Misinformation in Social Media: The Role of Verification Incentives[*]

Gonzalo Cisternas and Jorge Vásquez

May 11, 2023

Abstract

We develop a model in which the prevalence and sharing of misinformation endogenously arise from the interaction between (i) users' decisions to verify and share news of unknown truthfulness and (ii) producers' choices to generate fake content. We use the model to examine how policies intended to combat misinformation affect users' incentives to engage in *costly news verification*. Via this channel, unintended effects may emerge from: lowering verification costs borne by users; disrupting the supply of fake content; and introducing imperfect filters. We provide sensitivity measures, akin to demand elasticities, to evaluate these effects, and make predictions about market outcomes based on observable characteristics such as users' age and popularity, as well as deeper parameters such as users' gains and losses of sharing news.

[*] An earlier version of this paper was titled "Fake News in Social Media: A Supply and Demand Approach." Cisternas: Federal Reserve Bank of New York, gonzalo.cisternas@ny.frb.org. Vásquez: Smith College, Department of Economics, jvasquez@smith.edu. We thank Marco Cipriani, Steven Durlauf, Andrew Haughwout, Nathan Kaplan, Jorge Lemus, Jonathan McCarthy, Lones Smith, Marek Weretka, and Juanjuan Zhang for their useful comments and conversations, and Orrie Page for his excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

1 Introduction

The spread of misinformation online has recently gained substantial prominence. Its threats to societies are real—potentially affecting elections, markets, and disease spread[1]—and unlikely to disappear without action.[2] Social media platforms have therefore responded to the issue, taking important steps in at least three domains. First, in the area of fact-checking: platforms have promoted professional and independent verification of news, which has resulted in the advent of a network of third-party fact-checkers who verify the accuracy of content.[3] Second, in the area of fake content visibility: to weaken the incentives of fake new producers, news items that have been confirmed to be false are given less relevance, and repeated offenders are removed from platforms.[4] Third, in technology, such as the deployment of algorithms designed to detect misinformation.[5]

Underlying this comprehensive response, however, is the principle that it is users themselves who must both assess the veracity of each news item and ultimately decide how to act upon it. To empower users, therefore, suspicious content is now accompanied by fact-checkers' reports or related material that provides context to users.[6] The success of the aforementioned policy responses then rests on users' willingness to verify evidence on the news items encountered. However, such verification process is naturally a costly activity, even if the evidence is readily available.

This paper introduces a flexible model of misinformation that can be used to address a variety of questions regarding the fake news problem. For instance, how impactful is the appearance of fact-checking initiatives that facilitate news verification 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 filters that detect fake content before it reaches a platform's users? The distinctive aspect of our work is that we focus on examining how users' ***incentives to verify news*** vary with the introduction, or change in

[1] Allcott and Gentzkow (2017) estimate that 760 million interactions with fake news occurred on the web around the 2016 U.S. presidential election, while Guess et al. (2020b) show that online platforms facilitated traffic to untrustworthy websites. See Rapoza (2017) for an incident of the stock market's reaction to fake news, and DiResta and Garcia-Camargo (2020) for falsehoods regarding the COVID-19 pandemic.

[2] The World Economic Forum has labeled the fake news problem as a major global risk (Howell, 2013), and it highlights the use of artificial intelligence in the production of "deepfake" videos as a major long-term threat (World Economic Forum, 2020).

[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 Meta Business Help Center (2022). Some organizations are also rating websites' trustworthiness, which ultimately influences the latter's advertising revenue, e.g., `https://disinformationindex.org/`.

[5] See, e.g., `https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/`.

[6] Lyons (2017) argues that this approach lowers sharing behavior more than merely labeling news as false.

the intensity, of any of these policies—through this approach, we uncover some unintended risks that may arise, and we inform on the type of data and elasticity measures that can be gathered to better design interventions.

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, a user can first choose to learn its veracity at a cost and then decide whether to share it. Importantly, users differ in their benefits and losses from sharing news but share a commonality: they experience gains from sharing true news and suffer losses from sharing fake content. These assumptions—which are motivated and discussed in the applied relevance subsection below—have two implications. First, because verification reveals the truthfulness of each news item, and users dislike sharing fake content, fake news can be shared only by users who do not verify. Second, from an incentives viewpoint, verifying news entails incurring in an upfront payment that must be weighed against the ex-post loss of sharing fake content with some chance—those engaging in such unverified sharing decide to skip the former cost.

The chance that a news article is false is given by the endogenous proportion of them among the total entering the platform. When such fake news *prevalence* rises, fewer users share news articles without verifying; hence, a decreasing *misinformation pass-through curve* emerges, linking prevalence levels to the mass of users engaged in unverified sharing behavior. Conversely, the pass-through rate of misinformation increases as the proportion of these users increases, incentivizing the production of fake content; thus, an increasing *supply of fake news* ensues. Much like in traditional markets, an equilibrium in this market emerges when these two curves intersect, delivering an equilibrium prevalence and pass-through rate of misinformation; in turn, the *diffusion rate* of fake news—prevalence times pass-through—is a measure of fake news items dissemination. Equipped with this tractable characterization, we can examine how current interventions affect outcomes in this induced market for news.

Overview of the results. As a first step, we study the introduction of fact-checking services (Section 4). This policy effectively reduces the verification costs borne by users, ultimately affecting equilibrium outcomes through its impact on the pass-through curve. Users, however, will engage in verification only when their expected losses are high. As a result, lowering such costs does not act as a usual "demand shifter," ensuring a strict inward shift in the pass-through curve. Put differently, users' incentives are at the heart of rigidities in the misinformation pass-through curve that may result in equilibrium outcomes remaining unchanged—studying the nature of these rigidities becomes of central importance.

Specifically, we derive conditions on primitives (benefits and losses) such that lowering verification costs effectively induces more users to verify news, yet the pass-through curve is *nowhere affected*. Indeed, such lower costs may induce verification—and hence sharing—only by users who were not originally sharing content, leaving the behavior of those already doing unverified sharing unchanged. At the other extreme, we find conditions for a strong form of sensitivity of the pass-through curve in the following sense: lowering the aforementioned costs always triggers verification first on those engaging in unverified sharing. We argue that our first set of conditions is most appropriate when we consider an "inter-cohort" interpretation of the model: users vary in their age, with each user type encoding an average "popularity" of the cohort in practice. The prediction is that most verification comes from younger users who where originally cautious of sharing news, and little from older users doing unverified sharing, so equilibrium outcomes can be unaffected. In turn, our second set of conditions are most appropriate under an "intra-cohort" interpretation: users of a given age who vary in their popularity, with more popular types exhibiting more propensity to share news. As verification is feasible, those users can be more prone to verify content to avoid large reputational losses, so the policy has the potential to affect outcomes at the outset.

Motivated by the use of policies that weaken the incentives to produce fake news, in Section 5 we examine supply effects, which we model as an inward shift of the supply curve at all possible levels. In this situation, while fake news prevalence decreases after the shift occurs, the equilibrium pass-through rate increases through two channels: some users cease verifying news, and others who were not originally sharing content begin to share without verifying. We provide conditions such that, through either of these channels, the pass-through curve is sufficiently elastic so that the diffusion rate of fake news increases after a reduction in prevalence; in other words, misinformation is more profusely diffused after the intervention. Equally important, we also provide conditions under which a reduction in verification costs results in a more elastic pass-through curve, highlighting circumstances under which joint policies can negatively reinforce one another.

In Section 6, we examine the use of detection algorithms. We focus on an algorithm that: (i) assesses news articles before these reach users, removing content if deemed fake, and (ii) makes type-II errors—mistakenly labeling a false article as truthful. In this context, we show that introducing such a filter can increase not only the diffusion of misinformation (as it happened under supply effects), but also its prevalence. Indeed, by affecting the inference made by users when they receive news, such filters can induce more unverified sharing, which can outweigh the benefit that a filter has on weakening producers' incentives via limiting the pass-through of fake news. Further, we provide conditions that guarantee a sufficiently elastic relaxation of users' incentives—those originally either verifying news or refraining from sharing unverified content—that underlies our finding.

The paper concludes with two extensions. First, in Section 7.1 we examine market power by considering 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; under these conditions, the only sequentially rational outcome is the competitive equilibrium. That said, a mild enrichment of the monopolist's toolbox can improve profits: if he can both segment the market trivially (i.e., target sub-populations that differ from the original one only in their size) and also supply truthful content, then the monopolist can achieve prevalance-pass-through rate pairs that dominate the competitive outcome. We also show how this technique can be profitable when lowering verification costs creates convexities in the misinformation pass-through curve.

Second, in Section 7.2, we examine network externalities, understood as individual choices depending on the aggregate choices of others—a central element in the success or failure of platforms. There, we show that the misinformation pass-through curve can in fact become a correspondence. Consequently, supply reductions as in Section 5 can lead not only to more diffusion and greater prevalence, but they can also refine the set of equilibria to a single outcome that is worse than the original one.

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); with more sharing, however, more users can be reached, and the likelihood of more clicks increases. In fact, 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] lowers search costs for users, it does not necessarily eliminate 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 infor-

[7] Fact-checking sites have grown from 44 in 2014 to almost 300 by 2020 (Stencel and Luther, 2020).

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

mation freely provided by platforms.

That said, our model abstracts from two elements that are seen as important in practice. First, news may have an ideological bent—i.e., a form of horizontal differentiation—with users potentially favoring certain types of news over others. Second, users may belong to social networks, and so they can get information from individuals they have chosen to "follow." One would expect, however, that verification incentives are weaker when biases are at play and/or news articles come from trusted sources. Not only that, these features may obscure policies' limitations that stem from incentives considerations exclusively, such as the types of rigidities that we uncover in Section 4. Since our goal is to focus on such incentives, we explicitly choose a setting in which (i) rational users encounter news that differ only along a "vertical" true/false dimension and (ii) encounters are random. At the end of Section 2 we also argue why the model is still meaningful from a positive standpoint despite not explicitly modeling these elements, especially if the goal is to study ***two-sided*** markets.

In summary, the recent actions taken to combat misinformation reinforce the idea that news verification is central to this issue. From this perspective, our model allows us to: provide concrete elasticity measures—akin to "demand elasticities"—to evaluate policy interventions; connect these measures to conditions on both primitives and equilibrium objects; and make predictions regarding market outcomes once controlling for observable user characteristics such as popularity and age. Provided sufficient data is gathered, our results could in principle be tested empirically.

Related literature Our paper is most closely related to studies whose main goal has been to assess the efficacy 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, which is consistent with our "popularity" specification. Regarding news inspection, 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 when an algorithmic filter is in place in our model. Finally, regarding policies aimed at limiting the sharing of misinformation, Ershov and Morales (2021) empirically examine the impact that increasing users' costs of sharing news has on the transmission of content from highly trustworthy outlets relative to news from less trustworthy counterparts; by contrast, we examine how interventions that would reduce the supply of fake content can lead to more sharing and diffusion of misinformation.

