpler, "semi-symmetric" form as shown in the next lemma.

Lemma 1. *All equilibria are semi-symmetric: for every equilibrium, there exist* $\{(b_\ell^*, b_\ell^{**})\}_{\ell=1}^k$ *such that* $b_i^* = b_{\ell_i}^*$ *and* $b_i^{**} = b_{\ell_i}^{**}$ *for all agents* i *in island* ℓ.

The simplification established in Lemma 1 will allow us to work with a lower dimensional cutoff vector (just two cutoffs for each island).

4.1 Comparative Statics: Homophily

The next theorem, characterizing the effects of homophily on the spread of misinformation, is our first main result:

Theorem 2. *There exist* $0 < \underline{r} < \bar{r} < 1$ *such that:*

(a) If $r < \underline{r}$*, an* increase *in homophily increases the virality of content.*

(b) If $r > \bar{r}$*, a* decrease *in homophily increases the virality of content.*

Theorem 2 shows how low-reliability content can spread virally in networks with high homophily. Intuitively, when content comes from a low-reliability source, only agents who (strongly) agree with the article's message share it. However, as homophily increases, users know that they will mostly share with other like-minded people, who will also be inclined to share this content. This creates a type of echo chamber: the likelihood of being called out for spreading misinformation is now lower, making users "less disciplined" or more likely to share lower-reliability content. Strategic complementarities then extend these incentives throughout the network. In this way, homophily leads to the viral spread of low-reliability articles that likely contain misinformation.

However, Theorem 2 shows that homophily can have non-monotone effects. This is because greater homophily also keeps an article circulating among the same group of like-minded users and reduces the likelihood that it will reach other communities. Theorem 2(a) establishes that the first effect of homophily, working through incentives to share low-reliability content, is more powerful than the second, "circulation effect", when we focus on particularly low-reliable content (with $r < \underline{r}$). This implies, in particular, that homophily's impact is to increase the virality of especially low-quality content, which is of course relevant for public policy (as we discuss in Section 6).

The results in Theorem 2 are in line with recent evidence highlighting the importance of echo chambers for the spread of misinformation. Törnberg (2018) and Vicario et al. (2016), among

Bakshy et al. (2015): "friend networks" on Facebook are ideologically segregated, with the median share of friends from the opposing ideology around only 20%. Mosleh et al. (2021b) provides evidence of similar homophily on Twitter.

[15] Because network density raises connectivity and can directly increase virality, we hold network density fixed in order to isolate the effects from homophily.

others, show that homophily in sharing behavior propagates ideologically-congruent ideas, with little incentive to question the veracity of this information, while Quattrociocchi et al. (2016) document how echo chambers on Facebook fuel conspiracy theories and the popularization of incorrect scientific ideas, for example, on vaccines. Levy (2021) provides evidence that "filter bubbles" generated by Facebook's algorithms are an important source of propagation of misinformation.

4.2 Comparative Statics: Divisive Content and Belief Polarization

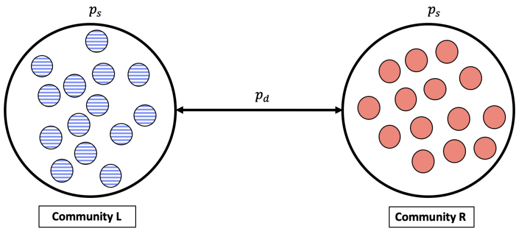


Figure 2. Two-Island Model.

In this subsection, we provide an additional comparative static with respect to polarization. For this result, we focus on the case of just two islands, a left-wing and a right-wing one with prior distributions H_L and H_R, respectively, as pictured in Figure 2. Moreover, we suppose that there is disjoint support of prior beliefs across communities. Formally, we assume H_R has support on $[\underline{b}_R, \bar{b}_R]$ and H_L has support on $[\underline{b}_L, \bar{b}_L]$, with $\bar{b}_L < 1/2 < \underline{b}_R$.

We say content with parameters (p', q') is *more divisive* than content with parameters (p, q) if $p \geq p'$ and $q \leq q'$. Divisive content has a message that is more tethered to the true state θ when it is truthful (and more likely to argue against θ if it is misinformation). In our case, we think of state θ as related to political ideology. Therefore, non-political content, such as wedding photos or cat videos, has little divisiveness relative to more political ones, such as "Obama Signs Executive Order Banning The Pledge of Allegiance in Schools Nationwide" (Fourney et al. (2017)). Note also that, from equation (2), prior beliefs about θ matter more for updating and the assessment of an article's veracity when content is more divisive.

We say H_2 is *more polarized* than H_1 if it satisfies the following single crossing property: $H_2^{-1}(\alpha) - H_1^{-1}(\alpha)$ is a nondecreasing function in α, crossing zero at $\alpha^* = 1/2$ with $H_1(1/2) = H_2(1/2) = 1/2$. An increase in polarization results in a "stretching" of the belief distribution around the most moderate user (i.e., $b = 1/2$) while preserving an equal distribution of left-wing and right-wing agents (meaning that $H(1/2) = H(1/2) = 1/2$, which is applied in the island model to the average distribution of beliefs, $H = \frac{1}{N}\sum_{\ell=1}^{N} N_\ell H_\ell$). The available evidence indicates that the US public has become more polarized (see Pew Research Center (2014) and Abramowitz (2010)), and an important question of debate has been whether this polarization has fueled the spread of misinformation on social media.

The next result studies how political divisiveness and polarization impact social media behavior and the spread of misinformation, as a function of the homophily in the sharing network.

Proposition 2. *There exist* $r^* \in (0, 1)$ *and* $p^* \in (0, 1)$ *such that:*

(a) If $r < r^*$ and $p_s/p_d > p^*$, then an *increase* in divisiveness or an increase in polarization leads to greater content virality.

(b) If $r > r^*$ and $p_s/p_d < p^*$, then a *decrease* in divisiveness or a decrease in polarization leads to greater content virality.

Proposition 2 is complementary to Theorem 2. When the content in question has high reliability ($r > r^*$) and homophily is limited, more divisive content or greater polarization tends to reduce content virality, because in a well-connected, non-homophilic network, controversial articles will solicit a wide range of reactions, disciplining those tempted to share misinformation. In contrast, when the article in question has low reliability and there is significant homophily, there are again echo chamber-like effects. More divisive content generates more divergent behavior from individuals with different ideologies, and greater polarization means there are sharper differences in terms of these ideologies. As a result, echo chambers matter especially for divisive content and in the presence of polarization. Strategic complementarities once again amplify this effect, as users recognize that others in their community will tend to share divisive content, and this makes them even more willing to share.

It is notable that Proposition 2, like Theorem 2, implies that greater divisiveness and polarization increase the virality of especially low-reliability content, which is most likely to contain misinformation. These two results together thus imply that echo chambers, greater political polarization and divisive content all exacerbate the circulation of misinformation on social media.

5 Platform Design and Filter Bubbles

We now turn to our second main result: how platform behavior affects misinformation. Consider a collection of social media users with beliefs distributed according to a distribution H. The platform can identify communities of users according to prior ideological beliefs, for example, based on content previously shared or affiliations with ideological groups. In particular, each user is binned into one of k communities, with each community ℓ having a belief distribution H_ℓ with support over $[b_\ell, b_{\ell+1}]$, and where $1 \geq b^{(1)} > b^{(2)} > \cdots > b^{(k)} > b^{(k+1)} \geq 0$ (with at least one left-wing and one right-wing community). The size of these bins depends on the platform's microtargeting technology at identifying users' ideological beliefs (see, for example, Papakyriakopoulos et al. (2018)). Formally, we let $\varepsilon \equiv \max_\ell(b_{\ell+1} - b_\ell)$, with the interpretation that lower values of ε correspond to better platform technology for identifying ideology.

The platform's objective is to maximize user engagement, which is equivalent to maximizing content virality (see the definition of content virality in Section 3.2).[16] The platform does not directly care about whether the content users are engaging with is truthful or contains misinformation.

[16] This objective is rooted in 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 Facebook's total revenue in 2011 was from advertising, and from 2017-2019, around 98% was (see Andrews (2012) and `https://www.nasdaq.com/articles/what-facebooks-revenue-breakdown-2019-03-28-0`).

Strictly speaking, we are modeling social media platform objectives before the more recent public backlash over misinformation. If the platform faces potential penalties from public backlash or regulators for spreading misinformation, its objective function will change, as we explore in greater detail in Section 6.

The platform chooses how content is shared across users. That is, for each article, the platform not only picks the seed agent at $t = 1$ to whom it recommends this article, but also chooses the sharing network—the matrix of link probabilities $\mathbf{P}$.[17] The platform's choice of $\mathbf{P}$ can be interpreted as its "algorithm" to determine how users are exposed to content circulating in the social media site. This algorithm choice is assumed to be common knowledge.

5.1 Optimality of Island Networks and Filter Bubbles

We remind the reader that the island networks of Section 4 are parameterized by three components: the within-island link probability (p_s), the across-island link probability (p_d), and the number of islands (k). As special cases, we have (i) an island model that has *maximal homophily*, where $p_s > 0$ but $p_d = 0$ (and thus there is extreme ideological segregation on the network); and (ii) an island model with *maximal connectivity*, where $p_s = p_d$ (and there is minimal homophily and no segregation by ideology). We next show that although the platform is allowed to design any sharing network $\mathbf{P}$, its profit-maximizing choice is within the class of island networks.

Theorem 3. *There exists $\bar{\varepsilon} > 0$ such that if $\varepsilon < \bar{\varepsilon}$, the platform's profit-maximizing sharing network is determined by a reliability threshold $r_P \in (0, 1)$ such that:*

(i) If $r < r_P$, the platform's profit-maximizing sharing network has maximal homophily.

(ii) If $r > r_P$, the platform's profit-maximizing sharing network has maximal connectivity.

Part (i) of the theorem shows that when articles are mostly unreliable—and likely to contain misinformation—the platform creates an extreme filter by designing its algorithms to achieve a sharing network with the greatest homophily. In contrast, part (ii) demonstrates that when articles have higher reliability, the platform refrains from introducing algorithmic homophily. This result highlights an important channel by which misinformation spreads: it is precisely when articles are likely to contain misinformation that the platform seeks to maximize engagement by creating (endogenous) echo chambers, or filter bubbles, where these articles spread virally within like-minded communities. Put differently, with low reliability content, neither the platform nor the users are disciplined about sharing misinformation, and so these news items spread virtually uninhibited.

We also remark that the threshold r_P parameterizes the extent of filter bubbles on the platform. Specifically, the most extreme left-wing agent will be exposed to all $m = L$ articles, regardless of their reliability, but only see right-wing articles with reliability $r \geq r_P$. Similarly, the most extreme right-wing agent will see all $m = R$ articles, but only $m = L$ articles with $r \geq r_P$.

It is further worth noting that this theorem builds on but also significantly strengthens Theorem 2. In Theorem 2, the effects of homophily are non-monotone and are ambiguous when an article is neither very low reliability nor very high reliability ($r \in (\underline{r}, \bar{r})$). In contrast, Theorem 3 gives a sharp characterization of the platform's algorithm: when $r > r_P$, the platform goes for maximal connectivity,

[17] Facebook's algorithms may induce different sharing networks depending on features of the article, such as whether it contains cat videos, wedding photos, or political content. See https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/.

and when $r < r_P$, it chooses maximal homophily, and in this case, misinformation spreads virally precisely because of the echo chambers that the platform has manufactured. In both cases, the island structure of the network considered in Section 4 arises endogenously as the profit-maximizing sharing network for the platform.

In addition, Theorem 3 shows that when a platform can shape the network topology through its recommendation algorithm, the echo chamber effect that arises from Theorem 2(a) is exactly what fuels misinformation (whereas in Theorem 2(b), echo chambers were harmless). This observation generalizes to cases where the platform has less precise microtargeting technology, albeit in a less sharp way than the result presented in Theorem 3.[18]

Remark—In Theorem 3, 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 amplify content across different links, the same result readily follows.

5.2 Comparative Statics on r_P: Divisiveness and Polarization

In Theorem 3, the threshold r_P fully summarizes the extent to which misinformation will spread virally on social media.

We next perform comparative statics for this threshold to understand the conditions under which the platform will create a filter bubble and propagate misinformation.

