Review of Economic Studies (2024) **00**, 1–34 doi:10.1093/restud/rdad111

Advance access publication 12 December 2023

A Model of Online Misinformation

Daron Acemoglu
Massachusetts Institute of Technology, NBER, and CEPR, USA
Asuman Ozdaglar
Massachusetts Institute of Technology, USA
and
James Siderius
Tuck School of Business at Dartmouth College, USA

First version received July 2022*; Editorial decision September* 2023*; Accepted November* 2023 *(Eds.)*

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*. Each agent starts with an ideological bias and gains utility from positive social media interactions but does not want to be called out for propagating misinformation. We characterize the (Bayesian–Nash) equilibria of this social media game and establish that it exhibits strategic complementarities. Under this framework, we study how a platform interested in maximizing engagement would design its algorithm. Our main result establishes that when the relevant articles have low-reliability and are thus likely to contain misinformation, the engagement-maximizing algorithm takes the form of a "filter bubble"—creating an echo chamber of like-minded users. Moreover, filter bubbles become more likely when there is greater polarization in society and content is more divisive. Finally, we discuss various regulatory solutions to such platform-manufactured misinformation.

Key words: Echo chambers, Fake news, Filter bubbles, Homophily, Misinformation, Networks, Social media
JEL codes: D83, D85, P16

1. INTRODUCTION

Social media has become an important source of information for citizens around the world. Leading up to the 2016 U.S. presidential election, 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 (Ro'ee, 2021). At the same time, there is growing concern about misinformation on social media platforms, including made-up news stories about U.S. political candidates, misleading Brexit reports, and photoshopped images during the Indian national elections, just to name a few.[1] Recent evidence suggests that

1. For example, see https://www.snopes.com/fact-check/mass-shootings-under-trump/ and Greene *et al.* (2021) and Garimella and Eckles (2020).

The editor in charge of this paper was Andrea Galeotti.

misinformation on social media has impacted critical decisions such as vaccinations against COVID-19 (Pennycook *et al.*, 2018, 2020).

Although there is 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 users communicate and share content with like-minded individuals (Lazer *et al.*, 2018; Sunstein, 2018). Törnberg (2018) and Del Vicario *et al.* (2016) show that echo chambers reinforce existing political viewpoints and tend to propagate misinformation. Social media allows individual users to choose who and what they listen to much more than traditional media, and thus echo chambers may be an unavoidable side effect of online interactions. There is also evidence that echo chambers are a result of "filter bubbles" that platform algorithms create by promoting content from like-minded users (see Pariser, 2011; Ro'ee, 2021 on Facebook). The second factor is the general political polarization in many countries, and especially the U.S.,[2] and there is some preliminary evidence that polarization has contributed to selective exposure to questionable content on social media as well (Guess *et al.*, 2018). Despite the importance 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 article, we develop a parsimonious model of online sharing behaviour in the presence of misinformation, and as a first step, we focus on the behaviour of fully Bayesian agents.[3] Our model is inhabited by a set of N social media users. Each agent (user) 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 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 users. 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 component.[4]

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 user. 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 calling out ("disliking") items that they believe contain misinformation.

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

3. Cognitive biases appear to play a major role in social media interactions, for example, via the "confirmation bias" (see *e.g.* Pennycook and Rand, 2019; Buchanan, 2020) or emotional outrage. Nevertheless, our Bayesian benchmark already generates a number of empirically relevant results, and incorporating realistic behavioural biases into this framework is an important next step.

4. Our focus in this article 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; Grinberg *et al.*, 2019; Guess *et al.*, 2019; Allen *et al.*, 2020). For example, according to this definition a news item that favourably 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.

We characterize the Bayesian–Nash equilibria of this sequential game. Equilibria exist and are in cutoff strategies—a user shares any item that she believes is truthful with a high probability and dislikes articles that she believes to contain falsehoods. Items with intermediate beliefs are ignored. Beliefs about the truthfulness of articles are formed on the basis of the article's reliability and message, and agents' ideology (prior). We first establish that our game exhibits strategic complementarities: 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 then turn our attention to the platform's algorithm design choices. We assume that platforms determine the sharing network of users in order to maximize engagement (which serves to increase revenues from advertisements).[5] Under this assumption, we first establish that the platform always selects a sharing network from a class of island networks, characterized by a partition of agents into distinct islands and two parameters, the probabilities of content being shared within an island and between islands. When islands contain like-minded agents, the gap between these two parameters defines the extent of homophily—the degree to which content is shared between users with similar ideologies.

Our main result shows that, in the presence of low-reliability articles (which are likely to contain misinformation), the engagement-maximizing algorithm choice is to induce maximal homophily. In contrast, with high-reliability articles, the platform prefers no homophily and maximal connectivity. This bifurcated pattern of algorithmic choices reflects a fundamental non-monotonicity in our model. All else equal, high homophily reduces the viral spread of articles and thus user engagement. Counteracting this, however, high homophily enables low-reliability content to escape scrutiny, because it is being shared among like-minded individuals who are predisposed to agree with its message and thus are likely to share it themselves and unlikely to dislike it. In contrast, low-reliability content is likely to receive dislikes and less likely to be shared in ideologically diverse networks. This non-monotonicity implies that, faced with low-reliability articles, the platform prefers to form filter bubbles: in these circumstances, echo chambers are attractive for platforms precisely because they create an environment in which it is safe for low-reliability content to spread. Consequently, filter bubbles and viral spread without counter-attitudinal voices become more prevalent when the relevant content is more likely to contain misinformation.

Equally concerning are the comparative statics of filter bubbles. We show that platforms are more likely to favour filter bubbles when society is politically polarized and articles are more divisive (generating greater disagreement among agents with different ideologies). The intuition for this comparative static is related to the logic of filter bubbles: with greater political polarization and divisiveness, low-reliability content is more likely to be disliked and stop spreading in ideologically diverse networks, making platforms opt for echo chambers.

If platform algorithms are propagating misinformation, can public policy discourage this type of behaviour?[6] In the last part of the article, we show that the answer is yes, but with

5. In practice, one's network on social media sites is formed both by their friends and acquaintances who have also joined the same platform, and the algorithmic choices of the platform that promote and make visible different users and posts. The sharing network should thus be interpreted as how (mis)information can potentially spread from platform recommendations.

6. As of August 2021, federal law in the U.S. protects social media platforms from being held responsible for content posted by its users, *even if* the platform determines how this content is recommended and shared (see Section 230 of the Communications Decency Act of 1996, discussed by https://hbr.org/2021/08/its-time-to-update-section-230).

some caveats. We define a social welfare metric related to information acquisition—the average distance between user beliefs and the true state—and establish that this is related to the spread of low-reliability content. We then 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.*, 2020) that contributes to the 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 due to a related implied truth effect, but this time 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 platforms are appropriately monitored and penalized for violations, strict targets can exacerbate the spread of misinformation. Lastly, we show how regulation of platform algorithms can reduce misinformation, but the non-monotone effects of homophily imply that such regulations need to be finely calibrated.

1.1. *Related literature*

Our article builds on a growing body of work that studies the emergence and spread 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, 2013), the existence of persuasive agents can impede information aggregation and enable misinformed beliefs to survive, and sometimes even become dominant, in the population (these works in turn build on prior models of learning with boundedly rational agents, such as Bala and Goyal, 1998; Golub and Jackson, 2010). In Mostagir *et al.* (2022) and Mostagir and Siderius (2023), a principal who wants to persuade agents of an incorrect belief can distort the learning process by leveraging social connections and echo chambers to propagate misinformation. Similarly, models of misinformation "contagion"—without Bayesian agents or strategic decisions—have been studied in Budak *et al.* (2011), Nguyen *et al.* (2012), and Törnberg (2018). Our work is distinguished by its focus on strategic behaviour by Bayesian agents and algorithmic choices of engagement-maximizing platforms, which are not featured in these papers.

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; 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 this content, while Chen and Papanastasiou (2021) and Keppo *et al.* (2022) consider more manipulative actions by platforms, including strategic seeding of information or "cheap talk" signals about quality. Relatedly, 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 article contributes to this literature by highlighting

Many policy debates in the U.S. have revolved around potential amendments to Section 230, largely motivated by legislation in Europe, such as the Online Safety Bill in the U.K.

the role of ideological leaning, strategic sharing decisions, homophily, and most importantly, the design of social media algorithms by platforms.

The most closely related work to ours is Papanastasiou (2020). Papanastasiou (2020) studies a model where agents hold heterogenous ideological beliefs, and digest and share a news article sequentially. There are three critical differences between his model and ours, however. First, Papanastasiou (2020) focuses on costly inspection, which makes sharing decisions strategic substitutes in his model (see also related work Merlino *et al.*, 2023). In contrast, our focus on share and dislike decisions by other agents leads to strategic complementarities, as individuals care about the reaction of their social network.[7] All of our results and formal analysis turn on strategic complementarities. Second, and relatedly, echo chambers play no role in Papanastasiou (2020).[8] Third, our analysis of algorithmic design choices for maximizing engagement 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.[9]

Finally, our work shares some similarities with Kranton and McAdams (2022) and Mostagir and Siderius (2022a, 2022b) in studying the influence of misinformation in networks. In Kranton and McAdams (2022), social connectivity determines the equilibrium production of misinformation from content creators, but they do not consider platforms' algorithm choices and policy responses. Mostagir and Siderius (2022a, 2022b) are subsequent to this article and also discuss some of the regulatory questions surrounding misinformation. In any case, the focus of those papers is limited to various behavioural learning rules agents adopt, with no strategic behaviour from users and platforms, and thus has little overlap with the insights our model generates.

The rest of the article 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 considers the algorithmic choices of the platform through an endogenously chosen sharing network that aims to maximize engagement. Section 5 discusses a range of regulations aimed at containing misinformation. Section 6 concludes, while all proofs are provided in Appendix A. Supplementary Appendices B and C, which contain additional results and microfoundations for some of the assumptions in the text, are available online.

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.

7. Our reading of the evidence is that strategic complementarities are more relevant for social media behaviour 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 behaviour and posting. See also Taylor and Eckles (2018) and Aral and Dhillon (2018), and the detailed discussion in Frenkel and Kang (2021).

8. 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).

9. 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, favours 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.

2.1. *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 (realized) neighbourhood $\mathcal{N}_i$ as the set of agents attached to her with an outgoing link (and $|\mathcal{N}_i|$ as its size). As discussed in Section 1, the platform's algorithms for promoting content fully determine this sharing network. The news item in question could be an article or a post by one of the users, and throughout we refer to it as an "article."

2.2. *Misinformation and news generation*

Each article has a three-dimensional type (r, m, v). 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, v 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* (as defined in footnote 4).

We assume that, at the beginning of the game, the type vector (r, v, m) of the article is sampled from the following process:

(i) The article has some given reliability score $r \in [0, 1]$.
(ii) The veracity of the article is drawn as $v = \mathcal{T}$ (contains truthful content) with probability $\phi(r)$ or as $v = \mathcal{M}$ (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 $v = \mathcal{T}$ (the article is truthful), then its message is generated as $m = \theta$ with probability $p > 1/2$. Conversely, if $v = \mathcal{M}$ (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 (*e.g.* the message m is directly observed and reliability depends on certain commonly observed characteristics such as source and headline), the third dimension, v, is unknown to all agents. We assume that agents update their beliefs about v using Bayes's rule given there prior about θ and the observables (r, m) of the article.

2.3. *Social media behaviour*

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 with the content contained in it (*e.g.* call it out as misinformation).