Our paper also belongs to a growing body of theoretical papers examining various aspects of the fake news problem. In a fully dynamic setting, Papanastasiou (2020) allows users to uncover the veracity of a piece of news upon paying a cost, but the main focus is on examining a platform's decision to engage in costly verification when fake-news cascades harm the platform; by contrast, we focus both on how users' incentives to inspect news vary with changes in the underlying heterogeneity and, by virtue of examining a two-sided market, on how supply forces respond to policy interventions. Costly verification also appears Cheng and Hsiaw (2022), where it is shown that misinformation and belief disagreement can persist in the long run when news items repeatedly come from a privately informed sender. Abstracting from verification incentives, 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, while Bowen et al. (2021) analyze a dynamic Bayesian learning model in which agents learn from what others share, and show that this could lead to belief disagreement in the long-run (belief polarization).

To conclude, our model also contributes to the matching literature in settings in which individuals choose to protect themselves at a cost from harm with endogenous intensity (e.g., Quercioli and Smith, 2015 and Vásquez, 2022). Two observations are useful in this regard. First, we do this in the context of social media platforms, and where utility is non-transferable; see Chade et al. (2017) for a comprehensive survey of matching models with and without transferable utility. Second, our analysis of segmentation strategies entails the choice of allocating an "average quality" to (i) identical sub-populations, (ii) when prices are absent, and (iii) when consumers do not know the quality of the good at hand—this is in great contrast with recent treatments of market segmentation in traditional product markets; see, e.g., Bergemann et al. (2015).

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. We introduce the main elements of the model first and then discuss how to interpret the latter at the end of the section.

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 in that the likelihood of encountering false news is identical across users and solely determined by the proportion of fake news circulating in the platform. Upon encountering a news item, each user can decide to determine its veracity by paying a fixed cost t 0; for instance, the search costs incurred 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 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 $b(\ )$ and $\ell(\ )$ are continuous in $[0,\bar{v}]$, and strictly positive almost everywhere (a.e.) in the interval. The underlying characteristic v is distributed according to an atomless cumulative distribution function (CDF) $G(\ )$, with support $[0,\bar{v}]$ and density $g(\ )$ which is differentiable.

Fake news producers. We assume that there is a fixed mass of uncertified news entering the platform, which we set at a level equal to 1. Among this mass, a proportion $\in [0,1]$ is false. We refer to 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 (which yields zero); these costs vary across producers according to an atomless CDF $F(\ )$ with support $[0,1]$ and a differentiable density $f(\ )$. Fake news producers then 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 rate of misinformation*, and we denote it by $\in [0,1]$.

We note that while our choice to fix the volume of uncertified news implies that increases in the mass of fake news production necessarily crowd out truthful content entering the platform and vice-versa, this assumption is inconsequential for our analysis: what we require is that fake news production is increasing in its pass-through rate, which operationally translates to an increasing supply curve of misinformation. Also, note that since users dislike passing on fake content and the verification technology is perfect, misinformation is shared only when it is not verified.[9] Thus, the fake news pass-through rate coincides with the mass of users who *engage in unverified sharing*.

Information and equilibrium concept. The pair $(\ ,\ )$ is determined by forces akin to supply and demand as we explain in the next section. Importantly, neither users nor producers directly observe these endogenous variables, as it happens in practice; rather, all platform participants correctly anticipate the values that and take in equilibrium. The

[9] Section 6 introduces a second layer of verification, albeit imperfect and in the form of a technology reminiscent of an "algorithmic filter."

model is "competitive" in that all platform participants are assumed to take (;) as given.

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

To conclude, a measure of misinformation dissemination in our model is Δ , which captures the number of fake news articles that are shared. This variable plays an important role in Section 5-6 and we refer to it as the *diffusion rate* of fake news articles.

Interpretation of the model. Let us conclude this section by briefly discussing how to interpret the model and some of our modeling choices.

1. User types and preferences. The general specification for gains and losses permits two interpretations of our population of users that are relevant in practice. First, v is a measure of *within-cohort popularity*: users correspond to individuals of a specific age (and potentially other observable characteristics—see next point) who in turn vary in terms of their popularity (e.g., number of followers). Naturally, we expect b to be increasing across types, but losses can be increasing too: reputational costs may grow as popularity increases. Second, v captures *age*: each type in our model is a (e.g., popularity) average of a different cohort of users in practice. We can then expect both b and ` to be decreasing across types: if average visibility decreases with age, so can the benefits and losses of sharing news. Crucially, as we will demonstrate in Section 4, these two interpretations/cases are not necessarily isomorphic via a relabeling of types; in fact, there can be important qualitative differences in the ways through which policies induce verification of news by users in both cases (Propositions 2 and 3).

2. Matching technology. The fact that the chance of encountering fake content is solely determined by the relative abundance of the latter implies that the matching technology is effectively random. But in reality, users' encounters with news are not completely accidental: as argued, users may rely on their networks, or they may exhibit biases that favor certain types of news (horizontal aspect). Allowing for randomness is nevertheless appropriate from a *two-sided market* perspective—i.e., producers of fake content interacting with consumers of news—because it reflects fake news producers' imperfect reach when targeting individuals of interest. In turn, this can happen because platforms offer an imperfect degree of targeting granularity in practice, or because algorithms mediate matches using information not possessed by news producers.

From this perspective, suppose that the population of interest is characterized by a common ideological position. In addition, fake news producers specialize in generating misinformation about actions by representatives of the other side of the ideological spectrum. Then, the news articles effectively share an element of homogeneity (e.g., ‘news about the opposite ideology’), ultimately making the vertical true/false dimension the main source of differentiation. In this “partial equilibrium” context, as long as there is some residual uncertainty regarding users’ popularity or age, perfectly channeling news through those who will easily pass it on is not possible—producers need to evaluate the non-trivial chance of reaching users who engage in unverified sharing.

3. Static nature. The model is static in that users share news (at most) once—but fake news diffusion clearly is a dynamic phenomenon. While certainly a simplification, this assumption reflects a situation in which the sharing decisions of individuals who encounter fake news early on can be good predictors of the overall diffusion of fake content. There are three factors that can favor this relationship. First, if news items arrive frequently to platforms: as new cohorts of news are placed more prominently in users’ news feeds, the subsequent sharing rate of old cohorts can be limited, thereby making the initial sharing rates the most relevant for producers’ payoffs. Second, if dynamic complementarities are at play: if news that are shared more diffusely early on are given more prominence and vice-versa, initial sharing rates can determine the fate of a piece of news. Third, if producers have some targeting power: news items are then likely to be directed to users perceived as good entry points.[10]

3 Unverified Sharing and Equilibrium Prevalence

Sharing misinformation. Consider the set of users who choose to skip verification and share news for each prevalence level   and verification cost t. This set is a subset of the users who find it optimal to share when verification is unavailable, or prohibitively costly (e.g., $t = +1$). Indeed, by revealed preferences, for any user who does not share in this latter case, sharing without verifying remains dominated when the possibility of verification is added.

To this end, imagine for a moment that verification is unavailable. Fixing prevalence   , type v will decide to share provided

$$(1 \qquad)b(v) \qquad `(v) \qquad 0;$$

[10] Grinberg et al. (2019) show that fake news sharing has been concentrated among older people, while Guess et al. (2019) that age is its main predictor even after controlling for partisanship and ideology.

where we have assumed that ties are broken in favor of sharing. Rearranging the above expression, we see that sharing is optimal if

$$\frac{b(v)/\ell(v)}{1+b(v)/\ell(v)} \qquad , \tag{1}$$

or when types v have a large enough *propensity to share* news (left-hand side of (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 misinformation versus engaging in unverified sharing. Critically, since costly verification is naturally paid irrespective of the verification outcome, unverified sharing dominates verified sharing whenever

$$(1 \quad)b(v) \quad \ell(v) > (1 \quad)b(v) \quad t,$$

(i.e., when sharing, we break ties in favor of verifying news). Thus, unverified sharing happens for types v with sufficiently high *propensity to skip verification* $t/\ell(v)$, i.e.,

$$t/\ell(v) > \quad . \tag{2}$$

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

$$\mathbf{U}(\ ;t) \quad \{v \in [0,\bar{v}] : \varphi(v;t) \quad \}, \text{ where } \varphi(v;t) \quad \min \quad \frac{b(v)/\ell(v)}{1+b(v)/\ell(v)}, \frac{t}{\ell(v)} \quad . \tag{3}$$

We can now state a central study object: the *misinformation pass-through curve*, which maps prevalence levels to the mass of users sharing unverified content:

$$\Sigma(\ ;t) \quad \int_{\mathbf{U}(\ ;t)} dG(v). \tag{4}$$

By definition of (3), the correspondence $\mapsto$ $\mathbf{U}(\ ;t)$ is weakly decreasing in the sense of set inclusion. Thus, the pass-through curve 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 vanishes, all users share news articles without verifying (i.e., $\mathbf{U}(0;t) = [0,\bar{v}]$) as sharing ceases to have a downside, i.e., $\Sigma(0;t) = 1$. Moreover, because $\ell > 0$ a.e., the set of users who share news has measure zero if only fake news articles circulate. Thus, $\Sigma(1;t) = 0$ with Σ possibly vanishing strictly before 1 in some specifications.

Remark 1. For fake news, distinguishing between the "pass-through rate" ( ) and the "pass-through curve" (Σ) is appropriate not only because of their dimensionality (scalar vs. function) but also because both notions may differ in more general specifications, such as when we study the use of internal filters in Section 6.

We focus on the case in which Σ is continuous. While this 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 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 $'(\ ;t)$ is differentiable a.e. with $j'^{0}(\ ;t)j > 0$ a.e, then Σ is continuous.*

That is, provided $'(\ ;t)$ does not fluctuate too wildly or exhibit flat regions, small changes in the environment do not have large effects in the aggregate. Otherwise, $'$ can take any form, allowing for kinks that can appear naturally due to the presence of a minimum. We assume the continuity of the pass-through curve Σ in the rest of this paper.[11]

Producing Fake News. We now turn to the supply of fake news entering the platform.

Given a pass-through rate of misinformation , a fake news producer with cost c chooses to produce if and only if c. Thus, the *supply of fake news*, i.e., the function that maps pass-through rates to prevalence levels, is given simply by

$$\Pi(\ ) \qquad F(\ ): \tag{5}$$

Due to the properties of the distribution $F(\ )$, the supply $\Pi(\ )$ 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. As argued in Section 2, normalizing the mass of news to 1 is without loss: all our qualitative results go through as long as the supply of fake content is increasing. Also, as in Remark 1, distinguishing between "prevalence" and "supply" is justified because they need not coincide in other specifications (e.g., when the total mass of news articles exceeds 1).

Altogether, a tuple $(\ ;\ )$ constitutes an equilibrium of this economy if and only if $= \Sigma(\ ;t)$ and $= \Pi(\ )$. Our assumptions guarantee the following result:

Proposition 1. *There exists a unique equilibrium $(\ ;\ )$.*

[11] This rules out the case in which losses ` are constant in v, wherein all users trivially verify news or not.