Proposition 3. *The reliability threshold r_P increases as message divisiveness and/or belief polarization increases.*

Proposition 3 mimics the conclusions of Proposition 2(a). As divisiveness or polarization increases, content is consumed more aggressively within echo chambers and scrutinized more aggressively outside of them. Under these conditions, filter bubbles become more advantageous to the platform, especially when the relevant content has low reliability. This is because communities with more extreme beliefs now feel more strongly about news in general, rarely second-guess politically-congruent news, but often doubt and dislike counter-attitudinal news. As a result, low-reliability content spreads virally inside the platform's filter bubbles. In contrast, outside of the filter bubble, this content would have been quickly disliked and stopped—which is the reason why the platform favors algorithms that induce such filter bubbles.

Proposition 3 also provides a possible (albeit of course speculative) interpretation for why accelerating political polarization and identity politics in the last two decades may have come with

[18] When $\varepsilon > \bar{\varepsilon}$, the platform still induces echo chamber-like environments to generate viral spread of low-reliability content, even if its choice does not take the form of an island network. For instance, if there are only three communities (with broad ideology spectra), a misinformation right-wing article can still spread virally via the usage of filter bubbles. However, the platform may now prefer a sharing network that lies outside the class of island networks considered in Section 4. Specifically, user engagement may be maximized if the left-wing community is completely disconnected from all other communities, but the moderate (middle) community has sparse connections to the right-wing community. This facilitates a strong echo chamber within the right-wing community but also allows the article to spread to the more moderate community (while receiving no discipline from the left-wing community).

more aggressive filter bubble algorithms from social media sites (Apprich et al. (2018)). 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, its 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...”.[19]

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 four distinct types of regulations that have been discussed in this context: (1) *censorship* or tagging of misinformation; (2) regulations that force platforms to reveal articles’ *provenance*; (3) *performance targets* that require the platform to keep misinformation below a given threshold; and, (4) *network regulations*, restricting the extent of ideological homophily or segregation introduced by platform algorithms intended to maximize engagement.

We consider the effects of these policies when the platform can optimally choose the sharing network in response to the public policy. For simplicity, we suppose the regulator’s objective is to decrease the virality of articles containing misinformation on the platform. We say a policy is more *effective* than another policy (or no policy) if it reduces the virality of misinformation (and is *most* effective if more effective than any other feasible policy). Throughout, we fix the reliability of the article and assume the most-sharing equilibrium before the regulation involves some agents sharing and some agents not sharing (i.e., $\mathbf{b}^* \neq \mathbf{0}$ and $\mathbf{b}^{**} \neq \mathbf{1}$), allowing for the possibility that regulation might backfire and increase the virality of low-reliability content, or potentially help by reducing the virality of such content likely to contain misinformation.

6.1 Censorship

We first consider a policy where the regulator can censor misinformation that appears on the platform (also known as “content moderation”).[20] Formally, we model this as the regulator being able to adopt a policy that removes at most $\delta \in (0, 1)$ fraction of the content containing misinformation (with each

[19] 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/.

[20] Alternately, it can “tag” the article in question as disputed by outside sources, with analogous implications. See, for example, Facebook’s policies leading up to the 2020 election on labeling suspected misinformation: https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/.

piece of misinformation removed with probability δ).[21] In other words, the regulator selects $\delta^* \leq \delta$, with δ^* proportion of misinformation removed at $t = 0$, before it is observed by any of the users.

Proposition 4. *There exist* $0 < \delta_1 < \delta_2 < \delta_3 < 1$ *such that:*

(a) If $\delta \in (0, \delta_1) \cup (\delta_3, 1)$*, then* $\delta^* = \delta$ *is the most effective policy;*

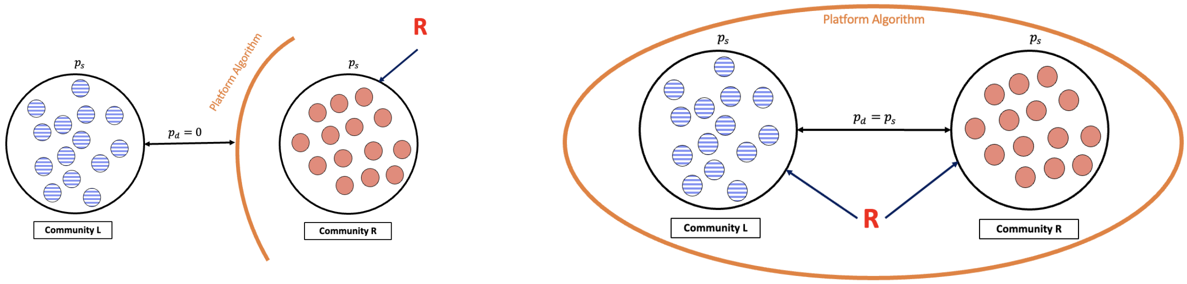
(b) If $\delta \in (\delta_1, \delta_2)$*, the most effective policy sets* $\delta^* < \delta$*.*

To understand this result, note that censorship has a two-pronged effect. On the one hand, it removes misinformation from circulation and prevents its potential to spread on the platform. On the other hand, it generates an "implied truth" effect for uncensored articles (as empirically identified in Pennycook et al. (2020a)). Bayesian users believe, correctly, that articles are more likely to be truthful when there is censorship of misinformation. In this case, the platform might naturally expand its recommendation filter bubble to generate more engagement, increasing the virality of any remaining misinformation. In some cases, this latter effect may more than offset any gains from the detection and elimination of misinformation.

In part (a), this implied truth effect is not sufficiently powerful, and as a result, both limited (small δ) and highly effective (large δ) censorship lead to better outcomes. Consequently, the policymaker should always censor as much as technologically feasible. In the small δ regime, the sharing network chosen by the platform remains constant and the censorship helps filter out a fraction of the misinformation. In the large δ regime, censorship can remove most of the misinformation, which is the most effective policy in any sharing network, including the one selected by the platform. In the intermediate censorship regime, however, more censorship might exacerbate the spread of misinformation. As we illustrate in the following example, intermediate censorship may create such a serious backlash that it exacerbates the spread of misinformation relative to no censorship.

Example 1. Let us consider the two-island setup depicted in Figure 2 from Section 4.2, where there are $N/2$ left-wing agents with belief $b_L = 5/12$ and $N/2$ right-wing agents with belief $b_R = 7/12$. Let us

[21] We think of δ as being a technology parameter related to how effective the regulator is in identifying misinformation. The assumption that the regulator may make type-I errors but not type-II errors (truthful articles are never misidentified, but misinformation is identified with some probability less than one) is adopted for simplicity.



(a) Low censorship.

(b) High censorship.

Figure 3. Optimal Platform Sharing Networks for Example 1 under Censorship Policies.

consider an article with a reliability score indicating it is equally likely to contain misinformation or to be truthful ($\phi(r) = 1/2$), but which is perfectly informative about the state θ ($p = 1$ and $q = 0$).

We assume that $u = c = 1$ and $\kappa = 1/(2N)$ so the payoff from sharing for agent i is given by $U_i = (2\pi_i - 1) + (S_i - D_i)/(2N)$, where π_i is agent i's posterior belief that the article is truthful conditional on reliability and message $m = R$. At the same time, we assume $\tilde{u} = 1$ and $\tilde{c} = 0$, so the payoff from disliking is $1 - \pi_i$ (which by nature of $\pi_i \geq 0$ is always a better response than ignoring).

With no censorship policy ($\delta = 0$), the optimal platform sharing network is given by Figure 3a, where the algorithm applies a filter bubble to the right-wing island, shielding the left-wing island from receiving the content. The article spreads among $N/2$ proportion of the population. Once a censorship policy is adopted, the implied truth effect will replace $\phi(r)$ with $\tilde{\phi}(r) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))}$, leading to a higher value for π_i on both the left and right-wing islands. Consider three separate regimes:

1. *Limited censorship*: With limited censorship ($\delta < 2/7$), the optimal platform sharing network remains the same as in Figure 3a. However, the virality of misinformation declines to $(1-\delta)N/2 < N/2$, and thus overall misinformation is reduced.

2. *Intermediate censorship*: With a more aggressive censorship policy ($2/7 < \delta < 1/2$), the optimal platform sharing network switches to the one shown in Figure 3b (maximal connectivity), but still does not filter out most of the misinformation. The platform selects a more expansive sharing network because, with censorship, platform users correctly believe that any given content is less likely to contain misinformation, even counter-attitudinal messages. The platform responds to this by moving from a maximally homophilic network to one with maximal connectivity (as per Theorem 3). The resulting virality of misinformation then becomes $(1 - \delta)N$, which is greater than $N/2$ for all $2/7 < \delta < 1/2$. In this range, censorship is worse than no censorship policy at all.

3. *Highly effective censorship*: With a censorship policy that can accurately detect most misinformation ($1/2 < \delta < 1$), the policy reduces misinformation, even though the platform again adjusts its algorithms in response to censorship in order to increase the virality of undetected misinformation. ∎

6.2 Provenance

Next, we consider a policy that requires the platform to reveal the original context or *provenance* of a piece of content. For example, provenance may point the user to a peer-reviewed medical study or the full discourse from which a quote was pulled. Such a policy allows users to verify (or "fact-check") social media content easily and quickly.

We model a provenance policy by allowing users to fact-check the article before making their share and dislike decisions. We assume that revealing provenance allows each agent to identify misinformation with (across-user independent) probability $\rho \in (0, 1)$; a truthful article is never misidentified as misinformation. Hence, more effective provenance policies allow a greater fraction of users to quickly identify misinformation.[22]

[22] In practice, certain demographic groups, such as users over 65 years old, appear more likely to accept (blatant)

Proposition 5. *There exist* $0 < \rho_1 < \rho_2 < \rho_3 \leq \delta_3 < 1$ *such that:*

(a) If $\rho \in (0, \rho_1) \cup (\rho_3, 1)$*, then* $\rho^* = \rho$ *is the most effective policy;*

(b) If $\rho \in (\rho_1, \rho_2)$*, then* $\rho^* < \rho$ *is the most effective policy.*

Moreover, a provenance policy with $\rho \in (\delta_3, 1)$ *is more effective than a censorship policy with* $\delta = \rho$*.*

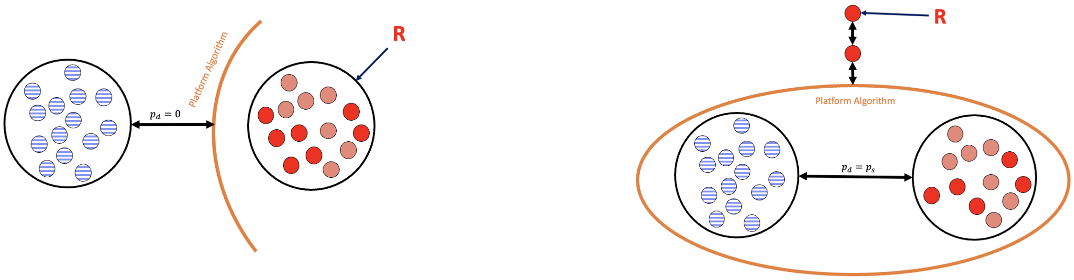
The result is similar to Proposition 4: soft and strong provenance policies are always effective, but moderate provenance policies can exacerbate the spread of misinformation.[23] The proposition also establishes that provenance policies are in some sense more effective than censorship policies when implemented well. Decentralized fact-checking reduces the likelihood of type-I errors (misidentifying misinformation as truthful) that can result in large share cascades similar to those in Example 1. Because multiple users are independently assessing veracity through the provenance channel, misinformation will tend to be stopped as it is checked along various paths in the sharing network. Strategic complementarities further amplify this effect: because users are aware that the provenance policy may allow others downstream to identify misinformation, they are also more cautious themselves in sharing low-reliability content. That being said, provenance policies are not always superior to censorship policies, as illustrated by the following example.

Example 2. Let us consider the setting of Example 1, with the slight amendment that the $N/2$ right-wing agents are split into $N/4$ extreme right-wing agents (with belief $b_{RR} = 3/4$) and $N/4$ moderate right-wing agents (with belief $b_R = 7/12$). It is straightforward to verify that the profit-maximizing sharing network for the platform with no policy is still the same as in Example 1 (Figure 4a), and that a censorship policy of $\delta = 3/16$ has the same effect as before (in particular, it is more effective than no policy because $\delta < 2/7$).