Although there are some differences in nomenclature across real social media platforms, online interactions on many platforms can broadly fit within this paradigm.[10]

10. For example, several social media platforms actually have an explicit "Share" button (*e.g.* Facebook/TikTok/Instagram), and others have an implicit way to Share (*e.g.* "Crossposting" on Reddit or "Retweeting" on Twitter). Actions similar to our "Dislike" are available to users by selecting a negative reaction (pressing the "downvote" button on Reddit or YouTube) or leaving a disparaging comment on Facebook.

2.4. *Timing*

Time is discrete $t = 1, 2, \ldots$. The article diffuses throughout the network as agents receive the article and then decide whether to share, ignore, or dislike it. At time $t = 1$, we assume that some initial seed agent receives the article. If the article is shared by agent i, it is passed to all $j \in \mathcal{N}_i$. If the article is ignored by her, it does not propagate past agent i (but may still propagate along other diffusion paths). If the article is disliked by agent i, the agent j who shared the article with i receives negative peer feedback (getting "called out"), as we describe below.

2.5. *Payoffs*

Let us denotes the number of shares after i by $S_i = |\{j \in \mathcal{N}_i : a_j = \mathcal{S}\}|$ 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

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

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 but not truthful ones.

(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 $\mathcal{S}$, $U_i^{(1)} = u \cdot \mathbf{1}_{v=\mathcal{T}} - c \cdot \mathbf{1}_{v=\mathcal{M}}$. Second, agents enjoy positive feedback from their peers (such as likes, or in our setting reshares) but are negatively affected by dislikes. This is represented by the second component of utility $U_i^{(2)} = \kappa \cdot S_i - d \cdot D_i$.[11] The total utility for agent i's sharing action is the sum of these two components, $U_i^{(1)} + U_i^{(2)}$.

11. The terms $\kappa \cdot S_i$ and $d \cdot D_i$ could be replaced by arbitrary functions $\varphi_S(\kappa, S_i)$ and $\varphi_D(d, D_i)$ with (weakly) increasing differences. This generalization would enable us to incorporate a more extensive set of social interactions. For example, an additive constant κ would mean that there is a higher propensity to share all else equal. These modifications lead to an essentially identical analysis, and we do not pursue them here to prevent the notation from getting more complex. We could also allow the share utility from peer reactions, $U_i^{(2)}$, to depend on the article's veracity as well. For example, the functional form $\tilde{U}_i^{(2)} = \pi_i \cdot \kappa \cdot S_i - (1 - \pi_i) \cdot d \cdot D_i$ (where π_i is the posterior belief of agent i that the article is truthful) would maintain strategic complementarity but would reward (resp. punish) the agent for reshares (resp. dislikes) only if the article was truthful (resp. contains misinformation). Our results apply identically under this alternative formulation as well.

2.6. *Information structure and solution concept*

Both the stochastic network $\mathbf{P}$ and the distributions $\{H_i\}_{i=1}^N$ of beliefs in the population are common knowledge, but the realized links and beliefs of other agents are not known with certainty. We focus on Bayesian–Nash equilibria, and refer to these as "equilibria" for short.[12]

To eliminate trivial and unrealistic equilibria, we assume κ 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 behaviour and information structure in social media. As documented in studies such as Pennycook *et al.* (2021), users want to share content that they believe does 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 *e.g.* 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 *e.g.* Kim *et al.*, 2020 for evidence in the context of 2018 midterm elections).

3. SOCIAL MEDIA EQUILIBRIA

In this section, we characterize the structure of equilibria for any sharing network structure $\mathbf{P}$. Without loss of generality, we fix the article's message as $m = R$ and reliability r for the remainder of the article.

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 behaviour of payoffs will lead to simple best-response decision rules, 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

12. Observe that the sharing process is Markovian and does not depend on the history of an article's spread (in contrast to models such as Papanastasiou, 2020). Specifically, sharing decisions depend only on the characteristics of the article (reliability and message), the network structure, and the belief distributions. Because our results do not depend on other aspects of agents' beliefs about sharing history, nor do agents maximizing expected utility need to observe the realized subsequent actions from others on the platform, our model is applicable to behaviour on most social media platforms, such as Facebook, Twitter, and Instagram.

$b_i < b_i^*(r)$. 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 $(\mathbf{b}^*, \mathbf{b}^{**}) = (b_1^*, b_1^{**}, \ldots, b_N^*, b_N^{**})$.

The social media game additionally 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, which also raises $U_j^{(2)}$. Strategic complementarities capture an important dimension of social media interactions: utility feedback from others' behaviour tends to encourage agents to cohere with those behaviours.

3.2. *Equilibrium structure*

Our next result characterizes the set of equilibria.

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.*

Moreover, greater r, greater κ*, or smaller d decrease the extremal equilibria (in the lattice order) and all agents share with higher probability.*

Theorem 1 shows that, despite equilibrium 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*). Moreover, comparative statics with respect to r, κ, and d are intuitive. The reliability parameter r is linked to the likelihood of online misinformation and thus our results imply that, all else equal, less reliable articles spread less virally (this also clarifies that virality of misinformation is not a mechanical result in our model). The parameter κ captures how *sensational* the article is: higher κ implies that agents receive greater social utility from sharing, including greater value from future shares, because these shares are associated with others paying more attention to the relevant posts (see Footnote 11 for a general discussion). Finally, the parameter d proxies for the importance of *reputational concerns*. Higher d means that dislikes are more damaging, which corresponds to the agent being more concerned about receiving negative reactions (*e.g.* loss in reputation, see Supplementary Appendix B for details).

Empirical mapping—Although stylized, our model's comparative statics are in line with a growing empirical literature on social media and misinformation. Our predictions are consistent with Pennycook and Rand (2019), who document that attentive social media users, regardless of partisanship, do not typically share low-reliability content such as those from Breitbart or Infowars. Despite this observation, other works such as Vosoughi *et al.* (2018) document that misinformation spreads farther, faster, deeper, and more broadly than truthful news on social media. Our results suggest that this may be because there is a set of articles with low r and high κ, which become viral despite their low reliability. This also coheres with the findings in the literature: Molina *et al.* (2021) and Kozyreva *et al.* (2020) observe that sensational content is often low-reliability, and Grinberg *et al.* (2019) establish that once the effects of sensational news items are controlled for, misinformation does not spread farther or faster than truthful content.

4. PLATFORM DESIGN AND FILTER BUBBLES

4.1. *Platform's problem*

The platform's profit objective is to maximize *user engagement*, defined as the expected fraction of agents in the network who share the article, and throughout we focus on the most-sharing equilibrium.[13]

The platform chooses how content is shared across users. That is, for each article, the platform not only picks a seed agent at $t = 1$ to whom it recommends this article but also chooses the sharing network—the matrix of link probabilities $\mathbf{P}$. The platform's choice of $\mathbf{P}$ can be interpreted as its "algorithm" to determine how users are exposed to content circulating in the social media site. While the platform has freedom to choose any sharing network is desires, this algorithm choice is assumed to be common knowledge.[14]

We assume there is 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 may depend on the platform's microtargeting technology at identifying users' ideological beliefs (see *e.g.* 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 capabilities for identifying ideology.

4.2. *Optimality of island networks*

Let us define the class of *island networks* (or equivalently, the stochastic block model), which are lower-dimensional than general networks we have considered so far. In an island network, agents are partitioned into k blocks of size $N_1, N_2, \ldots, N_k$, called *islands* each with 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 island she is in. Link probabilities are then given by $p_{ij} = p_s > 0$ if $\ell_i = \ell_j$ and $p_{ij} = p_d$ if $\ell_i \neq \ell_j$, with $p_s \geq p_d$. Our next result shows that profit-maximizing sharing algorithms for the platform are within the class of island networks.

13. This objective is rooted in the business model of media sites, like Facebook, which primarily rely on advertising revenue and depend on user engagement in order to be able to present digital ads (see https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/). For example, 85% of Facebook's total revenue in 2011 was from advertising, and from 2017 to 2019, around 98% was (see Andrews, 2012 and https://www.nasdaq.com/articles/what-facebooks-revenue-breakdown-2019-03-28-0). Although other online interactions, such as dwelling or clicking on content, involve some amount of "user engagement," within our model sharing appears as the best proxy for user engagement.

One implication of our assumptions is that platform does not directly care about whether the content is truthful or contains misinformation. This can be interpreted as the objective of social media platforms 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 5.

14. This assumption is made under the interpretation that the platform's choice of sharing network dictates how content and (mis)information propagate online (and is not necessarily tied to any underlying social network). Note further that $\mathbf{P}$ varies at the article-level, as Facebook's algorithms may induce different sharing networks depending on features of the article such as topic, latent positions of the viewers, action rate on the article, and so on (see Eckles, 2022 and https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/).

Lemma 1. *There exists $\bar{\varepsilon} > 0$ such that for $\varepsilon < \bar{\varepsilon}$, the profit-maximizing sharing network always has an island structure. That is, in the profit-maximizing sharing network there exists a partition of all users into $k \geq 2$ islands such that agents on the same island are connected with probability $p_s > 0$ and agents on different islands are connected with probability $p_d \leq p_s$.*

Lemma 1 establishes that platform profit incentives lead to a type of "assortative matching" according to ideology or, simply, to *homophily*: the sharing network induces an individual to interact more frequently with others that have common characteristics as herself.

4.3. *Homophily comparative statics*

To understand how homophily affects user engagement, we first study comparative statics with respect to an arbitrary, exogenously given island network. For our next result, we suppose that there are k islands of agents, where within-island link probabilities are given by p_s and across-island link probabilities are given by p_d. Moreover, we assume the prior distribution for agents on the same island ℓ is the same, denoted by H_ℓ, and each island ℓ has distribution H_ℓ with support on $[b^{(\ell)}, b^{(\ell+1)}]$, where $1 \geq b^{(1)} > b^{(2)} > \ldots > b^{(k)} > b^{(k+1)} \geq 0$.[15] This implies that lower-indexed islands have stronger right-wing beliefs. We say that an island network with (p_s, p_d) has *greater homophily* than an island network with (p'_s, p'_d) if the expected degrees (expected connections) of all agents remains constant, but where $p_s > p'_s$ and $p_d < p'_d$.[16]

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

(a) if $r < \underline{r}$, greater homophily increases user engagement;
(b) if $r > \bar{r}$, greater homophily decreases user engagement.

Theorem 2 establishes an inherent non-monotonicity from homophily. With high homophily, users know that they are mostly sharing 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—which we call the "discipline effect." At the same time, high homophily makes it more probable that an article will circulate among the same group of like-minded users and reduces the likelihood that it will reach other communities—which we refer to as the "circulation effect." Theorem 2 thus establishes that the "discipline effect" is stronger for low-reliability content likely to contain misinformation, whereas the "circulation effect" is stronger for more reliable content.

4.4. *Comparative statics: divisiveness and polarization*

We now provide an additional comparative static with respect to polarization. For this result, we focus on the case of two islands, a left-wing and a right-wing one with prior distributions H_L and H_R, respectively. 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 $\bar{b}_L < 1/2 < \underline{b}_R$.

15. 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.

16. Higher network density (greater expected degrees) can be responsible for increases in user engagement solely due to an increase in the potential for more re-shares (through the strategic complementarity channel). Thus, when studying the effects of homophily, we hold network density fixed.

We say content with parameters (p', q') is *more divisive* than content with parameters (p, q) if $p \leq p'$ and $q \geq q'$. 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. As a result, it tends to generate more disagreement among ideologically diverse people; for a fixed-reliability article, if $b_i < 1/2$ (resp. $b_i > 1/2$) then π_i is lower (resp. higher) for more divisive content. In our case, we think of state θ as related to political ideology (with apolitical content having little divisiveness). 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. The next result studies how political divisiveness and polarization impact social media behaviour and the spread of misinformation, as a function of the homophily in the sharing network.