Much like in a standard supply and demand framework, the opposing forces behind an increasing supply of fake news and a decreasing misinformation pass-through curve ensure the existence of a unique pair (;) that balances this "market for misinformation." Economically, therefore, 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 a number of issues related to the fake news problem. To obtain sharp economic insights, we make and maintain the following assumption.

Assumption 1. *The benefit function $v \mapsto b(v)$ and loss function $v \mapsto \ell(v)$ are strictly monotone, while their ratio $v \mapsto b(v)/\ell(v)$ is strictly increasing.*

A strictly increasing ratio of benefits to losses captures that higher types exhibit a stronger propensity to share news—ordering types v in this way imposes minimal structure on the problem. Equipped with this ranking of types, it suffices to assume that benefits b and losses ℓ are monotone to illustrate the key insights of the model.

4 Facilitating Verification by Users

To combat fake news, social media platforms have begun to offer more and better fact-checking services to users, while still allowing them to decide whether to share. These services can be seen as a decrease in the verification costs t that users face. Importantly, fully eliminating these costs may be infeasible: even if evidence is readily available, verifying information is a costly activity. Hence, in what follows we focus on reductions in verification costs with t remaining strictly positive.

Such reductions have the potential to affect outcomes via their impact on the pass-through curve $\Sigma(\ ; t)$: if verification is less costly, the extent of unverified sharing can fall, manifested in a contraction of $\Sigma(\ ; t)$. However, as Figure 1 demonstrates, the associated contractions are, generically, only *partial* (from $= 0.4$ onward in Figure 1). Hence, whether or not outcomes are affected depends on where the equilibrium lies via the location of the supply curve; see Figure 1. Thus, as a first observation, while $\Sigma(\ ; t)$ mimics a traditional demand function (it is decreasing in the quantity of the good of interest, , and it determines the revenue for producers) *reductions in t do not act as a demand shifter* that guarantees a strict inward shift at all levels of prevalence, and hence a drop in the equilibrium prevalence . This trait of the pass-through curve suggests an inherent lack of sensitivity to reductions in verification costs, ultimately impacting the effectiveness of fact-checking policies.

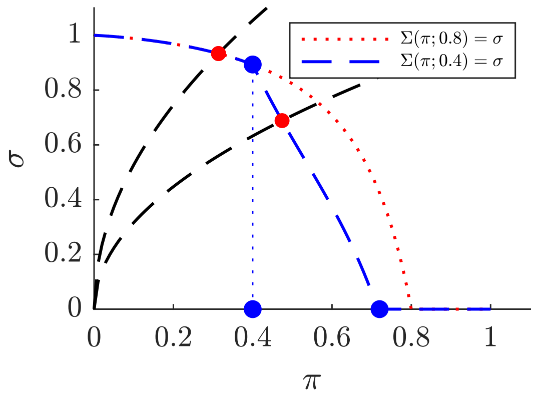

Figure 1: Parameter values: $\bar{v} = 1$, $G(v) = v$, $b(v) = 4v^{1.8}$, and $\ell(v) = v$. Verification cost $t \in \{0.4, 0.8\}$. The upward sloping curves represent two hypothetical misinformation supplies (5), namely, $\Pi(\sigma) = \sigma^2$ and $\Pi(\sigma) = 0.36\sigma^2$.

The remainder of this section is devoted to more deeply examining this issue of sensitivity of Σ to changes in t. At the core of our analysis lies the possibility that the choice to verify news segments users into three sets: users who engage in *unverified sharing* (**U**); who *verify and share* (**V**); and who simply *do not share* (**N**). We establish conditions under which lowering verification costs t implies a transfer of users from **U** to **V**, effectively leading to a (partial) contraction of the pass-through curve Σ. But we also state conditions that lead to a transfer of users from **N** to **V** exclusively, implying that although the policy does indeed induce more circulation of verified content, the amount of misinformation sharing is unchanged. Consequently, lack of sensitivity to reductions in verification costs need not occur only for low levels of prevalence, where it is difficult to incentivize verification due to the low chances of encountering misinformation—it can also arise at higher prevalence levels if "entry" by those not originally sharing takes place.[12]

4.1 Increasing benefits

Suppose that $b(\cdot)$ is strictly increasing. As we show below, in this case the pass-through curve Σ exhibits a *strong sensitivity* to verification costs in the following sense: for each level of prevalence π, entry to the (verified) sharing world (**N** → **V** transfer) cannot occur before the pass-through curve Σ contracts at that point (**U** → **V** transfer). If the opposite occurs (Section 4.2), we say that the pass-through curve has *weak sensitivity* to the same variable.

As a first step, we need to construct $\pi \mapsto \Sigma(\pi; t)$ for each π given t. To this end, recall that the set of users engaging in unverified sharing $\mathbf{U}(\pi; t)$ (dependence on π and t is now made explicit) is determined by the π-superlevel set of φ defined in (3), i.e., those types v

[12] Recall that, by revealed preferences, users switching from **N** to **U** is, for all π, not possible when t falls.

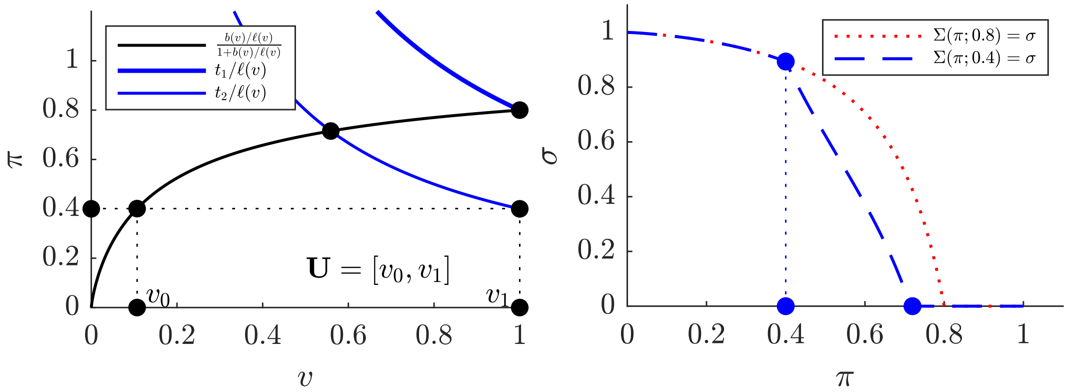

Figure 2: Parameters as in Figure 1: $\bar{v} = 1$, $G(v) = v$, $b(v) = 4v^{1.8}$, $\ell(v) = v$, $t_1 = 0.8$ and $t_2 = 0.4$.

for whom

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

To illustrate, suppose that loss $\ell(\cdot)$ is strictly increasing, and consider Figure 2 below. The left panel depicts an increasing propensity to share function $b/\ell/(1+b/\ell)$ as well as two decreasing propensity to skip verification functions t_1/ℓ and t_2/ℓ with $t_1 > t_2$. For values of $t > t_1$, no one engages in verification, as $\varphi \equiv b/\ell/(1 + b/\ell)$. Further, for levels of prevalence $\pi > 0.8$, no one shares news as $\varphi < 0.8$ a.e.; in the right panel, therefore, the pass-through curve $\Sigma(\cdot; t)$ for $t > t_1$ in red dots vanishes from prevalence $\pi = 0.8$ onward.

As verification cost t begins to fall, high types v find it profitable to start verifying news, provided prevalence π is sufficiently high. Indeed, consider $t_2/\ell(v)$ in Figure 2. Notice that $\varphi(\cdot; t_2)$ in (3) is hump-shaped, and thus $\varphi(v; t_2) \equiv t_2/\ell(v)$ for all types $v \geq \arg\max_v \varphi(v; t_2)$. Consequently, if prevalence $\pi > 0.7$, no one engages in unverified sharing (as $\varphi < 0.7$ a.e.); if prevalence $\pi \in [0.4, 0.7)$, types beyond a threshold (soon defined) begin verifying news; and if prevalence $\pi < 0.4$, no user doing unverified sharing has incentives to change her behavior (since $\varphi \equiv b/\ell/(1+b/\ell)$ for low prevalence). In the right panel, the new pass-through curve, $\Sigma(\cdot; t_2)$, is depicted in dashed blue. Observe that it exhibits a non-trivial contraction for mid prevalence ($\pi \in [0.4, 0.7]$) and it vanishes for high enough prevalence ($\pi \geq 0.7$).

In this case of increasing losses, therefore, the types that are more prone to sharing news articles (i.e., high types) are also those more inclined to verify them. Thus, the types engaging in unverified sharing must be in the mid range. Formally, for each prevalence $\pi \in [0, 1]$, the set of users doing unverified sharing is of the form $\mathbf{U} = [v_0(\pi), v_1(\pi; t))$, where

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

(with the convention $v_0(\ ) = \bar{v}$ and $v_1(\ ;t) = \bar{v}$ if their respective sets are empty.[13]) In particular, if $\mathbf{U}$ and $\mathbf{V}$ are non-empty, then a small reduction in verification costs t strictly lowers $v_1(\ ;t)$ but leaves $v_0(\ )$ unaffected, effectively implying that high types switch from sharing unverified news $\mathbf{U}$ to sharing verified information $\mathbf{V}$.

This brings us to the notion of sensitivity introduced earlier. Refer again to Figure 2, and consider $\ = 0{:}4$ in the figure. In this case $\mathbf{N} = [0; v_0(\ ))$, and mildly reducing t from t_2 segments users into three groups:

$$\mathbf{N} = [0; v_0(\ )); \ \mathbf{U} = [v_0(\ ); v_1(\ ;t_2)); \text{ and } \mathbf{V} = [v_1(\ ;t_2); \bar{v}]{:} \tag{7}$$

Importantly, observe that because benefits $b(v)$ increase in v, all types v below $v_0(\ )$—the type who is indifferent between doing unverified sharing and not sharing—necessarily obtain lower utility $((1\ \ \ )b(v)\ \ \ t)$ from verified sharing than the threshold type $v_0(\ )$. As t keeps falling from t_2, the threshold types v_0 and v_1 will, eventually, coincide (i.e., $v_1(\ ;t) = v_0(\ )$), implying that there will be no unverified sharing in the market. That is, the market will segment into two at that point—but critically, no types below $v_0(\ )$ have entered the sharing world yet, as entry emerges only when type v_0 turns to verify news. Thus, reductions in verification costs trigger entry only after that point—a strong form of sensitivity then ensues.

The next result collects these findings and also states how our notion of strong sensitivity manifests when the sharing loss $`$ is *decreasing*.

Proposition 2. *Suppose that b is increasing, and x prevalence $\ 2\ [0;1]$. Then:*

(i) If $`$ is increasing, entry cannot occur before $\mathbf{U}$ is exhausted, i.e., a transfer of mass from $\mathbf{N}$ to $\mathbf{V}$ cannot occur before all users in $\mathbf{U}$ have switched to $\mathbf{V}$.

(ii) If $`$ is decreasing, entry cannot occur before type $v_0(\ )$ exits $\mathbf{U}$, i.e., a transfer of mass from $\mathbf{N}$ to $\mathbf{V}$ cannot occur before some users in $\mathbf{U}$ switch to $\mathbf{V}$.