Let us now consider a provenance policy with $\rho = 3/16$ (which in this case is in the range (ρ_1, ρ_2)). Here, the following sharing network increases engagement relative to the network in Figure 4a: agent 1

misinformation, perhaps because of poor media interpretation skills (see Grinberg et al. (2019) and Guess et al. (2019)). Thus, provenance policies which provide a less clear pathway to fact-checking may lead certain social media users to make type-I errors.

[23] Soft provenance policies are closely related to accuracy nudging interventions, where users are prompted to think carefully about the accuracy of content before sharing. These have been empirically shown to have positive impacts on reducing the spread of misinformation, see Pennycook et al. (2021) and Pennycook et al. (2020b).



(a) No provenance.

(b) Intermediate provenance.

Figure 4. Optimal Platform Sharing Networks for Example 2 under Provenance Policies.

is connected to agent 2, who is connected to a clique of the other $N-2$ agents, as shown in Figure 4b.[24] For all agents $i \in \{3, \ldots, N\}$, conditional on the article reaching them, their belief $\tilde{\phi}(r)$ about the article's veracity is greater than under censorship (with $\delta = 3/16$), since two independent fact-checks with $\rho = 3/16$ each have not detected it as misinformation. As a result, all agents in the clique of Figure 4b will blindly share the article, correctly assuming that it has likely been already fact-checked. Expected user engagement with misinformation in this case is $(1-\rho)+(1-\rho)^2+(1-\rho)^3(N-2) > N/2$, for $\rho = 3/16$. Consequently, the provenance policy with $\rho = 3/16$ is worse than no policy at all, which is in turn worse than a censorship policy with $\delta = 3/16$. ∎

This example illuminates the potential weakness of provenance policies: when imperfectly implemented, they may be less robust than censorship. Because users presume others before them have fact-checked, they do not make independent judgments based on the reliability of the content. This observation is related to the literature on informational cascades and herding (for example, see Banerjee (1992), Bikhchandani et al. (2021), and Chen and Papanastasiou (2021)). When provenance policies are enacted, agents may excessively follow the sharing decisions of those before them instead of making independent inferences about content veracity before sharing. This "herding" opens the door for share cascades that may increase the virality of misinformation.

6.3 Performance Targets

Another possible regulation, which has recently been proposed by social media platforms (see Bickert (2020)), is to set *performance targets* that limit the amount of misinformation. However, since the monitoring and removal of misinformation is imperfect, a performance target allows the platform some leeway to have "bad" content on their site while still requiring a degree of accountability.

While in Sections 6.1 and 6.2 we considered policies where, respectively, the regulator and the users were responsible for removing content, a performance target transfers the burden of content removal to the platform itself. We assume that the platform can discard content (by not recommending it to any user), and in doing so, forfeits any potential engagement this content may have with the users.

We assume the regulator sets a performance target of λ, which requires the proportion of misinformation shares (to total shares) on the platform to fall below λ.[25] The regulator enforces this performance target by auditing the platform and sampling the content to verify that it meets the required standard. Formally, we assume that the regulator has an auditing technology $\alpha \in (0,1)$, which represents the probability of detecting that the platform has violated its performance target, and if detected, the platform incurs a cost C due to regulatory fines. Moreover, we assume $\phi(r) <$

[24] Notice that this sharing network is not in the class of island networks we have focused on so far. In terms of Proposition 5, when $\rho \in (0, \rho_1) \cup (\rho_3, 1)$, the platform's choices always lie within the class of island networks, but not necessarily when we are outside of this range.

[25] This metric for performance comes from Facebook's own statements on platform standards: "Regulators could say that internet platforms must publish annual data on the 'prevalence' of content that violates their policies, and that companies must make reasonable efforts to ensure that the prevalence of violating content remains below some standard threshold" (from Bickert (2020)), with the definition of prevalence being: "We care most about how often content that violates our standards is actually seen relative to the total amount of times any content is seen on Facebook" (from `https://about.fb.com/news/2019/05/measuring-prevalence/`).

$\max_{i^*} \mathbb{E}[\mathbf{S}_{i^*}]$, where the virality is with respect to no policy, otherwise the platform may be happy to comply with a performance target that removes all misinformation.

Our next result establishes how stricter performance targets affect the spread of misinformation:

Proposition 6. *There exists a performance target $\lambda^* \in (0, 1)$ such that:*

(a) If $\lambda > \lambda^*$, a stricter performance target (lower λ) is more effective;

(b) If $\lambda < \lambda^*$, a stricter performance target (lower λ) is less effective than λ^*.

This result establishes that when performance targets are lax, making them stricter (reducing λ) always curbs the spread of misinformation. In this region, when held more accountable, the platform removes some of the misinformation in circulation, foregoing the engagement that these contents would have generated. As a result, lower targets therefore align regulator and platform incentives to remove less reliable content.

However, with stricter targets, the incentives of the regulator and platform diverge. In particular, for targets stricter than λ^*, the platform needs to remove more and more content, with an increasingly larger sacrifice in engagement. In this case, the platform may prefer to violate the performance target and this implies that the tightening of the performance target actually backfires.

This analysis also implies that stricter performance targets need to be combined with better auditing or higher penalties for violation. This simple observation goes against the view that harsher punishments should be imposed when the platform fails to meet low targets (because there would be little excuse for violating them), and weaker punishments may be called for with stricter targets (because the platform may fail to meet them even when it tries). Instead, our analysis clarifies that stricter penalties may be necessary for stricter performance targets in order to prevent its incentives diverging from those of the regulator.

6.4 Network Regulations

As we saw in Theorem 3, when unregulated, the platform chooses the island model of Section 4 with parameters (p_s, p_d). Here, we consider limits on the ideological homophily induced by the platform's algorithm. Suppose the regulator can choose a homophily standard p^*, based on the ratio between within-island links to across-island links. In other words, this standard would force the platform to choose $p_s/p_d \leq p^*$.

Proposition 7. *There exists $\gamma < \infty$ such that for any $p^* \geq \gamma$, if the regulator imposes a homophily standard p^*, then (i) the platform chooses the island model with $p_s/p_d \leq p^*$; and (ii) the virality of misinformation is reduced.*

The regulator can thus reduce misinformation by imposing a homophily standard on the sharing network of the platform. This standard prevents the type of extreme homophily we saw in Theorem 3(a) and forces the platform to choose an algorithm that shares content across ideological groups. This policy is related to the "ideological segregation standard" proposed in Sunstein (2018), which aims to restrict the extent to which content is curated specifically to the ideology or interests of a specific group of users. Such standards ensure that echo chambers are broken and users of differing ideology interact more frequently, limiting the spread of misinformation.

We finally note that the regulation in Proposition 7 is not always binding for the platform. As Theorem 3(b) demonstrated, with highly reliable content, the platform maximizes engagement by implementing a maximally-connected sharing network, so the homophily constraint is moot. However, when the article is less reliable, the regulation will bind and the platform will be forced to maintain a minimum level of connectivity between different subgroups.

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 informative about an underlying state, or may contain misinformation, making them (weakly) anti-correlated with the underlying state. Upon receiving an article, an agent can decide to share it with others, ignore it, or actively call out another agent for propagating misinformation ("dislike"). Misinformation spreads when agents share articles expecting positive social media feedback and little negative reactions.

Though simple and parsimonious, the model encapsulates several rich strategic interactions. Agents receive utility from sharing truthful articles and not misinformation, but also enjoy peer engagement with shared content. The ideological congruence between an agent and those in her sharing network, which we capture with the notion of homophily, is critical for sharing decisions. Because individuals are more likely to dissent against articles that disagree with their prior beliefs, an agent will be more cautious in sharing articles that disagree with the views of those in her sharing network.

We provide several comparative static results. Some of those are 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 while misinformation typically spreads less than truthful content (holding all else constant), more sensational content tends to be shared more. Moreover, when misinformation is correlated with sensationalism, the rapid spread of misinformation can be problematic.

Of particular interest are comparative statics with respect to homophily. We show that when there is a highly-reliable article, an increase in homophily reduces the virality of content. Because this article is unlikely to contain misinformation, it is of broad appeal to a wide range of social media users, independent of ideology. An increase in homophily then reduces the extent to which this article can spread throughout the sharing network. The implications of homophily for low-reliability articles are very different, however: in a well-connected network, such articles will be disliked and stopped by users who disagree with their message, and anticipating this behavior and the loss of reputation they can suffer from spreading misinformation, even those who agree with their message would not share them widely. In contrast, high homophily creates echo chambers, where users share low-reliability messages aligned with their beliefs, because they understand that there are few negative reputational consequences from doing so. Misinformation contained in low-reliability articles can then spread virally in these echo chambers.

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). Our main result is a striking one. When an article is highly reliable, the platform chooses a sharing network with minimal homophily to maximize the spread and appeal of the content throughout the user community. In contrast to this case, when the relevant articles have lower reliability, the platform chooses a network with maximal homophily and recommends articles to users with aligned beliefs. These articles then spread 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. Content moderation, for example censoring low-reliability articles, can remove some misinformation. However, it also creates a Bayesian version of "false sense of security" and make agents more confident in the quality of remaining items. Similarly, revealing the provenance of a news item (for example, providing full context for a quote or clearer sources) can be useful, because this additional information allows users to more easily fact-check the content for veracity. However, this intervention can backfire, too, because it generates a type of information cascade: each agent expects others to have fact-checked and becomes more lax in his or her inspection. Performance standards that require platforms to remove a certain fraction of posts with misinformation can also backfire, this time because demanding targets can induce the platform to deviate from the standards, with the hope of not being detected. Finally, we show that regulation of platform algorithms, for example, in the form of ideological segregation standards, can be effective, though need to be well calibrated.

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 benchmark. In our setting, it brought out certain new strategic forces—highlighting how social media actions exhibit strategic complementarity and how the degree of homophily alters agents' strategic behavior. Although various behavioral biases and psychological factors appear to be important in social media behavior, we believe that the economic forces we have identified in this paper will continue to apply in the presence of most of these effects, and our Bayesian benchmark enabled us to isolate these forces in a transparent manner. Nevertheless, it remains true that misinformation can be more damaging when agents are boundedly rational, and incorporating such considerations is an important direction 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 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 other agents in their sharing network. Fully endogenizing reputational concerns, taking into account the network position of agents, would be another interesting direction for future research. 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 place an upper bound on the speed and possibility of 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 these information cascades on social media, is another important area for future research.

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.

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.

Allon, Gad, Kimon Drakopoulos, and Vahideh Manshadi (2021), "Information Inundation on Platforms and Implications." *Operations Research*, 69, 1784–1804.

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

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*.

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

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic (2015), "Exposure to ideologically diverse news and opinion on Facebook." *Science*, 348, 1130–1132.

Banerjee, Abhijit V. (1992), "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics*, 107, 797–817.

Bickert, Monika (2020), "Charting a Way Forward on Online Content Regulation."

Bikhchandani, Sushil, David Hirshleifer, Omer Tamuz, and Ivo Welch (2021), "Information Cascades and Social Learning." Working Paper 28887, National Bureau of Economic Research.

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]*.

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.

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.

Candogan, Ozan and Kimon Drakopoulos (2020), "Optimal Signaling of Content Accuracy: Engagement vs. Misinformation." *Operations Research*, 68, 497–515.

Centola, Damon (2010), "The spread of behavior in an online social network experiment." *science*, 329, 1194–1197.

Centola, Damon and Michael Macy (2007), "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology*, 113, 702–734.

Chen, Li and Yiangos Papanastasiou (2021), "Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation." *Management Science*, 67, 6734–6750.

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.

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.

Egelhofer, Jana Laura and Sophie Lecheler (2019), "Fake news as a two-dimensional phenomenon: a framework and research agenda." *Annals of the International Communication Association*, 43, 97–116.

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.

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.

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.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker (2019), "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." *Science Advances*, 5.

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.

Hsu, Chin-Chia, Amir Ajorlou, and Ali Jadbabaie (2020), "News Sharing, Persuasion, and Spread of Misinformation on Social Networks." SSRN Scholarly Paper ID 3391585.