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

(a) if $r < r^*$ *and* $p_s/p_d > p^*$*, then greater divisiveness and/or greater polarization lead to greater user engagement;*
(b) If $r > r^*$ *and* $p_s/p_d < p^*$*, then greater divisiveness and/or greater polarization lead to less user engagement.*

Proposition 1 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 user engagement, 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 behaviour from individuals with different ideologies, and greater polarization means there are sharper differences in terms of these ideologies. Thus, Proposition 1 shows that echo chambers matter especially for divisive content and with polarized beliefs.

4.5. *Platform's sharing network: filter bubbles*

Special cases of the island network structure are: (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 $p_s = p_d = 1$ (and there is minimal homophily and no segregation by ideology). We next show that when the platform is allowed to design any sharing network $\mathbf{P}$, its profit-maximizing choice will be one of these two special networks.

Theorem 3. *There exists* $\bar{\varepsilon} > 0$ *such that if* $\varepsilon < \bar{\varepsilon}$*, the platform's profit-maximizing sharing network has* $k = 2$ *islands and 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;*

Moreover, the reliability threshold r_P *increases as divisiveness and/or polarization increases.*

Part (i) of the theorem shows that when articles are mostly unreliable—and likely to contain misinformation—the platform creates an extreme filter bubble 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

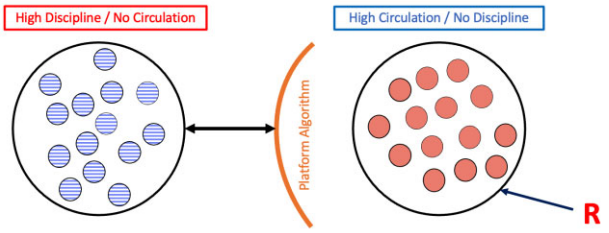

FIGURE 1
Platform filter bubble: the profit-maximizing sharing network

Notes: In the case where $r < r_P$, the platform's algorithm separates users into two islands. Those ideologically opposed to the article in question are placed on the left island. Those who are more favourably disposed to the article are placed on the right island and are disconnected from the left island, and thus can share without the disciplining influence of the reactions from those with different beliefs. Theorem 3 shows that this network structure maximizes user engagement among all possible sharing networks.

homophily. This result highlights an important channel via which misinformation spreads: it is precisely when articles are likely to contain misinformation that the platform seeks to maximize engagement by creating echo chambers, or filter bubbles, where these articles spread virally within like-minded communities. Put differently, with low reliability content, neither the platform nor users are disciplined about sharing misinformation, and so these news items spread virtually uninhibited.

The intuition for Theorem 3 lies in balancing the discipline and circulation effects from Theorem 2. As Figure 1 indicates, the platform chooses the number and composition of different islands, so that it achieves as much circulation as possible, while limiting the discipline effect. When $r > r_P$, this leads to maximal connectivity, because the discipline effect becomes moot for these high-reliability articles and the platform opts for the network structure that maximizes circulation. When $r < r_P$, the discipline effect is limited by placing users that are most ideologically opposed to and thus least likely to share the article in question into a separate island. The remaining agents are placed on their own island and are more likely to share without the disciplining role that interactions with ideologically opposed users induces.

It is worth noting that this theorem builds on but strengthens Theorem 2, where the effects of homophily are ambiguous for articles with intermediate reliability ($r \in (\underline{r}, \bar{r})$). In contrast, Theorem 3 gives a sharp characterization of the platform's algorithm depending simply on whether r is above or below the threshold value r_P.

We also note 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 $r \geq r_P$. Moreover, due to same effects as in Proposition 1, as message divisiveness and belief polarization increase, the platform r_P rises and selects more aggressive filter bubbles.

Remark—This last 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. For example, for higher values of ε ($\varepsilon > \bar{\varepsilon}$), balancing the discipline and circulation effects may require the platform to set $p_d \in (0, 1)$. However, the platform still typically induces echo chamber-like environments to generate the viral spread of low-reliability content ($p_d < p_s$).

5. REGULATION

Our analysis so far raises the question of what types of regulations might counter user engagement with misinformation and platform choices leading to excessive ideological homophily. We

now briefly discuss four distinct types of regulations that have been suggested 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 public policy.

We assume that the regulator's welfare objective is related to misinformation and learning, and to analyse this problem formally, we consider the updating of users' beliefs about the state θ. Specifically, we suppose that users who receive the platform's article update their prior beliefs b_i about θ to $\hat{b}_i$ using Bayes's rule. Users who do not receive the article allocate their attention off social media, to an offline content providing an i.i.d. binary signal $s_i \in \{L, R\}$ for agent i, whereby $s_i = \theta$ with probability $z \in (1/2, 1)$. Agents who engage with the offline content similarly update their beliefs b_i using Bayes's rule.[17] Given these beliefs, we take the exact welfare metric to be to minimize the expected average distance of agents' posteriors from the true state, $-\frac{1}{N}\sum_{i=1}^{N} |\hat{b}_i - \mathbf{1}_{\theta=R}|$. In evaluating this expectation, we assume that the regulator is *ex ante* uncertain about and has a uniform prior over the true state θ. We say a policy is *more effective* than another policy (or no policy) if it improves welfare, and is *most* effective if it is more effective than any other feasible policy. While the platform is interested in maximizing user engagement, as defined in Section 4, we demonstrate in Appendix A that the regulator's welfare measure depends on *content virality* or "total reach," defined as the fraction of agents in the network who receive the article.

Throughout, we fix the article's reliability r and signal probabilities (p, q), with $q = 1/2$, which implies that misinformation is just noise. This assumption is adopted for convenience and can be relaxed, so long as an article containing misinformation does not provide a strong signal arguing for the opposite of its message. Under this assumption, we obtain the following intermediate result.

Lemma 2. *There exists $r_{\mathcal{R}} > 0$ such that:*

(i) if $r < r_{\mathcal{R}}$, then welfare decreases whenever content virality increases;
(ii) if $r > r_{\mathcal{R}}$, then welfare increases whenever content virality increases.

Lemma 2 forms the backbone of our public policy analysis. When low-reliability content ($r < r_{\mathcal{R}}$) circulates on the platform, welfare decreases because agents' attention is drawn to this content, which provides them little useful information. Hence, in this case there is a major divergence between the objectives of the regulator and the platform. On the other hand, with higher-reliability content ($r > r_{\mathcal{R}}$), the regulator's incentives are aligned with the platform's, because individuals obtain useful information from the article on the platform. This observation implies that the main goal of the regulator will be to correct platform incentives to promote low-reliability content, especially when it tends to be sensational and associated with filter bubble algorithms.

For the remainder of this section, we assume that the most-sharing equilibrium before the regulation entails some agents sharing and some agents not sharing (*i.e.* $\mathbf{b}^* \neq \mathbf{0}$ and $\mathbf{b}^{**} \neq \mathbf{1}$).

17. We did not introduce this updating problem until now, because it has no impact on user and platform strategies. Instead, the updating process only influences the regulator's welfare metric.

In this context, one subtle issue relates to whether the users draw an inference from the lack of online content. In this section, we assume that they do not, and we show in Supplementary Appendix C that the same results as in this section apply even when they draw such inferences.

This allows 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.

5.1. *Censorship*

We first consider a "content moderation" policy where the regulator can directly censor misinformation that appears on the platform (or alternatively, "tag" such misinformation as in Clayton *et al.*, 2020). 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 piece of misinformation removed with probability δ).[18] 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 2. *There exist* $0 < r^*_{\mathcal{R}} \leq r_{\mathcal{R}}$ *and* $0 < \delta_1 < \delta_2 < \delta_3 < 1$ *such that:*

(a) if $r > r^*_{\mathcal{R}}$*, then* $\delta^* = \delta$ *is the most effective policy;*
(b) if $r < r^*_{\mathcal{R}}$ *and* $\delta \in (0, \delta_1) \cup (\delta_3, 1)$*, then* $\delta^* = \delta$ *is the most effective policy;*
(c) if $r < r^*_{\mathcal{R}}$ *and* $\delta \in (\delta_1, \delta_2)$*, 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: users believe, correctly, that articles are more likely to be truthful when there is censorship of misinformation (as empirically documented in Pennycook *et al.*, 2020). 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), we are in the regime of Lemma 2(ii), where the incentives of all players (users, the platform, and the regulator) are aligned: the implied truth effect increases sharing, which increases the reach of the platform's sharing network, which necessarily improves welfare. When content is high-reliability and unlikely to contain misinformation, more censorship is a more effective policy. This may at first appear paradoxical: more aggressive censorship policies are only guaranteed to be welfare-improving when the article is already sufficiently likely to be truthful.

However, with low-reliability content likely to contain misinformation, instead we lie in the regime of Lemma 2(i), where there are non-monotone welfare implications from additional censorship. In Part (b), the 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 remove 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.

18. 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.

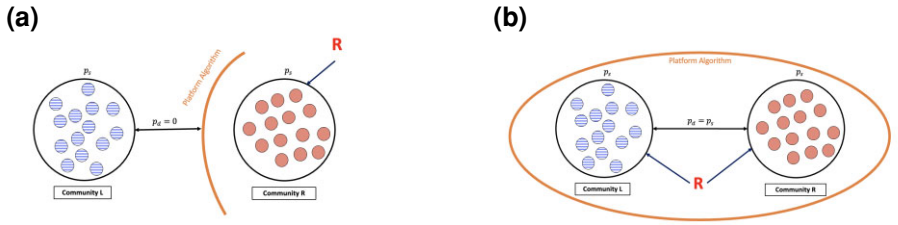

FIGURE 2
Optimal platform sharing networks for Example 1 under Censorship Policies (a) Little censorship and (b) more censorship

Notes: An intermediate censorship policy generates an "implied truth" effect for misinformation that remains undetected, making all agents more likely to share. In response, the platform increases the connectivity of the network. This increases the virality of content that is likely to contain misinformation and, perhaps counterintuitively, negatively affects welfare despite some of the misinformation being removed.

Example 1. There are two islands, one containing $N/2$ left-wing agents with belief $b_L = 1/4$, and the other $N/2$ right-wing agents with belief $b_R = 3/4$. Let us consider an article with a reliability score indicating it is equally likely to contain misinformation or to be truthful (ϕ is the identity and $r = 1/3$), but where truthful content is perfectly informative about the state θ ($p = 1$). We assume the offline content is only partially informative of θ ($z = 4/5$), which admits $0 < r^*_{\mathcal{R}} = 2/5 < r_{\mathcal{R}} = 3/5$. This puts us in regime (i) of Lemma 2 or Parts (b) and (c) of Proposition 2, where content virality reduces welfare.

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 $m = R$. We assume $\tilde{u} = 1$ and $\tilde{c} = 2/3$.

With no censorship policy ($\delta = 0$), the optimal platform sharing network is given by Figure 2a, 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 ($\delta < 1/4$), the optimal platform sharing network remains the same as in Figure 2a. However, the virality of content declines to $(1 - \delta)N/2 < N/2$, and thus welfare improves.
(2) *Intermediate censorship*: With a more aggressive censorship policy ($1/4 < \delta < 1/2$), the optimal platform sharing network switches to the one shown in Figure 2b (maximal connectivity), but still does not remove most of the misinformation. In response, the platform selects a more expansive sharing network because, in the presence of censorship, users correctly believe that any given content is less likely to contain misinformation. In particular, in this case, the platform moves the network from maximal homophily to maximal connectivity. The resulting virality of misinformation then becomes $(1 - \delta)N$, which is greater than $N/2$ for all $1/4 < \delta < 1/2$. In this range, censorship is worse than no censorship policy at all.
(3) *Highly effective censorship*: With a censorship policy that can detect most misinformation ($1/2 < \delta < 1$), the policy reduces misinformation, even though the platform again adjusts its algorithms in response to censorship, increasing the virality of undetected misinformation. □

5.2. *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 users to a peer-reviewed medical study or the full discourse from which a quote was pulled. Such a policy allows users to "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 (independent) probability up to $\rho \in (0, 1)$; a truthful article is never misidentified as misinformation. Hence, the platform can set $\rho^* \leq \rho$ with more impactful provenance policies allowing for a greater fraction of users to identify misinformation.

