Cisternas, Gonzalo; Vásquez, Jorge

Working Paper

Misinformation in social media: The role of verification incentives

Staff Report, No. 1028

Provided in Cooperation with:
Federal Reserve Bank of New York

Suggested Citation: Cisternas, Gonzalo; Vásquez, Jorge (2022) : Misinformation in social media: The role of verification incentives, Staff Report, No. 1028, Federal Reserve Bank of New York, New York, NY

This Version is available at:
https://hdl.handle.net/10419/266112

STAFF
REPORTS

NO. 1028
AUGUST 2022

Misinformation in Social Media: The Role of Verification Incentives

Gonzalo Cisternas | Jorge Vásquez

FEDERAL RESERVE BANK *of* NEW YORK

Misinformation in Social Media: The Role of Verification Incentives

Gonzalo Cisternas and Jorge Vásquez

Federal Reserve Bank of New York Staff Reports, no. 1028
August 2022
JEL classification: D40, L10, L50

Abstract

We develop a model of misinformation wherein users' decisions to verify and share news of unknown truthfulness interact with producers' choices to generate fake content as two sides of a market that balance to deliver an equilibrium *prevalence* and *pass-through* of fake news. We leverage the tractability of the model to examine the efficacy of various policies intended to combat misinformation that are in place currently, stressing how these may nontrivially interact with users' incentives: news verification is a costly activity. Our analysis emphasizes the importance of examining users' and producers' decisions jointly, as well as of evaluating how policies interact with one another. It also provides sensitivity measures that are key for policy evaluation.

Key words: misinformation, news verification, social media

Cisternas: Federal Reserve Bank of New York (email: gonzalo.cisternas@ny.frb.org). Vásquez: Smith College, Department of Economics (email: jvasquez@smith.edu). This paper was previously circulated under the title "Fake News in Social Media: A Supply and Demand Approach." The authors thank Marco Cipriani, Steven Durlauf, Jorge Lemus, Lones Smith, Marek Weretka, and Juanjuan Zhang for their useful comments, and Orrie Page for research assistance.

1 Introduction

The spread of misinformation online has gained substantial prominence in society. For instance, only in the 2016 U.S. presidential election, Allcott and Gentzkow (2017) estimate that 760 million interactions with fake news occurred on the web, while Guess et al. (2020b) show that online platforms were a key gateway for landing on untrustworthy websites. The tangible impact that fake news can have is then sizable, potentially affecting elections, markets, and disease spread,[1] with the World Economic Forum labeling this issue as a major global risk (Howell, 2013). The pace at which internet access, social media usage, and technologies evolve also indicates that this problem is unlikely to disappear.[2]

The response by key actors in the news world—most notably, social media platforms—can be synthesized based on three tenets. First, professional, independent fact-checking is required, which has resulted in the advent of a network of *third-party fact-checkers* who verify the accuracy of content.[3] Second, the incentives of fake news producers must be weakened; for instance, any news confirmed to be false is given less relevance and repeated offenders are removed from platforms,[4] while websites can be rated on their trustworthiness in an attempt to influence their advertising revenues.[5] Third, users must be empowered to assess the veracity of each news item to ultimately choose how to act upon it. For example, on some platforms, flagged content is now accompanied by fact-checkers' reports or related material that provides context so users can determine the accuracy of the material for themselves.[6]

In this paper, we introduce a flexible model of misinformation that can be used as a framework for addressing a variety of questions regarding the fake news problem, and also to shed light on potential policy interventions. In light of the aforementioned tenets, for instance, how impactful is the appearance of fact-checking that lowers the verification costs borne by users? How are key variables such as the prevalence and diffusion of fake news altered by policies that attempt to disrupt their production? How effective are algorithmic

[1] Rapoza (2017) documents a temporary decrease of $130 billion in stock value after a false tweet about an explosion injuring Barack Obama in 2013. More recently, DiResta and Garcia-Camargo (2020) examine how a video promoting falsehoods surrounding the COVID-19 pandemic went viral.

[2] As of January 2020, 4.54 billion people were estimated to be using the internet, with 3.8 billion active social media users, representing 7% and 9% yearly increases, respectively. As discussed in World Economic Forum (2020), a major long-term concern is the use of artificial intelligence for misinformation purposes in the form of "deepfake" videos, which substantially increases verification costs for laypeople.

[3] Some platforms partner with fact-checking organizations that adhere to the International Fact-checking Code of Principles (`https://www.ifcncodeofprinciples.poynter.org`).

[4] See, for instance, Meta Business Help Center (2022).

[5] See the Global Disinformation Index, `https://disinformationindex.org/`.

[6] This policy need not only reflect technological limitations such as an algorithm's inability to check all the content on a platform but also a vision of the role of social media in society. Lyons (2017) argues that this contextual approach lowers sharing behavior more than does merely using labels to flag news as false.

filters that detect fake content before it reaches the platform's users? At the center of our analysis is how users' incentives to verify information respond to the policy in place.

Model and equilibrium. In our model, a platform is a venue on which users encounter news that can be true or false. Upon contact with a news item/article, a user can first choose to learn its veracity at a cost and then decide whether to share it. Importantly, users experience gains from sharing true news but suffer losses from sharing fake content—verification then entails evaluating the payment of a cost with certainty versus a loss of sharing fake content with some chance. The latter chance is given by the endogenous news proportion that is false among the total entering the platform, or fake news *prevalence*. As prevalence rises, fewer users share news articles without verifying them; hence, a decreasing locus emerges, linking prevalence levels to the mass of users engaged in unverified sharing behavior.

Because verification reveals the truthfulness of each news item, and users dislike sharing fake content, fake news can be shared only if they reach users performing unverified sharing. As the proportion of these users increases, the pass-through of misinformation increases, thereby incentivizing the production of fake content. Thus, an increasing supply of fake news ensues. Coupled with the aforementioned decreasing locus capturing unverified sharing behavior, an equilibrium in this market corresponds to a level of fake news prevalence and a pass-through rate of fake content that occurs when these two curves intersect. Equipped with this tractable characterization, we turn to examining how current interventions affect outcomes in this induced market for news.

Overview of the results. *(1.) Unverified sharing behavior can be insensitive to reductions in verification costs.* For fake news producers, a decreasing unverified sharing locus constitutes a "demand for misinformation" emanating from the platform. Further, lowering verification costs can potentially lower unverified sharing. It is then natural to ask if a reduction in verification costs acts as a "demand shifter," always ensuring a drop in prevalence.

In Section 4 we show that, for any decrease in verification costs, there are always regions of prevalence at which unverified sharing behavior is insensitive to such changes. Further, those regions need not be only of low prevalence in which users' incentives to verify can be expected to be low: changing how the users' benefits and losses vary across the population can alter the location of such regions. Hence, depending on the status of fake news prevalence, equilibrium outcomes need not change after reductions in verification costs.

(2.) Supply interventions can increase the diffusion of fake news. Various policies currently intended to combat misinformation attempt to disrupt its production by making fake news creation less attractive. Clearly, the most optimistic case is when the supply of fake news shifts left/upwards at all possible levels, which we explore in Section 5.

Not surprisingly, fake news prevalence decreases. However, the pass-through of fake content increases because users' verification incentives weaken in the process. The diffusion rate of fake news—prevalence times pass-through—captures the proportion of fake content that is shared; hence, it acts as a measure of the diffusion of fake content. We provide conditions under which the unverified sharing locus is sufficiently elastic so more misinformation is transmitted after the intervention. Importantly, we also provide conditions under which a reduction in verification costs results in a more elastic unverified sharing locus.

These adverse effects can be more extreme in the context of network externalities, understood as individual choices depending on the aggregate choices of others. In Section 8, we show how the unverified sharing locus can become a correspondence; thus, a supply reduction can lead not only to more diffusion but also to greater prevalence. Moreover, such a policy can refine the set of equilibria to a single outcome that is worse than the original one.

(3.) The traditional exercise of market power has impediments, but simple segmentation strategies can be profitable. In Section 6, we consider the case of a single news producer that inherits the cost structure of the "competitive" case. Traditional market power would then consist of reducing "trade'—in our case, fake news prevalence—to obtain a larger per-unit revenue—pass-through in our model. However, the monopolist does not control users' sharing decisions directly, and fake news prevalence is in general not observed by users. This informational limitation implies that the only sequentially rational outcome is the competitive equilibrium already found, and uniform policies are rendered ineffective.

That said, a mild enrichment of the monopolists' toolbox can improve profits. Specifically, if the monopolist can (i) segment the market trivially, i.e., target sub-populations that only differ from the original one in terms of size, and (ii) supply truthful content too, then the monopolist can implement prevalence-pass-through pairs on the *concave closure* of the unverified sharing locus. This technique can be particularly profitable when decreasing verification costs creates convexities in the aforementioned locus, partially counteracting the decreases in prevalence and diffusion that would otherwise take place.

(4.) Detection algorithms that remove news for users can backfire. Finally, in Section 7 we evaluate the efficacy of internal filters that assess news articles before they reach users and that remove content if deemed fake. We focus on an algorithm that makes type-II errors, i.e., failing to recognize false news, as the social costs of type-I errors are more straightforward.

In this context, we show that introducing a filter can increase the prevalence and diffusion of misinformation over a large range of filter qualities. Indeed, by affecting the inference made by users when receiving news, such filters can induce more unverified sharing, and this equilibrium effect can outweigh the positive direct effect that a filter can have on reducing the production of fake news via limiting its pass-through, all else equal.

Applied relevance. From an institutional viewpoint, our model rests on three assumptions. First, *the production of fake content increases with sharing rates*. Indeed, whether ideologically or profit-driven, clicks are the main source of profitability for untrustworthy websites in the online market for fake news (e.g. Allcott and Gentzkow, 2017 and Tucker et al., 2018). However, with more sharing, more users can be reached, and the likelihood of more clicks increases. Even more so, the mere emergence of social media platforms represents an increase in sharing ability on par with an increase in the severity of the fake news problem.

Second, we chose a set up in which *users find it beneficial to share truthful content and dislike sharing misinformation*. From a modeling perspective, this assumption intentionally reduces the chances that misinformation will be transmitted. But it is consistent with evidence on users finding it important to share only accurate news (Pennycook et al., 2021), and worrying about their reputations when fake news is shared (Altay et al., 2022).

Third, *users bear costs to verify the information they encounter*. Indeed, while the growth in fact-checking outlets worldwide[7] represents lower search costs for users, it does not necessarily eliminates verification costs: to make an informed decision, users must review reports whether searching independently at specialized sites,[8] or accessing them as part of contextual information readily provided for free on platforms.

Our assumptions intend to stress that successful interventions require a joint examination of users' and producers' decisions, putting at the center users' verification incentives. Further, our analysis uncovers a variety of sensitivity measures that are useful for evaluating policies, just as in conventional markets. From this perspective:

(a) The way in which benefits and losses vary across populations of interest determines the prevalence levels in which verification incentives are (not) sensitive to verification costs—such relationships are key for assessing the efficacy of fact-checking initiatives.

(b) Policies that decrease unverified sharing should also evaluate their impact on the sensitivity of unverified sharing to changes in prevalence, as a reduction in misinformation may be accompanied by an increase in its diffusion.

(c) Segmentation strategies that seem "simplistic" a priori need not imply a lack of sophistication or negligible harm.

(d) Substitution effects between users' verification incentives and the use of internal filters can be sizable. Likewise, strategic complementarities arising from network externalities can make increases in prevalence consistent with less verification.

[7] Stencel and Luther (2020) shows that the number of fact-checking sites has grown from 44 in 2014 to almost 300 by the middle of 2020.

[8] See, for instance, `https://snopes.com` or `https://politifact.com`.

Related literature. Due to our policy-oriented spirit, this work is naturally related to papers that have studied, either empirically or experimentally, the implementation of policies intended to mitigate the fake news problem. Regarding fact-checking, Henry et al. (2022) document in experiments that users who are more prone to sharing news articles are also more prone to verifying them. In our setting, this finding suggests a positive relationship between benefits and losses across types, which we sometimes use as a leading example. Regarding verification mechanisms that are not user based, Pennycook et al. (2020) show, also through experiments, that labeling only a subset of false news articles leads users to believe that untagged articles are more accurate, which positively influences the sharing of the latter; a similar phenomenon arises in our model when users are aware of the presence of news filters. Finally, regarding policies aimed at limiting the sharing of misinformation, Ershov and Morales (2021) empirically examine the impact of an increase in the cost of sharing on a popular platform on the transmission of news from highly trustworthy outlets relative to news from less trustworthy counterparts. In contrast, we examine how supply interventions that would reduce sharing behavior, all else equal, can in fact lead to more sharing and diffusion of fake news due to a weakening of users' verification incentives.

Our paper also complements a growing body of theoretical papers examining various aspects of the fake news problem. Papanastasiou (2020) shows that news virality can result from a traditional rational cascade logic that is also facilitated by a costly signal acquisition. Bowen et al. (2021) show that selective sharing and selection neglect could result in belief polarization when agents learn from what others share, while Cheng and Hsiaw (2022) consider a signaling model in which uncertainty about a sender being benevolent or malevolent could lead to users' beliefs about an unknown state disagreeing in the long term. Finally, Acemoglu et al. (2022) develop a model in which users have heterogeneous priors about an unknown state, and find that homophily introduces a trade-off between virality and the emergence of eco chambers.

To conclude, our model of costly verification also contributes to the literature exploiting the tractability of matching models in settings in which individuals choose to protect themselves from harm with endogenous intensity: Quercioli and Smith (2015) examines the economics of counterfeiting, while Vásquez (2022) develops an equilibrium theory of crime and vigilance. In such a matching context, our segmentation analysis employing the concave closure of the unverified sharing function connects with the techniques in Kamenica and Gentzkow (2011) for Bayesian persuasion problems, albeit with an endogenous "prior" that is now determined via an optimal production decision; see Bergemann et al. (2015) and Haghpanah and Siegel (2022) for general treatments of market segmentation in traditional product markets.

2 Model

We develop a model of a platform on which a large number of users encounter fake content that originates from a large number of fake news producers.

News viewers A unit mass of infinitesimal risk neutral *users* have access to an online platform on which they encounter "uncertified" news articles, i.e., news articles for which truthfulness cannot be determined upon first contact (e.g., by reading the headline, the originating website, or even the whole article). These encounters are random from the perspective of all platform participants.[9] Upon encountering a news item, each user can decide to determine its veracity by paying a fixed cost $t \geq 0$; for instance, search costs when consulting specialized websites for fact checks, the time costs associated with reviewing related articles presented as part of "contextual information," or even attention costs.

After the verification decision is made, users can decide to share the news item. Not sharing yields a payoff of zero. The payoff of sharing nonetheless depends on the veracity of the news item, and it varies across users. Specifically, we assume that users are heterogeneous according to a one-dimensional characteristic v—or *type*—taking values in $[0, \bar{v}]$, such that sharing truthful news yields a *benefit* $b(v)$, while sharing fake news entails a *loss* $\ell(v)$. The functions $b(\cdot)$ and $\ell(\cdot)$ are continuous, and also strictly positive almost everywhere (a.e.). The underlying characteristic v is distributed according to an atomless cumulative distribution function (CDF) $G(\cdot)$, with support $[0, \bar{v}]$ and density $g(\cdot)$ which is differentiable.