Kamenica, Emir (2019), "Bayesian Persuasion and Information Design." *Annual Review of Economics*, 11, 249–272.

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.” *Working Paper, SSRN Scholarly Paper ID 3465030*.

Kim, Dam Hee, S. Mo Jones-Jang, and Kate Kenski (2020), “Why Do People Share Political Information on Social Media?” *Digital Journalism*, 0, 1–18.

Kozyreva, Anastasia, Stephan Lewandowsky, and Ralph Hertwig (2020), “Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools.” *Psychological Science in the Public Interest*, 21, 103–156.

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.

Lee, Chei Sian, Long Ma, and Dion Hoe-Lian Goh (2011), “Why Do People Share News in Social Media?” In *Active Media Technology*, Lecture Notes in Computer Science, 129–140, Springer.

Levy, Ro’ee (2021), “Social Media, News Consumption, and Polarization: Evidence from a Field Experiment.” *American Economic Review*, 111, 831–870.

Molina, Maria D., S. Shyam Sundar, Thai Le, and Dongwon Lee (2021), ““Fake News” Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content.” *American Behavioral Scientist*, 65, 180–212.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David Rand (2021a), “Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 182, 1–13.

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

Mostagir, Mohamed, Asu Ozdaglar, and James Siderius (2021), “When is society susceptible to manipulation?” *Management Science, forthcoming*.

Mostagir, Mohamed and James Siderius (2021), “Social inequality and the spread of misinformation.” *Management Science, forthcoming*.

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.

Papakyriakopoulos, Orestis, Simon Hegelich, Morteza Shahrezaye, and Juan Carlos Medina Serrano (2018), “Social media and microtargeting: Political data processing and the consequences for Germany.” *Big Data & Society*, 5.

Papanastasiou, Yiangos (2020), "Fake News Propagation and Detection: A Sequential Model." *Management Science*, 66, 1826–1846.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David G. Rand (2020a), "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.

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand (2018), "Prior exposure increases perceived accuracy of fake news." *Journal of Experimental Psychology. General*, 147, 1865–1880.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand (2021), "Shifting attention to accuracy can reduce misinformation online." *Nature*, 592, 590–595.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand (2020b), "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention." *Psychological Science*, 31, 770–780.

Pennycook, Gordon and David G. Rand (2019), "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition*, 188, 39–50.

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.

Quattrociocchi, Walter, Antonio Scala, and Cass R. Sunstein (2016), "Echo Chambers on Facebook." SSRN Scholarly Paper ID 2795110.

Sunstein, Cass R. (2018), *#Republic: Divided Democracy in the Age of Social Media*.

Tarski, Alfred (1955), "A lattice-theoretical fixpoint theorem and its applications." *Pacific Journal of Mathematics*, 5, 285–309.

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*, Computational Social Sciences, 289–322.

Topkis, Donald M. (1998), *Supermodularity and Complementarity*, first edition edition.

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

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.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018), "The spread of true and false news online." *Science*, 359, 1146–1151.

A Proofs

A.1 Auxiliary Lemmas

We define a (mixed-strategy) strategy σ_i for agent i to be a map from priors b_i to elements of the simplex $\Delta(\{\mathcal{D}, \mathcal{I}, \mathcal{S}\})$. In others words, σ_i specifies for each ideological prior b_i of agent i the probability that she will play each of the three actions, $\mathcal{D}$, $\mathcal{I}$, and $\mathcal{S}$. We let $\boldsymbol{\sigma}_{-i}$ denote the (vector of) strategies of all agents other than agent i.

Lemma A.1. *Given any set of strategies $\boldsymbol{\sigma}_{-i}$, agent i's best response is a cutoff strategy with cutoffs (b_i^*, b_i^{**}) such that if $b_i < b_i^*$ agent i dislikes ($\mathcal{D}$), if $b_i^* < b_i < b_i^{**}$ agent i ignores ($\mathcal{I}$), and if $b_i > b_i^*$ agent i shares ($\mathcal{S}$).*

Proof of Lemma A.1. When agent i receives an article, she forms (ex-post) belief π_i about the article's veracity which depends only on the observables (r, m). By Bayes' rule:

$$\pi_i \equiv \mathbb{P}[\nu = \mathcal{T} \,|,\, r, m = R] = \frac{\mathbb{P}[m = R \,|\, r, \nu = \mathcal{T}]\mathbb{P}[\nu = \mathcal{T} \,|\, r]}{\mathbb{P}[m = R \,|\, r, \nu = \mathcal{M}]\mathbb{P}[\nu = \mathcal{M} \,|\, r] + \mathbb{P}[m = R \,|\, r, \nu = \mathcal{T}]\mathbb{P}[\nu = \mathcal{T} \,|\, r]} \,.$$

By the law of total probability, we have:

$$\begin{aligned}
\mathbb{P}[m = R \,|\, r, \nu = \mathcal{T}] = \mathbb{P}[m = R \,|\, \nu = \mathcal{T}] &= \mathbb{P}[m = R \,|\, \theta = R, \nu = \mathcal{T}]\mathbb{P}[\theta = R] + \mathbb{P}[m = R \,|\, \theta = L, \nu = \mathcal{T}]\mathbb{P}[\theta = L] \\
&= pb_i + (1-p)(1-b_i) \,; \\
\mathbb{P}[m = R \,|\, r, \nu = \mathcal{M}] = \mathbb{P}[m = R \,|\, \nu = \mathcal{M}] &= \mathbb{P}[m = R \,|\, \theta = R, \nu = \mathcal{M}]\mathbb{P}[\theta = R] + \mathbb{P}[m = R \,|\, \theta = L, \nu = \mathcal{M}]\mathbb{P}[\theta = L] \\
&= qb_i + (1-q)(1-b_i) \,.
\end{aligned}$$

Putting these together we obtain equation (2). Moreover, π_i is monotone in b_i since

$$\frac{\partial \pi_i}{\partial b_i} = \frac{(1-\phi(r))\phi(r)(p-q)}{(1 - b_i + q(1-\phi(r))(2b_i - 1) - p(\phi(r) - 2\phi(r)b_i))^2} > 0.$$

Note that $U_i(\mathcal{I})$ and $U_i(\mathcal{D})$ is independent of $\boldsymbol{\sigma}_{-i}$, and in particular $\mathcal{D}$ is a better response to $\mathcal{I}$ if and only if $\pi_i < (\tilde{u} - \tilde{c})/\tilde{u}$. Because π_i is monotone in b_i, this implies there exists some $\tilde{b}$ where $\mathcal{D}$ is a better response to $\mathcal{I}$ if and only if $b_i < \tilde{b}$ (where $\tilde{b} = 1$ if disliking dominates ignoring and $\tilde{b} = 0$ if ignoring dominates disliking). Next, recall that the payoff to sharing is $U_i(\mathcal{S}) = U_i^{(1)} + U_i^{(2)}$, where $U_i^{(1)} = u\mathbf{1}_{\nu=\mathcal{T}} - c\mathbf{1}_{\nu=\mathcal{M}}$ and $U_i^{(2)} = \kappa S_i - dD_i$. Observe that, as before, $U_i^{(1)}$ is independent of $\boldsymbol{\sigma}_{-i}$ and has expected payoff $(u + c)\pi_i - c$, which is monotonically increasing in π_i. Moreover, $\mathbb{E}_{\mathbf{P}, \boldsymbol{\sigma}_{-i}}[\kappa S_i - dD_i]$ does not depend on b_i. Because π_i is monotone in b_i, we see that $U_i(\mathcal{S})$ is increasing in b_i, $U_i(\mathcal{I})$ is constant in b_i (it is always zero), and $U_i(\mathcal{D})$ is decreasing in b_i (it is equal to $\tilde{u}(1 - \pi_i) - \tilde{c}$). This implies that either (i) ignoring dominates sharing, (ii) sharing dominates ignoring, or (iii) $U_i(\mathcal{S}) = 0$ for some prior b':

(i) If ignoring dominates sharing, we set $(b_i^*, b_i^{**}) = (\tilde{b}, 1)$.

(ii) If sharing dominates ignoring, then either sharing dominates disliking (in which case set $(b_i^*, b_i^*) = (0, 0)$), disliking dominates sharing (in which case we set $(b_i^*, b_i^{**}) = (1, 1)$), or there exists some prior b'' where $U_i(\mathcal{S}) = U_i(\mathcal{D})$ (in which case set $(b_i^*, b_i^{**}) = (b'', b'')$).

(iii) Otherwise, if $\tilde{b} < b'$, set $(b_i^*, b_i^{**}) = (\tilde{b}, b')$; however, if $\tilde{b} \geq b'$, then we set $(b_i^*, b_i^{**}) = (b', b')$.

This is of the cutoff form claimed in the lemma. ∎

An immediate consequence of Lemma A.1 is that any Bayesian-Nash equilibrium must be in cutoff strategies for all agents. Hence, we can limit our attention to cutoff strategies (b_i^*, b_i^{**}) for every agent i, which can be represented as $(\mathbf{b}^*, \mathbf{b}^{**})$ in vector notation. This is a partially-ordered set according to the component-wise order $\succeq$. Hence, the cutoff space $\mathbf{B} = [0,1]^{2N}$ forms a *complete lattice*.[26]

Next, we define a map $\psi : \mathbf{B} \to \mathbf{B}$ that maps cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$ to best-response cutoffs $(\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$. This map is well-defined because (i) H is a continuous distribution, so we need not specify the strategies of agents precisely on the cutoffs, and (ii) by Lemma A.1, for any set of strategies $\boldsymbol{\sigma}_{-i}$ (including the cutoff strategies given by $(\mathbf{b}^*, \mathbf{b}^{**})$), all agents' best responses are in cutoff form.

Lemma A.2. *The map ψ preserves the component-wise order $\succeq$.*

Proof of Lemma A.2. Consider some $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**}) \succeq (\mathbf{b}^*, \mathbf{b}^{**})$. Fixing an article with observables (r, m), $U_i(\mathcal{D})$, $U_i(\mathcal{I})$ and $U_i^{(1)}$ are independent of $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ and $(\mathbf{b}^*, \mathbf{b}^{**})$. However, for $U_i^{(2)}$ we have:

$$\begin{aligned}
\mathbb{E}_{\mathbf{P},(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}[\kappa S_i - dD_i] &= \sum_{j=1}^{N} p_{ij}\left(\kappa \mathbb{P}_{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}[a_j = \mathcal{S}] - d\mathbb{P}_{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}[a_j = \mathcal{D}]\right) = \sum_{j=1}^{N} p_{ij}\left(\kappa \mathbb{P}_H[b_j > \hat{b}_j^{**}] - d\mathbb{P}_H[b_j < \hat{b}_j^{*}]\right) \\
&\leq \sum_{j=1}^{N} p_{ij}\left(\kappa \mathbb{P}_H[b_j > b_j^{**}] - d\mathbb{P}_H[b_j < b_j^{*}]\right) = \mathbb{E}_{\mathbf{P},(\tilde{\mathbf{b}}^*,\tilde{\mathbf{b}}^{**})}[\kappa S_i - dD_i]\,.
\end{aligned}$$

As a result, $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S})$. As in Lemma A.1, we define $\tilde{b}$ as the prior where $U_i(\mathcal{D}) = 0$ if such a $\tilde{b}$ exists, otherwise let $\tilde{b} = 0$ if ignoring dominates disliking and $\tilde{b} = 1$ if disliking dominates ignoring. Observe that $\tilde{b}$ is the same for both $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ and $(\mathbf{b}^*, \mathbf{b}^{**})$. We have three cases for the best-response cutoffs $(\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$ given other agents' cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$ (which we compare to $(\hat{\mathbf{b}}^{*,BR}, \hat{\mathbf{b}}^{**,BR})$ given other agents' cutoffs $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$):

(i) Ignoring dominates sharing for agent i (for given cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$). Then by virtue of $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S})$, ignoring dominates sharing with $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ as well. Thus, $(b_i^{*,BR}, b_i^{**,BR}) = (\hat{b}_i^{*,BR}, \hat{b}_i^{**,BR}) = (\tilde{b}, 1)$.