Proposition 3. *There exist* $0 < \tilde{r}_{\mathcal{R}} \leq r_{\mathcal{R}}$ *and* $0 < \rho_1 < \rho_2 < \rho_3 \leq \delta_3 < 1$ *such that:*

(a) if $r > \tilde{r}_{\mathcal{R}}$*, then* $\rho^* = \rho$ *is the most effective policy;*
(b) if $r < \tilde{r}_{\mathcal{R}}$ *and* $\rho \in (0, \rho_1) \cup (\rho_3, 1)$*, then* $\rho^* = \rho$ *is the most effective policy;*
(c) if $r < \tilde{r}_{\mathcal{R}}$ *and* $\rho \in (\rho_1, \rho_2)$*, then* $\rho^* < \rho$ *is the most effective policy.*

Moreover, for every $\delta \in (\delta_3, 1)$*, a censorship policy with technology* δ *is less effective than a provenance policy with technology* $\rho = \delta$*.*

The result is similar to Proposition 2: soft and strong provenance policies are always effective, but moderate provenance policies can exacerbate the spread of misinformation. In this case, soft provenance policies are closely related to accuracy nudging interventions, where users are prompted to think carefully about the accuracy of content before sharing. See Pennycook *et al.* (2021, 2020) on such policies and their impact on the spread of misinformation.

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 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. Consider the setting of Example 1, where the profit-maximizing sharing network for the platform with no provenance policy is again given by Figure 3a, and a censorship policy of $\delta = 3/16$ is more effective than no policy because $\delta < 1/4$. Note that $r < \tilde{r}_{\mathcal{R}} < r_{\mathcal{R}}$, so greater content virality reduces welfare.

Now consider a provenance policy with $\rho = 3/16 \in (\rho_1, \rho_2)$. Then, the following sharing network increases engagement relative to the network in Figure 3a: Agent 1 is connected to Agent 2, who is connected to a clique of the other $N - 2$ agents, as shown in Figure 3b. This sharing network is not in the class of island networks we have focused on so far. In terms of Proposition 3, when $\rho \in (0, \rho_1) \cup (\rho_3, 1)$, the platform's choices lie within the class of island networks, but not necessarily when we are outside of this range.

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 3b will share the article, because they know that it has been fact-checked twice. Expected user engagement with this article in this case is $(1 - \rho) + (1 - \rho)^2 +$

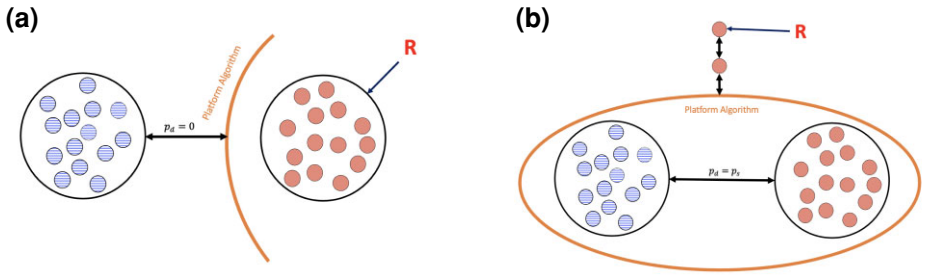

FIGURE 3
Optimal platform sharing networks for Example 2 under provenance policies (a) Little provenance and (b) more provenance

Notes: Intermediate levels of provenance policy make users detect some misinformation but also makes them more likely to share undetected misinformation. In response, the platform increases the level of connectivity of the network, contributing to more viral content likely to contain misinformation. (a) Little provenance and (b) more provenance

$(1-\rho)^3(N-2) > N/2$, 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, analogously to the literature on informational cascades (*e.g.* Banerjee, 1992; Bikhchandani *et al.*, 2021; Chen and Papanastasiou, 2021), they do not make independent judgments based on the reliability of the content and "follow the herd."

5.3. *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. A performance target transfers the responsibility of content moderation to the platform and simultaneously gives it some leeway to have "bad" content on their site, so long as this is below the specified target. In essence, the platform can discard content identified as misinformation by not recommending it to any user, and in doing so, forfeits any potential engagement this content may have had with the users. But the ability of the regulator to monitor the platform's content moderation is imperfect.

Specifically, 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.[19] 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 that the platform's unconstrained profit-maximizing sharing network does not satisfy this performance target.

19. 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/).

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

Proposition 4. *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 improves welfare by reducing 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 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 adequate auditing 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 higher 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.

5.4. *Network regulations*

As we saw in Lemma 1, when unregulated, the platform chooses an island model with parameters (p_s, p_d). Here, we consider limits on the ideological homophily induced by the platform's algorithm. Suppose the regulator can set a homophily standard p^*, based on the ratio between within-island links to across-island links, forcing the platform to choose $p_s/p_d \leq p^*$.[20] When $r > r_{\mathcal{R}}$, the regulator and platform have aligned incentives in the choice of sharing network, so the regulator should abstain from imposing any regulation. However, for less reliable articles, we establish the following result.

Proposition 5. *Suppose that* $r < r_{\mathcal{R}}$*. There exists* $\gamma < \infty$ *such that for any* $p^* \geq \gamma$*, if the regulator imposes a homophily standard* p^**, then the platform's sharing network becomes the island model with* $p_s/p_d \leq p^*$*, and welfare improves.*

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 prevent content being curated specifically to the ideology of a specific group of users. In our model, such standards break up echo chambers and ensure that ideologically diverse users interact more often, which limits the spread of misinformation. However, Proposition 5 also highlights that the standards need to be well-calibrated to the types of ideological divides in the user community and the degree of reliability of the relevant content.

20. There are several suggestions in the literature on how this class of policies can be implemented in practice. See, for example, Yeung and Lodge (2019), Yeung (2018), and Cen and Shah (2021).

6. CONCLUSION

This article develops 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 state. Upon receiving an article, an agent can decide to share it with others, ignore it, or 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. Individuals are more likely to dissent against articles that disagree with their prior beliefs, so an agent will be more cautious in sharing articles that disagree with the views of those in her sharing network.

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 echo chambers.

We also study regulations aimed at improving the welfare of platform users. Content moderation, for example censoring low-reliability articles, can remove some misinformation. However, it also creates a (Bayesian) "false sense of security" and make agents more confident in the quality of remaining items. Similarly, revealing the provenance of a news item (*e.g.* 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 targets that require platforms to remove a certain fraction of posts with misinformation are generally effective, but can backfire when demanding targets induce the platform to deviate from the targets, 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 a powerful tool against filter bubbles and misinformation, but 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 complementarities and how the degree of homophily alters agents' strategic behaviour. Although various behavioural biases and psychological factors appear to be important in social media behaviour, we believe that the economic forces we have identified in this article will continue to apply in the presence of

most of these effects, and our Bayesian benchmark enables 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 could be used to study how agents update their initial political views over time. 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.

APPENDIX

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:

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

By the law of total probability, we have:

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

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 $U_i^{(1)} = u\mathbf{1}_{v=\mathcal{T}} - c\mathbf{1}_{v=\mathcal{M}}$ 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 $(b_i^*, b_i^{**}) = (b', b')$.

□

An immediate consequence of Lemma A.1 is that any Bayesian–Nash equilibrium must be in cutoff strategies. 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*. 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.

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 strategies at the cutoffs are immaterial; 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, $\hat{b}_i^{*,BR} = \tilde{b} \geq b_i^{**,BR} = b_i^{*,BR}$.

(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:

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

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). □

Lemma A.5. *All equilibria are semi-symmetric in an island network. In other words, for every equilibrium, there exist $\{(b_\ell^*, b_\ell^{**})\}_{\ell=1}^k$ such that $b_i^* = b_{\ell_i}^*$ and $b_i^{**} = b_{\ell_i}^{**}$ i in island ℓ.*

Proof of Lemma A.5. 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, $b_i^{**}=b_j^{**}$. □

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 $\geq$ (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 $\mathbf{b}^{**}$). This completes the proof of (i)–(iii) for Theorem 1.

For the comparative statics, 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.

We next prove that the social media game is supermodular and has increasing differences in reliability. Note that for all $r'\geq r$:

$$\begin{aligned}\big[U_i(\mathcal{S},r')-U_i(\mathcal{I},r')\big]-\big[U_i(\mathcal{S},r)-U_i(\mathcal{I},r)\big]&=U_i(\mathcal{S},r')-U_i(\mathcal{S},r)\\&=U_i^{(1)}(r')-U_i^{(1)}(r)=(u+c)(\pi_i(r')-\pi_i(r)),\end{aligned}$$

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

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

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

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

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 $d' \leq d$:

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

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. We prove the stronger claim, which is the first half of Theorem 3 (excluding the comparative statics). First, let us define $\tilde{\sigma}_{\mathbf{P}}$ to be the most-sharing equilibrium under a given sharing network $\mathbf{P}$, and $\mathbb{P}^{\varepsilon}_{\tilde{\sigma}_{\mathbf{P}}}$ to be the probability measure over action set $\{\mathcal{D}, \mathcal{I}, \mathcal{S}\}$, given microtargeting technology $\varepsilon > 0$ and the most-sharing equilibrium induced on network $\mathbf{P}$. The proof is constructed in five parts:

(i) Consider the sets $\mathcal{A}_\varepsilon = \{i : \exists \mathbf{P}' \text{s.t. } \mathbb{P}^{\varepsilon}_{\tilde{\sigma}_{\mathbf{P}'}}[a_i = \mathcal{S}] > 0\}$ (the set of agents who share with positive probability under some sharing network) and $\mathcal{A}^c_\varepsilon = \{i : \forall \mathbf{P}', \mathbb{P}^{\varepsilon}_{\tilde{\sigma}_{\mathbf{P}'}}[a_i = \mathcal{S}] = 0\}$ (the set of agents who share with probability 0 under all sharing networks), with $\mathcal{A}_\varepsilon \cup \mathcal{A}^c_\varepsilon$ spanning the entire set of agents in the network.[21] We show there exists $\bar{\varepsilon} > 0$ such that for all $\varepsilon < \bar{\varepsilon}$, the sharing network $\mathbf{P}^*_\varepsilon$ constructed by setting $p_{ij} = 1$ for all $i, j \in \mathcal{A}_\varepsilon$ and $p_{ij} = 0$ otherwise, satisfies $\mathbb{P}^{\varepsilon}_{\tilde{\sigma}_{\mathbf{P}^*_\varepsilon}}[a_i = \mathcal{S}] \cdot \mathbb{P}^{\varepsilon}_{\tilde{\sigma}_{\mathbf{P}^*_\varepsilon}}[a_i = \mathcal{D}] = 0$ for all agents $i \in \mathcal{A}_\varepsilon$. For Parts (ii)–(v), we will fix some $\varepsilon < \bar{\varepsilon}$ and suppress the dependence on ε for simplicity of notation.