Figure 3 illustrates this phenomenon. There, we plot the masses $\Sigma_{\mathbf{U}}(= \Sigma)$, $\Sigma_{\mathbf{V}}$ and $\Sigma_{\mathbf{N}} = 1\ \ \ \Sigma_{\mathbf{U}}\ \ \ \Sigma_{\mathbf{V}}$ of $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{N}$, respectively, as a function of verification cost t (horizontal axis) for a fixed prevalence $\ $. In the left panel, sharing loss $`$ is increasing. Thus, as t decreases, i.e., as we move from right to left along the horizontal axis, all the masses remain constant for large t. At some point, however, $\Sigma_{\mathbf{U}}$ begins to decrease, while

[13] Clearly, these threshold types are unique because each corresponds to the generalized inverse of a monotone function. When interior, $v_0(\ )$ is the user type who is indifferent between sharing news without verifying and not sharing, while $v_1(\ ;t)$ is the user type who is indifferent between verified and unverified news sharing. The left panel of Figure 2 depicts these thresholds types for prevalence $\ = 0{:}4$. Finally, if losses $`$ are decreasing, then type $v_1(\ ;t)$ must be redefined to $v_1(\ ;t)\ \ \ \inf f v\ 2\ [0;\bar{v}] : t{=}`(v)\ \ \ \ g.$

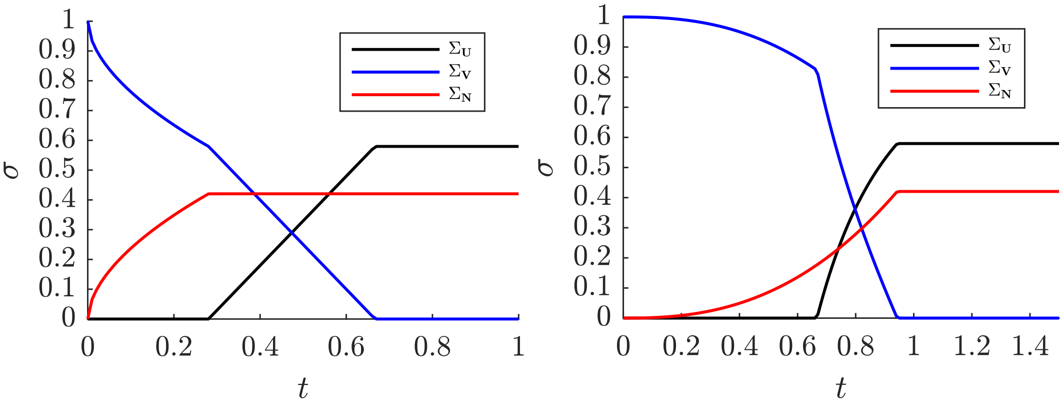

Figure 3: Masses $\Sigma_{\mathbf{U}}; \Sigma_{\mathbf{V}}$ and $\Sigma_{\mathbf{N}}$. The left panel considers increasing losses, using $\ell(v) = v$ and $b(v) = 4v^{1:8}$ as in Figure 2. The right panel considers decreasing losses, using $\ell(v) = 1=(v + \ )^{0:4}$ and $b(v) = 4v^{0:8}=(v + \ )^{0:4}$, with $\ = 0{:}01$ (the use of any $\ > 0$ renders $\ell(v)$ continuous at $v = 0$, only for consistency with our assumption). In both panels, $b(v)=\ell(v) = 4v^{0:8}$ (so the propensity to share is fixed across these two cases), while $\ = 2=3$ and $v \ \ U[0; 1]$.

$\Sigma_{\mathbf{V}}$ increases by the same magnitude. Eventually, $\Sigma_{\mathbf{U}}$ vanishes while $\Sigma_{\mathbf{V}}$ keeps increasing but now at the expense of $\Sigma_{\mathbf{N}}$.

The right panel considers a decreasing loss function ℓ. As in the previous case, for any given prevalence , entry for low types cannot occur unless $v_0(\ )$ is incentivized to verify. However, unlike in the previous case, $v_0(\ )$ is the first type in $\mathbf{U}$ that will engage in verified sharing as t falls, as its loss is the highest among those users. Consequently, as verification cost t falls from high values, $\Sigma_{\mathbf{V}}$ takes off from zero at the same time that both $\Sigma_{\mathbf{N}}$ and $\Sigma_{\mathbf{U}}$ begin decreasing, reflecting Proposition 2-(*ii*). In particular, whenever the market segments into three sets, it is the intermediate types (i.e., those around $v_0(\ )$) that verify news; in turn, low types do not share, while high types engage in unverified sharing—as seen in Figure 9 in Appendix A.6.

4.2 Decreasing benefits

With a propensity to share that increases across types (Assumption 1), decreasing benefits b necessarily imply that losses ℓ must decay too, albeit at a higher rate. We can now establish our main result on weak sensitivity of the pass-through curve Σ: lowering verification costs can have a positive and non-trivial effect on the circulation of true content, but need not decrease the sharing of misinformation.

Proposition 3. *Fix prevalence $\ \ 2\ [0; 1]$ and suppose b is decreasing (and so ℓ is decreasing too). Then, entry occurs strictly before* $\mathbf{U}$ *is reduced, i.e., as verification cost t falls from high values, a transfer of mass from* $\mathbf{N}$ *to* $\mathbf{V}$ *occurs before users in* $\mathbf{U}$ *switch to* $\mathbf{V}$.

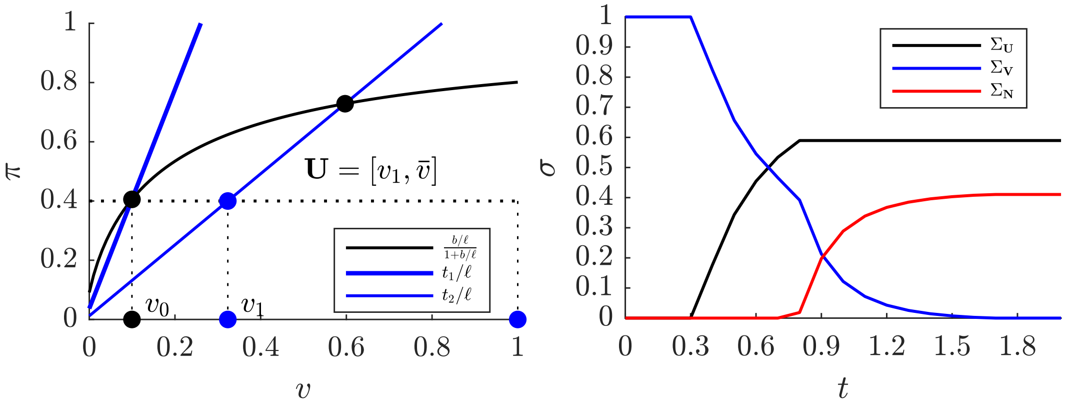

Figure 4: Propensities to share news and skip verification (left), and masses $(\Sigma_{\mathbf{V}}; \Sigma_{\mathbf{U}}; \Sigma_{\mathbf{N}})$ (right) when types $v \quad U[0;1]$ and losses and benefits are decreasing according to $`(v) = 0{:}5{=}(v + \ \ )$ and $b(v) = 2{=}(v + \ \ )^{0:2}$, with $\ \ = 0{:}01$. As in the previous figure, $\ \ = 2{=}3$ in the right panel. (Without loss, the loss ratio $b(v){=}`(v) = 4(v + \ \ )^{0:8}$ shifts up compared to the previous figures.)

In the left panel of Figure 4, lowering verification costs results in a clock-wise rotation of the increasing straight lines encoding $t{=}`$. In the right panel, this generates entry and no reductions in unverified sharing for large t: $\Sigma_{\mathbf{V}}$ grows by the same amount that $\Sigma_{\mathbf{N}}$ falls. To understand why, recall that both *losses and benefits* are decreasing. Types below the threshold $v_0(\ \ )$—the first type doing unverified sharing—then decide against doing unverified sharing because their potential (large) losses outweigh their potential (large) benefits. Since benefits are decreasing, some low types find it optimal to enter the market as verification costs fall: verifying news enables then to skip such potential losses and enjoy their relatively high benefits from sharing truthful content. But since losses are decreasing too, type $v_0(\ \ )$ will not change his behavior initially: since this type's losses are smaller relative to those of lower types, $v_0(\ \ )$ will demand a bigger drop in verification costs to switch from unverified to verified sharing. Eventually, as seen in the right panel, as t continues dropping, this happens and $\Sigma_{\mathbf{U}}$ begins falling only after a substantial amount of entry has taken place.

Propositions 2-3 can be a useful starting point for understanding the effectiveness of fact-checking initiatives, beyond the role that behavioral biases and/or relationships may have in shaping the incentives to check news. Further, these findings yield testable predictions. To this end, let us return to our two interpretations, in which v reflects either popularity or age.

Start with the latter: if a user's number of followers falls with the user's age on average, we can expect the benefits of sharing truthful content to be lower for older individuals on average. But the associated losses can decay with age too, and at a higher rate: lower reputational losses can originate from both fewer followers and the shorter horizons over which the losses are experienced. In this case, if verification is infeasible, older users engage in unverified sharing, while younger ones act cautiously by refraining from sharing news. By

Proposition 3, as fact-checking services are introduced, younger cohorts immediately begin sharing content that they have certified to be true, but this need not change the equilibrium pass-through rate of misinformation and prevalence if older users continue doing unverified sharing. Our analysis indicates that platforms should see more news being circulated early on largely driven by true content, but a rather constant fake news prevalence.

By contrast, suppose that v is a popularity index within a cohort, i.e., users of a given age are ranked by their number of followers—in this case, both benefits and losses can be increasing across types. If more popular types are more likely to share news (benefits dominate losses), those types do unverified sharing if verification is infeasible. By Proposition 2-(i), the strongest form of sensitivity arises: as fact-checking is introduced, the most popular types begin verifying news, and it is only until unverified sharing stops that less popular types enter the market. The prediction is that, conditional on a change in equilibrium prevalence (recall Figure 1), a platform should experience a decrease in both fake news prevalence and overall news circulation as the service becomes available early on.

Let us conclude with two observations. First, as argued, an increasing propensity to share function $b{=}`{=}(1 + b{=}`)$ is simply a natural way of ordering types. However, we have also imposed that b and $`$ are monotone: this is conceptually useful—it allows us to uncover the notions of weak and strong sensitivity discussed—and it constitutes a reasonable approximation when thinking about age and popularity. Importantly, the technical advantage of this assumption is minimal. For instance, Figure 10 in Appendix A.6 shows that, if losses are non-monotone yet the ranking of types is fixed according to their propensity to share, the pass-through curve exhibits qualitatively similar contractions as t falls; the main difference is that $\mathbf{U}$ ceases to be a single interval—how users segment is more subtle.[14]

Second, our last discussion on overall news circulation brings us to the topic of *diffusion of fake news*; in particular, its interplay with verification incentives, which we examine in the next section. To offer the maximum scope for verification effects affecting outcomes, we assume that both b and $`$ are increasing in what follows. In this way, the misinformation pass-through curve necessarily contracts when verification is at play.