(ii) Sharing dominates ignoring for agent i (for given cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$). Then either sharing dominates disliking (in which case $(b_i^{*,BR}, b_i^{**,BR}) = (0,0) \preceq (\hat{b}_i^{*,BR}, \hat{b}_i^{**,BR})$ trivially), or there exists some prior b'' where $U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = U_i(\mathcal{D})$ denoted by b'' and $(b_i^{*,BR}, b_i^{**,BR}) = (b'', b'')$. Moreover, because $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = U_i(\mathcal{D})$ at prior b'', for an agent with prior b'', playing $\mathcal{D}$ is a (weakly) better response than sharing when other agents play according to cutoffs $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$. By monotonicity of $U_i(\mathcal{S})$ and $U_i(\mathcal{D})$ in prior b_i, this implies that $b_i^{**,BR} \leq \hat{b}_i^{**,BR}$. If ignoring is never a best response when $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$, then $\hat{b}_i^{*,BR} = \hat{b}_i^{**,BR}$. Otherwise, $\hat{b}_i^{*,BR} = \tilde{b} \geq b_i^{**,BR} = b_i^{*,BR}$.

[26] Note that for any collection of cutoffs $\{(\mathbf{b}^{*,(1)}, \mathbf{b}^{**,(1)}), (\mathbf{b}^{*,(2)}, \mathbf{b}^{**,(2)}), \ldots, \}$ in the cutoff space, there is a greatest lower bound given by the component-wise infimum and a least upper bound given by the component-wise supremum.

(iii) $U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = 0$ for some prior b' for agent i. Then $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = 0$ implies that for an agent with prior b' playing $\mathcal{I}$ is a (weakly) better response than sharing when other agents play according to $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$. By monotonicity of $U_i(\mathcal{S})$ in prior b_i, this implies that $b_i^{**,BR} \leq \hat{b}_i^{**,BR}$. If $\tilde{b} < b'$, then $b_i^{*,BR} = \hat{b}_i^{*,BR} = \tilde{b}$; otherwise, if $\tilde{b} \geq b'$, $b_i^{*,BR} = b_i^{**,BR} = b' \leq \hat{b}_i^{*,BR}$.

This establishes that $(\hat{\mathbf{b}}_i^{*,BR}, \hat{\mathbf{b}}_i^{**,BR}) \succeq (\mathbf{b}_i^{*,BR}, \mathbf{b}_i^{**,BR})$, so the order $\succeq$ is preserved by ψ. ∎

Lemma A.3. *An increase in polarization of beliefs can 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 A.3. Let H_2 be more polarized than H_1. 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). ∎

Lemma A.4. *If $\kappa \leq \bar{\kappa} \equiv (c\tilde{c} - u(\tilde{u} - \tilde{c}))/(\tilde{u}N)$, then for any agent i:*

(i) If $b_i^* > 0$ and $b_i^{**} < 1$, then $b_i^{**} > b_i^*$;

(ii) For all $\bar{b} < 1$, there exists $\tilde{r} > 0$ such that agent i plays $\mathcal{D}$ in any equilibrium for an article with $r < \tilde{r}$ and on any sharing network $\mathbf{P}$, provided that $b_i < \bar{b}$.

Proof of Lemma A.4. For part (i), by way of contradiction suppose that $b_i^* = b_i^{**}$. Then for an agent with prior b_i^* (and corresponding ex-post belief π_i^* that the article is truthful), it must be the case that:

$$\tilde{u}(1 - \pi_i) - \tilde{c} = u\pi_i - c(1 - \pi_i) + \mathbb{E}[\kappa S_i - dD_i] \geq 0 \, .$$

Re-arranging we get that $\pi_i = \frac{\tilde{u} - \tilde{c} + c - \mathbb{E}[\kappa S_i - dD_i]}{\tilde{u} + u + c}$. Substituting into the payoff for action $\mathcal{D}$, we see that:

$$U_i(\mathcal{D}) = \tilde{u}\left(\frac{u + \tilde{c} + \mathbb{E}[\kappa S_i - dD_i]}{\tilde{u} + u + c}\right) - \tilde{c} \leq \tilde{u}\left(\frac{u + \tilde{c} + \kappa N}{\tilde{u} + u + c}\right) - \tilde{c} < \tilde{u}\left(\frac{u + \tilde{c} + \bar{\kappa} N}{\tilde{u} + u + c}\right) - \tilde{c} \leq 0 \, .$$

By assumption, $U_i(\mathcal{S}) = U_i(\mathcal{D}) < 0$, but since $U_i(\mathcal{I}) = 0$, ignoring is the best response at prior b_i^*, which is a contradiction.

For part (ii), notice by equation (2), for a fixed $b < 1$, as $r \to 0$, $\pi_i \to 0$, and therefore:

$$U_i(\mathcal{S}) = u\pi_i - c(1 - \pi_i) + \mathbb{E}[\kappa S_i - dD_i] < u\pi_i - c(1 - \pi_i) + \bar{\kappa} N \leq u\pi_i - c(1 - \pi_i) + \frac{c}{N} N \overset{r \to 0}{=} -c + c = 0 \, .$$

where the last inequality follows from the observation that:

$$\bar{\kappa} \equiv \frac{c\tilde{c} - u(\tilde{u} - \tilde{c})}{\tilde{u}N} < \frac{c\tilde{c}}{\tilde{u}N} < \frac{c}{N} \, ,$$

because $\tilde{u} > \tilde{c}$. Thus, as $r \to 0$, ignoring is a better response than sharing. But note that $U_i(\mathcal{D}) = \tilde{u}(1 - \pi_i) - \tilde{c} \overset{r \to 0}{=} \tilde{u} - \tilde{c} > 0$, so as $r \to 0$, disliking is a better response than ignoring. As a result, disliking is a best response for any fixed $b < 1$ as $r \to 0$. The claim in (ii) thus follows from continuity of equation (2). ∎

A.2 Proofs from Section 3

Proof of Theorem 1. Claim (ii) follows directly from Lemma A.1 and establishes that the Bayesian-Nash equilibria are the fixed points of the map ψ. Clearly the cutoff space $\mathbf{B}$ is convex and compact (it is defined by $[0,1]^{2N}$). To see that ψ is continuous, notice that for $\psi : (\mathbf{b}^*, \mathbf{b}^{**}) \mapsto (\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$, $\mathbb{E}_{\mathbf{P},(\mathbf{b}^*,\mathbf{b}^{**})}[U_i^{(2)}]$ is continuous because H is continuous (and $U_i(\mathcal{D})$, $U_i(\mathcal{I})$, and $U_i^{(1)}$ do not depend on $(\mathbf{b}^*, \mathbf{b}^{**})$). Moreover, by the same reasoning as in Lemma A.2, $U_i^{\mathbf{P},(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S})$ and $U_i^{\mathbf{P},(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) - U_i(\mathcal{D})$ are monotone and continuous. Because these expressions are continuous in $(\mathbf{b}^*, \mathbf{b}^{**})$, the corresponding best-response cutoffs, $(\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$ are also continuous in $(\mathbf{b}^*, \mathbf{b}^{**})$. By Brouwer's fixed-point theorem, there exists a Bayesian-Nash equilibrium, proving (i).

Finally, noting that the cutoff space $\mathbf{B}$ is a complete lattice and ψ preserves the component-wise order $\succeq$ (by Lemma A.2), Tarski's fixed-point theorem establishes that the set of equilibrium cutoffs forms a lattice (see Tarski (1955)). By definition of a lattice order, there exists a least-sharing equilibrium (largest $\mathbf{b}^{**}$) and a most-sharing equilibrium (smallest $\mathbf{b}^{**}$). ∎

Proof of Proposition 1. Recall that π_i is given by equation (2) and provides the (ex-post) belief of the article's veracity conditional on observables (r, m). Also observe that:

$$\frac{\partial \pi_i}{\partial r} = \frac{(1 - b_i + p(2b_i - 1))(1 - b_i + q(2b_i - 1))}{(1 - b_i + q(1 - \phi(r))(2b_i - 1) - p(\phi(r) - 2\phi(r)b_i))^2} \phi'(r) \,.$$

Because $\phi'(r) > 0$, it is clear that when $b_i > 1/2$, $\partial\pi_i/\partial r > 0$. When $b_i < 1/2$, $1 - b_i + p(2b_i - 1)$ is minimized when $p = 1$, in which case it is equal to $b_i \geq 0$ (and with this inequality strict whenever $p < 1$). Similarly, when $b_i < 1/2$, $1 - b_i + q(2b_i - 1)$ is minimized when $q = 1/2$, in which case it is equal to $1/2 > 0$. Thus, $\partial\pi_i/\partial r > 0$ for all b_i.

Similarly, when $\phi' \geq \phi$, a reliability score r with misinformation structure ϕ' can be translated into a higher reliability score $r' \geq r$ under misinformation structure ϕ (because both ϕ, ϕ' are monotonically increasing). As a consequence, a decrease in misinformation is isomorphic to greater reliability of the articles. It is thus sufficient to prove the latter leads to uniformly more sharing in both the least and the most sharing equilibria.

Note that the social media game is supermodular and has increasing differences in reputability. To see this, note that for all $r' \geq r$:

$$\left[U_i(\mathcal{S}, r') - U_i(\mathcal{I}, r')\right] - \left[U_i(\mathcal{S}, r) - U_i(\mathcal{I}, r)\right] = U_i(\mathcal{S}, r') - U_i(\mathcal{S}, r) = U_i^{(1)}(r') - U_i^{(1)}(r) = (u + c)(\pi_i(r') - \pi_i(r)),$$

which is non-negative via the above observation that $\frac{\partial \pi_i}{\partial r} > 0$. Similarly, for all $r' \geq r$:

$$\big[U_i(\mathcal{S}, r') - U_i(\mathcal{D}, r')\big] - [U_i(\mathcal{S}, r) - U_i(\mathcal{D}, r)] = \big[U_i(\mathcal{S}, r') - U_i(\mathcal{S}, r)\big] + \big[U_i(\mathcal{D}, r') - U_i(\mathcal{D}, r)\big]$$
$$= (u + c)(\pi_i(r') - \pi_i(r)) + \tilde{u}(\pi_i(r') - \pi_i(r)) \, ,$$

which is non-negative via the same observation. Finally, for all $r' \geq r$:

$$\big[U_i(\mathcal{I}, r') - U_i(\mathcal{D}, r')\big] - [U_i(\mathcal{I}, r) - U_i(\mathcal{D}, r)] = U_i(\mathcal{D}, r') - U_i(\mathcal{D}, r) = \tilde{u}(\pi_i(r') - \pi_i(r)) \, ,$$

which, again, is non-negative. Thus, via Topkis's monotone comparative statics theorem (see Topkis (1998)), there is uniformly more sharing.

Similarly, the social media game is supermodular and has increasing differences in sensationalism and (the negative of) reputational concerns. To see this, note for all $\kappa' \geq \kappa$ and $d' \leq d$:

$$\big[U_i(\mathcal{S}, \kappa', d') - U_i(\mathcal{I}, \kappa', d')\big] - [U_i(\mathcal{S}, \kappa, d) - U_i(\mathcal{I}, \kappa, d)] = U_i^{(2)}(\kappa', d') - U_i^{(2)}(\kappa', d') = (\kappa' - \kappa)S_i + (d - d')D_i$$

which is non-negative. Moreover, note that comparing $\mathcal{S}$ and $\mathcal{D}$ is identical to comparing $\mathcal{S}$ and $\mathcal{I}$ because parameters (κ, d) affect both $\mathcal{I}$ and $\mathcal{D}$ identically (they only factor into the payoff of action $\mathcal{S}$). For this same reason, we note that $[U_i(\mathcal{I}, \kappa', d') - U_i(\mathcal{D}, \kappa', d')] - [U_i(\mathcal{I}, \kappa, d) - U_i(\mathcal{D}, \kappa, d)] = 0$. Thus, via Topkis's theorem, there is uniformly more sharing. ∎

A.3 Proofs from Section 4

Proof of Lemma 1. To obtain a contradiction, suppose that there exists an agent i and an agent j with $\ell_i = \ell_j$ but either (i) $b_i^* \neq b_j^*$ or (ii) $b_i^{**} \neq b_j^{**}$.