(ii) Let us define the complete sharing network as $\mathbf{P}^\circ \equiv \mathbf{1}_{N\times N} - \mathbf{I}$. We show that if $\mathbb{P}_{\tilde{\sigma}_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] = 0$ for all agents i, then user engagement is highest under $\mathbf{P}^\circ$, *i.e.* $\max_{i^*} \mathbb{E}_{\tilde{\sigma}_{\mathbf{P}^\circ}}[\mathbf{S}_{i^*}] \geq \max_{i^*} \mathbb{E}_{\tilde{\sigma}_{\mathbf{P}'}}[\mathbf{S}_{i^*}]$ (where $\mathbf{S}_{i^*}$ is the number of shares conditional on seed i^*), for any other arbitrary sharing network $\mathbf{P}'$. In other words, $\mathbf{P}^\circ$ is the profit-maximizing sharing network.

(iii) We show that if $\mathbb{P}_{\tilde{\sigma}_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] > 0$ for some agent i, then $\mathcal{A}^c$ is non-empty.

(iv) We show that when $\mathcal{A}^c$ is non-empty, then the platform's profit-maximizing sharing network is $\mathbf{P}^*$, which sets $p_{ij} = 1$ for all $i, j \in \mathcal{A}$ and $p_{ij} = 0$ otherwise.

(v) We show there exists some $r_P \in (0, 1)$ such that $\mathbb{P}_{\tilde{\sigma}_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] = 0$ for all i if and only if $r > r_P$. Moreover, a profit-maximizing sharing network takes the form of Theorem 3(ii) in Case (ii) described above when $r > r_P$, and the form of Theorem 3(i) in Case (iv) described above when $r < r_P$, completing the proof.

□

Proof of Claim (i): Let us partition the agents into the set of "active agents," $\mathcal{A}_\varepsilon = \{i : \exists \mathbf{P}' \text{s.t. } \mathbb{P}^{\varepsilon}_{\tilde{\sigma}_{\mathbf{P}'}}[a_i = \mathcal{S}] > 0\}$, and the set of "inactive agents," $\mathcal{A}^c_\varepsilon = \{i : \forall \mathbf{P}', \mathbb{P}^{\varepsilon}_{\tilde{\sigma}_{\mathbf{P}'}}[a_i = \mathcal{S}] = 0\}$. As before, we define the sharing network $\mathbf{P}^*_\varepsilon$ as setting $p_{ij} = 1$ for all agents $i, j \in \mathcal{A}$ and setting $p_{ij} = 0$ for all other agents. Note that for any agent $i \in \mathcal{A}$, her

21. Note that these sets are indexed by ε because the probability of an agent's action depends on the platform's targeting technology $\varepsilon > 0$, which allows the platform to assess agent i's prior b_i up to $\pm\varepsilon$.

equilibrium strategy is identical to her strategy in a complete network on $|\mathcal{A}|$ agents, which means by Lemma A.5, that all agents $i \in \mathcal{A}$ employ the same cutoffs $(b_\varepsilon^*, b_\varepsilon^{**})$ in the most-sharing equilibrium under sharing network $\mathbf{P}^*$.

Consider $\varepsilon = 0$, which occurs when the platform technology endows it with perfect knowledge of each agent's b_i. Note that $\mathbb{P}^0_{\bar{\sigma}_{\mathbf{P}_\mathbf{0}^*}}[a_i = \cdot] \in \{0, 1\}$ because each agent's action is known with certainty. There are three cases to consider.

(1) $b_0^* = 0$, so $\mathbb{P}^0_{\bar{\sigma}_{\mathbf{P}_0^*}}[a_i = \mathcal{D}] = 0$ for all i. There are two subcases. If $b_0^{**} = 0$ also, then $\mathcal{A}_0^c = \varnothing$ and because $\mathcal{A}_\varepsilon^c \subset \mathcal{A}^c$ (the inactive set monotonically decreases in ε), we know that $\mathcal{A}_\varepsilon^c = \varnothing$ and $\mathbb{P}^\varepsilon_{\bar{\sigma}_{\mathbf{P}_\varepsilon^*}}[a_i = \mathcal{D}] = 0$ for all i. If $b_0^{**} > 0$, then set $\varepsilon < \bar{\varepsilon} \equiv b_0^{**}/4$. Note that while $\mathcal{A}_\varepsilon$ (and $\mathcal{A}_\varepsilon^c$) may change, no agent with belief $b_i \leq b_0^{**} - \varepsilon$ is in $\mathcal{A}_\varepsilon$, and all agents in $\mathcal{A}_\varepsilon \backslash \mathcal{A}_0$ choose ignore with probability 1, which does not affect the payoffs of any agents. Thus, $b_\varepsilon^* = b_0^*$ and $b_\varepsilon^{**} = b_0^{**}$, and so for all $i \in \mathcal{A}_\varepsilon$, we also obtain that $\mathbb{P}^\varepsilon_{\bar{\sigma}_{\mathbf{P}_\varepsilon^*}}[a_i = \mathcal{D}] = 0$.

(2) $b_0^{**} = 1$, so $\mathbb{P}^0_{\bar{\sigma}_{\mathbf{P}_0^*}}[a_i = \mathcal{S}] = 0$ for all i. Then $\mathcal{A}_0 = \varnothing$. However, this immediately implies that $\mathcal{A}_\varepsilon = \varnothing$ for any $\varepsilon > 0$ as no agent would share in any sharing network $\mathbf{P}'$ (by definition of $\mathcal{A}_\varepsilon$). Therefore, independent of $\varepsilon > 0$, we have for all $\mathbf{P}'$ that $\mathbb{P}^\varepsilon_{\bar{\sigma}_{\mathbf{P}'_\varepsilon}}[a_i = \mathcal{S}] = 0$ for all agents i.

(3) $b_0^*, b_0^{**} \in (0, 1)$. By Lemma A.4(i), this implies that $b_0^* < b_0^{**}$. Take $\varepsilon < \bar{\varepsilon} \equiv (b_0^{**} - b_0^*)/4$. Identical to Case (1), no agents with $b_i \leq b_0^{**} - \varepsilon$ will be in $\mathcal{A}_\varepsilon$, and all agents in $\mathcal{A}_\varepsilon \backslash \mathcal{A}$ will ignore, implying that $b_\varepsilon^* = b_0^*$ and $b_\varepsilon^{**} = b_0^{**}$, so for all $i \in \mathcal{A}_\varepsilon$, we obtain $\mathbb{P}^\varepsilon_{\bar{\sigma}_{\mathbf{P}_\varepsilon^*}}[a_i = \mathcal{D}] = 0$.

Taking $\bar{\varepsilon}$ for the appropriate case (or $\bar{\varepsilon} = 1$ in Case (2)), we see that for all $\varepsilon < \bar{\varepsilon}$, we have $\mathbb{P}^\varepsilon_{\bar{\sigma}_{\mathbf{P}_\varepsilon^*}}[a_i = \mathcal{S}] \cdot \mathbb{P}^\varepsilon_{\bar{\sigma}_{\mathbf{P}_\varepsilon^*}}[a_i = \mathcal{D}] = 0$ for all agents $i \in \mathcal{A}_\varepsilon$, establishing the claim.

Proof of Claim (ii): Suppose under the complete sharing network $\mathbf{P}^\circ$, we have $\mathbb{P}_{\bar{\sigma}_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] = 0$ for all agents i. We focus on a modified social media game that only allows agents to ignore ($\mathcal{I}$) or share ($\mathcal{S}$). Notice this necessarily (weakly) increases user engagement for any sharing network $\mathbf{P}'$—because $U_i(\mathcal{S})$ weakly increases for all agents i with no dislike action, Topkis's theorem guarantees a higher probability of sharing for all agents in equilibrium, which implies an increase in user engagement. At the same time, this does not increase user engagement for the complete sharing network $\mathbf{P}^\circ$, because no agent dislikes in the equilibrium of the original game (by assumption), so would choose the same actions (either $\mathcal{I}$ or $\mathcal{S}$) in the modified game. Thus, a sufficient condition to prove Claim (i) is to show that $\mathbf{P}^\circ$ maximizes user engagement among all other networks $\mathbf{P}'$ for modified game.

Let $\hat{\mathbf{B}} \subset [0, 1]^{2N} = \{(\mathbf{b}', \mathbf{b}^\circ) \,|\, \mathbf{b}^\circ \preceq \mathbf{b}'\}$ denote the partial cutoff space where the second dimension has lower cutoffs (uniformly more sharing) than the first dimension (in the component-wise lattice order). We will define the map $\psi : \hat{\mathbf{B}} \to \hat{\mathbf{B}}$ to be the one that maps an arbitrary fixed set of cutoff strategies $(\mathbf{b}', \mathbf{b}^\circ)$ to best-response cutoff strategies $(\mathbf{b}^{BR'}, \mathbf{b}^{BR\circ})$, given the current strategy being played is $(\mathbf{b}', \mathbf{b}^\circ)$ under the sharing networks $(\mathbf{P}', \mathbf{P}^\circ)$, respectively.

First, we show this map is well-defined. To do this, we just need to guarantee that indeed $\mathbf{b}^\circ \preceq \mathbf{b}' \Longrightarrow \mathbf{b}^{BR\circ} \preceq \mathbf{b}^{BR'}$ under the mapping ψ. This is immediate by letting $U_j^\circ(\mathcal{S})$ be the utility for agent j from sharing under the complete network and $U_j'(\mathcal{S})$ the utility from sharing under $\mathbf{P}'$. Then,

$$U_j^\circ(\mathcal{S}) - U_j'(\mathcal{S}) = \kappa \sum_{\tilde{j} \neq j} p'_{j\tilde{j}} \left(H\left(b_{\tilde{j}}^{**'}\right) - H\left(b_{\tilde{j}}^{**\circ}\right)\right) + \left(1 - p'_{j\tilde{j}}\right)\left(1 - H\left(b_{\tilde{j}}^{**\circ}\right)\right) \geq 0.$$

Thus, an agent shares under $\mathbf{P}'$ (and $\mathbf{b}'$) only if she shares under $\mathbf{P}^\circ$ (and $\mathbf{b}^\circ$) and indeed $\mathbf{b}^{BR\circ} \preceq \mathbf{b}^{BR'}$ must hold. This shows ψ is well-defined.

Consequently, by Tarski's fixed point theorem (leveraging Lemma A.2), the set of fixed points of ψ form a lattice in $\hat{\mathbf{B}}$. Moreover, we note that every fixed point of ψ represents one Bayesian–Nash equilibrium (BNE) on sharing network $\mathbf{P}^\circ$ and one BNE on sharing network $\mathbf{P}'$. Moreover, for every BNE, $\mathbf{b}^{**'}$, on sharing network $\mathbf{P}'$, there is a fixed point of ψ of the form $(\mathbf{b}^{**\circ}, \mathbf{b}^{**'})$, where $\mathbf{b}^{**\circ}$ is simultaneously a BNE of $\mathbf{P}^\circ$. In particular, note we can pick the most-sharing BNE $\mathbf{b}^{**'}$ for network $\mathbf{P}'$, and because ψ maps into $\hat{\mathbf{B}}$, we must have that $\mathbf{b}^{**\circ} \preceq \mathbf{b}^{**'}$ for the most-sharing equilibrium in $\mathbf{P}^\circ$ relative to the most-sharing equilibrium in $\mathbf{P}'$.

Finally, it is easy to see this implies that $\mathbf{P}^\circ$ induces more user engagement than $\mathbf{P}'$ given equilibrium strategies satisfy $\mathbf{b}^{**\circ} \preceq \mathbf{b}^{**'}$ in the most-sharing equilibrium. For every prior realization and seed agent i^*, $\mathbf{S}_{i^*}$ is larger in the complete sharing network $\mathbf{P}^\circ$ than in any other sharing network $\mathbf{P}'$ (and thus in expectation it is also higher). This can be seen by considering every possible diffusion path from seed agent i^* through intermediate agents $j_1, j_2, \ldots, j_z$ eventually reaching target agent j. The probability of a share (and a link) between j_s and j_{s+1} is $1 - H(b_{j_s}^{**\circ})$ in

$\mathbf{P}^\circ$ and $p_{j_s j_{s+1}}(1 - H(b^{**'}_{j_s}))$ in $\mathbf{P}'$, which is less for every link. Thus, $\mathbf{P}^\circ$ maximizes user engagement (and is the profit-maximizing sharing network).