5 Supply Interventions and Fake News Diffusion

Among the variety of responses to combat misinformation, platforms have taken actions that aim to reduce the profitability of fake content. This section uncovers when and why

[14] Lack of sensitivity of $\Sigma(\ ;t)$ to t for sufficiently low prevalence rates   is a generic property in the model because $`$ is finite, so $t{=}` > 0$ for all t 0. Thus, for sufficiently low values of , namely, $t{=}`(\ ) >$, no user type has incentives to engage in news verification.

the resulting *weaker* producers' incentives can have the downside of increasing the *diffusion* of misinformation.

Diffusion of fake news. As a starting point, note that $\Pi(\ ) = F(\ )$ is effectively a traditional supply function. Thus, interventions that limit the supply of fake news can be modeled as standard inward supply shifts that capture weaker incentives to produce fake content. For instance, banning repeat offenders can be seen as a cost for producing fake news articles $f \ 2 \ (0; 1)$ that is orthogonal to that of content generation (e.g., change of identity and website appearance), resulting in a new supply

$$\Pi(\ ) = F(\qquad f):$$

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 $\ 2 \ (0; 1)$ such that

$$\Pi(\ ) = F(\quad):$$

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 5 depicts this phenomenon.

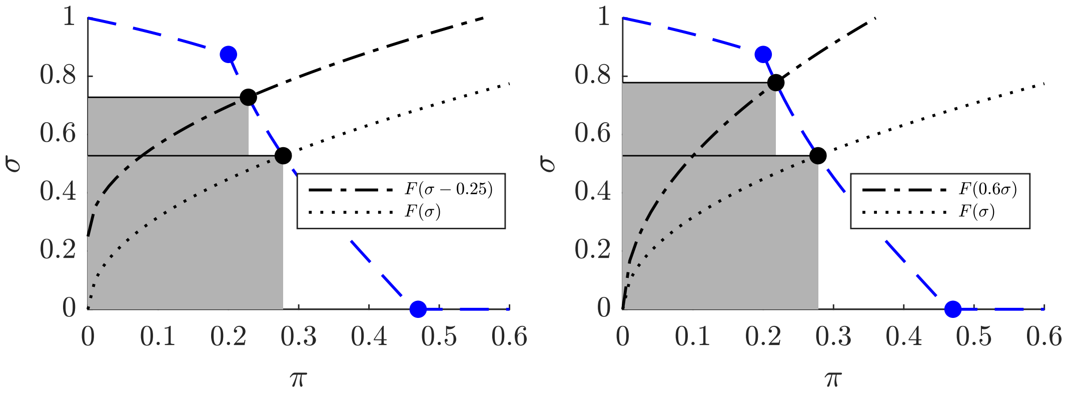

Figure 5: Left panel: $\Pi(\qquad f)$, $f \ = 0{:}25$. Right panel: $\Pi(\quad)$, $\quad = 0{:}6$. Both panels consider parameter values: $\bar{v} = 1$, $G(v) = v$, $F(c) = c^2$, $b(v) = v^2$, $`(v) = v{=}2$, and $t = 0{:}1$.

The novelty is that the shaded areas in Figure 5 capture a key measure of interest: the amount of fake news in circulation, or the *diffusion rate* of fake news articles: $\Delta \qquad\qquad :$ In the remainder of this section, we seek conditions under which a reduction in prevalence leads to an increase in diffusion Δ , exactly as depicted in the shaded areas in Figure 5 when the supply of fake news contracts.

Elasticities. Our approach consists of deriving local *sensitivity measures* analogous to traditional elasticities that are key in this respect. To this end, let us denote

$$E_z(\ ) := \frac{@\log(\ )}{@\log z}$$

for the elasticity of a differentiable function with respect to the variable z. Equipped with this, let $\Delta(\ ) := \Sigma(\ ;t)$ denote the extent of diffusion as a function of prevalence. It is then easy to verify that, when b and ` are differentiable—which we assume in what follows—$\Delta(\ )=@\ < 0$ if and only if

$$jE\ (\Sigma)j > 1: \tag{8}$$

In other words, diffusion Δ increases after a decrease in prevalence , as long as the misinformation pass-through curve Σ is elastic. (Recall that $@\Sigma=@\quad 0$ always.)

Next, as defined in Section 4, $\Sigma(\ )\quad \Sigma_{\mathbf{U}}(\ )$, $\Sigma_{\mathbf{V}}(\ )$, and $\Sigma_{\mathbf{N}}(\ )$ denote the mass of users who share unverified news, verified news, and do not share, respectively, given prevalence . Furthermore, since benefits and losses are increasing (Section 4.1), then, provided the pass-through curve Σ is positive, it follows by (7) that

$$\Sigma_{\mathbf{N}}(\ ) = G(v_0(\ )); \ \Sigma(\ ) = G(v_1(\ )) \quad G(v_0(\ )); \text{ and } \Sigma_{\mathbf{V}}(\ ) = 1 \quad G(v_1(\ ));$$

where $v_0(\ )$ and $v_1(\ )$ are defined in (6). Thus, since $\Sigma = 1 \quad \Sigma_{\mathbf{N}} \quad \Sigma_{\mathbf{V}}$, we deduce that

$$jE\ (\Sigma)j = \quad \frac{\Sigma_{\mathbf{N}}}{\Sigma}\quad E\ (\Sigma_{\mathbf{N}}) + \quad \frac{\Sigma_{\mathbf{V}}}{\Sigma}\quad E\ (\Sigma_{\mathbf{V}}); \tag{9}$$

where all the terms on the right-hand side are non-negative. This is intuitive: an increase in fake news prevalence can lead to a strong (percentage) reduction in unverified sharing Σ because many users decide to "exit" the sharing world ($E\ (\Sigma_{\mathbf{N}})$ term) and/or because many of them decide to begin verifying news ($E\ (\Sigma_{\mathbf{V}})$ term).

The advantage of (9) is twofold. First, from an empirical perspective, it depends on users' decisions to share and verify news, which is data readily available to platforms. Second, from a conceptual perspective, identity (9) can be used to inform under what circumstances (8) holds by finding intuitive conditions on primitives that guarantee that either $\Sigma_{\mathbf{N}}$ or $\Sigma_{\mathbf{V}}$ are sufficiently elastic themselves. We state the conditions in the next result.

Proposition 4. *Suppose that* b *and* $`$ *are increasing and that Assumption 1 holds.*

(a) Suppose that the equilibrium $(\ \ ;\ \ )$ *is such that* $\quad \Sigma_{\mathbf{N}}(\ \ )$*. Then,* $jE\ (\Sigma)j >$ $E\ (\Sigma_{\mathbf{N}})$*. A sufficient condition for* $E\ (\Sigma_{\mathbf{N}}) > 1$ *is that* $b(v)=`(v)$ *is concave and the density* $g(v)$ *is non-decreasing.*

(b) Suppose that the equilibrium $(\ \ ;\ \ )$ *is such that* $\quad \Sigma_{\mathbf{V}}(\ \ )$*. Then,* $jE\ (\Sigma)j >$ $E\ (\Sigma_{\mathbf{N}})$*. A sufficient condition for* $E\ (\Sigma_{\mathbf{V}}) > 1$ *is that* $1=`(\ )$ *is concave and the density* $g(v)$ *is non-increasing.*

Consider (a). From (9), it is straightforward to conclude that the elasticity of $\Sigma_{\mathbf{N}}$ is a lower bound for the elasticity of Σ. The rest of the conditions in turn ensure that $\Sigma_{\mathbf{N}}$ has an elastic response to an increase in . Indeed, since $\Sigma_{\mathbf{N}}(\ \ ) = G(v_0(\ \ ))$, it follows that

$$E\ (\Sigma_{\mathbf{N}}) = E_v(G(v_0(\ \ ))) \quad E\ (v_0(\ \ ));$$

which means that it suffices to have both an elastic distribution G and an elastic threshold type v_0—the last (popularity-wise) type doing unverified sharing, which is the "marginal" type in this case. In turn, these are guaranteed by a convex distribution G and a concave benefit-to-loss ratio $b=`$, respectively. Intuitively, as more popular types—i.e., those more prone to sharing news—become increasingly more abundant (convex G), an increase in prevalence can lead to a strong exit of users from the sharing world for a fixed change in v_0. And if the propensity to share has slow growth (as $b=`$ is concave), any increase in prevalence leads to a large increase in the marginal type v_0, analogous to a sharp reduction in the quantity demanded after a price increase in the case of an elastic demand curve.

Similarly, the conditions in (b) ensure a significant exodus of users from $\mathbf{U}$ to $\mathbf{V}$ after a (percentage) increase in prevalence. Indeed, since $\Sigma_{\mathbf{V}}(\ \ ) = 1 \quad G(v_1(\ \ ))$, by symmetry relative to case (a), this happens when G is concave and v_1 (which is decreasing in) is elastic; but the latter is guaranteed when $1=`$ falls at decreasing rates due to the (verification) marginal type satisfying $t=`(v_1) = \ \ :$ as with $b=`$ in (a), this reflects an elastic $t=`$, ensuring that a large number of less popular types become willing to verify news after increases.

One advantage of the elasticity test in part (a) in Proposition 4 is that can be applied to situations in which the equilibrium prevalence lies on the "rigid" part of the misinformation pass-through curve—namely, the prevalence region where $\Sigma_{\mathbf{V}} = 0$ for small changes in t—as the elasticity of Σ is determined by that of $\Sigma_{\mathbf{N}}$ in those regions.[15] On the other hand, (b) is applicable only to situations in which verification does take place in equilibrium, as it

[15] Of course, the exact region will depend on the hypothetical change in t considered. But so long as $\Sigma_{\mathbf{V}} = 0$, it follows that $\Sigma = 1 \quad G(v_0(\ \ ))$, where $v_0(\ \ )$ is independent of t if $\Sigma > 0$.

demands $\Sigma_{\mathbf{V}} > 0$. (In those regions, test (a) could be used as well, as users can be segmented into three non-trivial sets as argued.) Two observations are instructive in this regard. First, Proposition A.1 in the Appendix shows that (a)-(b) admit minimal modifications so the conditions on the density g are not mutually exclusive, thereby allowing for the possibility of a non-monotone density g for instance; in turn, this means that (a) and (b) are potentially applicable depending on where specifically the equilibrium prevalence lies.[16] Second, less stringent tests can be easily obtained using (9): for instance, after dropping the requirement that $< \Sigma_{\mathbf{N}}(\ )$ in (a), one can check ex post that $E\ (\Sigma_{\mathbf{N}})(\ ) > \Sigma(\ )=\Sigma_{\mathbf{N}}(\ )$ to verify an elastic pass-through curve Σ at .