As an interpretation, if v is a measure of user popularity (e.g., number of followers), and reputation matters for users, we can expect b and ℓ to be increasing over $[0, \bar{v}]$. Our analysis is, however, general in that we only impose the conditions previously stated. That said, because users dislike passing on fake news (i.e., $\ell > 0$ a.e.) and the verification technology is perfect, fake articles can be shared only when they are not verified.

Fake news producers We assume that a fixed mass of uncertified news enters the platform, which we set at a level equal to 1. Among this mass, a proportion $\pi \in [0, 1]$ is false. We refer to π as the *fake news prevalence*. This news type originates from a pool of potential *fake news producers*, each facing the choice of producing a fake news article upon paying an opportunity *cost* $c \in [0, 1]$, or not producing at all. Cost c varies across producers according to an atomless CDF $F(\cdot)$ with support $[0, 1]$ and a differentiable density $f(\cdot)$.

[9] This assumption is not intended to reflect that the encounters are truly accidental but that the news outlets' have limited reach when targeting individuals. In practice, this situation arises because platforms' services offer an imperfect degree of granularity at the user level, and because algorithms mediate matches using information not possessed by news producers. Targeting is then imperfect within populations of interest.

Fake news producers receive a payoff of 1 when a news item is shared, and 0 otherwise. Thus, the expected revenue for any producer is given simply by the probability with which a fake news item is ultimately shared. We refer to this probability as the *pass-through of fake news*, and we denote it by $\sigma \in [0, 1]$. Since in this baseline model only users verify news, and fake news can be shared only when users choose not to verify, the fake news pass-through coincides with the mass of users who engage in unverified sharing (given their conjecture of fake news prevalence).

Equilibrium concept The model is "competitive" in that all platform participants are assumed to take the variables (π, σ) as given.

Definition 1 (Equilibrium). *An equilibrium consists of a prevalence π^* and a pass-through rate σ^* such that: (i) given prevalence π^*, sharing and verification choices are optimal for all users, and (ii) given σ^*, potential fake news producers' choices are optimal.*

The model can be seen as the steady state of a platform with a *rapidly evolving* news influx. Specifically, each period can be divided into two stages. First, a new cohort of news articles enters the platform, each encountering a user. Second, if the user clicks share, this news item becomes visible to a subset of individuals, yielding revenues to the untrustworthy website. In the next period, a new cohort of news articles enters the platform and is favored by the platform's algorithm by being placed more prominently in the users' news feed. As a result, the subsequent sharing rate of the old cohort is limited or simply eliminated, thereby making the initial rate the most relevant for payoffs.

3 Unverified Sharing and Equilibrium Prevalence

Sharing misinformation Our first task is to find the user set that chooses to skip verification and share the news articles for each prevalence level π and verification cost t. This set must be a subset of the users who find it optimal to share when verification is unavailable, or prohibitively costly for everyone (e.g., $t = +\infty$). Indeed, by revealed preferences, for any user who does not share in this case, sharing without verifying remains dominated when the possibility of verification is added.

Suppose that verification is unavailable. Fixing prevalence π, type v will decide to share provided

$$(1 - \pi)b(v) - \pi\ell(v) \geq 0,$$

where we have assumed that ties are broken in favor of sharing. That is, sharing will occur for

those types that exhibit a high *propensity to share* $b(v)/\ell(v)$, as subsumed in the condition

$$\frac{b(v)/\ell(v)}{1+b(v)/\ell(v)} \geq \pi. \tag{1}$$

If the platform introduces a costly verification technology, types v satisfying (1) will trade off paying the verification cost t and saving the loss ℓ from sharing fake news versus engaging in unverified sharing. Critically, since costly verification is naturally paid irrespective of the verification outcome, unverified sharing dominates verified sharing whenever

$$(1-\pi)b(v) - \pi\ell(v) \geq (1-\pi)b(v) - t,$$

or when their propensity to skip verification, $t/\ell(v)$, is sufficiently high:

$$t/\ell(v) \geq \pi. \tag{2}$$

Altogether, given prevalence π and verification cost t, the set of users $\mathcal{V}(\pi;t)$ who share unverified news articles are those whose type v satisfies inequalities (1) and (2), namely:

$$\mathcal{V}(\pi;t) \equiv \{v \in [0,\bar{v}] : \varphi(v;t) \geq \pi\}, \text{ where } \varphi(v;t) \equiv \min\left\{\frac{b(v)/\ell(v)}{1+b(v)/\ell(v)}, \frac{t}{\ell(v)}\right\}. \tag{3}$$

We can now state a central study object: the *sharing rate of unverified news*, which maps prevalence levels to the mass of users sharing unverified content:

$$\Sigma(\pi;t) \equiv \int_{\mathcal{V}(\pi;t)} dG(v). \tag{4}$$

By definition of (3), the correspondence $\pi \mapsto \mathcal{V}(\pi;t)$ is weakly decreasing in the sense of set inclusion. Thus, the sharing rate $\pi \mapsto \Sigma(\pi;t)$ is non-increasing, reflecting that as fake news become more prevalent, fewer users share news articles without verifying them. In addition, if the production of fake news ceases, all users share news articles without verifying (i.e., $\mathcal{V}(0;t) = [0,\bar{v}]$) as sharing ceases to have a downside, i.e., $\Sigma(0;t) = 1$. Moreover, the set of users who share news has measure zero if only fake news articles circulate because $\ell > 0$ a.e. Thus, $\Sigma(1;t) = 0$ with Σ possibly vanishing strictly before 1 in some specifications.

Remark 1. Distinguishing between the "pass-through of fake news" (σ) and the "sharing rate of unverified news" (Σ) is appropriate not only because of their dimensionality (scalar vs. functional) but also because both notions may differ in more general specifications, such as when we study the use of internal filters in Section 7.

We focus on the case in which Σ is continuous. While this choice is largely for convenience—i.e., it avoids qualifying our results to ensure that an equilibrium exists—it is also natural. Specifically, it rules out the possibility of marginal reductions in perceived prevalence π prompting a large number of users to become active in unverified sharing at arbitrary levels of prevalence. The next result outlines conditions that ensure continuity of Σ.

Lemma 1. *If $\varphi(\cdot;t)$ is differentiable a.e. with $|\varphi'(\cdot;t)| > 0$ a.e, then Σ is continuous.*

That is, provided $\varphi(\cdot;t)$ does not fluctuate too wildly or exhibit flat regions, small changes in the environment do not have large effects in the aggregate. Otherwise, φ can take any form, allowing for kinks that can appear naturally due to the presence of a minimum. We assume the continuity of Σ in the rest of this paper.

Producing Fake News. We now turn to the supply of fake news entering the platform. Given a pass-through of fake news σ, a fake news producer with cost c chooses to produce if and only if $\sigma \geq c$. Thus, the *supply of fake news*, i.e., the function that maps pass-through rates σ to prevalence levels, is given simply by

$$\Pi(\sigma) \equiv F(\sigma). \tag{5}$$

Due to the properties of the distribution $F(\cdot)$, the supply $\Pi(\cdot)$ is continuous, satisfies $\Pi(0) = 0$ and $\Pi(1) = 1$, and is non-decreasing. In other words, more fake content is generated when the pass-through rate increases, as the platform becomes more attractive for producers.

Remark 2. Normalizing the mass of news articles to 1 (thereby implying that the supply of truthful content shrinks as more fake news articles enter the market) is without loss of generality. Indeed, any production side that yields inflows of fake content that positively respond to pass-throughs would deliver. Also, as in Remark 1, distinguishing between "prevalence" and "supply" is justified because neither notion coincides in other specifications (e.g., when the total mass of news articles exceeds 1).

Altogether, the existence and uniqueness of an equilibrium is always guaranteed.

Proposition 1. *There exists a unique equilibrium (π^*, σ^*).*

As in traditional competitive analyses, the opposing forces behind the supply of fake news and the unverified sharing locus ensure the existence of a unique pair (π^*, σ^*) that balances this "market for misinformation." Economically, the model rests ultimately on two natural assumptions: unverified sharing behavior decreases as fake news articles become more abundant, and the supply of fake news articles increases along with the pass-through rate.

In the following sections, we leverage the model structure to illuminate on a number of issues related to the fake news problem. Further, to simplify the exposition, we assume that types are ranked by their propensity to share, with higher types being more likely to share. In other words, to illustrate the economics, we make the following assumption:

Assumption 1. *The function $v \mapsto b(v)/\ell(v)$ is strictly increasing.*

4 Facilitating Verification by Users

We now examine the effects of verification cost t on the sharing of unverified news $\Sigma(\pi; t)$. The motivation is the recent push by social media platforms to offer more and better fact-checking services to users, while allowing them to decide what to share. These efforts can be seen as a decrease in the verification costs that users face.

While $\Sigma(\pi; t)$ certainly could be seen a supply of fake news to the followers of those who share content without verifying it, its resemblance to a traditional demand function is apparent: it is strictly decreasing in the quantity (π) of the good of interest, and it determines the revenue for its producers. Reductions in t that shrink $\Sigma(\pi; t)$ by fostering verification are expected to lower the equilibrium prevalence of fake news, π^*, in analogy with *demand shifts* that induce a lower equilibrium quantity in a traditional supply and demand framework. Figure 1 illustrates the change in Σ as verification costs decrease, and the resulting decrease in the prevalence and pass-through of fake news in equilibrium.

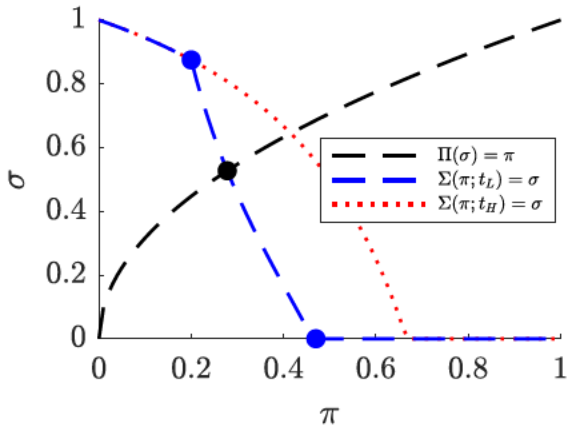

Figure 1: Parameter values: $\bar{v} = 1$, $G(v) = v$, $F(c) = c^2$, $b(v) = v^2$, and $\ell(v) = v/2$. The downward sloping red-dotted and blue-dashed curves represent the unverified sharing function for verification cost $t_H = 0.4$ and $t_L = 0.1$, respectively. The upward sloping curve is the supply of fake news.

Our main observation is that, generically, verification costs *do not act as traditional demand shifters*: as seen in Figure 1, there are regions of π—in this case, low values—

wherein the sharing of unverified news $\Sigma(\pi; t)$ is *insensitive* to the change in verification cost t. Furthermore, the locations of these regions may vary across specifications.

In the remainder of the section, we use the case of increasing sharing losses $\ell(\cdot)$ as the leading example to (i) show how to construct $\Sigma(\pi; t)$ for each given t, and (ii) formally establish the existence of such regions that lack sensitivity. Then, we explain how our results and insights can be generalized without major conceptual changes.

Increasing losses $\ell(\cdot)$. With losses that increase with user types, those who are more prone to share—higher types, by Assumption 1—will also be more prone to verify because $t/\ell(\cdot)$ is decreasing.

To compute $\Sigma(\pi; t)$, we must determine the set $\mathcal{V}(\pi; t)$ of user types that share without verifying. Recall that this set is determined by the π-superlevel set of φ defined in (3), i.e., those types v for whom

$$\varphi(v; t) = \min \left\{ \frac{b(v)/\ell(v)}{1 + b(v)/\ell(v)}, \frac{t}{\ell(v)} \right\} \geq \pi.$$

With an increasing propensity to verify (first argument) and a decreasing propensity to skip verification (second argument), φ is quasiconcave. Naturally, if t is sufficiently large, there is no verification *for any* π. We collect and formalize these observations in Lemma 2.

Lemma 2. *φ is continuous and quasiconcave. Further, if $t \geq \frac{b(\bar{v})}{1+b(\bar{v})/\ell(\bar{v})}$, φ is increasing; hence, there is no verification for any $\pi \in [0, 1]$. Otherwise, φ is eventually decreasing.*

Verification is said to be *feasible* if $t < \frac{b(\bar{v})}{1+b(\bar{v})/\ell(\bar{v})}$, as it is only in that region that lowering verification costs can have an effect on outcomes. We focus on this case in what follows.

Let us now explain how Σ in Figure 1 is obtained. Consider Figure 2. In the left panel, the propensity-to-share function $\frac{b(v)/\ell(v)}{1+b(v)/\ell(v)}$ corresponds to the solid increasing function starting at the origin; the decreasing dashed-dotted curve is the propensity-to-skip-verification function, $t/\ell(v)$, when verification is feasible (e.g. $t < \frac{b(\bar{v})}{1+b(\bar{v})/\ell(\bar{v})}$; lower curve) or infeasible (upper curve). Two observations are instructive. First, since the propensity-to-share function is below 0.5, no sharing behavior—verified or unverified—can occur for prevalence $\pi \geq 0.5$. This explains why Σ, in dotted red in the right panel of Figure 2, constructed when verification is infeasible, vanishes from $\pi = 0.5$ onward. Second, t/ℓ begins crossing the propensity to share from above at $v = \bar{v}$, i.e., where the latter function is maximized. Thus, Σ becomes sensitive to verification costs only for high prevalence levels π. This phenomenon is depicted in the right panel of Figure 2, where unverified sharing (i) ceases for any $\pi \in [0.4, 0.5]$, and (ii) decreases but remains positive when $\pi \in [0.25, 0.4]$.

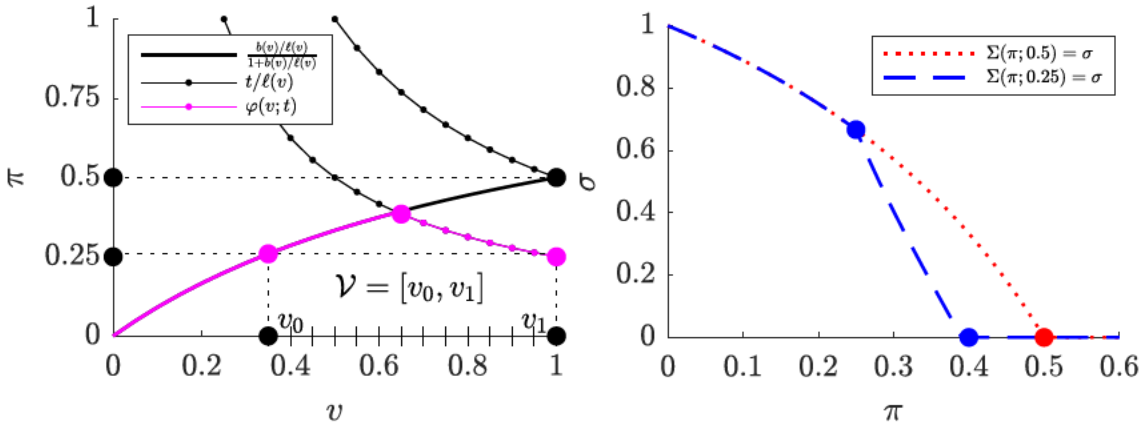

Figure 2: Both panels considers the same parametrization of Figure 1 but with a loss function $\ell(v) = v$ and verification costs $t \in \{0.25, 0.5\}$. The left panel depicts the propensities both to share and skip verification, as well as the minimum φ between these two maps. The right panel depicts the sharing rate of unverified news Σ for $t = 0.5$ and $t = 0.25$, respectively.