Proof of Claim (iii): Next, we consider the case where the complete sharing network $\mathbf{P}^\circ$ induces most-sharing equilibrium cutoffs (b^*, b^{**}), but where $b^* > 0$ (by our assumption that there exists an agent i such that $\mathbb{P}_{\bar{\sigma}_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] > 0$). To prove that $\mathcal{A}^c \neq \varnothing$, we show that there must exist an open interval $(0, \bar{b})$ with $\bar{b} > 0$ where agents with priors $b_i \in (0, \bar{b})$ will *never* share, regardless of the sharing network $\mathbf{P}'$. To show this, we argue by contradiction that there cannot exist a sharing network $\mathbf{P}'$ where *all* agents either share or ignore, *i.e.* where the equilibrium cutoffs are determined solely by $\mathbf{b}^{**'}$ (and $\mathbf{b}^{*'} = \mathbf{0}$). Because the platform's choice set of sharing networks is compact, this guarantees that $\mathbf{b}^*$ is bounded away from $\mathbf{0}$ under any selection of $\mathbf{P} \in [0, 1]^{N \times N}$.

Let us consider the same mapping ψ from the proof of Claim (ii); note that in general, the fixed points of this map are not equilibria, because we are restricting $\mathbf{b}^* = \mathbf{0}$, but action $\mathcal{D}$ may be a best response for some agents under a given fixed point $(\mathbf{0}, \mathbf{b}^{**})$ of ψ. However, we know the most-sharing fixed point (lowest $\mathbf{b}^{**'}$) *is*, in fact, an equilibrium in sharing network $\mathbf{P}'$, by assumption. We can also show that the most-sharing fixed point from ψ for sharing network $\mathbf{P}^\circ$ is an equilibrium using Topkis's theorem: under strategy profile $(\mathbf{0}, \mathbf{b}^{**'})$, we have $U^\circ_j(\mathcal{I}) - U'_j(\mathcal{I}) = U^\circ_j(\mathcal{D}) - U'_j(\mathcal{D}) = 0$, but $U^\circ_j(\mathcal{S}) - U'_j(\mathcal{S}) = \kappa \sum_{\tilde{j} \neq j}(1 - p_{j\tilde{j}})(1 - H(b^{**'}_{\tilde{j}})) \geq 0$. Therefore, in the most-sharing equilibrium, we must have $\mathbf{b}^{*\circ} \preceq \mathbf{0}$, implying $\mathbf{b}^{*\circ} = \mathbf{0}$. Finally, as with Claim (i), because ψ maps into $\hat{\mathbf{B}}$, the most-sharing equilibrium under $\mathbf{P}^\circ$ involves $\mathbf{b}^{**\circ} \preceq \mathbf{b}^{**'}$. This yields a contradiction, so in fact $\mathcal{A}^c \neq \varnothing$.

Proof of Claim (iv): Let $\mathcal{A}^c \neq \varnothing$ and consider the sharing network $\mathbf{P}^*$ that sets $p_{ij} = 1$ for all $i, j \in \mathcal{A}$ and sets $p_{ij} = 0$ otherwise. We can, without loss of generality, assume that $\mathcal{A} \neq \varnothing$; otherwise no agent shares under any sharing network, and user engagement is zero for all $\mathbf{P}'$ (including $\mathbf{P}^*$), making it a trivial optimum (which is non-unique) for profit maximization. Note that by construction, agents in $\mathcal{A}^c$ share under no sharing network $\mathbf{P}'$, so fully disconnecting them from the seed agent $i^* \in \mathcal{A}$ cannot increase user engagement under any other arbitrary sharing network $\mathbf{P}'$. At the same time, if all agents $i \in \mathcal{A}$ never choose dislike, *i.e.* $\mathbb{P}_{\bar{\sigma}_{\mathbf{P}^*}}[a_i = \mathcal{D}] = 0$, it must be true that setting $p_{ij} = 1$ for all i, j is profit-maximizing because of Claim (ii). Thus, we can assume that $\mathbb{P}_{\bar{\sigma}_{\mathbf{P}^*}}[a_{i^*} = \mathcal{D}] > 0$ for some agent $i^* \in \mathcal{A}$. We will derive a contradiction.

By definition of $\mathcal{A}$, there must exist some sharing network $\tilde{\mathbf{P}}$ where $\mathbb{P}_{\bar{\sigma}_{\tilde{\mathbf{P}}}}[a_{i^*} = \mathcal{S}] > 0$ (and it is without loss to assume $\tilde{\mathbf{P}}$ is defined only on agents in $\mathcal{A}$). Suppose we amend the payoff of agents such that $d_{i^*} = 0$ for agent i^* (*i.e.* dislikes do not penalize i^*'s utility). From Topkis's theorem, it must be the case that $\mathbb{P}_{\bar{\sigma}_{\tilde{\mathbf{P}}}}[a_{i^*} = \mathcal{S}] > 0$. Moreover, sharing utility is only monotonically increasing in i^*'s neighbourhood, so under this utility transformation, it must necessarily be the case that $\mathbb{P}_{\bar{\sigma}_{\mathbf{P}^*}}[a_{i^*} = \mathcal{S}] > 0$. Applying Claim (i) again, this means that given $d_{i^*} = 0$, $\mathbb{P}_{\bar{\sigma}_{\mathbf{P}^*}}[a_{i^*} = \mathcal{D}] = 0$.

We repeat this process for all such i^* of the original game on $\mathbf{P}^*$. This yields a network $\mathbf{P}^*$ of a modified game where all agents elect $\mathbb{P}_{\bar{\sigma}_{\mathbf{P}^*}}[a_{i^*} = \mathcal{D}] = 0$. This is an equilibrium of the original game, because in equilibrium, no agent dislikes with positive probability. In fact, the most-sharing equilibrium must have $(\mathbf{b}^*, \mathbf{b}^{**}) \preceq (\tilde{\mathbf{b}}^*, \tilde{\mathbf{b}}^{**})$ for any other equilibrium $(\tilde{\mathbf{b}}^*, \tilde{\mathbf{b}}^{**})$. Because an equilibrium exists where $\tilde{\mathbf{b}}^* = \mathbf{0}$, this must be the same for the most-sharing equilibrium. This is a contradiction, and so in fact $\mathbf{P}^*$ is the profit-maximizing sharing network.

Proof of Claim (v): When $r = 1$, an all-share equilibrium exists because the article contains misinformation with probability 0, so in particular, $\mathbb{P}_{\sigma_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] = 0$, for all i. When $r = 0$, Lemma A.4 guarantees that there exists some agent i with $\mathbb{P}_{\sigma_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] > 0$. By the monotone comparative statics of Theorem 1, sharing is monotone in reliability, as is the payoff to disliking, so by continuity (and IVT) there exists some cutoff $r_P \in (0, 1)$ such that for $r > r_P$, the complete sharing network admits only shares and ignores (*i.e.* satisfies $\mathbb{P}_{\sigma_{\mathbf{P}^\circ}}[a_i = \mathcal{D}] = 0, \forall i$), whereas when $r < r_P$, some agent dislikes with positive probability.

When $r > r_P$, the network has maximal connectivity as in Theorem 3(ii) with $p_s = p_d = 1$. When$r < r_P$, we can form a network with maximal homophily and $(p_s, p_d) = (1, 0)$, which has the same user engagement as $\mathbf{P}^*$. For this, we can construct two islands, one of which consists of agents in $\mathcal{A}$ and one of which consists of agents in $\mathcal{A}^c \neq \varnothing$ (by Claim (iii)). Within island links are $p_s = 1$, but the network is fully disconnected between islands, with $p_d = 0$. This is the same sharing network as in $\mathbf{P}^*$ for agents in $\mathcal{A}$, and because the article reaches agents in $\mathcal{A}^c$ with probability 0, the user engagement is the same. □

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 on every island 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 Lemma A.5) 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 lattice, 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 on island 1, ζ. 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, $\partial U_1(\mathcal{S})/\partial p_s = \kappa N_1 (1 - H(\hat{b}_1^{**})) + d N_1 (1 - H(\hat{b}^*)) > 0$. 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 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 user engagement, 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 user engagement 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 user engagement increases following the increase in homophily, because engagement 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 user engagement 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 $\bar{\pi} = \max\{\frac{c}{u+c}, \frac{\tilde{u}-\tilde{c}}{\tilde{u}}\} < 1$, we note that whenever $r \geq \phi^{-1}(\frac{(1-q)\bar{\pi}}{(p-q)\bar{\pi}+(1-p)}) \equiv \bar{r} \in (0,1)$, $\pi_i \geq \bar{\pi}$. 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 $r > \bar{r}$, user engagement is 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, engagement 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 1. 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 total user engagement. 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 $(\hat{b}_R^{*,BR}, \hat{b}_R^{**,BR})$. 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 $U_R^{(2)}(\mathcal{S}) = U_R^{(2)'}(\mathcal{S})$ 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 $U_R^{(1)}(\mathcal{S})$ and $U_R^{(1)'}(\mathcal{S})$. Specifically, the difference in share payoff depends only on the change in π_i following the increase in divisiveness or polarization. Moreover,

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

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 $U_R^{(1)'}(\mathcal{S}) > U_R^{(1)}(\mathcal{S})$, 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, user engagement 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 engagement, naturally). □

Proof of Theorem 3. Given the proof of Lemma 1, all that remains is to prove the comparative statics in the second half of Theorem 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. Therefore, $U_i(\mathcal{D})$ 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 1), π_i decreases for this agent (making $a_i = \mathcal{D}$ a best response). In both cases, we see that r_P increases. □

A.4. *Proofs from Section 5*

Proof of Lemma 2. Let us call the "effective signal strength" of an article with reliability r, $\zeta \equiv p\phi(r) + (1-\phi(r))/2$ (corresponding to the probability that the message argues for θ). Note that the offline content has an effective signal strength of z. We show that welfare improves when the effective signal strength of the article increases. The claim of Lemma 2 then follows from noting the effective signal strength is monotonically increasing in r, so there is either (i)

a unique crossing point $r_{\mathcal{R}} \in (0,1)$ where $p\phi(r_{\mathcal{R}}) + (1-\phi(r_{\mathcal{R}}))/2 = z$, or (ii) $p\phi(r) + (1-\phi(r))/2 < z$ for all $r \in (0,1)$, in which case the result holds by setting $r_{\mathcal{R}} = 1$.