Part (b) is also useful for examining how the overall elasticity of Σ responds to changes in t. Specifically, recall from Section 4.1 that the pass-through curve Σ exhibits *strong sensitivity* to verification costs, in the sense that a partial reduction in unverified sharing resulting from a drop in t necessarily occurs via users in **U** switching to **V** exclusively (i.e., $@\Sigma_{\mathbf{N}}=@t = 0$). As a result, it suffices to show that $\Sigma_{\mathbf{V}}$ becomes more elastic with respect to as verification cost t falls—i.e., that $@E\ (\Sigma_{\mathbf{V}})=@t < 0$—to ensure that the pass-through curve Σ inherits the same property.[17] A mild strengthening of the conditions on primitives in (b) guarantees this to happen.

Proposition 5. *Suppose that the hazard rate $g(v)=(1\ \ G(v))$ is non-increasing and that $1=`(v)$ is concave. Then, as long as $\Sigma_{\mathbf{V}} > 0$, the elasticity of the pass-through curve $jE\ (\Sigma)j$ strictly increases as verification cost t decreases.*

Proposition 5 is useful because it can inform, in conjunction with Proposition 4-(b), on when ***joint policies*** can potentially reinforce each other negatively. Specifically, by making the misinformation pass-through curve more sensitive to reductions in prevalence, improvements in fact-checking can result in supply interventions acting as a catalyst for increasing the diffusion of misinformation in a platform.

Relative to Proposition 4-(b), the conditions on primitives ensure that the aforementioned exodus of users from **U** to **V** after an increase in prevalence in fact becomes *stronger* as verification cost t decreases. In turn, this is guaranteed to happen whenever

$$E\ (\Sigma_{\mathbf{V}}) = E_v(1\ \ G(v_1(\ )))\quad E\ (v_1(\ )) = \frac{g(v_1)}{1\ \ G(v_1)}\qquad \frac{@v_1}{@} \tag{10}$$

[16] Condition (a) for g is weakened to hold over $[0; v_0(\ )]$ only, while the analogous one in (b) over $[v_1(\ ); \bar{v}]$, where $v_0 < v_1$. Consider now Figure 1: in the low prevalence equilibrium, $v_1 = \bar{v}$ (no one verifies news), so v_0 matters for the elasticity of Σ. Conversely, in the high prevalence equilibrium, the fact that $v_1(\ ) < \bar{v}$ makes (b) applicable. Trivially, a decreasing concave $1=`$ does not preclude an increasing concave $b=`$.

[17] Using (9), this follows from $(\Sigma_{\mathbf{N}}=\Sigma)$ and $(\Sigma_{\mathbf{V}}=\Sigma)$ growing after a fall in t, and the fact that $@\Sigma_{\mathbf{N}}=@t = 0$.

is decreasing in t. Indeed, since the "marginal type" engaging in verification, v_1, is increasing in t and decreasing in , it suffices to have a decreasing hazard rate and a supermodular v_1. In other words, either g must decay sufficiently fast as popularity increases or, as verification cost t falls, type v_1 must fall faster following an increase in prevalence .

6 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 as a form of censorship (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 2 $[0; 1]$. Because of the public announcements that platforms have made on this topic, we assume that changes in , 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 from zero.

Our analysis from Section 3 admits a direct adaptation to this case. To this end, suppose that a mass 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

$$(\quad;\quad)\qquad\frac{(1\qquad)}{1};$$

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

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

$$(1\qquad)\Sigma(\;(\;;\;))): \tag{11}$$

Each potential fake news producer then takes as given the (candidate) equilibrium pass-through rate, 2 $[0; 1]$, which leads to a supply curve $\Pi(\;)=F(\;)$ as in (5). Also, as in

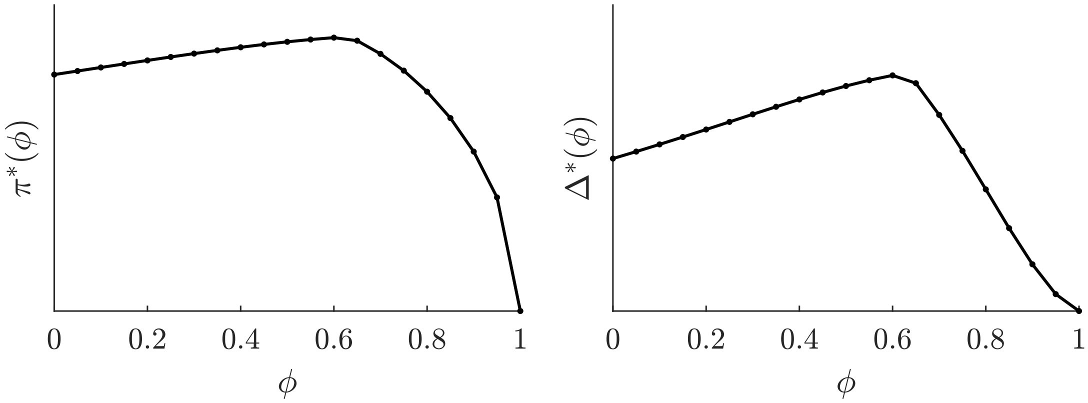

Figure 6: Parameter values: $\bar{v} = 1$, $G(v) = v$, $F(c) = {}^{p}\bar{c}$, $b(v) = v^2$, $`(v) = v{=}2$, and $t = 0{:}1$.

Section 3, the unique equilibrium (;) is given by the intersection of the pass-through of fake news and the supply curve, i.e., $\Pi((1 \quad)\Sigma(\quad (\quad ; \quad))) = \quad$ and $\quad = (1 \quad)\Sigma(\quad (\quad ; \quad)))$.

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 6. *(a) Suppose that $b(v){=}`(v)$ is concave and the density $g(v)$ is increasing. If $\quad \Sigma_{\mathbf{N}}(\quad)$ when $\quad = 0$, both the equilibrium production, , and the rate of diffusion, Δ , are increasing in at $\quad = 0$ and eventually decrease in the same variable.*

(b) Suppose that $1{=}`(v)$ is concave and that the hazard rate $g(v){=}(1 \quad G(v))$ is non-increasing. If $\quad \Sigma_{\mathbf{V}}( \quad )$ when $\quad = 0$, and $g(0) > `^0(0){=}`(0)$, then there exists a verification cost $\hat{t} > 0$ such that for all $0 < t < \hat{t}$ the same conclusion holds.

The presence of a filter implies that encountering content is "good news:" each user is now more optimistic about the veracity of the news item, encapsulated in $(\quad ; \quad) <$ for all $\quad 2 \ (0; 1)$. With greater optimism, there is more entry to the unverified sharing world: users in **N** switch to **U**. Part (a) in the proposition shows that the requirements of Proposition 4-(a)—which guarantee an elastic $\Sigma_{\mathbf{N}}$ in the absence of a filter, and hence a substantial **N** ! **U** switch—are enough to ensure an increase in unverified sharing that more than outweighs the extra layer of protection that a filter provides—at least for low values of . Figure 6 shows how this effect can be strong for a wide range of levels of filter quality.

But such optimism can also lead to a weakening of the incentives of those verifying: users in **V** switch to **U**. Similarly to (a), part (b) in the proposition shows that the requirements of Proposition 5 can be exploited to generate a sufficiently elastic $\Sigma_{\mathbf{V}}$, but this time provided verification costs are sufficiently low: in this case, there is enough verification taking place in equilibrium that the addition of a filter can generate a strong relaxation of verification incentives. This strong relaxation is in the form of $\Sigma_{\mathbf{V}}$ satisfying that, when $\quad = 0$,

$$E\ (\Sigma_{\mathbf{V}}) \quad \frac{1}{1};$$

which is more demanding than the unit elasticity analog, in a reflection of the additional protection that the filter provides around $\quad = 0$. Whenever there is enough elasticity around the lowest type $v = 0$, i.e., $g(0)`(0)=`^{0}(0) > 1$, the above will be satisfied as t falls due to v_1 becoming small—and we can show that the requirement $\quad < \Sigma_{\mathbf{V}}(\quad)$, which ensures that $E\ (\Sigma) > E\ (\Sigma_{\mathbf{V}})$, becomes automatically satisfied for low enough verification costs.[18]

7 Extensions

In this section, we extend our model to allow for (i) market power and (ii) network externalities. The former can be relevant for at least three reasons: supply interventions pressure the industry to shrink over time, at least in the short term; the costs of producing digital fake content are largely fixed, and they increase as technologies become more sophisticated (e.g., "deep fakes"); and recent work has documented the existence of parent companies that own several untrustworthy sites (Sydell, 2016). As for network externalities, it is well-recognized that the value of platforms (in a broad sense) critically depends on how many individuals use it—in addition, in the particular case of social media, such platforms are a natural setting where elements of social influence likely shape the behavior of users.

7.1 Market Power

Consider a monopolistic fake news producer. For consistency with our previous "competitive" analysis, we assume that the monopolist inherits the cost structure of the competitive case, encapsulated in the distribution F: the monopolist is either a single producer who experiences marginal cost $F^{\ \ 1}(\quad)$ when $\quad$ fake articles have been produced, 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 $c \quad F^{\ \ 1}(\quad)$ would be active when fake articles are to be produced, and hence the same marginal cost interpretation ensues.

Uniform policies. The monopolist must choose the mass $\quad 2\ [0;1]$ of fake news articles to be produced—since the choice variable is a one-dimensional scalar, we say that the policy

[18] It is interesting that part (a) here does not require stronger conditions relative to Proposition 4-(a) to satisfy the analog condition for $\mathbf{N}$, namely, $E\ (\Sigma_{\mathbf{N}}) > \frac{1}{1}$. This is because v_0 is inherently more elastic than v_1: changes in $\quad$ affect both margins in the unverified sharing constraint $(1 \quad)b > \quad `$ which pins down $v_0$, but only one margin in the no-verification constraint $t > \quad `$ that determines v_1.

is *uniform*. If the pass-through rate of fake content is given by when prevalence takes value , the monopolist's profits then given by

$$\int_0 \Pi\ ^1(\ ^0)d\ ^0; \qquad (12)$$

where $\Pi(\ ) = F(\ )$ denotes the supply of fake news.[19] In other words, profits are captured by the area between and the (inverse) supply $\Pi\ ^1$ over the interval $[0;\ ]$.

In traditional market power, we would use the fact that $\ = \Sigma(\ )$ to optimize (12) and find an optimum; call it ^{M}. Since Σ is downward sloping, $^M <$. Implicit 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. From section 2, however, neither nor are directly observable by users, but rather *anticipated in equilibrium*. Indeed, this is a key institutional feature that distinguishes our setting from markets for traditional goods.

This informational asymmetry renders the situation as one of simultaneous moves from a strategic viewpoint. From this perspective, a simple inspection of (12) reveals that the equilibrium of Section 3 is unchanged: $(\ ;\ )$ is the only tuple such that (i) the monopolist behaves sequentially rational given a fixed misinformation pass-through rate, and (ii) users hold correct beliefs given the monopolist's choice of fake news prevalence. A takeaway, therefore, is that policies that foster platform transparency, such as revealing estimates to users, may facilitate the monopolist's ability to exercise market power. Conversely, platform *opacity* can be used to harm "large" fake news producers by lowering their profits.