The intuition is as follows. For a high prevalence level, only high types, if any, are prone to share. However, at the same time, those are the types most prone to verify, because the losses that they suffer when sharing fake news articles are substantial. For high prevalence levels, therefore, high types decide to verify—thus avoiding a large potential loss—and safely share. Conversely, for lower prevalence levels, many user types are willing to share news articles. However, the reduction in t required to induce those who are more prone to verify—again, high types—to actually do so is too large, so no one engages in news verification.

The technical advantage of increasing losses ℓ is mild: $\mathcal{V}(\pi; t)$ is a closed interval consequence of φ being quasiconcave. Using the left panel of Figure 2 as a guide again, the latter set is found by the points of intersection between the horizontal line at level π and both propensity functions, if any. Denote these points by $v_0(\pi)$ and $v_1(\pi)$, defined by

$$v_0(\pi) \equiv \inf \left\{ v \in [0, \bar{v}] : \frac{b(v)/\ell(v)}{1 + b(v)/\ell(v)} \geq \pi \right\} \quad \text{and} \quad v_1(\pi) \equiv \inf \left\{ v \in [0, \bar{v}] : \frac{t}{\ell(v)} \leq \pi \right\}, \quad (6)$$

with the convention $v_0(\pi) = \bar{v}$ and $v_1(\pi) = \bar{v}$ if their respective sets are empty. Clearly, when interior, these types are found by imposing equality on the corresponding expressions.[10] The left panel of Figure 2 depicts the thresholds $v_0(\pi)$ and $v_1(\pi)$ for prevalence $\pi = 0.25$.

Equipped with these thresholds $v_0(\cdot)$ and $v_1(\cdot)$, we can formalize our previous discussion. We omit the explicit dependence of $\mathcal{V}$ on t unless needed.

Proposition 2. *Suppose that verification is feasible. (a) There exist prevalence levels* $0 <$

[10] In this case, $v_0(\pi)$ is the user type who is indifferent between sharing news without verifying and not sharing, while $v_1(\pi)$ is the user type who is indifferent between verified and unverified news sharing.

$\underline{\pi} < \overline{\pi} < 1$ *such that* $\mathcal{V}(\pi) = [v_0(\pi), \bar{v}]$ *if* $\pi \leq \underline{\pi}$; $\mathcal{V}(\pi) = [v_0(\pi), v_1(\pi)]$ *if* $\pi \in [\underline{\pi}, \bar{\pi}]$; *and* $\mathcal{V}(\pi) = \emptyset$ *if* $\pi > \bar{\pi}$. *(b) Moreover, if* $t' < t''$, *then there exists prevalence* $\underline{\pi}_t \in (0,1)$ *such that* $\mathcal{V}(\pi; t') = \mathcal{V}(\pi; t'')$ *for all prevalence* $\pi \leq \underline{\pi}_t$.

Proposition 2 shows that users who share news articles without verifying, if any, are those with mid valuations. In particular, when fake news prevalence takes intermediate values, the users are segmented into three sets: low valuation users $v < v_0(\pi)$ neither share nor verify; mid valuation users $v \in [v_0(\pi), v_1(\pi)]$ share without verification; and high valuation users $v > v_1(\pi)$ verify and thus share truthful news only. Proposition 2 also confirms that (i) there is no unverified sharing when prevalence is too high (so any sharing after that point, if any, is verified) and (ii) for any reduction in verification costs t, there are always regions of sufficiently low prevalence that do not respond to it.

The previous result suggests that verification costs must decrease sufficiently low into the feasibility region to affect equilibrium outcomes. The next result confirms this intuition, and Figure 3 depicts a typical situation.

Proposition 3. *There exists* $t' \in (0, \frac{b(\bar{v})}{1+b(\bar{v})/\ell(\bar{v})})$ *such that equilibrium prevalence* π^* *and unverified sharing* σ^* *can be reduced if and only if* $t < t'$.

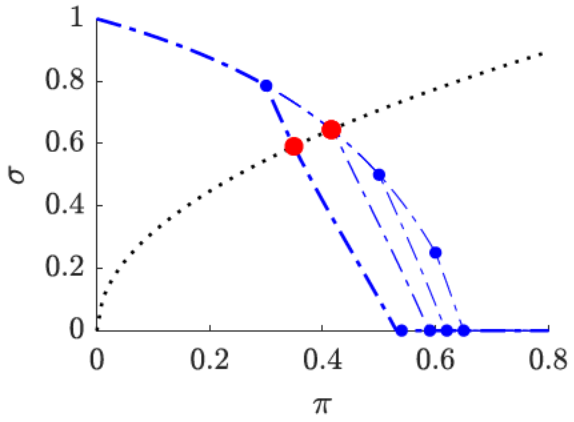

Figure 3: Parameter values: $\bar{v} = 1$, $G(v) = v$, $F(c) = c^2$, $b(v) = v^2$, $\ell(v) = v/2$. The unverified sharing locus $\Sigma(\pi; t)$ is depicted for verification cost $t \in \{0.1, 0.2, 0.25, 0.3\}$, moving from thick to thin dashed lines, and the equilibrium (π^*, σ^*) is affected only when $t < 0.2$.

To the best of our knowledge, there is scarce evidence on the effectiveness of fact-checking improvements as a tool for lowering the verification costs borne by users. One exception is Friggeri et al. (2014), who examine users who are granted access to chains of comments containing evidence that a news item is false. They argue that fact-checking can fail because "not all comments [are] read by users before sharing, either due to lack of interest, or because

other, more recent comments are more easily viewable." From this perspective, our analysis points at the importance of evaluating such cost considerations in relation to users' perception of the severity of the fake news problem: it is the interplay of verification costs vis-à-vis perceived prevalence that dictates the size of any successful intervention.

Decreasing losses ℓ and the general case. One may feel tempted to conclude that verification always has an effect when π high. However, focusing only on prevalence neglects how losses and benefits vary across users. To illustrate, we examine losses that are decreasing.[11]

Consider Figure 4, where the left panel depicts an increasing propensity-to-share function, as in Figure 2, and two propensity-to-skip-verification functions t/ℓ that are increasing and straight. For high prevalence levels (vertical axis), only high types, if any, would be willing to share news articles. Unlike in the previous case, however, those types now face low losses from sharing fake content. In the left panel, the reduction in verification costs (downward shift in the straight line) does not induce types $v \geq 0.75$ to engage in verification.

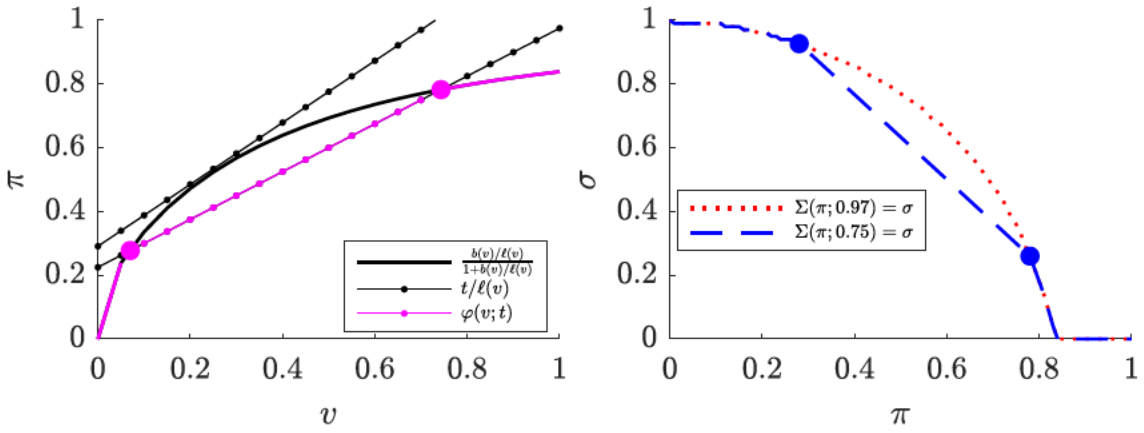

Figure 4: Decreasing losses. Parameter values: $\bar{v}=1$, $G(v)=v$, $b(v)=4\sqrt{v}$, $\ell(v)=1/(0.3+v)$, and $t \in \{0.75, 0.97\}$; in particular, $t/\ell(v)$, is increasing, and so is φ (left panel). Reducing t lowers the sharing rate of unverified news $\Sigma(\pi;t)$ only for intermediate value of π (right panel).

The result is that Σ is insensitive to t for high prevalence levels: in the right panel, both functions coincide after $\pi = 0.8$. Overall, verification impacts behavior only for *intermediate prevalence levels*, because only then are intermediate types willing to share; but since they face higher costs relative to high types, they benefit more from engaging in news verification. Users again segment into three groups, with high types now engaging in unverified sharing.

In the general case, multiple crossing points between both propensity functions can arise. As a result, (i) there could be multiple disconnected intervals of verification, or (ii) existing intervals could feature more types engaging in verification.

[11] Losses can be decreasing if they are in monetary terms, and higher types—e.g., those with more income—have better defense mechanisms (e.g., legal protection).

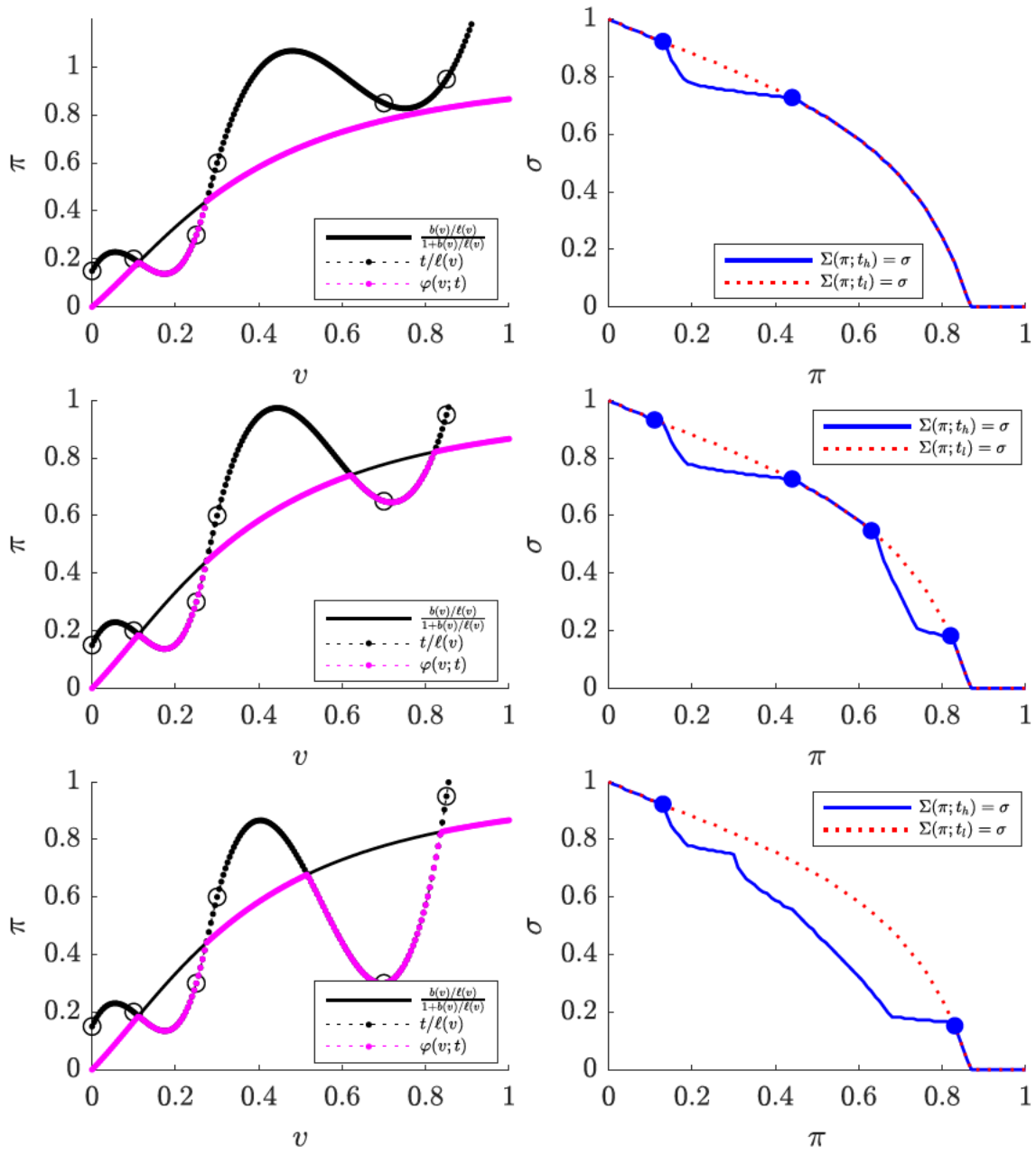

Figure 5: Non-monotone losses. Multiple intersections could lead to multiple disconnected verification regions (middle panel), and existing ones could exhibit more verification and become connected (bottom panel).

Without loss of generality, Figure 5 preserves the monotonicity of the propensity-to-share function but introduces a non-monotonic t/ℓ: in the top row, two crossing points for low prevalence levels (left panel) generate a nonlinear analog of Figure 4 (right panel); in the middle row, with a second crossing point, a second disconnected region of sensitivity appears for high π; and in the bottom panel, as high types exhibit larger losses and t/ℓ decreases

more, both regions connect and a subset of the original one exhibits more verification.[12]

Having understood how the propensity functions shape users' incentives to verify, we next examine policies other than reductions in verification costs, and how they may interact with the latter. To avoid confounding effects, therefore, it is convenient to fix a specification for the family of unverified sharing functions parametrized by t, $\{\Sigma(\cdot;t) : t \geq 0\}$. For consistency with our leading example in this section, we consider the case of increasing losses ℓ hereafter.

5 Supply Interventions and Fake News Diffusion

Among the variety of responses to combat fake news, platforms have attempted to reduce its production. This section uncovers when and why this practice can have the downside of increasing the *diffusion* of fake news.

Diffusion of fake news. As a starting point, note that $\Pi(\sigma) = F(\sigma)$ is effectively a traditional supply function. Thus, interventions that limit the supply of fake news can be modeled as standard *supply shifts* that capture a weakening in producer's incentives. From this perspective, banning repeat offenders can be seen as a cost for producing fake news articles $f \in (0,1)$ that is orthogonal to that of content generation (e.g., change of identity and website appearance to pass screening tests), resulting in a new supply

$$\Pi^f(\sigma) = F(\sigma - f).$$

Alternatively, curtailing fake news producers' ability to attract advertisers results in an overall lower return after sharing occurs (e.g., because fewer advertisers place ads on untrustworthy websites), which can be captured via a scalar $\alpha \in (0,1)$ such that

$$\Pi^\alpha(\sigma) = F(\alpha\sigma).$$

As in standard competitive analyses, the prevalence of fake news decreases after a left shift of the supply curve. However, by moving upward along the "demand curve," the pass-through rate increases: with a lower prevalence of fake news, fewer users are willing to verify. Figure 6 depicts this phenomenon.