Without loss of generality, suppose that $b_i^* < b_j^*$. By way of contradiction suppose $b_i^{**} > b_i^*$, and consider priors $\tilde{b} \in (b_i^*, \min\{b_j^*, b_i^{**}\})$ where agent i would ignore but agent j with that same prior would dislike. However, both agents with prior $\tilde{b}$ receive payoff $\tilde{u}(1 - \pi(\tilde{b})) - \tilde{c}$ from disliking and payoff of 0 from ignoring. Thus, one of them must not be playing a best response. This establishes that $b_i^{**} = b_i^*$.

Thus, when agents i and j both have some prior $b' \in (b_i^*, b_j^*)$, agent i shares and agent j dislikes. By symmetry of agent i and j's network positions, it is clear that for agent i and agent j with prior b' that $U_j(\mathcal{S}) - U_i(\mathcal{S}) = p_s(\kappa + d)$. Similarly, $U_j(\mathcal{D}) - U_i(\mathcal{D}) = 0$. But in this case,

$$[U_j(\mathcal{S}) - U_i(\mathcal{S})] - [U_j(\mathcal{D}) - U_i(\mathcal{D})] = [U_j(\mathcal{S}) - U_j(\mathcal{D})] + [U_i(\mathcal{D}) - U_i(\mathcal{S})] = p_s(\kappa + d) > 0 \, .$$

This implies that either $[U_j(\mathcal{S}) - U_j(\mathcal{D})] > 0$ or $[U_i(\mathcal{D}) - U_i(\mathcal{S})] > 0$ (or both). This yields a contradiction because at prior b', it is supposed to be a best response for agent j to play $\mathcal{D}$ and a best response for agent i to play $\mathcal{S}$. Thus, $b_i^* = b_j^*$.

Without loss of generality, suppose that $b_i^{**} < b_j^{**}$. If $b_i^{**} \leq b_j^*$, then for priors $b'' \in (b_i^{**}, b_j^*)$, agent i shares and agent j dislikes. Via the same reasoning as in the previous paragraph, this is a contradiction, so $b_j^* < b_i^{**} < b_j^{**}$. Let us consider some prior $\hat{b} \in (b_i^{**}, b_j^{**})$, where agent i shares and agent j ignores. By symmetry of agent i and j's network positions, it is clear that for agent i and agent j with prior $\hat{b}$ that $U_j(\mathcal{S}) - U_i(\mathcal{S}) = p_s\kappa$. Similarly, $U_j(\mathcal{I}) - U_i(\mathcal{I}) = 0$. Then notice that:

$$[U_j(\mathcal{S}) - U_i(\mathcal{S})] - [U_j(\mathcal{I}) - U_i(\mathcal{I})] = [U_j(\mathcal{S}) - U_j(\mathcal{I})] + [U_i(\mathcal{I}) - U_i(\mathcal{S})] = p_s\kappa > 0\,.$$

This implies that either $[U_j(\mathcal{S}) - U_j(\mathcal{I})] > 0$ or $[U_i(\mathcal{I}) - U_i(\mathcal{S})] > 0$ (or both). However, this is a contradiction because at prior b', it is supposed to be a best response for agent j to play $\mathcal{I}$ and a best response for agent i to play $\mathcal{S}$. Thus, $b_i^{**} = b_j^{**}$. ▮

Proof of Theorem 2. For part (a), let us consider belief $b^{(2)} < 1$ and a reliability threshold $\underline{r}$ such that for all $\mathbf{P}$, all agents with $b < b^{(2)}$ choose $\mathcal{D}$ in every equilibrium (including the most-sharing equilibrium) whenever the article has reliability $r < \underline{r}$. Such an $\underline{r}$ exists by Lemma A.4(ii). Thus, for all $r < \underline{r}$, every agent on an island $\ell \geq 2$ dislikes in the most-sharing equilibrium, regardless of $\mathbf{P}$.

Next, we consider an increase in homophily (while holding expected degree fixed). By our choice of $\underline{r}$, all agents on islands $\ell \geq 2$ still dislike in the most-sharing equilibrium whenever $r < \underline{r}$. We can thus consider the social media game that only involves island 1, treating islands 2 through k as automata that always dislike. Before the shift in homophily, consider the equilibrium cutoffs (b_1^*, b_1^{**}) for island 1 in the most-sharing equilibrium (the same for all agents on island 1, per Lermma 1) and let $\mathbf{B}_1$ denote the modified cutoff space defined by all cutoffs $(\hat{b}_1^*, \hat{b}_1^{**}) \preceq (b_1^*, b_1^{**})$. Finally we define a map $\varphi : \mathbf{B}_1 \to \mathbf{B}_1$ that maps cutoffs in $\mathbf{B}_1$, $(\hat{b}_1^*, \hat{b}_1^{**})$, to best-response cutoffs $(\hat{b}_1^{*,BR}, \hat{b}_1^{**,BR})$, given that agents on island 1 play according $(\hat{b}_1^*, \hat{b}_1^{**})$. By the arguments in Lemma A.2, φ preserves $\succeq$ and $\mathbf{B}_1$ is a complete sublattice, provided that the map φ is well-defined in that it always maps to an element in $\mathbf{B}_1$.

To establish this, consider the utility $U_1(\mathcal{S})$ of sharing on island 1 with homophily parameters (p_s, p_d), holding fixed the cutoff strategy $(\hat{b}_1^*, \hat{b}_1^{**})$ and the expected degree of each agent on island 1, ζ_1. Thus, we can write $p_d = (\zeta - N_1 p_s)/(N - N_1)$ and observe then that

$$U_1(\mathcal{S}) = U_1^{(1)} + \kappa N_1 p_s (1 - H(\hat{b}_1^{**})) - d\left(N_1 p_s H(\hat{b}_1^*) + \frac{\zeta - N_1 p_s}{N - N_1} \cdot (N - N_1)\right)\,,$$

and in particular, $\partial U_1(\mathcal{S})/\partial p_s = \kappa N_1(1 - H(\hat{b}_1^{**})) + dN_1(1 - H(\hat{b}^*)) > 0$. Therefore, if we compare utility $U_1'(\mathcal{S})$ after the increase in homophily to $U_1(\mathcal{S})$ before the increase in homophily (leaving $(\hat{b}_1, \hat{b}_1^{**})$ fixed), we see that $U_1'(\mathcal{S}) \geq U_1(\mathcal{S})$. Hence, φ necessarily maps any cutoffs in $\mathbf{B}_1$ into $\mathbf{B}_1$. Applying Again Tarski's fixed-point theorem, the set of fixed points (and thus Bayesian-Nash equilibria) form a lattice within the space of cutoffs $\mathbf{B}_1$. Moreover, there is a most-sharing equilibrium in $\mathbf{B}_1$, which is also the most-sharing equilibrium in $\mathbf{B}$. We denote this equilibrium by $(b_1^{*'}, b_1^{**'})$ and note that $(b_1^{*'}, b_1^{**'}) \preceq (b_1^*, b_1^{**})$ (because it lies in $\mathbf{B}_1$). In particular, this means $b_1^{**'} \leq b_1^{**}$, and more agents share on island 1 in the most-sharing equilibrium following the rise in homophily.

To measure the change in virality, we first observe that the seed agent i^* (that maximizes $\mathbb{E}[\mathbf{S}_{i^*}]$) is chosen from the agents on island 1. We consider the virality of the article when agents on island 1 share with probability $1 - H(b_1^{**})$ under the stronger homophily structure (p_s', p_d') versus (p_s, p_d) (and all other agents kill the article). This is sufficient to show that virality increases following the increase in homophily, because virality with $b_1^{**'} < b_1^{**}$ (but the same network $\mathbf{P}$) is strictly higher, given that agents on island 1 share more often, that is, $(1 - H(b^{**'}) > 1 - H(b^{**}))$.

We consider the diffusion process of an article on the (p_s', p_d') network that starts with an agent on island 1. Let us define a *path* of the diffusion process to be a chain $i^* \to i_1 \to i_2 \to \ldots \to i_z$ representing a sequence of agents who receive the article in this process, with i^* being the seed agent, i_1 through i_{z-1} all being agents who shared it, and agent i_z being an agent who either ignored or disliked the article. There may be many such paths for the diffusion of the article (by assumption, all agents with the possible exception of agent i_z must be on island 1).

For each path, we define an alternative path (generated randomly) as follows. For any links to agents other than to agent i_z (i.e., links within island 1), with probability $(p_s' - p_s)/p_s'$, the link instead goes to one of islands $2, \ldots, k$ (chosen in proportion to their population) and otherwise remains the same. Applying this to all paths, we define an isomorphic diffusion process to one on a sharing network with weaker homophily parameters (p_s, p_d). However, note that the length of every path cannot increase following this transformation. Because any transition to islands $2, \ldots, k$ is necessarily the end of the path, paths can only shorten. Moreover, the number of paths must weakly decrease. As a result, the fraction of agents who receive the article, $\mathbf{S}_{i^*}$, must be lower, and virality is less under the (p_s, p_d) sharing network. This establishes part (a).

For part (b), we first note that there exists $\bar{r}$ such that the most-sharing equilibrium when $r > \bar{r}$ is all-share ($b_\ell^{**} = 0$ for all islands ℓ) regardless of $\mathbf{P}$. Notice that equation (2) is minimized when $b_i = 0$, and in particular, for all agents i (regardless of their prior) $\pi_i \geq \frac{(1-p)\phi(r)}{(1-q)(1-\phi(r))+(1-p)\phi(r)}$. Then, letting $\bar{\pi} = \max\left\{\frac{c}{u+c}, \frac{\tilde{u}-\tilde{c}}{\tilde{u}}\right\} < 1$, we note that whenever $r \geq \phi^{-1}\left(\frac{(1-q)\bar{\pi}}{(p-q)\bar{\pi}+(1-p)}\right) \equiv \bar{r} \in (0,1)$, $\pi_i \geq \bar{\pi}$. Of course, when all other agents (other than i) share and $r > \bar{r}$, $U_i(\mathcal{S}) \geq u\pi_i - c(1-\pi_i) \geq 0$ and $U_i(\mathcal{D}) = \tilde{u}(1-\pi_i) - \tilde{c} \leq 0$, so $a_i = \mathcal{S}$ is a best response for agent i. Thus, the most-sharing equilibrium is all-share (because it *is* an equilibrium and no other strategy profile can have more sharing).

Observe that when $r > \bar{r}$, virality is measured simply by the expected size of the connected component (formed by $\mathbf{P}$) containing the seed agent i^*. Regardless of the homophily parameters, the seed agent i^* will be chosen from the largest island (call this island ℓ^*). This is immediate from the fact that all agents share in equilibrium, agents on island ℓ^* have the most connections to any other arbitrary island ℓ' (in expectation), and are connected to all agents on their own island.