Segmentation. That said, the monopolist can attempt to improve upon the competitive outcome by resorting to segmentation strategies. To see how this can be done, consider Figure 7: there, the competitive outcome is denoted by E, and we have depicted a segment $\overline{AB}$ that intersects the supply of fake news at the point C. Clearly, if the latter point is achievable, it will yield more profits than E (recall the area representation of profits).

Point $C := (\ _C;\ _C)$ can be implemented by segmenting the original market into two submarkets. Indeed, letting $_A$ and $_B$ denote the projections of A and B on the axis, there is $\ 2\ (0;1)$ such that $_C = \ _A + (1\ \ )\ _B$. The monopolist could then create two submarkets—call them A and B—with intended prevalence levels $_A$ and $_B$, as follows:

1. Send $_C$ fake news articles to segment A, and $(1\ \ )\ _C$ fake news articles to B;
2. If $!$ and $1\ \ !$ are the sizes of A and B, respectively, send $(1\ \ _C)(1\ \ !)$ truthful news articles to B, the *high prevalence* market;

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

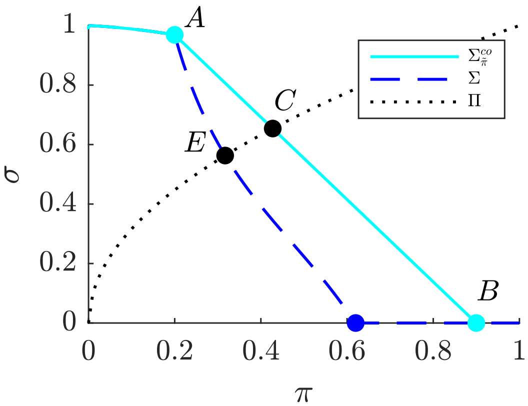

Figure 7: User segmentation as the concave closure of the pass-through cruve Σ over $[0; \tilde{}]$. The figure considers the parametrization of Figure 1 with verification cost $t = 0{:}1$ and $\tilde{} = 0{:}9$. By segmenting the user population, the monopolist can reverse the effects of verification.

3. Supplement the remaining volume of truthful news $(1 \quad {}_C)!$ to be sent to submarket A—the ***low prevalence*** submarket—with additional news;
4. This procedure generates levels of prevalence $\frac{{}_C}{{}_C + (1 \quad {}_C)! +}$ and $\frac{(1 \quad) {}_C}{(1 \quad) {}_C + (1 \quad {}_C)(1 \quad !)}$ for A and B, respectively. The parameters $! \ 2 \ (0; 1)$ and > 0 are then chosen so that

$$ {}_A = \frac{{}_C}{{}_C + (1 \quad {}_C)! +} \qquad \text{and} \qquad {}_B = \frac{(1 \quad) \ {}_C}{(1 \quad) \ {}_C + (1 \quad {}_C)(1 \quad !)}: \tag{13}$$

The monopolist can always create a high prevalence submarket by shrinking its size given a fixed supply of fake content: in (13), there is a unique $! \ 2 \ (0; 1)$ such that the second equality holds. However, by doing so, the size of the other submarket becomes fixed, so an extra degree of freedom is needed to achieve low prevalence in the second market: in (13), there is a unique volume > 0 of additional truthful news such that the first equality holds. Altogether, the mass ${}_C$ of fake items produced generates a revenue of

$$ {}_C\Sigma(\ {}_A) + (1 \quad) \ {}_C\Sigma(\ {}_B) = \ {}_C[\ \Sigma(\ {}_A) + (1 \quad)\Sigma(\ {}_B)]:$$

That is, point C in $\overline{AB}$ is implemented, and fake news profits grow.[20]

Towards the formal result, let us consider ***partial concavifications*** of Σ. Namely, given $0 < {}_{\tilde{}} < \tilde{} < 1$, let $\Sigma^{co}_{\tilde{};\tilde{}}$ denote the infimum of the concave functions that are weakly

[20] Notice that as a consequence of adding truthful content to submarket A, fake news prevalence in the *whole platform* becomes ${}_C{=}(1+ \ )$. Also, since the costs of producing truthful content by an unverified news provider are fairly low—untrustworthy outlets can free ride on established ones—a payoff criterion based on false content exclusively can be a good approximation if in addition truthful content diffuses more slowly than misinformation does, a feature that has empirical support recently (e.g., Vosoughi et al., 2018).

larger than Σ over $[\underset{\sim}{} ; \tilde{}]$. Partial concavifications eliminate the possibility that $_A = 0$ or $_B = 1$ in our previous discussion—i.e., segmentations in which the low-prevalence segment either receives an infinite volume of truthful content ($= \mathit{1}$) or has negligible size ($! = 0$), respectively, both of which are clearly unrealistic.

Proposition 7. *Given* $0 < \underset{\sim}{} < \tilde{} < 1$, *any point* $\Sigma^{co}_{\underset{\sim}{};\sim}$ *can be implemented by creating at most two subpopulations that only differ in their size, and where one sub-market is supplied with additional truthful content. Further, if the competitive equilibrium is such that* $\Sigma^{co}_{\underset{\sim}{};\sim}(\;) > \Sigma(\;)$, *then segmenting the market can be profitable.*

The type of segmentation just described is *trivial* in that the sub-populations do not differ on observable characteristics, but only on their size.[21] The strategy may seem naive, especially in light of the granularity that some platforms offer in their targeting services: our analysis uncovers that basic forms of segmentation need not be associated with negligible harm. That said, the implementation of this form of segmentation requires users in each group not only correctly anticipating the total number of fake content to be created, but also the corresponding prevalence, which is more demanding than under uniform policies.

Finally, returning to our general model in sections 2–3, note that since b and ` are continuous, it suffices that ` is bounded away from zero to ensure that an increasing propensity to share b=`=$(1+b$=`) is strictly below 1 over $[0; \bar{v}]$. In this case, $\Sigma(\,; t)$ vanishes strictly before 1, and hence the ability to construct partial concavifications is a generic property: a convexity in Σ around the vanishing point can be exploited with a $\tilde{}$ sufficiently close to 1 (in which case $_B$ coincides with $\tilde{}$). Two observations are useful in this respect. First, by shifting this convex region inward (see Figure 1), segmentation strategies become an attractive option to partially offset the reduction in prevalence that would have occurred otherwise. Second, if the monopolist wanted to implement higher prevalence-pass-through rate pairs along the supply curve as described above, this is likely to become increasingly difficult: the concavification involved would exhibit a larger $\tilde{}$—i.e., a more extreme segmentation—which means that the high/low prevalence segment must be larger/smaller in size. In particular, a larger volume of news must be directed to a shrinking low-prevalence segment.[22]

[21] Population sizes are easy to control when targeting in online platforms; for instance, by setting different monetary budgets in (similar) locations of interest.

[22] By contrast, the ability of exploiting concavifications with $\underset{\sim}{}$ very close to zero is not a generic property. For instance, this does not happen when $b(0)$=`$(0) > 0$, in which case @Σ=@ $= 0$ in a neighborhood of zero.

7.2 Network Externalities

We conclude the paper by showing how our framework can easily incorporate elements of "social influence," and how these can lead to nontrivial changes to our original pass-through curve. 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

$$\tilde{`}(v;\ ) := \frac{`(v)}{n(\ )};$$

where corresponds to the misinformation pass-through rate, while n is a differentiable function satisfying $n^{0} > 0$ 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 justify such behavior if many others are engaging in the practice).

To isolate how a relaxation of verification incentives affects the sharing of unverified news, we simply assume the benefits of sharing truthful news articles $b(\ )$ scale in the same manner; in this way, the propensity to share remains unchanged.[23] Specifically, recalling Section 3, the mass of users who share without verifying obeys $\tilde{\Sigma}(\ ;\ )\quad G(v_1(\ ;\ ))\quad G(v_0(\ ))$, where

$$v_0(\ ) = \inf\quad v\ 2\ [0;\bar{v}] : \frac{b(v)=`(v)}{1+b(v)=`(v)}\qquad \text{and}\quad v_1(\ ;\ ) = \inf\quad v\ 2\ [0;\bar{v}] : \frac{tn(\ )}{`(v)}\qquad :$$

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

In equilibrium, users' beliefs about the pass-through rate, , must be correct, so the pass-through curve, $\Sigma(\ )$, must solve the fixed point $\tilde{\Sigma}(\ ;\ ) =$ for each , just as the aggregate quantity demanded must satisfy a fixed-point in traditional models of network externalities. The top panels of Figure 8 illustrate a typical situation: the left panel plots the fixed points for a given value of , while the right panel, the resulting Σ, which now becomes a *correspondence*.[24]

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 veri-

[23] For instance, users may expect other users to eventually leave the platform if unverified sharing behavior were to become more frequent. With fewer users, the value of sharing truthful content decreases (popularity interpretation). Thus, dividing b by n can be seen as a penalty associated with those long-term losses.

[24] 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 phenomenon arises in Kranton and McAdams (2020), where users are embedded in a network structure; there, however, verification incentives are absent.

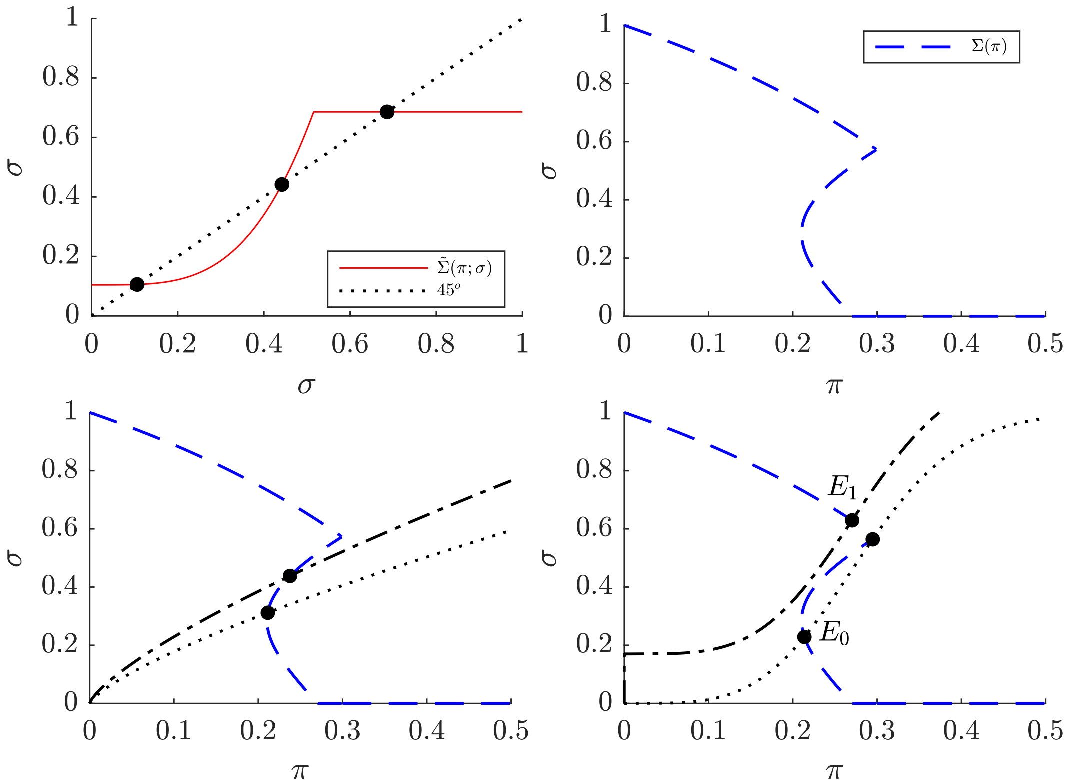

Figure 8: Top left: fixed points of $\tilde{\Sigma}(\ ;\ )$ for prevalence $\ = 0{:}25$. Top right: pass-through correspondence $\Sigma(\ )$. The bottom panels examine supply interventions: the baseline supply falls to $\Pi(\ ) = F(0{:}78\ )$ in the left panel with $F(c) = c^{4=3}$; while in the right panel, the supply falls to $\Pi(\ ) = F(\ \ 0{:}17)$ with F being the CDF of a beta distribution with shape parameters $(0{:}25{;}\,0{:}1)$. Parameter values: $\bar{v} = 1$, $G(v) = v$, $b(v) = v^2$, $`(v) = v$, $t = 0{:}1$, and $n(\ ) = 1 + 10\ \ ^4(3\ \ \ 2\ )$.

fication incentives, making those expectations self-fulfilling. Supply interventions intended to reduce the supply of fake news, as in Section 5, can have stronger unintended effects: in the bottom left panel, this results in a higher prevalence and pass-through rate of fake news, and hence to a larger diffusion rate Δ. In the bottom right panel, these interventions have adverse "refining" effects: they shrink 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).