[12]The insensitivity of $\Sigma(\pi;t)$ to t for sufficiently low prevalence rates π is a generic property in the model because ℓ is finite, so $t/\ell > 0$ for all $t \geq 0$. Thus, for sufficiently low values of π, namely, $t/\ell(\cdot) > \pi$, no user type has incentives to engage in news verification.

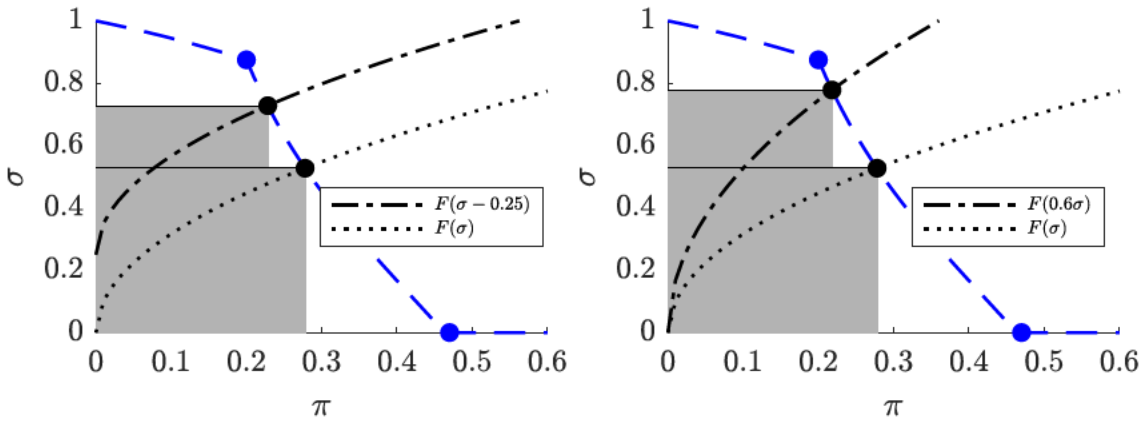

Figure 6: Left panel: $\Pi(\sigma - f)$, $f = 0.25$. Right panel: $\Pi(\alpha\sigma)$, $\alpha = 0.6$. Both panels use the parametrization of Figure 1 with $t = 0.1$.

Increased diffusion. The novelty is that the shaded areas in Figure 6 capture a key measure of interest: the *diffusion rate* of fake news articles,

$$\Delta^* \equiv \pi^* \sigma^*,$$

or the expected number of fake news articles that are transmitted downstream, which is a measure of misinformation dissemination in our model.[13] In the remainder of this section, we seek conditions under which a contraction in the fake news supply leads to an increase in diffusion Δ^*, exactly as depicted in the shaded areas in Figure 6.

Our analysis highlights sensitivity measures analogous to traditional elasticities that are key in this respect. Specifically, because supply interventions shift Π along Σ, it is natural to define the function $\Delta(\pi) := \Sigma(\pi; t)\pi$ which, under sufficient regularity, increases if and only if $\partial\Delta(\pi)/\partial\pi < 0$. Equivalently, when sharing function Σ is (absolutely) *elastic*:

$$|\mathcal{E}_\pi(\Sigma)| \equiv \left| \frac{\partial \log(\Sigma)}{\partial \log \pi} \right| > 1.$$

To find conditions such that this inequality holds in equilibrium, we first impose sufficient regularity on the benefits and losses from sharing news.

Assumption 2. *The benefit $b(\cdot)$ and loss $\ell(\cdot)$ functions are twice differentiable.*

Second, we denote by $\Sigma_V(\pi)$ and $\Sigma_N(\pi)$ the mass of users who share *verified* news and who do *not* share, respectively, given prevalence π. Consequently, $\Sigma(\pi) + \Sigma_V(\pi) + \Sigma_N(\pi) = 1$

[13] This variable is especially relevant when those who first come in contact with fake news articles have many followers. Further, Serra-Garcia and Gneezy (2021) found that people may be more inclined to believe a news item if one was previously shared, which could exacerbate the initial sharing of misinformation.

and, by Proposition 2, $\Sigma_V(\pi) = 1 - G(v_1(\pi))$ and $\Sigma_N(\pi) = G(v_0(\pi))$ whenever $\Sigma(\pi) > 0$. The subsequent result leverages these relationships.

Proposition 4. *Consider an equilibrium (π^*, σ^*) and suppose that one of the following conditions holds:*

(a) The propensity to share $b(v)/\ell(v)$ is concave, the density $g(v)$ is increasing, and the pass-through of fake news obeys $\sigma^* \leq \Sigma_N(\pi^*)$;

(b) The function $1/\ell(\cdot)$ is concave, the density $g(v)$ is decreasing, and the pass-through of fake news obeys $\sigma^* \leq \Sigma_V(\pi^*)$.[14]

Then, the unverified sharing function Σ is elastic, i.e. $|\mathcal{E}_\pi(\Sigma)| > 1$; thus, a supply intervention (leftward shift of supply Π) will lead to a greater diffusion of fake news Δ^*.

Finding conditions that ensure an elastic response of Σ to changes in π is challenging. This is because Σ depends in general on thresholds v_0 and v_1 via $\Sigma(\pi) = G(v_1(\pi)) - G(v_0(\pi))$ in general, and these thresholds depend on primitives in complicated ways. We bypass this obstacle by exploiting the fact that, whenever $\Sigma(\pi; t) > 0$, the identity

$$|\mathcal{E}_\pi(\Sigma)| = \left(\frac{\Sigma_N}{\Sigma}\right) \mathcal{E}_\pi(\Sigma_N) + \left(\frac{\Sigma_V}{\Sigma}\right) \mathcal{E}_\pi(\Sigma_V) \tag{7}$$

can be used to provide a lower bound for $\mathcal{E}_\pi(\Sigma)$.

For example, consider condition (a). Using (7), we can show that a 1% increase in prevalence π decreases unverified sharing Σ by at least $\mathcal{E}_\pi(\Sigma_N)$%, provided the equilibrium pass-through σ^* is no greater than $\Sigma_N(\pi^*)$. Moreover, the conditions on primitives imply that $\Sigma_N(\cdot)$ is an elastic convex function, thereby ensuring an elastic response of $\Sigma(\cdot)$ in equilibrium. Specifically, when the propensity to share b/ℓ is concave in v, then the first type willing to share, $v_0(\pi)$ in (6), increases at decreasing rates in π. Thus, an increase in prevalence π increases the mass of users who do not share news, Σ_N, more when prevalence is high than when it is low, provided the density g is increasing.[15] Part (b) follows a similar logic that instead uses $\Sigma_V(\cdot)$.

[14] If the equilibrium prevalence is in a region in which no active verification occurs, i.e., $\Sigma_V(\pi^*) = 0$, only condition (a) applies as (b) is never satisfied.

[15] For an illustration, consider $b(v) = v^2$, $\ell(v) = v$, and a uniform density $g(v) = 1/\bar{v}$, which corresponds to Figure 1's parametrization and satisfies the conditions given in Proposition 4-(a). For mid-range prevalence levels, $v_0(\pi) = \pi/(1-\pi)$ and $v_1(\pi) = t/\pi$, and so it can be checked that $\Sigma(\pi) \leq \Sigma_N(\pi)$ if and only if $4\pi \geq -t + \sqrt{t^2 + 8t} \in [0, 1]$. The right-hand side of this inequality decreases with t. Moreover, for $t = 0.1$ the right-hand side equals 0.2; hence, Σ is elastic for all $\pi \in (0.2, 0.5)$. Thus, as depicted in Figure 6, if equilibrium has active verification, i.e., $\pi^* \in (0.2, 0.5)$, then $\sigma^* \leq \Sigma_N(\pi^*)$. Hence, by Proposition 4-(a), a leftward supply shift leads to a greater diffusion of fake news even if the content creation is effectively reduced.

Two observations about Proposition 4 are instructive. First, while the conditions on the primitives b, ℓ and g hold pointwise, and hence are strong, they have the advantage of being easy to check. Critically, they can be transformed to local conditions if mild extra regularity on b/ℓ is imposed (see Propositions A.1-A.2 in the Appendix).

Second, the conditions $\sigma^* \leq \Sigma_N(\pi^*)$ and $\sigma^* \leq \Sigma_V(\pi^*)$ are endogenous, in a reflection of both sides of the market non-trivially adjusting to potential interventions, and hence encoding the importance of examining users' and producers' decisions jointly. Equally important, these conditions have the advantage that they rely on observable variables that can be obtained from the data; further, observe that they become *weaker* as sharing news articles without verifying them becomes less common.

The latter observation brings us to the next topic: examining how verification costs t impact the elasticity of unverified sharing, $|\mathcal{E}_\pi(\Sigma)|$. Intuitively, although reducing verification costs can lower the pass-through of fake news (Proposition 3), it can also make the unverified sharing function Σ *more elastic*. We next provide conditions for this to happen.

Proposition 5. *Suppose that hazard rate $g(v)/(1-G(v))$ is decreasing and that $\ell'(v)/\ell(v)$ is increasing. Then, unverified sharing elasticity $|\mathcal{E}_\pi(\Sigma)|$ rises as verification cost t falls.*

To understand Proposition 5, consider the elasticity identity (7). An increase in verification cost t has no impact on the $v_0(\pi)$ (see (6)), and so no impact on $\Sigma_N = G(v_0(\pi))$. However, t (weakly) lowers the amount of both unverified and verified sharing, $\Sigma(\cdot)$ and Σ_V. Thus, increasing t lowers the unverified sharing elasticity $\mathcal{E}_\pi(\Sigma)$ whenever it lowers the elasticity measure $\mathcal{E}_\pi(\Sigma_V)$. The conditions on primitives guarantee this result.[16]

Together, Propositions 4 and 5 can inform whether *joint policies* can reinforce each other *negatively*. Specifically, by making unverified sharing behavior more sensitive to reductions in prevalence, fact-checking can result in supply interventions acting as a catalyst for increasing the diffusion of fake news in a network.[17]

[16] Consider the last user type who shares without verifying—$v_1(\pi;t)$ defined in (6)—which increases with lower prevalence π or higher verification cost t. If sharing losses $\ell(\cdot)$ increase relatively more for high types (i.e., $(\ell'/\ell)' \geq 0$), then an increase in prevalence decreases v_1 less when verification cost t is high (i.e., $\partial^2 v_1(\pi;t)/(\partial\pi\partial t) \geq 0$). Thus, elasticity $\mathcal{E}_\pi(\Sigma_V)$ decreases as t increases if, additionally, distribution G has a decreasing hazard rate, since $\Sigma_V(\pi;t) = 1 - G(v_1(\pi;t))$.

[17] Notice that if $1/\ell(v)$ is concave and $g(v)/(1-G(v))$ is decreasing, then the conditions given in both Propositions 4-(b) and Proposition 5 hold. Likewise, the conditions given in Proposition 5 do not exclude the conditions given in the weaker form of Proposition 4-(a) in the Appendix (Proposition A.1).

6 Market Power

In this section, we explore the extent to which market outcomes are affected by the possibility of concentration on the production side.[18] There are two takeaways in this section. First, due to inherent informational asymmetries, traditional "uniform pricing" under market power need not operate as smoothly as with traditional goods. Second, segmentation strategies that appear unsophisticated can in fact be profitable for fake news producers.

Uniform policies. A monopolistic fake news producer must decide the mass $\pi \in [0,1]$ of fake news to be produced. For consistency with our previous "competitive" analysis, the monopolist's cost structure is given by distribution F: the monopolist is either of a single producer who experiences marginal cost $F^{-1}(\pi)$ at level π or it corresponds to a "parent" company who owns a large number of smaller producers with costs distributed according to F—in this case, only those producers with $c \leq F^{-1}(\pi)$ would be active.

If the pass-through of fake news takes value σ when a mass π of news has been produced, the monopolist's profits are then given by

$$\sigma \cdot \pi - \int_0^{\pi} \Pi^{-1}(\pi')d\pi', \tag{8}$$

where our notation $\Pi(\cdot) = F(\cdot)$ represents the supply of fake news.[19] As usual, profits are given by the area between σ and the supply locus over $[0,\pi]$. (The appearance of an inverse function Π^{-1} is due to our convention $\pi = \Pi(\sigma)$.)

In traditional market power with uniform pricing, we would use the fact that $\sigma = \Sigma(\pi)$ to optimize (8) and find an optimum; call it π^M. Since Σ is downward sloping, $\pi^M < \pi^*$, as usual. Implicitly in this logic, however, is that the monopolist can move along Σ which, in traditional markets, is trivially guaranteed by the observability of the price. The situation is more subtle in our context: neither σ nor π are directly observable by users.

This informational asymmetry effectively makes the situation one of simultaneous moves. A simple inspection of (8) shows that the equilibrium of Section 3 is unchanged: (σ^*, π^*) is the only tuple consistent with sequentially rational behavior by the monopolist taking σ as given, and assuming that users have correct beliefs. Policies that foster platform transparency,

[18] From an institutional standpoint, there are at least three reasons that justify such an examination. First, supply interventions (previous section) pressure the industry to shrink over time. Second, the costs of producing digital fake content are largely fixed, and they increase as technologies become more sophisticated (e.g., "deep fakes"). Finally, it has been documented recently the existence of parent companies that own several untrustworthy websites (Sydell, 2016).

[19] In the parent company interpretation, total costs obey $\int_0^{F^{-1}(\pi)} cF(dc)$, but the change of variables $c(\pi') = \Pi^{-1}(\pi')$ yields $\int_0^{F^{-1}(\pi)} cf(c)dc = \int_{\pi}^{0} \Pi^{-1}(\pi')[f(\Pi^{-1}(\pi'))/f(F^{-1}(\pi'))]d\pi = \int_0^{\pi} \Pi^{-1}(\pi')d\pi'$, as desired.

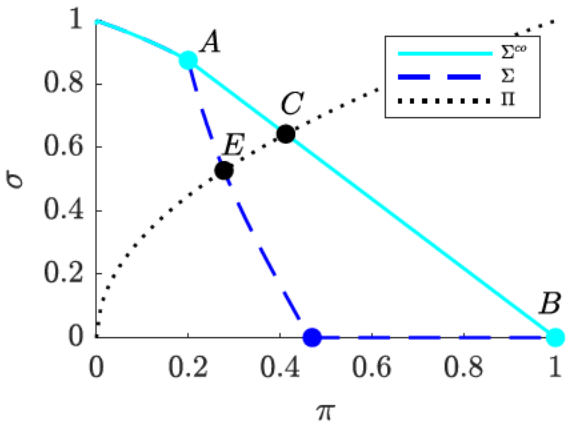

Figure 7: User segmentation as the concave closure of the unverified sharing function Σ. The figure considers the parametrization of Figure 1 with verification cost $t = 0.1$. By segmenting the user population, the monopolist can reverse the effects of verification.

such as revealing π estimates to users, can then enable the monopolist to move along Σ. Conversely, platform *opacity* can be useful if the goal is to harm "large" fake news producers by lowering their profits.

Segmentation strategies. Suppose that the monopolist is at (σ^*, π^*) due to her inability to convey changes in π in real time. Can she do better than in the competitive outcome?