A straightforward application of Bayes's rule shows that when agent i engages with the platform's article, then:

$$\hat{b}_i = \begin{cases} \pi_i \dfrac{pb_i}{pb_i + (1-p)(1-b_i)} + (1-\pi_i)b_i, & \text{if } m = R \\ \pi_i \dfrac{(1-p)b_i}{(1-p)b_i + p(1-b_i)} + (1-\pi_i)b_i, & \text{if } m = L. \end{cases} \tag{3}$$

When $\theta = R$, then

$$\hat{b}_i - b_i = \zeta\left(\pi_i \frac{pb_i}{pb_i + (1-p)(1-b_i)} + (1-\pi_i)b_i\right) + (1-\zeta)\left(\pi_i \frac{(1-p)b_i}{(1-p)b_i + p(1-b_i)} + (1-\pi_i)b_i\right) - b_i$$

and when $\theta = L$, then

$$b_i - \hat{b}_i = b_i - (1-\zeta)\left(\pi_i \frac{pb_i}{pb_i + (1-p)(1-b_i)} + (1-\pi_i)b_i\right) - \zeta\left(\pi_i \frac{(1-p)b_i}{(1-p)b_i + p(1-b_i)} + (1-\pi_i)b_i\right)$$

Summing these (and dropping the platform's prior of $1/2$ for each of the two possible states), we see that welfare $\mathcal{W}_i$ of agent i is proportional to:

$$\begin{aligned} \mathcal{W}_i \propto\ & \zeta\left(\frac{pb_i}{pb_i + (1-p)(1-b_i)} - \frac{(1-p)b_i}{(1-p)b_i + p(1-b_i)}\right) \\ & + (1-\zeta)\left(\frac{(1-p)b_i}{(1-p)b_i + p(1-b_i)} - \frac{pb_i}{pb_i + (1-p)(1-b_i)}\right) \end{aligned}$$

Observe that $\frac{pb_i}{pb_i+(1-p)(1-b_i)} > \frac{(1-p)b_i}{(1-p)b_i+p(1-b_i)}$ because $p > 1/2$, and so welfare for agent i is increasing in ζ (because $\mathcal{W}_i$ is the convex combination of a positive and negative term).

Lastly, note that if user engagement increases, via Theorem 3 we must in fact have a stronger relationship when the platform can choose the sharing network $\mathbf{P}$: if $\mathcal{E}$ is the set of engaged agents (those who share) before and $\mathcal{E}'$ is the set of engaged agents after, then $\mathcal{E} \subset \mathcal{E}'$. Thus, summing (or averaging) over the individual welfare of all agents yields the result. □

Proof of Proposition 2. For Part (a), take the smallest value of r (in the infimum sense) such that $\delta^* = \delta$ is the most effective policy for all $\delta \in [0,1]$ (such a value must exist because $r = 1$ satisfies this property). Call this value $\hat{r}$. If $\hat{r} < r_{\mathcal{R}}$, set $r^*_{\mathcal{R}}$ to $\hat{r}$ (in which case (a) holds by construction). If $\hat{r} > r_{\mathcal{R}}$, then set $r^*_{\mathcal{R}} = r_{\mathcal{R}}$. Note by Theorem 3, virality is always increasing in r, and by Lemma 2 we know that welfare is increasing in virality whenever $r > r_{\mathcal{R}}$. Moreover, because the regulator only makes type-I errors (it only ever removes misinformation which is welfare-reducing relative to the offline content), it is necessarily the case that $\delta^* = \delta$ is the most effective policy whenever $r < r_{\mathcal{R}}$.

For Parts (b) and (c), 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 Lemma 1). We will let $V(\delta)$ denote the virality of content when censorship policy δ is applied, *conditional* on the content not being censored (with total virality including censorship given by $(1-\delta(1-\phi(r)))V(\delta)$).

For Part (b), first consider any agent j residing on island B. Because the platform approximates the belief distribution H by a generic multinomial distribution,[22] 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 $\tilde{\phi}(r,\delta) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))}$ for all $\delta \in (0,\underline{\delta})$ achieves both: (i) leaving the profit-maximizing sharing network for the platform unchanged, because the strict inequality above still holds under any chosen sharing network and (ii) $\phi^{-1}(\tilde{\phi}(r,\delta)) < r_{\mathcal{R}}$,

22. 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 4.

so its virality is welfare-reducing by Lemma 2. The former implies that $V(0) = V(\delta)$ for all $\delta \in (0, \underline{\delta})$, and at the same time, we have that $(1 - \delta(1 - \phi(r)))V(\delta) < V(\delta) = V(0)$ for all δ in this range (because the platform makes no type-II errors). Moreover, the content circulating under censorship policy $\delta \in (0, \bar{\delta})$ has higher "effective" reliability $\tilde{\phi}(r, \delta)$, so is welfare-improving. Therefore, 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 $\tilde{\phi}(r, \delta) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))} = 1$ when $\delta = 1$. It is immediate then that for sufficiently high values of δ, the profit-maximizing sharing network has maximal connectivity, all agents share in equilibrium, and $\phi^{-1}(\tilde{\phi}(r, \delta)) > r_{\mathcal{R}}$. Let us consider $\bar{\delta}$ which is the smallest value (in the infimum sense) such that the profit-maximizing sharing network has maximal connectivity (note that $\bar{\delta} < 1$). We know that $V(\delta) = N$ for all $\delta \in (\bar{\delta}, 1)$ by construction. Moreover, because the platform never makes type-II errors, higher values of δ remove misinformation articles only (which by Lemma 2 is welfare-improving) and increase the "effective" reliability of remaining articles (which by Lemma 2 is also welfare-improving). Moreover, the virality of such articles that satisfy Lemma 2(ii) is maximized at $V(\delta) = N$, so any $\delta^* = \delta$ when $\delta \in (\bar{\delta}, 1)$ is more effective than any other policy with $\delta^* < \delta$.

Finally, for Part (c), let us construct $0 < \delta_1 < \delta_2 < \delta_3 < 1$ when $r < r^*_{\mathcal{R}}$ to conclude. We can assign $\delta_3 = \bar{\delta}$ from the previous paragraph. 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 $\delta \in (0, \delta_1)$. By construction of $r^*_{\mathcal{R}}$, we know that such a δ_1 exists and it is strictly less than $\bar{\delta}$. By the same reasoning as before, because the platform approximates H as a multinomial distribution, there always exists an open interval $(\delta_1, \hat{\delta})$ where the virality of content satisfies $V(\delta) \geq V(\delta_1) + 1/N$ and $\phi^{-1}(\tilde{\phi}(r, \delta)) < r_{\mathcal{R}}$ for all $\delta \in (\delta_1, \hat{\delta})$. At the same time, we know that for all $\delta \in (\delta_1, \min\{\hat{\delta}, (1 - \delta_1 NV(\delta_1)(1 - \phi(r)))/((NV(\delta_1) + 1)(1 - \phi(r)))\})$, we have $(1 - \delta(1 - \phi(r)))V(\delta) \geq (1 - \delta_1(1 - \phi(r)))V(\delta_1)$, so the total virality of content is lower under δ_1 than under any larger $\delta \in (\delta_1, \tilde{\delta})$ where $\tilde{\delta} \equiv \min\{\hat{\delta}, (1 - \delta_1 NV(\delta_1)(1 - \phi(r)))/((NV(\delta_1) + 1)(1 - \phi(r)))\}$. Because there is a discontinuity in the neighbourhood of δ_1 in virality, taking δ_2 to be sufficiently close to δ_1 allows $\tilde{\phi}(r, \delta_1) \approx \tilde{\phi}(r, \delta_2)$ but where $V(\delta_2) \geq V(\delta_1) + 1/N$. By Lemma 2(i), we see that the $\delta^* = \delta_1$ policy is more effective than any $\delta^* \in (\delta_1, \delta_2)$, and so in particular some $\delta^* < \delta$ is most effective. □

Proof of Proposition 3. We take a similar approach as in the proof of Proposition 2. For Part (a), observe by Lemma A.4 that no agent who fact-checks (through the provenance channel) and finds an article to contain misinformation, chooses to share it (regardless of the sharing network). The construction of $\tilde{r}_{\mathcal{R}}$ then follows exactly as in the proof of Proposition 2. We will let $V(\rho)$ denote the virality of content with provenance policy ρ is applied, *conditional* on the content not being fact-checked by *any* agent in the sharing network.

For Parts (b) and (c), once again 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 for all agents j on island B, and the profit-maximizing sharing network will be the same as $\rho = 0$ for all $\rho \in (0, \underline{\rho})$. This implies that $V(0) = V(\rho)$ for all $\rho \in (0, \underline{\rho})$, and the argument then follows identically as in the proof of Proposition 2.

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 (and where $\frac{\phi(r)}{\phi(r)+(1-\rho)(1-\phi(r))} > r_{\mathcal{R}}$). Once again, let us consider $\bar{\rho}$ which is the smallest value (in the infimum sense) such that the profit-maximizing sharing network has maximal connectivity (with $\bar{\rho} < 1$).[23] The rest of the argument follows similarly as in the proof of Proposition 2, because agents never make type-II errors and therefore only choose to not share if the content contains misinformation with probability 1, which is welfare-improving by Lemma 2. The construction of $0 < \rho_1 < \rho_2 < \rho_3 < 1$ then also follows exactly in the same way as from the last paragraph in Proposition 2.

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

23. In contrast to the censorship case, it is now less obvious that such a $\bar{\rho}$ is well-defined because the profit-maximizing sharing network does not necessarily satisfy Theorem 3. However, it is never the case that there exist $\rho' < \rho'' < \rho'''$ where the profit-maximizing sharing network is not maximally connected when ρ' and ρ''' but is maximally connected when ρ''. It is immediate to show that if maximally connected is profit-maximizing for ρ'', it must be profit-maximizing for ρ''' as well.

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 $\delta = \rho$. Under the provenance policy, truthful content has at least weakly higher "effective" reliability ($\tilde{\phi}(r,\cdot)$) so necessarily leads to higher virality, which is welfare-improving by Lemma 2. At the same time, misinformation is detected with probability δ under the censorship policy but detected with probability $1-(1-\rho)^N > \rho$. Thus, $\rho_3 \leq \delta_3$ and $\rho = \delta$ is a more effective policy in this parameter region. □

Proof of Proposition 4. By Lemma 2, the regulator always prefers a lower performance target conditional on the platform actually choosing to abide by it. If the platform removes ψ fraction of misinformation, then its performance metric is given by

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

where $V(\psi)$ is the content virality when the platform removes ψ fraction of misinformation and optimally chooses the sharing network, but notice that $V(\psi)$ 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 $\Pi = V(\psi)\phi(r)/(1-\lambda)$ whereas the payoff from not hitting it is $\Pi' = (1-\alpha)V(\psi^*) - \alpha C$, where ψ^* is the self-imposed target by the platform that maximizes engagement, *i.e.* $\psi^* = \arg\max_\psi V(\psi) < 1$ by nature of $\phi(r)$ not exceeding the platform's profit. For any $\psi < \psi^*$, the platform meets the performance target. For any $\psi > \psi^*$, we can define $V^*(\psi) = \max_{\psi' \geq \psi} V(\psi)$, which is a monotonically decreasing function in ψ. Thus, the platform compares $\Pi = V^*(\psi)\phi(r)/(1-\lambda)$ with a constant $\Pi' = (1-\alpha)V(\psi^*) - \alpha C$. Note that Π is monotonically increasing in λ: $1/(1-\lambda)$ is increasing in λ, and $V^*(\psi)$ is decreasing in ψ, and therefore increasing in λ. Thus, there exists some cutoff λ^* such that when $\lambda > \lambda^*$, $\Pi > \Pi'$, but when $\lambda < \lambda^*$, $\Pi < \Pi'$. Because the platform never makes type-II errors (only certain misinformation is removed), Parts (a) and (b) follow by noting that virality of misinformation is proportional to $V(\psi)$ where ψ is chosen by the platform. □

Proof of Proposition 5. 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 neighbourhood $\eta > 0$ of $\bar{b}$ (where $\bar{b}$ is the same $\bar{b}$ constructed in Theorem 3 Claim(iii)). 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 low-reliability content ($r < r_P$), showing (ii) via Lemma 2. To prove (i), we note that agents in this neighbourhood 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 $p_s/p_d \leq p^*$. □

Acknowledgments. This article 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, Andrea Galeotti (editor), Matthew Gentzkow, Ali Makhdoumi, Azarakhsh Malekian, Mohamed Mostagir, Ro'ee Levy, Francesca Parise, Davy Perlman, Alexander Wolitzky, three anonymous referees, and numerous participants at INFORMS, the Janeway Institute Networks Seminar, the MIT theory lunch, the Network Science & Economics conference, the SAET annual conference, as well as NSF and the Sloan Foundation.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

ABRAMOWITZ, A. I. (2010), *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy* (New Haven, CT: Yale University Press).