Lastly, we note that the probability that island ℓ has any connections to island ℓ' is given by $\tilde{p}_{\ell,\ell'} = 1 - (1-p_d)^{N_\ell N_\ell'}$ *before* the decrease in homophily and $\tilde{p}_{\ell,\ell'}' = 1 - (1-p_d')^{N_\ell N_\ell'}$ *after* the decrease in homophily, with $\tilde{p}_{\ell,\ell'}' > \tilde{p}_{\ell,\ell'}$ for all pairs of islands (ℓ, ℓ') because $p_d' > p_d$. Using the same terminology as in the argument for part (a), we map the diffusion paths of an article under the less homophilic sharing network with (p_s', p_d'). Consider cycles between islands $\ell^* \to \ell_1 \to \ell_2 \ldots \to \ell_z$, where ℓ_z is the same island as one of $\ell^*, \ell_1, \ldots, \ell_z$ (in which case, no additional engagement is obtained thereafter the article returns to island ℓ_z). Before the decrease in homophily (where $p_d < p_d'$), we can construct an isomorphic diffusion process where an article remains within the same island (instead of switching to a different one) with probability $(p'_d - p_d)/p_d$. By construction of the cycle, whenever such an event occurs, the cycle becomes complete and the islands reached thereafter in the (p'_s, p'_d) sharing network are not (for that given cycle). Measuring across all cycles that occur in the (p'_s, p'_d) model, (weakly) more islands are reached than under the more homophilic (p_s, p_d) model. Consequently, virality is higher under the (p'_s, p'_d) sharing network than with the (p_s, p_d) sharing network, which has more homophily. This establishes part (b). ∎

Proof of Proposition 2. Let us define r^* as

$$r^* \equiv \phi^{-1}\left(\max\left\{\frac{(1-q)(\tilde{u}-\tilde{c})}{(p-q)(\tilde{u}-\tilde{c})+(1-p)\tilde{u}}, \frac{c}{u+c}\right\}\right) \in (0,1)\,.$$

For part (a), first consider the case of $r < r^*$ and $p_d = 0$ (by continuity, the result extends to the case of sufficiently large p_s/p_d). In the most-sharing equilibrium, the seed agent most conducive to the article's spread is on the right-wing island, and given that $p_d = 0$, the equilibrium on the left-wing island is immaterial to its virality. Let us denote the right-wing island cutoffs by (b^*_R, b^{**}_R). Similar to the proof of Theorem 2(a), we define a cutoff space $\mathbf{B}_R$ such that $(\hat{b}^*_R, \hat{b}^{**}_R) \in \mathbf{B}_R$ if and only if $(\hat{b}^*_R, \hat{b}^{**}_R) \preceq (b^*_R, b^{**}_R)$. Similarly, we define the map $\varphi : \mathbf{B}_R \to \mathbf{B}_R$ which maps an arbitrary cutoff $(\hat{b}^*_R, \hat{b}^{**}_R)$ to best-response cutoffs $(\hat{b}^{*,BR}_R, \hat{b}^{**,BR}_R)$. To show the map is well-defined, consider $U_R(\mathcal{S})$ *before* the increase in divisiveness or polarization and $U'_R(\mathcal{S})$ *after* the increase in divisiveness or polarization. Because the network structure is fixed, note that $U^{(2)}_R(\mathcal{S}) = U^{(2)'}_R(\mathcal{S})$ when the cutoffs $(\hat{b}^*_R, \hat{b}^{**}_R)$ are taken as given, so the difference $U'_R(\mathcal{S}) - U_R(\mathcal{S})$ depends only on the difference between $U^{(1)}_R(\mathcal{S})$ and $U^{(1)'}_R(\mathcal{S})$. Specifically, the difference in share payoff depends only on the change in π_i following the increase in divisiveness or polarization. Moreover,

$$\begin{aligned}
\frac{\partial \pi_i}{\partial p} &= \frac{(2b_i-1)(1-\phi(r))\phi(r)(1-q-b_i(1-2q))}{(b_i(2p\phi(r)+2q(1-\phi(r))-1)-p\phi(r)-q(1-\phi(r))+1)^2} > 0\,;\\
\frac{\partial \pi_i}{\partial q} &= \frac{(2b_i-1)(1-\phi(r))\phi(r)(1-p+b_i(2p-1))}{((2b_i-1)\phi(r)(p-q)+2b_iq-b_i-q+1)^2} > 0\,,
\end{aligned}$$

whenever $b_i > 1/2$. Likewise, as we showed in Lemma A.1, $\partial\pi_i/\partial b_i > 0$ for all b_i and greater polarization increases ideological priors for agents with $b_i > 1/2$ (by Lemma A.3). By virtue of $\underline{b}_R > 1/2$, we observe that $U^{(1)'}_R(\mathcal{S}) > U^{(1)}_R(\mathcal{S})$, and so $U'_R(\mathcal{S}) > U_R(\mathcal{S})$. Thus, as in the proof of Theorem 2(a), φ is well-defined. Applying the Tarski fixed-point theorem, we find that the most-sharing equilibrium leads to more sharing in the right-wing island. Because the network structure $\mathbf{P}$ remains constant and there is a uniform shift in sharing, our weaker notion of virality also increases.

For part (b), consider $r \geq r^*$. Note that for $r \geq r^*$, ignoring is a better response to disliking for any agent, regardless of prior and sharing is a better response to ignoring for all $b_i > 1/2$. The former follows from noting $\pi_i \geq \frac{\tilde{u}-\tilde{c}}{\tilde{u}}$ for an agent with prior $b_i = 0$ and the latter from noting $\pi_i \geq \frac{c}{u+c}$ for agents with $b_i > 1/2$ *and* observing that disliking is a dominated strategy. Therefore, the right-wing island always shares, whereas the left-wing island has equilibrium cutoffs $(0, b_L^{**})$. Using the same approach as in part (a), it is enough to show that $U_L(\mathcal{S})$ increases following a decrease in divisiveness or polarization. Furthermore, for $b_i < 1/2$, we see that $\partial\pi_i/\partial p < 0$ and $\partial\pi_i/\partial q < 0$, and by Lemma A.3, decreasing polarization means that all agents on the left-wing island also have an increase in b_i. Thus, there is more sharing in the most-sharing equilibrium following a decrease in divisiveness or polarization. By strategic complementarity, the right-wing island remains at all-share, and sharing uniformly increases (and so does virality, naturally). ∎

A.4 Proofs from Section 5

Proof of Theorem 3. Consider the complete sharing network where $\mathbf{P} = \mathbf{1}_{N\times N} - \mathbf{I}$. We claim that if the most-sharing equilibrium involves all agents choosing $\mathcal{S}$ or $\mathcal{I}$ (with probability 1) and agents never choosing $\mathcal{D}$ under this configuration, then this is the platform's profit-maximizing sharing network. By Lemma 1, all agents employ the same cutoffs $(b^*, b^{**}) = (0, b^{**})$ and $1 - H(b^{**})$ determines the proportion who share in the most-sharing equilibrium.

We focus on a modified social media game that only allows agents to ignore or share, which necessarily increases virality of content for any sharing network $\mathbf{P}'$ but does not increase the virality for the complete sharing network, by assumption. We show that for any other sharing network $\mathbf{P}'$, the largest fixed point (the most-sharing equilibrium), must necessarily be above $b^{**}\mathbf{1}$ (in the order $\preceq$). To do this, we consider the largest fixed point under $\mathbf{P}'$ (call this $\mathbf{b}^{**'}$), and use the same mathematical arguments as before, only disregarding the dislike cutoff. Let $\mathbf{B}'$ be the cutoff space where $\hat{\mathbf{b}}^{**}$ satisfies $\hat{\mathbf{b}}^{**} \preceq \mathbf{b}^{**'}$ and let the map $\varphi : \mathbf{B}' \to \mathbf{B}'$ map fixed cutoff strategies $\hat{\mathbf{b}}^{**}$ to best-response sharing cutoff strategies under the complete sharing network. It only remains to prove that φ indeed maps into $\mathbf{B}'$. To do this, let $U_j^c(\mathcal{S})$ be the utility from sharing under the complete network and $U_j'(\mathcal{S})$ as sharing under $\mathbf{P}'$, and note that $U_j^c(\mathcal{S}) - U_j'(\mathcal{S}) = \kappa \sum_{\tilde{j}=1}^{N} (1 - p'_{j\tilde{j}})(1 - H(\hat{b}_{\tilde{j}}^{**})) \geq 0$.

Thus, by Tarski's fixed-point theorem, we once again obtain that $b^{**}\mathbf{1} \preceq \mathbf{b}^{**'}$. Finally, observe that this necessarily implies that $\mathbf{P}'$ is less viral, because for every prior realization and seed agent i^*, $\mathbf{S}_{i^*}$ is larger in the complete network than in any other sharing network, provided that $b^{**}\mathbf{1} \preceq \mathbf{b}^{**'}$ and $b^* = 0$ (no agent dislikes). By Proposition 1, (uniformly more) sharing is monotone in reliability, so there exists some r_P such that for $r > r_P$, the complete sharing network admits only shares and ignores, whereas when $r < r_P$, agents dislike with positive probability. When $r > r_P$, the network takes the form of part (ii) by setting $p_s = p_d = 1$.

Next, we consider the case where $r < r_P$, and so (b^*, b^{**}) are the cutoffs in the most-sharing equilibrium with a complete sharing network, but where $b^* > 0$. First, notice that there must exist an open interval where agents with priors $b_i \in (0, \bar{b})$ will *never* share, regardless of the sharing network $\mathbf{P}$. To see this, suppose that there exists some $\mathbf{P}'$ where *all* agents either share or ignore, so the equilibrium cutoffs are determined by $\mathbf{b}^{**'}$ (and $\mathbf{b}^{*'} = \mathbf{0}$). Using the reasoning as in the previous paragraph (but extending it to the full cutoff space $(b^*\mathbf{1}, b^{**}\mathbf{1})$), we conclude that the cutoffs in the most-sharing equilibrium of the complete sharing network must satisfy $(b^*\mathbf{1}, b^{**}\mathbf{1}) \preceq (\mathbf{b}^{*'}, \mathbf{b}^{**'})$, and in particular, $b^*\mathbf{1} \preceq \mathbf{b}^{*'} = \mathbf{0}$. This implies that all agents share or ignore in the complete network, yielding a contradiction. Therefore, such an interval $(0, \bar{b})$ must exist, and in particular, we choose the largest such $\bar{b}$ (in the supremum sense) where agents with priors in $(0, \bar{b})$ *never* share in any sharing network $\mathbf{P}$ in the most-sharing equilibrium.

Next, we consider disconnecting (and removing) all agents in any community ℓ with $b^{(\ell)} < \bar{b}$, but leaving all other communities connected in a (partial) complete network. We call this network the *active network*. We claim that when ε is sufficiently small, all of the remaining agents in the active network either share or ignore. By definition, an agent with $\bar{b}$ would share under some sharing network $\mathbf{P}^*$ but any agent with $\bar{b} - \epsilon$ would ignore (for arbitrarily small ϵ) under $\mathbf{P}^*$ (and by leveraging Lemma A.4(i), not all agents with $b < \bar{b}$ dislike). This implies that $U_i(\mathcal{D}) < 0$ for an agent with prior $\bar{b}$ (by monotonicity), and thus an agent with this prior either shares or ignores in the active network. Moreover, for agents with priors in a small half-open neighborhood around $\bar{b}$ (i.e., an interval $(\bar{b} - \eta, \bar{b}]$ for some $\eta > 0$) ignoring is a better response to disliking in the active network. Thus, for sufficiently small ε, we obtain a partial complete network (the active network) with agents who only share and ignore (with probability 1) and never dislike (with probability 0).

Finally, with these two observations, we claim that the profit-maximizing sharing network takes the form of part (i). First, consider all communities who participated in the active network described above (call these the active communities) and suppose that the agents in communities outside of this active network are non-existent in our model (call these the inactive communities). When the active communities are arranged in a complete sharing network, we showed in the previous paragraph that all agents either share or ignore. By the exact argument in the first two paragraphs then, engagement (and virality) are maximized (amongst only the active communities) when these communities are arranged in a complete sharing network. Second, by construction of the active network (and the active communities), all agents in inactive communities *never* share under *any* sharing network $\mathbf{P}$. Therefore, removing these agents is without loss to potential virality. Hence, whenever virality is maximized amongst only agents in active communities, it is also maximized in general.

Lastly, we note that we can form a (partial) complete network among the inactive communities, but provide no connections to the (partial) complete network of active communities who only share and ignore (with probability 1). By our previous observations, this is a profit-maximizing sharing network for the platform. At the same time, it is exactly the form of a two-island model with $(p_s, p_d) = (1, 0)$, which has maximal homophily. ∎

Proof of Proposition 3. Note by Theorem 3 that an agent i with prior $b_i = b^{(k+1)}$ is indifferent between ignoring and disliking when $r = r_P$ (but strictly prefers to either share or ignore for all $r > r_P$), so r_P increases if and only if this agent (strictly) prefers to dislike following a shift in parameters. Because $b^{(k+1)} < 1/2$, an increase in polarization means that agent i's prior decreases (see Lemma A.3), and given that $\partial\pi_i/\partial b_i > 0$ (see Lemma A.1), π_i decreases for this agent. As a consequence $U_i(\mathcal{D})$ increases but $U_i(\mathcal{I})$ remains the same, so agent i (strictly) prefers to dislike. Similarly, because $\partial\pi_i/\partial p < 0$ and $\partial\pi_i/\partial q < 0$ for $b_i < 1/2$ (see Proposition 2), π_i decreases for this agent (making $a_i = \mathcal{D}$ a best response). In both cases, we see that r_P increases. ∎

A.5 Proofs from Section 6

Proof of Proposition 4. Consider the profit-maximizing sharing network before any censorship policy is enacted ($\delta = 0$). By Theorem 3 and the assumption that $\mathbf{b}^* \neq \mathbf{0}$ and $\mathbf{b}^{**} \neq \mathbf{1}$, it must be the case the profit-maximizing sharing network has maximal homophily with two islands, one with the optimal seed agent (island A) and one without it (island B). By construction of the profit-maximizing sharing network, no agent on island B would share if connected fully to island A or under any other sharing network configuration (see the proof of Theorem 3).