Consider Figure 7, where competitive outcome (σ^*, π^*) corresponds to the intersection between the dotted increasing supply $\Pi(\cdot)$ and the dashed decreasing unverified sharing function $\Sigma(\cdot)$. Importantly, Figure 7 also depicts the *concave closure* of the unverified sharing function, $\Sigma^{co}(\cdot)$, namely, the infimum among all the concave functions weakly above Σ. Clearly, if the intersection point between Σ^{co} and Π is achievable, then it yields more profits than (σ^*, π^*), given the area interpretation of profits behind (8).

We now turn to the question of the *implementation* of points in Σ^{co}. For this purpose, note that points A and B are key in that they generate nontrivial pairs (π, σ) such that $\Sigma^{co} \neq \Sigma$. We denote their corresponding prevalence levels π_A and π_B.[20] For a given $\lambda \in [0, 1]$, the goal then is to implement $\Sigma^{co}(\pi) = \lambda\Sigma(\pi_A) + (1 - \lambda)\Sigma(\pi_B)$ in the segment $\overline{AB}$ using exactly $\pi = \lambda\pi_A + (1 - \lambda)\pi_B$ fake news articles, as any other point on the locus Σ^{co} satisfies $\Sigma^{co} = \Sigma$. Supporting these convex combinations are, however, *trivial* segmentations of the market, namely, subpopulations that differ from the original only in terms of their sizes. The monopolist could then aim to create two submarkets, A and B, with intended prevalence levels π_A and π_B understood as the ratio of the fake to the total content in each. In turn, this procedure can be executed by appropriately (i) distributing news across segments and

[20]This property is generic. By Carathéodory's Theorem (e.g., Theorem 17.1 in Rockafellar, 1970), the concavification of Σ is the convex combination of sharing rates Σ for, at most, *two* prevalence levels.

(ii) fine-tuning the segment sizes.

Specifically, suppose that the sizes of segments A and B are ω and $1-\omega$, respectively. Further, suppose that $\lambda\pi$ fake news articles are sent to segment A, and $(1-\lambda)\pi$ fake news articles to B. All else equal, $(1-\pi)\omega$ and $(1-\pi)(1-\omega)$ truthful news articles would reach segment A and B, respectively. However, choosing segment sizes is only partially successful: segment A requires having sufficiently low prevalence compared to the original market (i.e., $\pi_A < \pi$), which is not possible by resorting to fake content exclusively. If additional τ truthful news articles are supplied to segment A, margins τ and ω could be chosen so

$$\pi_A = \frac{\lambda\pi}{\lambda\pi + (1-\pi)\omega + \tau} \qquad \text{and} \qquad \pi_B = \frac{(1-\lambda)\pi}{(1-\lambda)\pi + (1-\pi)(1-\omega)}$$

hold. That is, the monopolist can always induce high prevalence in a submarket by shrinking its size relative to a fixed supply of fake content. However, as a byproduct of the process, the size of the other submarket becomes fixed; the monopolist can then reduce the prevalence in this latter submarket by making truthful content more abundant there.[21]

Proposition 6. *Suppose the monopolist can target identical subpopulations and supply them with additional truthful news. Then, any point on the concave closure of unverified sharing function Σ can be implemented.*

The next question we ask is what is the best production level that can be implemented, call it π^{**}. To this end, notice that under this type of segmentation, the revenue exclusively associated with the production of π fake news satisfies $\lambda\pi\Sigma(\pi_A) + (1-\lambda\pi)\Sigma(\pi_B) = \Sigma^{co}(\pi)\pi$. Given the informational asymmetries discussed in relation to uniform policies, the segments must correctly anticipate the monopolist strategy and, in particular, the total production π^{**} to be implemented in equilibrium. Thus, the monopolist's per unit revenue, $\Sigma^{co}(\pi^*)$, is also fixed in her objective. If the monopolist cares only about the benefits and costs from fake news production, she will choose π to maximize

$$\Sigma^{co}(\pi^{**})\pi - \int_0^{\pi} \Pi^{-1}(\pi')d\pi'. \tag{9}$$

Sequential rationality and correct beliefs then yield an equilibrium characterization of $\Sigma^{co}(\pi^{**}) = \Pi^{-1}(\pi^{**})$, i.e., to a version of the competitive equilibrium condition now involving the concave closure of the unverified sharing function Σ (point C in Figure 7). This equilibrium entails more creation, diffusion, and sharing of fake news articles.

[21] Alternatively, as λ becomes fixed in the first state in the implementation exercise, one degree of freedom remains, namely, ω,to target two endogenous variables, π_A and π_B. Furthermore, with additional τ truthful news articles being supplied, the fake news prevalence in the *whole platform* becomes $\pi/(1+\tau)$.

Proposition 7. *The segmentation that implements π^{**} leads to more creation, diffusion, and sharing of fake news articles compared to the competitive case. Further, if the producer cares only about fake content, this segmentation is optimal among the ones studied.*

We conclude with three observations. First, the types of segmentations behind this strategy are rather basic, as they condition only on size.[22] Additionally, given the inherent free-riding problem to which public information is subject, the cost of producing truthful content by an unverified news provider is fairly low. A payoff criterion based exclusively on false information is then a good approximation when either truthful content diffuses slowly compared to fake news,[23] or when the outlet is malicious in that it cares only about the spread of fake content.

Second, this type of segmentation is not equivalent to randomizing between two prevalence levels applied to the whole population: π^{**} must be produced with probability 1 so the cost of the last unit of fake news produced is exactly $\Pi^{-1}(\pi^{**})$. Similarly, the optimization over π in (9) considers deviations from total production π^{**} but not from ways of "splitting" π^{**}. That is, this implementation requires the monopolist to overcome the temptation to send its production to the segment with the highest sharing rate.[24]

Finally, our analysis in Section 4 shows that lowering verification costs for users materializes in the unverified sharing locus decreasing only in subregions of prevalence. In particular, *convexities* in Σ can be created or exacerbated—as seen in Figure 3. However, it is exactly in those regions that the described segmentation strategies become profitable. From this perspective, our analysis identifies simple forms of segmentation as profitable to fake news producers. This is an important observation in light of the sometimes high degree of granularity that certain platforms' targeting services offer in practice: very basic forms of segmentation need not be associated with negligible harm.

7 Internal Filters

We now enrich the model to allow for detection algorithms. A key concern regarding such platform *filters* has been their potential use for removing content before it reaches users, a practice that some studies document can be perceived by individuals as a form of censorship

[22] And they can be easily implemented, as population sizes are easy to control when targeting online in platforms; for instance, by setting different monetary budgets in (similar) locations of interest.

[23] Vosoughi et al. (2018) found that fake news spread faster, deeper and broader than true news.

[24] This situation is analogous to the standard commitment assumption widely used in the persuasion/information design literature (e.g., Kamenica and Gentzkow, 2011).

(e.g., Lazer et al., 2018). In contrast, we offer an economic rationale for the cautious use of such algorithms based on users' verification incentives.

We consider the case in which an algorithm screens news articles imperfectly as they enter the platform, and before they reach consumers. Clearly, eliminating truthful news articles carries social costs. Thus, we focus on the more interesting case in which truthful news articles always survive, but fake news articles are detected with probability $\phi \in [0,1]$. Because of the public announcements that platforms have made on this topic, we assume that changes in ϕ, which measures the filter quality, are observable to both users and producers. Also, we focus on the effects of *introducing* such filters, captured by increasing ϕ from zero.

Our analysis from Section 3 admits a direct adaptation to this case. To this end, suppose that a mass π of news enters the platform. By Bayes' rule, the *posterior chance* that a user identifies a news item as false upon encountering it is given by

$$\psi(\pi;\phi) \equiv \frac{(1-\phi)\pi}{1-\phi\pi},$$

which is the right measure of fake news prevalence in this context. This variable decreases as ϕ increases and as π decreases.

The methods from Section 3 then admit minimal modifications. Specifically, the set of users who share unverified news is now $\mathcal{V}(\psi(\pi,\phi);t)$ (see (3)), while the unverified sharing locus becomes $\Sigma(\psi(\pi;\phi))) = \int_{\mathcal{V}(\psi(\pi,\phi);t)} dG(v)$. However, from the producers' perspective, the relevant variable is the *pass-through of fake news*, which also incorporates the filter's effect:

$$(1-\phi)\Sigma(\psi(\pi;\phi))). \tag{10}$$

Each potential fake news producer then takes as given the (candidate) equilibrium pass-through rate, $\sigma \in [0,1]$, which leads to a supply curve $\Pi(\sigma) = F(\sigma)$ as in (5). Also, as in Section 3, the unique equilibrium (π^*, σ^*) is given by the intersection of the pass-through of fake news and the supply curve, i.e., $\Pi((1-\phi)\Sigma(\psi(\pi^*;\phi))) = \pi^*$ and $\sigma^* = (1-\phi)\Sigma(\psi(\pi^*;\phi)))$.

Equipped with this reformulation, we establish the conditions under which increasing the filter precision could lead to more creation and diffusion of fake news.

Proposition 8. *Suppose that $b(v)/\ell(v)$ is concave and the density $g(v)$ is increasing. If $\sigma^* \leq \Sigma_N(\pi^*)$ when $\phi = 0$, then, as ϕ increases, both the equilibrium prevalence, π^*, and the rate of diffusion, Δ^*, initially increase and then, eventually decrease. The pass-through σ^* increases with ϕ.*

The presence of a filter implies that encountering content is "good news" from the perspective of any user, i.e., the user is now more optimistic about the veracity of the news

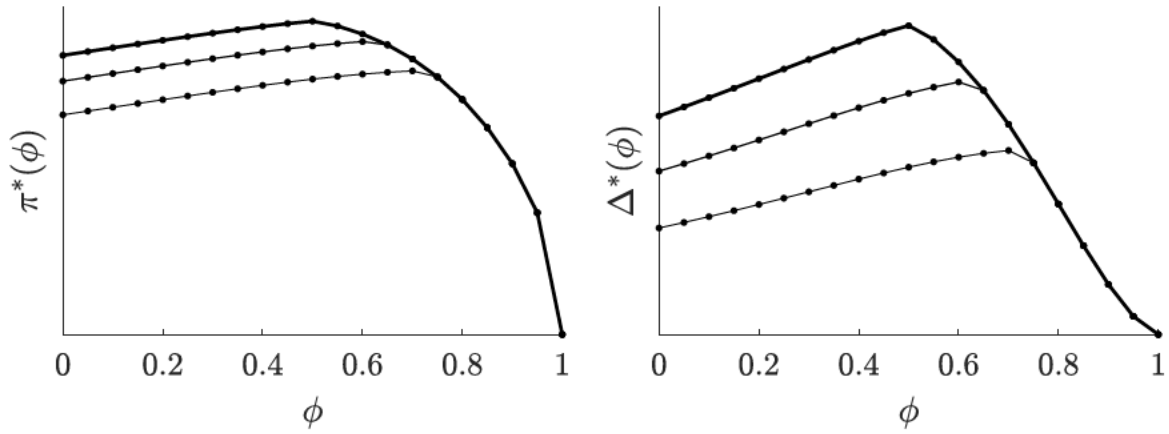

Figure 8: Both panels consider the parametrization of Figure 1 but with $F(c) = \sqrt{c}$. Verification costs range from $t = 0.1, 0.15, 0.2$ moving downwards from thick to thin curves.

item, encapsulated in $\psi(\pi, \phi) < \pi$ for all $\pi \in (0, 1)$. With greater optimism, more unverified sharing occurs conditional on the encounter taking place, as reflected in the set of users, $\mathcal{V}(\psi(\pi, \phi))$, increasing (in the sense of set inclusion) with ϕ. In other words, the users' verification incentives *relax*. Importantly, this latter effect can outweigh the negative direct effect that the filter has on the incentives to produce news—term $1 - \phi$ in the pass-through (10)—resulting in both a higher prevalence and a larger diffusion of fake news. Depending on primitives, this effect can prevail over extensive regions of filter quality, as seen in Figure 8.

A sufficiently elastic Σ again underlies our finding. Indeed, a marginal increase in the accuracy of the filter from $\phi = 0$ increase the equilibrium prevalence π^* if and only if the pass-through increases for each π, i.e., $\partial[(1 - \phi)\Sigma(\psi(\pi; \phi))]/\partial\phi|_{\phi=0} > 0$. Simple calculations show that this condition holds when the unverified sharing curve Σ is *sufficiently elastic*, or

$$|\mathcal{E}_\pi(\Sigma)| > \frac{1}{1 - \pi}.$$

Claim A.1 in the Appendix shows that this inequality holds under the conditions in Proposition 8. The logic is similar to that of Proposition 4: using (7) and $\sigma^* \leq \Sigma_N(\pi^*)$, it follows that $\mathcal{E}_\pi(\Sigma) \geq \mathcal{E}_\pi(\Sigma_N)$; the conditions on primitives then ensure that Σ_N is sufficiently elastic.

Finally, by the same logic given in Section 5, a decrease in verification cost t could make the unverified sharing function Σ more elastic (Proposition 5), thereby magnifying the unintended effects of news filtering through the interaction of these two policies.

8 Network Externalities

The value of a platform stems partly from how many other individuals use it. Further, social media platforms are natural venues where individual behavior can be influenced by others. Large number of users, publicly available information, and possibilities of imitating others all render network externalities a likely factor shaping the gains and losses that users experience by sharing news on social media platforms.

This section shows how elements of "social influence" can lead to nontrivial effects on the sharing of unverified news. To this end, we follow the approach of Becker (1991) by allowing individual choices to depend on aggregate variables. Specifically, we now consider the case of losses given by

$$\tilde{\ell}(v,\sigma) := \frac{\ell(v)}{n(\sigma)},$$

where σ corresponds to the mass of users sharing unverified news, while n is a differentiable function satisfying $n' > 0$ and $n(0) = 1$. That is, as a larger mass of users shares without verifying, the loss that each type v suffers from sharing fake content decreases (e.g., because it is easier to blame the situation on others). Other examples can be studied too.

To isolate how a relaxation of verification incentives affects the sharing of unverified news, we simply assume the benefits of sharing truthful news articles $b(\cdot)$ scale in the same manner; in this way, the propensity to share remains unchanged.[25] Specifically, recalling Section 3, the mass of users who share without verifying obeys $\tilde{\Sigma}(\pi;\sigma) \equiv G(v_1(\pi;\sigma)) - G(v_0(\pi))$, where

$$v_0(\pi) = \inf\left\{v \in [0,\bar{v}] : \frac{b(v)/\ell(v)}{1+b(v)/\ell(v)} \geq \pi\right\} \quad \text{and} \quad v_1(\pi;\sigma) = \inf\left\{v \in [0,\bar{v}] : \frac{tn(\sigma)}{\ell(v)} \leq \pi\right\}.$$

Our normalization then implies that all the changes occur via the margin $v_1(\pi;\sigma)$, which adjusts because of its explicit dependence on σ.

In equilibrium, users' beliefs about the pass-through of fake news, σ, must be correct, so the sharing rate of unverified news, $\Sigma(\pi)$, must solve the fixed point $\tilde{\Sigma}(\pi;\sigma) = \sigma$ for each π, just as the aggregate quantity demanded must satisfy a fixed-point in traditional models of network externalities. The top panels of Figure 9 illustrate a typical situation: the left panel plots the fixed points for a given value of π, while the right panel, the resulting Σ, which now becomes a *correspondence*.[26]

[25] One possibility is that, as users expect unverified sharing behavior to become more frequent, they anticipate other users eventually leaving the platform. With fewer users, the value of sharing truthful content is likely to decrease. In this interpretation, dividing b by n acts as a penalty associated with those long-term losses.