ACEMOGLU, D., CHERNOZHUKOV, V. and YILDIZ, M. (2016), "Fragility of Asymptotic Agreement Under Bayesian Learning", *Theoretical Economics*, **11**, 187–225.

ACEMOGLU, D., COMO, G., FAGNANI, F., *et al.* (2013), "Opinion Fluctuations and Disagreement in Social Networks", *Mathematics of Operations Research*, **38**, 1–27.

ACEMOGLU, D., OZDAGLAR, A. and PARANDEHGHEIBI, A. (2010), "Spread of (Mis)information in Social Networks", *Games and Economic Behavior*, **70**, 194–227.

ALLCOTT, H. and GENTZKOW, M. (2017), "Social Media and Fake News in the 2016 Election", *Journal of Economic Perspectives*, **31**, 211–36.

ALLEN, J., HOWLAND, B., MOBIUS, M., *et al.* (2020), "Evaluating the Fake News Problem at the Scale of the Information Ecosystem", *Science Advances*, **6**, eaay3539.
ALLON, G., DRAKOPOULOS, K. and MANSHADI, V. (2021), "Information Inundation on Platforms and Implications", *Operations Research*, **69**, 1784–1804.
ALTAY, S., HACQUIN, A.-S. and MERCIER, H. (2020), "Why Do So Few People Share Fake News? It Hurts Their Reputation", *New Media & Society*, **24**, 1303–1324.
ANDREWS, L. (2012), "Facebook is Using You", *The New York Times*, 4.
ARAL, S. and DHILLON, P. S. (2018), "Social Influence Maximization Under Empirical Influence Models", *Nature Human Behaviour*, **2**, 375–382.
BALA, V. and GOYAL, S. (1998), "Learning From Neighbours", *The Review of Economic Studies*, **65**, 595–621.
BANERJEE, A. V. (1992), "A Simple Model of Herd Behavior", *The Quarterly Journal of Economics*, **107**, 797–817.
BICKERT, M. (2020), "Charting a Way Forward on Online Content Regulation".
BIKHCHANDANI, S., HIRSHLEIFER, D., TAMUZ, O., *et al.* (2021), "Information Cascades and Social Learning" (Working Paper 28887, National Bureau of Economic Research).
BUCHANAN, T. (2020), "Why Do People Spread False Information Online? The Effects of Message and Viewer Characteristics on Self-Reported Likelihood of Sharing Social Media Disinformation", *PLoS One*, **15**, e0239666.
BUDAK, C., AGRAWAL, D. and EL ABBADI, A. (2011), "Limiting the Spread of Misinformation in Social Networks", in *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, Hyderabad, India (New York, NY: Association for Computing Machinery) 665–674.
CANDOGAN, O. and DRAKOPOULOS, K. (2020), "Optimal Signaling of Content Accuracy: Engagement vs. Misinformation", *Operations Research*, **68**, 497–515.
CEN, S. and SHAH, D. (2021), "Regulating Algorithmic Filtering on Social Media", *Advances in Neural Information Processing Systems*, **34**, 6997–7011.
CENTOLA, D. (2010), "The Spread of Behavior in an Online Social Network Experiment", *Science*, **329**, 1194–1197.
CENTOLA, D. and MACY, M. (2007), "Complex Contagions and the Weakness of Long Ties", *American Journal of Sociology*, **113**, 702–734.
CHEN, L. and PAPANASTASIOU, Y. (2021), "Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation", *Management Science*, **67**, 6734–6750.
CLAYTON, K., BLAIR, S., BUSAM, J. A., *et al.* (2020), "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media", *Political Behavior*, **42**, 1073–1095.
DEL VICARIO, M., BESSI, A., ZOLLO, F., *et al.* (2016), "The Spreading of Misinformation Online", *Proceedings of the National Academy of Sciences*, **113**, 554–559.
ECKLES, D. (2022), "Algorithmic Transparency and Assessing Effects of Algorithmic Ranking". *SocArXiv*.
ECKLES, D., KIZILCEC, R. F. and BAKSHY, E. (2016), "Estimating Peer Effects in Networks With Peer Encouragement Designs", *Proceedings of the National Academy of Sciences*, **113**, 7316–7322.
EGELHOFER, J. L. and LECHELER, S. (2019), "Fake News as a Two-Dimensional Phenomenon: A Framework and Research Agenda", *Annals of the International Communication Association*, **43**, 97–116.
FIORINA, M. P., ABRAMS, S. A. and POPE, J. C. (2008), "Polarization in the American Public: Misconceptions and Misreadings", *The Journal of Politics*, **70**, 556–560.
FRENKEL, S. and KANG, C. (2021), *An Ugly Truth: Inside Facebook's Battle for Domination* (UK: Hachette).
GARIMELLA, K. and ECKLES, D. (2020), "Images and Misinformation in Political Groups: Evidence From Whatsapp in India". arXiv preprint arXiv:2005.09784.
GENTZKOW, M. and SHAPIRO, J. M. (2006), "Media Bias and Reputation", *Journal of Political Economy*, **114**, 280–316.
GOLUB, B. and JACKSON, M. O. (2010), "Naive Learning in Social Networks and the Wisdom of Crowds", *American Economic Journal: Microeconomics*, **2**, 112–149.
GREENE, C. M., NASH, R. A. and MURPHY, G. (2021), "Misremembering Brexit: Partisan Bias and Individual Predictors of False Memories for Fake News Stories Among Brexit Voters", *Memory*, **29**, 587–604.
GRINBERG, N., JOSEPH, K., FRIEDLAND, L., *et al.* (2019), "Fake News on Twitter During the 2016 U.S. Presidential Election", *Science*, **363**, 374–378.
GUESS, A., NAGLER, J. and TUCKER, J. (2019), "Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook", *Science Advances*, **5**, eaau4586.
GUESS, A., NYHAN, B. and REIFLER, J. (2018), "Selective Exposure to Misinformation: Evidence From the Consumption of Fake News During the 2016 US Presidential Campaign", *European Research Council*, **9**, 4.
HSU, C.-C., AJORLOU, A. and JADBABAIE, A. (2020), "News Sharing, Persuasion, and Spread of Misinformation on Social Networks" (SSRN Scholarly Paper ID 3391585).
KAMENICA, E. (2019), "Bayesian Persuasion and Information Design", *Annual Review of Economics*, **11**, 249–272.
KAMENICA, E. and GENTZKOW, M. (2011), "Bayesian Persuasion", *American Economic Review*, **101**, 2590–2615.
KEPPO, J., KIM, M. J. and ZHANG, X. (2022), "Learning Manipulation Through Information Dissemination", *Operations Research*, **70**, 3490–3510.
KIM, D. H., JONES-JANG, S. M. and KENSKI, K. (2020), "Why Do People Share Political Information on Social Media?", *Digital Journalism*, **9**, 1–18.
KOZYREVA, A., LEWANDOWSKY, S. and HERTWIG, R. (2020), "Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools", *Psychological Science in the Public Interest*, **21**, 103–156.

KRANTON, R. and MCADAMS, D. (2022), "Social Connectedness and the Market for Information" (Technical Report, Working Paper).
LAZER, D. M. J., BAUM, M. A., BERINSKY, A. J., *et al.* (2018), "The Science of Fake News", *Science*, **359**, 1094–1096.
LEE, C. S., MA, L. and GOH, D. H.-L. (2011), "Why Do People Share News in Social Media?", in Active Media Technology, Lecture Notes in Computer Science (Lanzhou, China: Springer) 129–140.
MERLINO, L. P., PIN, P. and TABASSO, N. (2023), "Debunking Rumors in Networks", *American Economic Journal: Microeconomics*, **15**, 467–496.
MOLINA, M. D., SUNDAR, S. S., LE, T., *et al.* (2021), ""Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content", *American Behavioral Scientist*, **65**, 180–212.
MOSTAGIR, M., OZDAGLAR, A. and SIDERIUS, J. (2022), "When is Society Susceptible to Manipulation?" *Management Science*, **68**, 7153–7175.
MOSTAGIR, M. and SIDERIUS, J. (2022a), "Learning in a Post-Truth World", *Management Science*, **68**, 2860–2868.
MOSTAGIR, M. and SIDERIUS, J. (2022b), "When Do Misinformation Policies (Not) Work?" (Technical Report, Working Paper).
MOSTAGIR, M. and SIDERIUS, J. (2023), "Social Inequality and the Spread of Misinformation", *Management Science*, **69**, 968–995.
NGUYEN, N. P., YAN, G., THAI, M. T., *et al.* (2012), "Containment of Misinformation Spread in Online Social Networks", in *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, Evanston, Illinois (New York, NY: Association for Computing Machinery) 213–222.
PAPAKYRIAKOPOULOS, O., HEGELICH, S., SHAHREZAYE, M., *et al.* (2018), "Social Media and Microtargeting: Political Data Processing and the Consequences for Germany", *Big Data & Society*, **5**, 205395171881184.
PAPANASTASIOU, Y. (2020), "Fake News Propagation and Detection: A Sequential Model", *Management Science*, **66**, 1826–1846.
PARISER, E. (2011), *The Filter Bubble: What The Internet Is Hiding From You* (UK: Penguin Books Limited).
PENNYCOOK, G., BEAR, A., COLLINS, E. T., *et al.* (2020), "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings", *Management Science*, **66**, 4944–4957.
PENNYCOOK, G., CANNON, T. D. and RAND, D. G. (2018), "Prior Exposure Increases Perceived Accuracy of Fake News", *Journal of Experimental Psychology. General*, **147**, 1865–1880.
PENNYCOOK, G., EPSTEIN, Z., MOSLEH, M., *et al.* (2021), "Shifting Attention to Accuracy Can Reduce Misinformation Online", *Nature*, **592**, 590–595.
PENNYCOOK, G., MCPHETRES, J., ZHANG, Y., *et al.* (2020), "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention", *Psychological Science*, **31**, 770–780.
PENNYCOOK, G. and RAND, D. G. (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, M. (2013), "Media and Political Polarization", *Annual Review of Political Science*, **16**, 101–127.
RO'EE, L. (2021), "Social Media, News Consumption, and Polarization: Evidence From a Field Experiment", *American Economic Review*, **111**, 831–870.
SUNSTEIN, C. R. (2018), "#Republic: Divided Democracy in the Age of Social Media".
TARSKI, A. (1955), "A Lattice-Theoretical Fixpoint Theorem and its Applications", *Pacific Journal of Mathematics*, **5**, 285–309.
TAYLOR, S. J. and ECKLES, D. (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* (Cham: Springer) 289–322.
TOPKIS, D. M. (1998), *Supermodularity and Complementarity* (1st edn) (Princeton, NJ: Princeton University Press).
TÖRNBERG, P. (2018), "Echo Chambers and Viral Misinformation: Modeling Fake News as Complex Contagion", *PLoS One*, **13**, e0203958.
VOSOUGHI, S., ROY, D. and ARAL, S. (2018), "The Spread of True and False News Online", *Science*, **359**, 1146–1151.
YEUNG, K. (2018), "Algorithmic Regulation: A Critical Interrogation", *Regulation & Governance*, **12**, 505–523.
YEUNG, K. and LODGE, M. (2019), *Algorithmic Regulation* (Oxford, UK: Oxford University Press).