Consider any agent j residing on island B. Because the platform approximates the belief distribution H by a generic multinomial distribution,[27] it must be the case that $U_j(\mathcal{S}) < \max\{U_j(\mathcal{I}), U_j(\mathcal{D})\}$ under any sharing network configuration for agent j. Hence, there exists some $\underline{\delta} > 0$ such that substituting $\phi(r)$ with $\tilde{\phi}(r) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))}$ for all $\delta \in (0, \underline{\delta})$ leaves the profit-maximizing sharing network for the platform unchanged, because the strict inequality above still holds under any chosen sharing network. At the same time, if $E \equiv \max_{i^*} \mathbb{E}[\mathbf{S}_{i^*}|\nu = \mathcal{M}]$ is the engagement with misinformation before the policy, engagement with misinformation after the policy is $(1-\delta)E < E$. Thus, the policy for $\delta^* \in (0, \underline{\delta})$ is more effective than $\delta = 0$, and in fact higher values of $\delta^* \in (0, \underline{\delta})$ are more effective.

Next, we note that $\tilde{\phi}(r) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))} = 1$ when $\delta = 1$. It is immediate then that for sufficiently high values of δ, the profit-maximizing sharing network has maximal connectivity and all agents share in equilibrium. Let us consider $\bar{\delta}$ which is the largest value (in the supremum sense) such that the profit-maximizing sharing network does not have maximal connectivity. Under censorship policy $\bar{\delta}$, the virality of misinformation is at most $(1-\bar{\delta})(N-1)/N$, but for any $\zeta > 0$, the censorship policy with $\bar{\delta}$ has virality $1 - \bar{\delta} - \zeta$. Letting $\zeta < (1-\bar{\delta})/N$, we see that a censorship policy with $\bar{\delta}$ is more effective than any policy with $\bar{\delta} + \zeta$.

Finally, let us construct $0 < \delta_1 < \delta_2 < \delta_3 < 1$ to conclude. Let us take δ_1 to be the largest value of δ (in the supremum sense) such that $\delta^* = \delta$ is the most effective policy for all $\delta \in (0, \delta_1)$. We know that that such a δ_1 exists and is strictly less than 1 by the arguments in the two above paragraphs. Using a similar argument as in the second paragraph, there always exists an open interval (δ_1, δ_2) where the $\delta^* = \delta_1$ policy is more effective than any $\delta^* \in (\delta_1, \delta_2)$, and so in particular $\delta^* < \delta$ is optimal. Lastly, we show (i) any censorship policy bounded away from 1 can always be beat by one sufficiently close to 1, and (ii) the one that beats it can beat by any δ greater than it. This proves that there exists δ_3 whenever $\delta \in (\delta_3, 1)$, $\delta^* = \delta$ is the most effective policy. For (i) suppose that some policy $\tilde{\delta}$ achieves

[27] Formally, given that the platform has microtargeting technology $\varepsilon > 0$, the optimally chosen sharing network (for the platform) with prior distribution H is equivalent to the platform's optimally chosen sharing network for a multinomial distribution consisting of a number of atoms chosen "generically" (each atom chosen at random from an interval of size ε) in each of the prior regions $[b^{(k)}, b^{(k+1)}]$, as described in Section 5.

misinformation engagement $\tilde{E} > 0$; then, because engagement with unidentified misinformation cannot exceed N, any policy $\delta^* > (N - \tilde{E})/N$ is strictly more effective. For (ii) we know there exists some $\hat{\delta}$ sufficiently close to 1 such that profit-maximizing sharing network has maximal connectivity and thus is the same for all $\delta^* \in (\hat{\delta}, 1)$; therefore, the virality of misinformation is $(1-\delta^*)$ for all $\delta^* \in (\hat{\delta}, 1)$ and higher values of δ^* are always more effective. ∎

Proof of Proposition 5. We take a similar approach as in the proof of Proposition 4. Once again, we consider islands A and B which are guaranteed by Theorem 3 before any provenance policy has been enacted ($\rho = 0$) and note that $U_j(\mathcal{S}) < \max\{U_j(\mathcal{I}), U_j(\mathcal{D})\}$ for all agents j on island B regardless of the sharing network chosen. With the introduction of a provenance policy, however, the profit-maximizing sharing network may not take the form of Theorem 3. Despite this, we can still upper bound the ex-ante likelihood of an article being truthful by $\frac{\phi(r)}{\phi(r)+(1-\rho)^N(1-\phi(r))}$, which holds independent of the sharing network chosen. Once again, for small enough $\underline{\rho} > 0$ the strict inequality $U_j(\mathcal{S}) < \max\{U_j(\mathcal{I}), U_j(\mathcal{D})\}$ will still hold in any sharing network, and the profit-maximizing sharing network will be the same as $\rho = 0$ for all $\rho^* \in (0, \underline{\rho})$. In this region, virality is given by $\left[(1-\rho^*) + (1-\rho^*)^2\right]\beta$ where β is the fraction of agents (relative to N) on island A, which is monotonically decreasing in ρ^*.

At the same time, we can lower bound the ex-ante likelihood of an article being truthful by $\frac{\phi(r)}{\phi(r)+(1-\rho)(1-\phi(r))}$, which is equal to 1 when $\rho = 1$. Thus, note that for sufficiently high values of ρ, the profit-maximizing sharing network will fit the form of Theorem 3 with a maximally connected network because all agents share in equilibrium achieving maximal virality. Once again, let us consider $\bar{\rho}$ which is the largest value (in the supremum sense) such that the profit-maximizing sharing network is something *other* than maximal connectivity. For $\rho = \bar{\rho}$, there must be at least one agent who would not share under any chosen sharing network. As in Proposition 4, for any $\zeta < (1-\bar{\rho})/N$, a provenance policy with $\bar{\rho}$ is more effective than any policy with $\bar{\rho} + \zeta$. The construction of $0 < \rho_1 < \rho_2 < \rho_3 < 1$ then follows exactly in the same way as from the last paragraph in Proposition 4.

Finally, we show that $\rho_3 \leq \delta_3$ and in this region the provenance policy is more effective than censorship. Recall that δ_3 was chosen such that it is the minimum value of δ where the profit-maximizing sharing network is maximally connected, and for $\rho = \delta_3$, it must also be maximally connected, because the perceived ex-ante likelihood of truth is lower bounded by $\frac{\phi(r)}{\phi(r)+(1-\rho)(1-\phi(r))}$, which is the ex-ante likelihood of truth for a censorship policy where $\delta = \rho$. The expected virality in the censorship regime for $\delta^* > \delta_3$ is $(1-\delta^*)$, but it is only $[(1-\rho^*) + (1-\rho^*)^2(N-1)]/N < (1-\rho^*)$ when $\rho^* > \rho_3$. ∎

Proof of Proposition 6. If the platform removes ψ fraction of misinformation, then its performance metric is given by

$$\frac{(1-\psi)E(\psi)(1-\phi(r))}{(1-\psi)E(\psi)(1-\phi(r)) + E(\psi)\phi(r)} = \frac{(1-\psi)(1-\phi(r))}{(1-\psi)(1-\phi(r)) + \phi(r)}$$

where $E(\psi)$ is the user engagement when the platform removes ψ fraction of misinformation and optimally chooses the sharing network, but notice that $E(\psi)$ does not affect the performance metric.

If the platform hits the performance target λ, then it chooses ψ according to $\psi = \frac{1-\lambda-\phi(r)}{(1-\lambda)(1-\phi(r))}$. Observe that ψ is monotonically decreasing in λ, with the strictest target ($\lambda = 0$) yielding $\psi = 1$ and the loosest target ($\lambda = 1 - \phi(r)$) yielding $\psi = 0$. The payoff from hitting the performance target exactly is given by $V = E(\psi)\phi(r)/(1-\lambda)$ whereas the payoff from not hitting it is $V' = (1-\alpha)E(\psi^*) - \alpha C$, where ψ^* is the self-imposed target by the platform that maximizes engagement, i.e., $\psi^* = \max_\psi E(\psi) < 1$ by nature of $\phi(r) < \max_{i^*} \mathbb{E}[\mathbf{S}_{i^*}]$. For any $\psi < \psi^*$, the platform of course meets the performance target. For any $\psi > \psi^*$, we can define $E^*(\psi) = \max_{\psi' \geq \psi} E(\psi)$, which is of course a monotonically decreasing function in ψ. Thus, the platform compares $V = E^*(\psi)\phi(r)/(1-\lambda)$ with a constant $V' = (1-\alpha)E(\psi^*) - \alpha C$. Note that V is monotonically increasing in λ: $1/(1-\lambda)$ is increasing in λ, and $E^*(\psi)$ is decreasing in ψ, and therefore increasing in λ. Thus, there exists some cutoff λ^* such that when $\lambda > \lambda^*$, $V > V'$, but when $\lambda < \lambda^*$, $V < V'$. The claim follows by noting that virality of misinformation is proportional to $E(\psi)$ where ψ is chosen by the platform. ∎

Proof of Proposition 7. The network regulation does not bind for an article with $r > r_P$, so we need only consider $r < r_P$. Take some agent i with prior $b_i \in (\bar{b}, \bar{b} + \eta)$ in a small neighborhood $\eta > 0$ of $\bar{b}$ (where $\bar{b}$ is the same $\bar{b}$ constructed in Theorem 3). Following the same line of reasoning as in Theorem 2(a), agents with priors in this interval elect to ignore instead of share following the network regulation (and when η is sufficiently small), and this necessarily reduces the virality of misinformation, showing (ii). To prove (i), we note that agents in this neighborhood around $\bar{b}$ also do not share in the most-sharing equilibrium under *any* sharing network $\mathbf{P}'$ (following the network regulation), per the construction of $\bar{b}$ in Theorem 3. Therefore, the platform cannot generate additional engagement by departing from the class of island models (specifically, two-island models) while maintaining $p_s/p_d \leq p^*$. ∎

B Endogenous Reputation Loss

In the model of Section 2, we assume that each agent cares about getting called out for sharing potential misinformation, in the form of exogenous punishments for each dislike she receives. In this section, we show that this formulation can be microfounded by an endogenous reputational concern.

Suppose that at $t = 0$, every agent is born as either a *careless* agent or a *normal* agent. The careless type is behavioral and shares every article, whereas a normal user is a fully rational agent as in Section 2. The user is born careless with probability $\mu > 0$ which is i.i.d. across all agents and immutable. For each dislike agent i receives, there is some probability $\zeta > 0$ that the share by agent i receives public scrutiny and it becomes common knowledge to the population that i shared the article. Conditional on such a broadcast, the population updates its beliefs μ to $\hat{\mu}_i$ about the likelihood that agent i is actually the careless type.

We assume each agent i intrinsically values $1 - \hat{\mu}_i$, the public belief that she is not a careless user, for example, because this might affect their other social relations or economic prospects. We can represent the strength of this concern (relative to other sources of utility) with a parameter like d in Section 2. For example, a doctor may place much more weight on reputation than a social media troll trying to "stir the pot."

When there is no public broadcast about agent i, we have $\hat{\mu}_i \leq \mu_i$. In contrast, when there is a broadcast about i's share, $\hat{\mu}_i > \mu_i$, and this increase is larger for articles with lower reliability. This reasoning therefore introduces an endogenous reputational concern originating from dislikes that the agent receives. Specifically, it is straightforward to see that the agent's utility will now include a term $\psi_D(d, D_i) = d\Delta\hat{\mu}_i(1 - (1 - \zeta)^{D_i})$, where $\Delta\hat{\mu}_i$ is the difference between i's reputation after a public broadcast of her share and i's reputation without a broadcast, and $\psi_D(d, D_i)$ exhibits increasing differences. By the same observation as in footnote 9, all of our results apply without modification.