[26] Our correspondence is indeed a nonmonotonic function as in Becker (1991), but in the reversed coordinate system: we look for fixed points on the vertical axes while he does so on the horizontal one. A similar

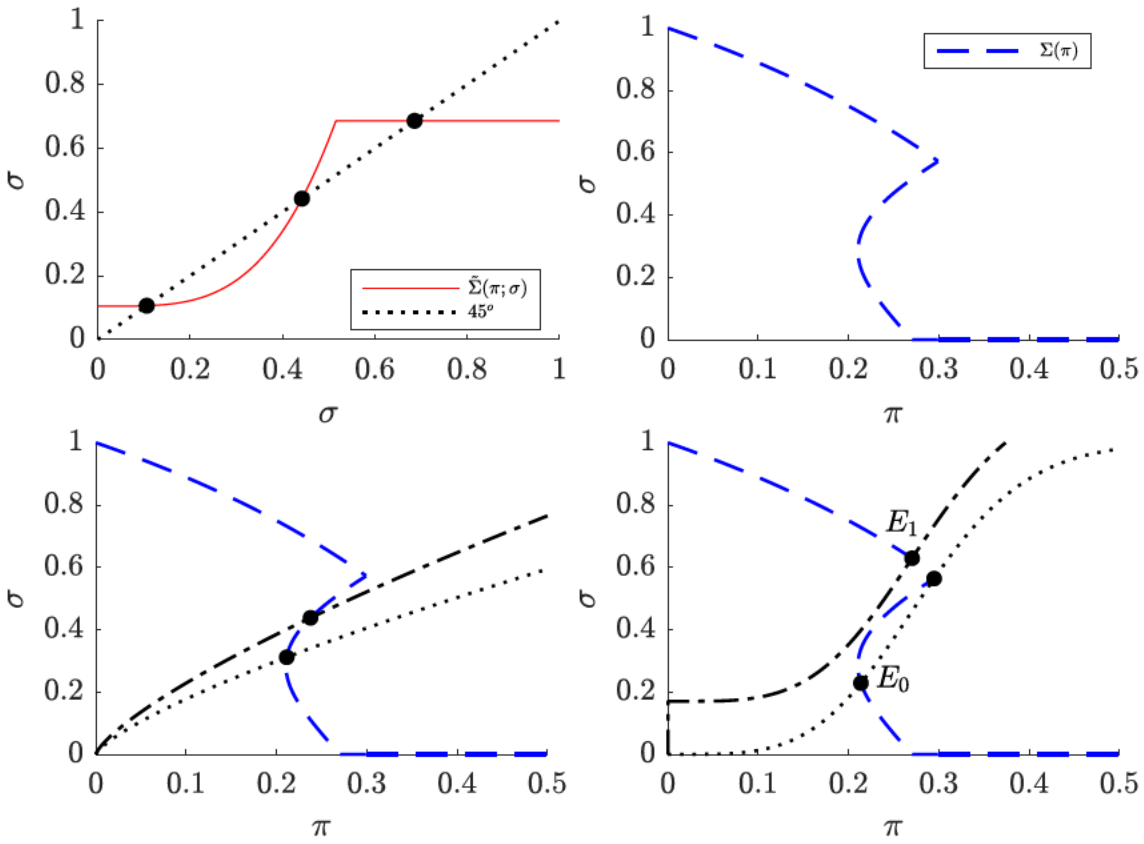

Figure 9: Top left: fixed points of $\tilde{\Sigma}(\pi;\cdot)$ for prevalence $\pi = 0.25$. Top right: unverified sharing rate locus $\Sigma(\pi)$. The bottom panels examine supply interventions: the baseline supply falls to $\Pi(\sigma) = F(0.78\sigma)$ in the left panel with $F(c) = c^{4/3}$; while in the right panel, the supply falls to $\Pi(\sigma) = F(\sigma - 0.17)$ with F being the CDF of a beta distribution with shape parameters $(0.25, 0.1)$. Parameter values: $\bar{v} = 1$, $G(v) = v$, $b(v) = v^2$, $\ell(v) = v$, $t = 0.1$, and $n(\sigma) = 1 + 10\sigma^4(3 - 2\sigma)$.

Key to our analysis is the presence of regions in which networks effects dominate, resulting in Σ being *increasing*: higher prevalence levels can be consistent with larger unverified sharing rates because users' expectations of high unverified sharing behavior reduces verification incentives, making those expectations self-fulfilling. Supply interventions intended to reduce the supply of fake news, as in Section 5, can then backfire: in the bottom left panel, such interventions lead to a higher prevalence and pass-through of fake news, and hence to a larger diffusion rate Δ. Furthermore, in the bottom right panel, these interventions are shown to have adverse "refining" effects by shrinking the set of equilibria to a single "bad" equilibrium as the unique possible stable outcome (transition from E_0 to E_1 in the figure).

phenomenon arises in Kranton and McAdams (2020) in which users are embedded in a network structure.

9 Concluding Remarks

This paper develops a model of fake news creation and diffusion intended to examine the efficacy of real-world policies deployed to combat misinformation, while stressing how these may interact with users' incentives to verify news. A distinctive aspect of our approach is that it is operationalized in a *framework* featuring forces akin to those of supply and demand. From this perspective, our work emphasizes the importance of sensitivity analyses for assessing policy interventions, and highlights key nuances relative to traditional markets. Next, we briefly discuss some of our assumptions and future work.

First, the static nature of the model is clearly a simplification. While news transmission is a dynamic process, the proportion of individuals who share without verifying remains key from the perspective of producers. Moreover, since producers can target users with some degree of granularity, it is reasonable to expect that the discounted benefits of fake news diffusing through a network are particularly sensitive to that "initial" proportion of individuals, in which case the model captures a key variable shaping the supply of fake news.

Second, while users' preferences (b, ℓ) are general, they depict a simplified version of a more general world in which heterogeneity (b, ℓ) is two dimensional. Two observations are instructive in this regard. First, as stated in the introduction, a positive relationship between propensities to share and verify has been documented experimentally: in our setting, this relationship can be modeled with losses ℓ and benefits b increasing across types, with benefits increasing at a faster rate. Second, our analysis exploits, for the most part, the interplay between and a continuous supply and a continuous unverified sharing rate of news. Clearly, the latter would hold with bi-dimensional user heterogeneity (b, ℓ), absent any atoms.

Third, we assume that the user is fully rational, ignoring behavioral/cognitive aspects that could play an important role in this market. That said, a sizable fraction of the efforts by platforms, fact-checking organizations, and journalist associations have been devoted to educational programs aimed at fostering user literacy in evaluating fake news.[27] The rational or "sophisticated" benchmark, therefore, need not be considered too distant.[28]

Finally, our policy analyses have been guided by their practical importance, but others are available. Example include making news transmission more costly (e.g., via additional clicks) or using algorithms that need not remove news articles but can instead route them to different individuals based on past behaviors, or on signals about news that the algorithms select. These and other topics are left for future research.

[27] See Lyons (2018), Guess et al. (2020a), and the News Integrative Initiative at https://www.journalism.cuny.edu/centers/tow-knight-center-entrepreneurial-journalism/news-integrity-initiative/.

[28] A similar trend towards educating consumers has emerged in response to *privacy* considerations. See Bonatti and Cisternas (2020) for an application to price discrimination.

A Omitted Proofs

A.1 Proofs of Section §3

A.1.1 Proof of Lemma 1

Let λ denote the Lebesgue measure, and take an arbitrary continuous function $\varphi : \mathbb{R} \to \mathbb{R}$.

Lemma A.1. *If $\varphi(\cdot;t)$ is differentiable a.e. with $|\varphi'(\cdot;t)| > 0$ a.e, then $\lambda(\varphi^{-1}(\{\pi\})) = 0$ for any $\pi \in \mathbb{R}$.*

Proof: Assume that $\lambda(\{v \in \mathbb{R} : \varphi'(v) = 0 \vee \varphi'(v) \text{ does not exist}\}) = 0$, and consider $\varphi^{-1}(\{\pi\})$ for $\pi \in [0,1]$. It is well-known that the set of isolated points of the previous set, $ISO(\varphi^{-1}(\{\pi\}))$, is countable under the usual topology in $\mathbb{R}$, and hence of Lebesgue measure equal to zero. Consider now a point v^* that it is not isolated and where the derivative exists. Then, since set $\varphi^{-1}(\{\pi\})$ is closed, there exists and approximating sequence $(v_n)_n$ with $\varphi(v_n) = \pi$ for all n, and so $\varphi'(v^*) = 0$. Consequently, the set

$$\left(\varphi^{-1}(\{\pi\}) \setminus ISO(\varphi^{-1}(\{\pi\}))\right) \cap \{v \in \mathbb{R} : \varphi'(v) \text{ exists}\}$$

has Lebesgue measure zero, and hence so does $\varphi^{-1}(\{\pi\})$. □

Proof of Lemma 1: Consider the Lebesgue-Stieltjes measure $B \mapsto \mu(B) \equiv \int_B dG$ for all Borel sets $B \subseteq [0, \bar{v}]$. We'll show that $\Sigma(\cdot)$ is left continuous. Take $\pi \in [0,1]$ and an increasing sequence $(\pi_n)_n$ with $\pi_n \uparrow \pi$ as $n \to \infty$. Since $\pi_n < \pi$, it follows that $\mathcal{V}(\pi) \subseteq \mathcal{V}(\pi_n)$. Thus,

$$\Sigma(\pi_n) - \Sigma(\pi) = \int_{\mathcal{V}(\pi_n)} dG - \int_{\mathcal{V}(\pi)} dG = \int_{\mathcal{V}(\pi_n)\setminus\mathcal{V}(\pi)} dG = \mu(A_n),$$

where $A_n \equiv \mathcal{V}(\pi_n) \setminus \mathcal{V}(\pi) = \{v : \varphi(v) \geq \pi_n \text{ and } \varphi(v) < \pi\}$. Clearly, A_n is measurable, since $\varphi : [0, \bar{v}] \to [0,1]$ defined in (3) is continuous as it is the minimum of two continuous functions. Next, notice that $A_{n+1} \subseteq A_n$ for all $n = 1, 2 \ldots$, namely, $(A_n)_n$ is a decreasing set sequence with $\lim_{n\to\infty} A_n = \{v : \varphi \geq \pi\} \cap \{v : \varphi(v) < \pi\} = \emptyset$. Thus, by continuity of the measure μ, $\lim_{n\to\infty} \mu(A_n) = \mu(\lim_{n\to\infty} A_n) = \mu(\emptyset) = 0$. Consequently, $\lim_{n\to\infty} \Sigma(\pi_n) = \Sigma(\pi)$.

We now show that $\Sigma(\cdot)$ is, in addition, right-continuous. Consider $\pi \in [0,1]$ and a decreasing sequence $(\pi_n)_n$ with $\pi_n \downarrow \pi$ as $n \to \infty$. Since $\mathcal{V}(\pi_n) \subseteq \mathcal{V}(\pi)$, as $\pi_n > \pi$, it follows that

$$\Sigma(\pi) - \Sigma(\pi_n) = \int_{\mathcal{V}(\pi)} dG - \int_{\mathcal{V}(\pi_n)} dG = \int_{\mathcal{V}(\pi)\setminus\mathcal{V}(\pi_n)} dG = \mu(B_n),$$

where $B_n \equiv \mathcal{V}(\pi) \setminus \mathcal{V}(\pi_n) = \{v : \varphi(v) \geq \pi \text{ and } \varphi(v) < \pi_n\}$. Moreover, $B_n \subseteq B_{n+1}$ for all n, i.e., $(B_n)_n$ is an increasing sequence with $\lim_{n\to\infty} B_n = \{v : \varphi(v) = \pi\} = \{v : \varphi \geq \pi\} \cap \{v :$

$\varphi(v) \leq \pi\}$. Thus, $\lim_{n\to\infty} \mu(B_n) = \mu(\lim_{n\to\infty} B_n) = \mu(\varphi^{-1}\{\pi\})$. But, since μ is absolutelty continuous with respect to the Lebesgue measure (λ), and $\lambda(\varphi^{-1}\{\pi\}) = 0$ by Lemma A.1, it follows that $\mu(\varphi^{-1}\{\pi\}) = 0$, and so $\lim_{n\to\infty} \Sigma(\pi_n) = \Sigma(\pi)$. □

A.1.2 Proof of Proposition 1

Consider the composite function $\Sigma \circ \Pi : [0,1] \to [0,1]$. We'll show that $\Sigma(\Pi(\cdot))$ has a unique fixed point. First, it is clear that $\Sigma(\Pi(\cdot))$ is continuous, since it is the composition of two continuous functions. Second, when $\sigma = 0$ we have $\Pi(0) = 0$ and so $\Sigma(\Pi(0)) = 1$. Conversely, when $\sigma = 1$, we have $\Pi(1) = 1$. We now argue that $\Sigma(1) = 0$. Indeed, notice that $\varphi(v) \leq \frac{b(v)/\ell(v)}{1+b(v)/\ell(v)} < 1$ a.e., since $b(v)/\ell(v)$ is finite a.e. Thus, $\mathcal{V}(1) = \{v : \varphi(v) = 1\}$ has Lebesgue measure zero, and so $\Sigma(1) = \int_{\mathcal{V}(1)} dG = 0$. Altogether, by the Intermediate Value Theorem, there exists $\sigma^* \in (0,1)$ with $\Sigma(\Pi(\sigma^*)) = \sigma^*$.

Finally, we show that σ^* is unique. Suppose not. Then, without loss of generality, there exists another fixed point $\tilde{\sigma} < \sigma^*$. Since $\Pi(\cdot)$ is increasing, it follows that $\Pi(\tilde{\sigma}) \leq \Pi(\sigma^*)$. Thus, $\Sigma(\Pi(\tilde{\sigma})) \geq \Sigma(\Pi(\sigma^*))$, since $\Sigma(\cdot)$ is decreasing. But then, $\tilde{\sigma} \geq \sigma^*$, since each is a fixed point, which is a contradiction. This completes the proof. □

A.2 Proofs of Section §4

A.2.1 Proof of Lemma 2

First, the continuity of φ follows directly from being the minimum of continuous functions. Second, φ is quasi-concave since the upper level set of the minimum of monotone functions is convex by being the intersection of convex sets. Finally, we turn to examine the shape of φ. Notice that $\mathcal{V}(\pi; t)$ is unaffected by t when $\varphi(v) = \frac{b(v)/\ell(v)}{1+b(v)/\ell(v)}$ for all $v \in [0, \bar{v}]$. This is the case when $\frac{b(v)/\ell(v)}{1+b(v)/\ell(v)} \leq t/\ell(v)$ for all $v \in [0, \bar{v}]$, or $t \geq \frac{b(\bar{v})}{1+b(\bar{v})/\ell(\bar{v})} =: \hat{t}$. Otherwise, φ eventually decreases. Indeed, for $t < \hat{t}$ we have $t/\ell(\bar{v}) < \hat{t}/\ell(\bar{v}) = \frac{b(\bar{v})/\ell(\bar{v})}{1+b(\bar{v})/\ell(\bar{v})}$, and thus φ is decreasing and equal to $\varphi(v) = t/\ell(v)$ for v close to $\bar{v}$. □

A.2.2 Proof of Proposition 2

Proof (a): Let $\hat{v} \equiv \arg\max_v \varphi(v)$. By Lemma 2, $\hat{v}$ is unique as φ is either monotone or hump-shaped. Since $\frac{b(v)/\ell(v)}{1+b(v)/\ell(v)} < 1 \ \forall v$, we have $\bar{\pi} \equiv \varphi(\hat{v}) < 1$ and $\mathcal{V}(\pi) = \emptyset$ for $\pi > \bar{\pi}$.

Now, consider $\pi \leq \bar{\pi}$. Then, we have that $\frac{b(v)/\ell(v)}{1+b(v)/\ell(v)} \geq \pi$ iff $v \in [v_0(\pi), \bar{v}]$, where $v_0(\cdot)$ is defined in (6). Next, we separate into two cases. If $\pi \leq t/\ell(\bar{v}) \equiv \underline{\pi}$ then $t/\ell(v) \geq \pi$ for all $v \in [0, \bar{v}]$; therefore, $\mathcal{V}(\pi) = [v_0(\pi), \bar{v}] \cap [0, \bar{v}] = [v_0(\pi), \bar{v}]$. Now if $\pi \geq t/\ell(\bar{v})$ then $t/\ell(v) \geq \pi$ iff $v \in [0, v_1(\pi)]$, where $v_1(\cdot)$ is defined in (6), and so $\mathcal{V}(\pi) = [v_0(\pi), v_1(\pi)] \ni \hat{v}$. □

Proof (b): Consider $t' < t'' < \hat{t}$. Then, by Proposition 2-(a), we have $\mathcal{V}(\pi; t') = [v_0(\pi), \bar{v}]$ for $\pi \leq t'/\ell(\bar{v})$, and also $\mathcal{V}(\pi; t'') = [v_0(\pi), \bar{v}]$ for $\pi \leq t''/\ell(\bar{v})$. Since $t' < t''$ it follows that $\mathcal{V}(\pi; t') = \mathcal{V}(\pi; t'') = [v_0(\pi), \bar{v}]$ for all $\pi \leq t'/\ell(\bar{v}) < t''/\ell(\bar{v})$. □

A.2.3 Proof of Proposition 3

First, following the proof of Proposition 2, for each cost t, define the level of prevalence $\underline{\pi}(t) \equiv t/\ell(\bar{v})$. Notice that this threshold $\underline{\pi}$ rises as verification cost t rises. Moreover, for $t = 0$ and $t = \hat{t}$ we have $\underline{\pi}(0) = 0$ and $\underline{\pi}(\hat{t}) < 1$, given the expression for $\hat{t}$ in the proof of Lemma 2. Now, consider $\bar{\pi}(t)$ from the proof of Proposition 2. Since $\bar{\pi}(t) = \max_{v \in [0,\bar{v}]} \varphi(v; t)$ and $\varphi(\cdot; t)$ rises in t, it follows that $\bar{\pi}(t)$ weakly rises in t. In fact, $\underline{\pi}(t) = \bar{\pi}(t) = \varphi(\bar{v})$ when $t = \hat{t}$, and thus the unverified sharing rate must obey $\Sigma(\underline{\pi}(\hat{t})) = 0$, given Proposition 2.

Next, let π^o denote the equilibrium prevalence when the possibility of verification is not available (or, alternatively, when $t \geq \hat{t}$), and let σ^o denote the associated equilibrium sharing rate. Clearly, $\pi^o, \sigma^o > 0$. We then define our verification cost of interest as

$$t' \equiv \inf\{t \leq \hat{t} \mid \underline{\pi}(t) = \pi^o\}.$$

By continuity, this infimum is attained, and is strictly less than $\hat{t}$, since $\sigma^o > 0$.

Finally, let $\sigma_S(\pi) \equiv F^{-1}(\pi)$ denote the inverse supply function which, critically, is unaffected by changes in t. For $t < t'$, therefore, we have that $\pi^o > \underline{\pi}(t)$, and so by Proposition 2 we deduce that $\Sigma(\pi^o) = \max\{\int_{v_0(\pi^o)}^{v_1(\pi^o)} dG, 0\} < \int_{v_0(\pi^o)}^{\bar{v}} dG = \sigma^o$, namely, a positive mass of users will now find it optimal to share and verify news. Consequently, $\Sigma(\pi^o) < \sigma_S(\pi^o)$, and thus the excess of supply is cleared with a lower prevalence rate, and so a lower unverified sharing rate. The diffusion rate falls as there is less production and sharing of unverified news.

For $t > t'$, however, $\pi^o < \underline{\pi}(t)$ and hence $\Sigma(\pi; t) = \Sigma(\pi; t')$ over $[0, \underline{\pi}(t)]$. Thus, the equilibrium continues to be (π^o, σ^o). This concludes the proof. □

A.3 Proofs of Section §5

A.3.1 Proof of Proposition 4

First, we show that the conclusion is implied by condition (a).

STEP 1: $v_0(\pi)$ IS CONVEX FOR $\pi \leq \bar{\pi}_0$. Let $p(v) \equiv b(v)/\ell(v)$ with $p(0) = \lim_{v \downarrow 0} b(v)/\ell(v)$. By (6), it follows that $v_0(\pi) \in (0, \bar{v})$ is determined by $p(v_0(\pi)) \equiv \pi/(1-\pi)$. Also, since $p : [0, \bar{v}] \to [p(0), p(\bar{v})]$ is strictly increasing and concave, its inverse $p^{-1} : [p(0), p(\bar{v})] \to [0, \bar{v}]$ is strictly increasing and convex. Thus, for $\underline{\pi}_0$ and $\bar{\pi}_0$ respectively solving $p(0) = \underline{\pi}_0/(1-\underline{\pi}_0)$

and $p(\bar{v}) = \bar{\pi}_0/(1-\bar{\pi}_0)$, we have $v_0(\pi) = p^{-1}\left(\frac{\pi}{1-\pi}\right)$ for $\pi \in [\underline{\pi}_0, \bar{\pi}_0]$. For $\pi < \underline{\pi}_0$, we have $v_0(\pi) = 0$ by (6). Since the map $\pi \mapsto \pi/(1-\pi)$ is increasing and convex, it follows that $v_0(\pi)$ is increasing and convex for $\pi \leq \bar{\pi}_0$.

STEP 2: IF $\sigma^* \leq \Sigma_N(\pi^*)$ THEN $|\mathcal{E}_\pi(\Sigma)| \geq \mathcal{E}_\pi(\Sigma_N)$. Let $\sigma_V^* = \Sigma_V(\pi^*)$ and $\sigma_N^* = \Sigma_N(\pi^*)$. Since $\Sigma(\pi) = 1 - (\Sigma_V(\pi) + \Sigma_N(\pi))$ for all $\pi \leq \bar{\pi}$ (Proposition 2), it follows that for $\pi = \pi^*$:

$$|\mathcal{E}_\pi(\Sigma)| = \mathcal{E}_\pi(\Sigma_V)(\sigma_V^*/\sigma^*) + \mathcal{E}_\pi(\Sigma_N)(\sigma_N^*/\sigma^*) \geq \mathcal{E}_\pi(\Sigma_N).$$

The inequality holds because both $\Sigma_V(\cdot)$ and $\Sigma_N(\cdot)$ are increasing in π, since the thresholds $v_0(\cdot)$ and $v_1(\cdot)$ are, respectively, increasing and decreasing in π; also, $\sigma^* \leq \sigma_N^*$.

STEP 3: $\mathcal{E}_\pi(\Sigma_N) \geq 1$. First, notice that $\Sigma_N(\pi) = G(v_0(\pi))$ is convex for $\pi \leq \bar{\pi}_0$, since $g(v)$ is increasing in v, and $v_0(\pi)$ is increasing and convex. Also, $\Sigma_N(0) = 0$ because $v_0(0) = 0$, and so $\Sigma_N(\pi)$ rises from the origin at increasing rates. Thus, Σ_N must have an increasing secant: $(\Sigma_N(\pi)/\pi)' \geq 0$. But then, $\mathcal{E}_\pi(\Sigma_N) \geq 1$. Altogether, $|\mathcal{E}_\pi(\Sigma)| \geq \mathcal{E}_\pi(\Sigma_N) \geq 1$. □

Following similar logic, we now show that condition (b) also implies the desired result.

STEP 1: IF $1/\ell(v)$ IS CONCAVE THEN $v_1(\pi)$ IS CONCAVE. By (6), $v_1(\pi) \in (0, \bar{v})$ is characterized by $\tilde{\ell}(v_1(\pi)) \equiv \pi/t$, where $\tilde{\ell}(v) \equiv 1/\ell(v)$ is monotone decreasing. Let $\underline{\pi}_1$ and $\bar{\pi}_1$ solve $\tilde{\ell}(0) = \bar{\pi}_1/t$ and $\tilde{\ell}(\bar{v}) \equiv \underline{\pi}_1/t$, respectively. Then, for $\pi \in [\underline{\pi}_1, \bar{\pi}_1]$, we have $v_1(\pi) = \tilde{\ell}^{-1}(\pi/t)$. Note that for $\pi < \underline{\pi}_1$, $v_1(\pi) = \bar{v}$, given (6). Thus, for $\pi \leq \bar{\pi}_1$, $v_1(\pi)$ is decreasing and concave, since $\tilde{\ell}^{-1}$ is the inverse of a monotone decreasing concave function.

STEP 2: IF $\sigma^* \leq \Sigma_V(\pi^*)$ THEN $|\mathcal{E}_\pi(\Sigma)| \geq \mathcal{E}_\pi(\Sigma_V)$. Let $\sigma_V^* = \Sigma_V(\pi^*)$ and $\sigma_N^* = \Sigma_N(\pi^*)$. Since $\Sigma(\pi) = 1 - (\Sigma_V(\pi) + \Sigma_N(\pi))$ for all $\pi \geq \bar{\pi}$, it follows that for $\pi = \pi^* \in (0, \bar{\pi})$:

$$|\mathcal{E}_\pi(\Sigma)| = \mathcal{E}_\pi(\Sigma_V)(\sigma_V^*/\sigma^*) + \mathcal{E}_\pi(\Sigma_N)(\sigma_N^*/\sigma) \geq \mathcal{E}_\pi(\Sigma_V),$$

where the inequality follows by the same reasons given in Step 2 above, but using $\sigma^* \leq \sigma_V^*$.

STEP 3: $\mathcal{E}_\pi(\Sigma_V) \geq 1$. First, notice that Σ_V is convex. Indeed, since $G(v)$ is concave, $1-G(v)$ is convex. Thus, $\Sigma_V(\pi) = 1 - G(v_1(\pi))$ is increasing and convex, since it is the composition of a decreasing concave function $v_1(\pi)$, and a decreasing convex function $1-G(v)$. Moreover, $\Sigma_V(0) = 0$ for all $\pi \leq \underline{\pi}_1$, since $v_1(\pi) = \bar{v}$. Altogether, $\Sigma_V(\pi)$ weakly rises from the origin at increasing rates, and thus Σ_V is superadditive and its secant must rise: $(\Sigma_V/\pi)' \geq 0$. But then, Σ_V must be elastic, i.e., $\mathcal{E}_\pi(\Sigma_V) \geq 1$. All told, $|\mathcal{E}_\pi(\Sigma)| \geq \mathcal{E}_\pi(\Sigma_V) \geq 1$. □

Proposition A.1. *Assume that $b(\cdot)/\ell(\cdot)$ is of class $\mathcal{C}^2$, and consider an equilibrium (π^*, σ^*). Suppose that for $v^* = v_0(\pi^*)$ we have $(b(v^*)/\ell(v^*))'' \leq 0$ and $v^* g(v^*)/G(v^*) \geq 1$. Then, if the passthrough $\sigma^* \leq \Sigma_N(\pi^*)$ then the unverified sharing function is elastic, i.e., $|\mathcal{E}_\pi(\Sigma)| \geq 1$.*

Proof: First, as shown in the proof of Claim A.1, the elasticity of $\Sigma_N(\pi) = G(v_0(\pi))$ obeys

$$\mathcal{E}_\pi(\Sigma_N) = \frac{g(v_0)}{G(v_0)} \left(\frac{b'}{b} - \frac{\ell'}{\ell} \right)^{-1} \frac{1}{1-\pi}.$$

Now, evaluate this expression at $\pi = \pi^*$. Next, by assumption, $g(v^*)/G(v^*) \geq 1/v^*$. Also, $(b(v^*)/\ell(v^*))'' \leq 0$, and thus b/ℓ is concave in a neighborhood about v^*, since b, ℓ are of class $\mathcal{C}^2$. This implies that b/ℓ is subadditive about v^*, and thus it has a decreasing secant $[b(v^*)/(\ell(v^*)v^*)]' \leq 0$, namely, $v^*(b'(v^*)/b(v^*) - \ell'(v^*)/\ell(v^*)) \leq 1$. Altogether, using the expression above for $\mathcal{E}_\pi(\Sigma_N)$, we conclude that $\mathcal{E}_\pi(\Sigma_N) \geq 1/(1-\pi^*) > 1$.

Finally, as shown in the proof of Proposition 4, if the equilibrium pass through of fake news obeys $\sigma^* \leq \Sigma_N(\pi^*)$, then identity (7) implies $|\mathcal{E}_\pi(\Sigma) \geq |\mathcal{E}_\pi(\Sigma_N) > 1$. □

Proposition A.2. *Consider an equilibrium (π^*, σ^*). Suppose that for $v^* = v_1(\pi^*)$ we have*

$$\frac{g(v^*)}{1-G(v^*)} \geq \frac{\ell'(v^*)}{\ell(v^*)}.$$

Then, if $\sigma^* \leq \Sigma_V(\pi^*)$, the unverified sharing function is elastic in equilibrium: $|\mathcal{E}_\pi(\Sigma)| \geq 1$.

Proof: First, assume the equilibrium entails active verification (otherwise the result is vacuously true). Then, $\Sigma_V(\pi) = 1 - G(v_1(\pi))$, by Proposition 2, with $v_1(\pi)$ uniquely solving $\pi\ell(v) = t$. Next, we compute $\mathcal{E}_\pi(\Sigma_V)$ and evaluate it at $\pi = \pi^*$:

$$\mathcal{E}_\pi(\Sigma_N) = \frac{-\pi^* g(v^*) v_1'(\pi^*)}{1-G(v^*)} = \frac{g(v^*)}{1-G(v^*)} \times \frac{\ell(v^*)}{\ell'(v^*)},$$

where we have used that $\pi v_1'(\pi) = -\ell(v_1)/\ell'(v_1)$. Thus, $\mathcal{E}_\pi(\Sigma_N) \geq 1$ provided the condition specified in the proposition holds. Finally, as in the proof of Proposition 4, if the passthrough of fake news obeys $\sigma^* \leq \Sigma_V(\pi^*)$, then identity (7) implies $|\mathcal{E}_\pi(\Sigma) \geq |\mathcal{E}_\pi(\Sigma_V) \geq 1$. □

A.3.2 Proof of Proposition 5

We will show that each term in expression (7) is decreasing in t. First, notice that $\Sigma_N(\pi) = G(v_0(\pi))$ is not affected by t, and so an increase in t reduces ratio Σ_N/Σ, since it raises the amount of unverified sharing. Similarly, the ratio (Σ_V/Σ) falls as t rises, since t raises Σ but it lowers Σ_V. It remains to verify that $\mathcal{E}_\pi(\Sigma_V)$ falls in t. To this end, we'll first show that v_1 is supermodular in (π, t). Indeed, differentiate $\pi\ell(v_1(\pi)) = t$ in π to get:

$$\frac{\partial v_1}{\partial \pi} = \frac{-1}{\pi\ell'(v_1)/\ell(v_1)} < 0.$$

Next, differentiate the above expression in t to obtain:

$$\frac{\partial^2 v_1}{\partial t \partial \pi} = \left[\frac{\pi \ell'(v_1)}{\ell(v_1)}\right]^2 \times \frac{\partial}{\partial v}\left[\frac{\pi \ell'(v)}{\ell(v)}\right]\Bigg|_{v=v_1} \times \frac{\partial v_1}{\partial t} \geq 0,$$

where the inequality holds because $\partial v_1/\partial t > 0$, and also $(\ell'(v)/\ell(v))' \geq 0$.

Finally, we show that $\mathcal{E}_\pi(\Sigma_V)$ falls in t. Differentiate $\mathcal{E}_\pi(\Sigma_V) = -\frac{\pi v_1' g(v_1)}{1-G(v_1)}$ in t to get:

$$\frac{\partial \mathcal{E}_\pi(\Sigma_V)}{\partial t} = -\frac{\partial}{\partial v}\left(\frac{g(v)}{1-G(v)}\right)\Bigg|_{v=v_1} \times \frac{\partial v_1}{\partial t} \times \frac{\partial v_1}{\partial \pi} - \frac{g(v_1)}{1-G(v_1)} \times \frac{\partial^2 v_1}{\partial t \partial \pi}.$$

The above expression is negative since $[g(v)/(1-G(v))]' \leq 0$, $\partial v_1/\partial t > 0 > \partial v_1/\partial \pi$, and v_1 is supermodular in (π, t). □

A.4 Proofs of Section §6

A.4.1 Proof of Proposition 6

Take $\pi \in (0,1)$ and suppose that $\Sigma(\pi) < \Sigma^{co}(\pi)$. By Carathéodory's Theorem, there exists $\lambda \in (0,1)$ and prevalence $\pi_A, \pi_B \in [0,1]$ such that $\lambda \pi_A + (1-\lambda)\pi_B = \pi$ and $\Sigma^{co}(\pi) = \lambda \Sigma(\pi_A) + (1-\lambda)\Sigma(\pi_B)$. Without loss of generality, assume $\pi_A < \pi_B$. The monopolist can induce π_A and π_B by choosing segment sizes ω and $1-\omega$ for A and B, respectively; and also, by sending an extra amount τ of truthful news to segment A, with $\tau \geq 0$. Following the logic explained in the main text, the prevalence in A and B obey

$$\pi_A = \frac{\lambda \pi}{\lambda \pi + (1-\pi)\omega + \tau} \qquad \text{and} \qquad \pi_B = \frac{(1-\lambda)\pi}{(1-\lambda)\pi + (1-\pi)(1-\omega)}.$$

Since the variables $(\lambda, \pi_A, \pi_B, \pi)$ are already fixed by the concavification, it follows that the pair (ω, τ) adjusts to induce the desired prevalence (π_A, π_B), given (λ, π).

Finally, any prevalence π for which $\Sigma(\pi) = \Sigma^{co}(\pi)$ can be trivially implemented by setting segment sizes $\omega = \lambda$ and $\tau = 0$ so that $\pi_A = \pi_B = \pi$. □

A.4.2 Proof of Proposition 7

Suppose $\Sigma^{co}(\pi^*) > \Sigma(\pi^*)$ (otherwise, the result is trivially true). Then, $\Pi(\Sigma^{co}(\pi^*)) > \Pi(\Sigma(\pi^*)) = \pi^*$. This implies that the monopolistic equilibrium π^{**} must be strictly greater than π^*, since $\Pi(\Sigma^{co}(\cdot))$ is strictly decreasing. Consequently, $\sigma^{**} = \Sigma^{co}(\pi^{**}) > \Sigma(\pi^*) = \sigma^*$. Since the monopolistic equilibrium induces strictly more creation and sharing, the diffusion

rate must rise, i.e., $\Delta^{**} > \Delta^*$. Finally, as discussed in the main text, since the monopolist profits are given by the area between σ and the supply curve, the segmentation above is optimal among the ones considered. □

A.5 Proofs of Section §7

A.5.1 Proof of Proposition 8

To prove Proposition 8, we use the following claim:

Claim A.1. *Suppose that the propensity to share function $b(v)/\ell(v)$ is concave, and the density $g(v)$ increasing. If, in equilibrium, $\sigma^* \leq \Sigma_N(\pi^*)$, then $\mathcal{E}_\pi(\Sigma) \geq 1/(1-\pi^*)$.*

Proof: First, recall that $\Sigma_N(\pi) \equiv G(v_0(\pi))$, where $v_0(\pi)$ obeys (6). Next, as in the proof of Proposition 4-(a), we can use identity (7) to show that, in equilibrium, $|\mathcal{E}_\pi(\Sigma)| \geq \mathcal{E}_\pi(\Sigma_N)$, provided $\sigma^* \leq \Sigma_N(\pi^*)$. Since $\sigma^* \in (0,1)$, $v_0(\pi)$ in (6) must be interior, and thus differentiable in π, given Assumption 2 and the Implicit Function theorem. Consequently, $\mathcal{E}_\pi(\Sigma_N) = \pi g(v_0(\pi))v_0'(\pi)/G(v_0(\pi))$, where $v_0(\pi)$ is determined by $b(v_0)/\ell(v_0) = \pi/(1-\pi)$, given (6). Log-differentiate this last expression in π, and then solve for $v_0'(\pi)$ to get:

$$v_0'(\pi) = \frac{1}{\pi(1-\pi)}\left(\frac{b'}{b} - \frac{\ell'}{\ell}\right)^{-1} > 0,$$

which is positive since b/ℓ is strictly increasing in v, by Assumption 1. Therefore,

$$\mathcal{E}_\pi(\Sigma_N) = \frac{g(v_0)}{G(v_0)}\left(\frac{b'}{b} - \frac{\ell'}{\ell}\right)^{-1}\frac{1}{1-\pi}.$$

Now, notice that $G(v)$ is strictly increasing and convex, with $G(0) = 0$. Thus, $G(\cdot)$ has an increasing secant $[G(v)/v]' \geq 0$, or $vg(v)/G(v) \geq 1$. Similarly, $b(v)/\ell(v)$ is strictly increasing and concave; hence, $b(v)/\ell(v)$ must have a decreasing secant $[b(v)/\ell(v)v]' \leq 0$, namely, $v(b'/b - \ell'/\ell) \leq 1$. Putting these observations together, $\mathcal{E}_\pi(\Sigma_N) \geq 1/(1-\pi)$. □

Proof of Proposition 8: Let $\pi^*(\phi)$ denote the equilibrium prevalence given a filter of quality ϕ. As argued in the main text of §7, this value is the unique solution to the equation

$$\underbrace{(1-\phi)\Sigma(\psi(\pi^*(\phi);\phi))}_{\Sigma^e(\psi(\pi^*(\phi),\phi))\equiv} = \Pi^{-1}(\pi^*(\phi)),$$

where Π^{-1} is the the upward sloping inverse supply function. Totally differentiating the

above equality with respect to ϕ, we get:

$$-\Sigma + (1-\phi)\Sigma' \left[\frac{\partial \psi}{\partial \pi}[\pi^*]'(\phi) + \frac{\partial \psi}{\partial \phi}\right] = [\Pi^{-1}]'[\pi^*]'(\phi).$$

Solving for $[\pi^*]'(\phi)$ yields:

$$[\pi^*]'(\phi) = \frac{-\Sigma + (1-\phi)\Sigma' \frac{\partial \psi}{\partial \phi}}{[\Pi^{-1}]' - (1-\phi)\Sigma' \frac{\partial \psi}{\partial \pi}}$$

Since $[\Pi^{-1}]' > 0 > \Sigma'$ and $\partial\psi/\partial\pi > 0$, it follows that the sign of $[\pi^*]'$ is fully determined by the sign of the numerator of the above expression:

$$\chi(\phi) \equiv -\Sigma(\psi(\pi^*(\phi);\phi)) + (1-\phi)\Sigma'(\psi(\pi^*(\phi);\phi))\frac{\partial \psi}{\partial \phi}(\psi(\pi^*(\phi),\phi)).$$

Taking the limit of the above expression as $\phi \to 0$,

$$\chi(0) = -\Sigma(\pi^*(0)) - \Sigma'(\pi^*(0))\pi^*(0)(1-\pi^*(0)),$$

where we have used that $\psi(\pi, 0) = \pi$ and $\frac{\partial \psi}{\partial \phi}(\pi, 0) = -\pi(1-\pi)$. By Claim A.1, it follows that $\chi(0) > 0$, since $|\mathcal{E}_\pi(\Sigma)| \geq 1/(1-\pi^*(0))$. This implies that a small increase in the filter's quality leads to an increase in the equilibrium prevalence $\pi^*(\phi)$. Because a change in the filter has no *direct* effect on the supply function $\Pi(\sigma^e)$, the new equilibrium is the result of an *upward* movement along the supply curve, and hence the new equilibrium displays a higher effective sharing σ^{e*} and, thus, a higher diffusion of fake news, $\Delta^*(\phi) \equiv \pi^*(\phi)\sigma^{e*}(\phi)$. Intuitively, the small increase in ϕ shifts the effective sharing $(1-\phi)\Sigma(\cdot)$ up in the (π, σ^e)-space around the point studied. Conversely, as the filter becomes perfect, i.e., $\phi \to 1$, Σ^e shifts left towards the origin, with $(\pi^*(\phi), \sigma^{e,*}(\phi)) \to (0,0)$ and $\Delta^*(\phi) = \sigma^{e*}(\phi)\pi^*(\phi) \to 0$.

Finally, we show that equilibrium sharing σ^* is monotone increasing. To see this, consider the (π, σ)-space. There, an increase in filter ϕ lowers the posterior $\psi(\pi, \phi)$ and so it raises the unverified sharing $\Sigma(\psi(\pi, \phi))$ for every π. At the same time, an increase in ϕ lowers supply $\Pi((1-\phi)\sigma)$ at every σ. Thus, the sharing rate σ^* unambiguously rises. □

References

Acemoglu, D., A. E. Ozdaglar, and J. Siderius (2022): "A model of online misinformation," *CEPR Discussion Paper No. DP16932*.

Allcott, H. and M. Gentzkow (2017): "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, 31.

Altay, S., A.-S. Hacquin, and H. Mercier (2022): "Why do so few people share fake news? It hurts their reputation," *New Media & Society*, 24, 1303–1324.

Becker, G. S. (1991): "A note on restaurant pricing and other examples of social influences on price," *Journal of Political Economy*, 99, 1109–1116.

Bergemann, D., B. Brooks, and S. Morris (2015): "The limits of price discrimination," *American Economic Review*, 105, 921–957.

Bonatti, A. and G. Cisternas (2020): "Consumer scores and price discrimination," *The Review of Economic Studies*, 87, 750–791.

Bowen, R., D. Dmitriev, and S. Galperti (2021): "Learning from shared news: when abundant information leads to belief polarization," *NBER working paper*.

Cheng, I.-H. and A. Hsiaw (2022): "Bayesian doublespeak," *Available at SSRN*.

DiResta, R. and I. Garcia-Camargo (2020): "Virality Project (US): Marketing meets Misinformation," *Stanford Internet Observatory*, https://cyber.fsi.stanford.edu/io/news/manufacturing–influence–0.

Ershov, D. and J. S. Morales (2021): "Sharing news left and right: The effects of policies targeting misinformation on social media," Tech. rep., Collegio Carlo Alberto.

Friggeri, A., L. Adamic, D. Eckles, and J. Cheng (2014): "Rumor cascades," *Eighth International AAAI Conference on Weblogs and Social Media*.

Guess, A. M., M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. Sircar (2020a): "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India," *Proceedings of the National Academy of Sciences*, 117, 15536–15545.

Guess, A. M., B. Nyhan, and J. Reifler (2020b): "Exposure to untrustworthy websites in the 2016 US election," *Nature human behavior*, 4, 472–480.

Haghpanah, N. and R. Siegel (2022): "The limits of multi-product price discrimination," *American Economic Review: Insights*, forthcoming.

Henry, E., E. Zhuravskaya, and S. Guriev (2022): "Checking and sharing alt-fact," *American Economic Journal: Economic Policy*, forthcoming.

Howell, L., ed. (2013): *Global Risks 2013, Eight Edition*, World Economic Forum.

Kamenica, E. and M. Gentzkow (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615.

Kranton, R. and D. McAdams (2020): "Social networks and the market for news," Tech. rep., Duke University.

Lazer, D. M., M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. (2018): "The science of fake news," *Science*, 359, 1094–1096.

Lyons, T. (2017): "Replacing Disputed Flags With Related Articles," *Facebook Newsroom*, https://about.fb.com/news/2017/12/news–feed–fyi–updates–in–our–fight–against–misinformation/.

——— (2018): "Hard Questions: How Is Facebook's Fact-Checking Program Working," *Facebook Newsroom*.

Papanastasiou, Y. (2020): "Fake news propagation and detection: A sequential model," *Management Science*, 1826–1846.

Pennycook, G., A. Bear, E. T. Collins, and D. G. Rand (2020): "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*.

Pennycook, G., Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand (2021): "Shifting attention to accuracy can reduce misinformation online," *Nature*, 592, 590–595.

Quercioli, E. and L. Smith (2015): "The economics of counterfeiting," *Econometrica*, 83, 1211–1236.

Rapoza, K. (2017): "Can 'Fake News' Impact the Stock Market?" *Forbes*, https://www.forbes.com/sites/kenrapoza/2017/02/26/can–fake–news–impact–the–stock–market/#335703a52fac.

Rockafellar, R. T. (1970): *Convex analysis*, vol. 18, Princeton university press.

Serra-Garcia, M. and U. Gneezy (2021): "Mistakes, overconfidence, and the effect of sharing on detecting lies," *American Economic Review*, 111, 3160–83.

Stencel, M. and J. Luther (2020): "Annual census finds nearly 300 fact-checking projects around the world," *Duke Reporters' Lab*, https://reporterslab.org/latest–news/.

Sydell, L. (2016): "We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned," *NPR*, https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr–finds–the–head–of–a–covert–fake–news–operation–in–the–suburbs.

Meta Business Help Center (2022): "About Fact-Checking on Facebook," *Meta Business Help Center*, https://www.facebook.com/business/help/2593586717571940?id=673052479947730.

World Economic Forum (2020): *Global Risks 2020, Fifteenth Edition.*

Tucker, J. A., A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan (2018): "Social media, political polarization, and political disinformation: A review of the scientific literature," *William and Flora Hewlett Foundation.*

Vásquez, J. (2022): "A theory of crime and vigilance," *American Economic Journal: Microeconomics*, forthcoming.

Vosoughi, S., D. Roy, and S. Aral (2018): "The spread of true and false news online," *Science*, 359, 1146–1151.