8 Concluding Remarks

This paper has developed a model of misinformation intended to examine the efficacy of real-world policies deployed to combat the fake news problem, focusing on the ways in which these interventions affect users' incentives to verify news. This issue is of central importance because a common theme in the response of social media platforms to fight misinformation has been to *empower* users: facilitating users' ability to determine the veracity of content without taking away their choice to ultimately decide whether to share.

Our framework is conceptually useful for two reasons. First, by isolating "verification constraints," our model uncovers the possibility of fundamental limitations to the success of several prominent interventions in place today, even when fully rational agents are present; that is, even if behavioral elements stemming from information coming from trusted sources, or from biases linked to ideologies, are not at play.[25] Even more so, our results offer sharp predictions regarding how observable traits such as age and popularity, as well as deeper parameters such as the distribution of gains and losses across those observable traits, are likely to shape outcomes after the introduction of these policies—sometimes in unintended ways, demonstrating that policies perceived to have positive ***direct*** effects may create some risks via channels linked to incentives.

Second, our model has the advantage that it reduces to a ***framework*** featuring forces akin to those of supply and demand. In particular, it identifies prevalence and pass-through rates as analogs of "prices" and "quantities" in a competitive market, giving us a familiar setting to perform policy analyses. We leverage this feature to obtain key elasticity measures that complement our qualitative findings by quantifying the extent to which the potential unintended risks studied can arise, which we believe are a critical aspect of policy evaluation in this area. A potential path going forward is to exploit the tractability of our approach for estimating our elasticity measures under a richer model that focuses not only on verification incentives. In their current form, however, these measures still constitute useful guidance regarding the type of data and sources of variation that are needed to empirically evaluate key interventions in social media platforms. They can also be seen as conservative estimates regarding the unintended risks that can arise.

Finally, our policy exercises have been guided by their practical importance, but others are available. Examples 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.

[25] The rational or "sophisticated" benchmark need not be considered too distant. Indeed, 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—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/`. 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 $\lambda(\varphi^{-1}(\{\theta\})) = 0$ for any $\theta \in \mathbb{R}$.*

Proof: Assume that $\lambda(\{v \in \mathbb{R} : \varphi'(v) = 0 \vee \varphi'(v) \text{ does not exist}\}) = 0$, and consider $\varphi^{-1}(\{\theta\})$ for $\theta \in [0, 1]$. It is well-known that the set of isolated points of the previous set, $ISO(\varphi^{-1}(\{\theta\}))$, 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 $\varphi^{-1}(\{\theta\})$ is closed, there exists and approximating sequence $(v_n)_n$ with $\varphi(v_n) = \theta$ for all n, and so $\varphi'(v^*) = 0$. Consequently, the set

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

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

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 $\theta \in [0, 1]$ and an increasing sequence $(\theta_n)_n$ with $\theta_n \uparrow \theta$ as $n \to \infty$. Since $\theta_n < \theta$, it follows that $V(\theta) \subseteq V(\theta_n)$. Thus,

$$\Sigma(\theta_n) - \Sigma(\theta) = \int_{V(\theta_n)} dG - \int_{V(\theta)} dG = \int_{V(\theta_n) \setminus V(\theta)} dG = \mu(A_n);$$

where $A_n \equiv V(\theta_n) \setminus V(\theta) = \{v : \varphi(v) \geq \theta_n \text{ and } \varphi(v) < \theta\}$. 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 $\lim_{n \to \infty} A_n = \{v : \varphi \geq \theta\} \cap \{v : \varphi(v) < \theta\} = \emptyset$. 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, $\lim_{n \to \infty} \Sigma(\theta_n) = \Sigma(\theta)$.

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

$$\Sigma(\theta) - \Sigma(\theta_n) = \int_{V(\theta)} dG - \int_{V(\theta_n)} dG = \int_{V(\theta) \setminus V(\theta_n)} dG = \mu(B_n);$$

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

$'(v) \quad g$. Thus, $\lim_{n\,!\,1} \quad (B_n) = \quad (\lim_{n\,!\,1} B_n) = \quad (' \;^{1}f \;\; g)$. But, since   is absolutelty continuous with respect to the Lebesgue measure ( ), and $(' \;^{1}f \;\; g) = 0$ by Lemma A.1, it follows that $(' \;^{1}f \;\; g) = 0$, and so $\lim_{n\,!\,1} \Sigma(\;_n) = \Sigma(\;)$.

A.1.2 Proof of Proposition 1

Consider the composite function $\Sigma \quad \Pi : [0;1] \;\; ! \;\; [0;1]$. We'll show that $\Sigma(\Pi(\;))$ has a unique fixed point. First, it is clear that $\Sigma(\Pi(\;))$ is continuous, since it is the composition of two continuous functions. Second, when $= 0$ we have $\Pi(0) = 0$ and so $\Sigma(\Pi(0)) = 1$. Conversely, when $= 1$, we have $\Pi(1) = 1$. We now argue that $\Sigma(1) = 0$. Indeed, notice that $'(v) \quad \frac{b(v)=`(v)}{1+b(v)=`(v)} < 1$ a.e., since $b(v)=`(v)$ is finite a.e. Thus, $V(1) = fv : '(v) = 1g$ has Lebesgue measure zero, and so $\Sigma(1) = R_{V(1)} dG = 0$. Altogether, by the Intermediate Value Theorem, there exists $2\ (0;1)$ with $\Sigma(\Pi(\quad)) = \quad$.

Finally, we show that is unique. Suppose not. Then, without loss of generality, there exists another fixed point $\tilde{} <$. Since $\Pi(\;)$ is increasing, it follows that $\Pi(\tilde{}) \quad \Pi(\quad)$. Thus, $\Sigma(\Pi(\tilde{})) \quad \Sigma(\Pi(\quad))$, since $\Sigma(\;)$ is decreasing. But then, $\tilde{} \qquad$, 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 Proposition 2

Proof (i): We'll show that if not all types $v\ 2$ **U** migrate to **V**, then a transfer from **N** to **V** cannot occur. To this end, first notice that **U** has a positive measure as long as verification cost $t > \quad `(v_0(\;))$. Indeed, since $`$ and $b=`$ are each increasing in $v$, a user with type $v$ finds it optimal to engage in unverified sharing when both $v \quad v_0(\;)$ and $v \quad v_1(\;;t)$, where the threshold types, $v_0(\;)$ and $v_1(\;;t_2)$, are defined in (6). Now, if $t > \quad `(v_0(\;))$ then $v_1(\;;t) > v_0(\;)$, and thus $\mathbf{U} = [v_0(\;);v_1(\;;t)]$ is a non-empty interval with positive mass. Thus, to prove the result we consider a decrease in verification cost from t_1 to $t_2 > \quad `(v_0(\;))$.

First, as previously argued, $\mathbf{U} = [v_0(\;);v_1(\;;t_2)]$ remains non-empty and with positive measure after the cost reduction. Second, consider the user types that find it optimal not to share, i.e., $v\ 2$ **N**. These types prefer not sharing to doing unverified sharing; thus, any $v\ 2$ **N** is bounded above by $v_0(\;)$. But since $v_0(\;) < v_1(\;;t_2)$, it follows that $v < v_1(\;;t_2)$ for all $v\ 2$ **N**. Finally, since the sharing loss function $`(\;)$ is increasing, this means that every $v\ 2$ **N** prefers unverified sharing to verified sharing, implying that no type $v\ 2$ **N** will find it optimal to switch to **V**. Thus, a transfer from **N** to **V** cannot occur before all $v\ 2$ **U** have migrated to **V**.

Proof (ii): We'll show that if no type $v \in \mathbf{U}$ switches to $\mathbf{V}$, then there cannot be a transfer from $\mathbf{N}$ to $\mathbf{V}$. To this end, first notice that because the loss function $\ell(\ )$ is decreasing, the set of users who engage in unverified sharing is now given by $\mathbf{U} = [\max\{v_0, v_1\}, \bar{v}]$. Indeed, since b/ℓ is increasing, type v prefers unverified sharing to not sharing when $v \quad v_0(\ )$; however, since ℓ is decreasing, the same type prefers unverified sharing to verified sharing when $v \quad v_1(\ , t)$, where $v_0(\ )$ and $v_1(\ , t_2)$ are given by (6). Therefore, $\mathbf{U}$ is unaffected by verification cost t as long as $v_0(\ ) > v_1(\ , t)$, or equivalently $t > \quad \ell(v_0(\ ))$.

Next, consider a reduction in verification cost from t_1 to $t_2 > \quad \ell(v_0(\ ))$. As discussed above, no type $v \in \mathbf{U}$ will switch after the verification cost falls, and thus we must have $v_0(\ ) \quad v_1(\ , t_2)$. We now show there cannot be a transfer from $\mathbf{N}$ to $\mathbf{V}$. To this end, let us define $v_2(\ , t)$ as the type that is indifferent between verified sharing and not sharing:

$$v_2(\ , t) \quad \inf\{v \in [0, \bar{v}] : (1 \quad)b(v) \quad t \quad 0\}, \tag{A.1}$$

with the convention that $v_2(\ , t) = \bar{v}$ if the set is empty. Next, we prove that $v_0(\ ) \quad v_2(\ , t_2)$. Indeed, type $v_0(\ )$ obtains a zero payoff from doing unverified sharing, and prefers unverified sharing to verified sharing; thus, $(1 \quad)b(v_0) \quad t_2 \quad 0 \quad (1 \quad)b(v_2) \quad t_2$, where the last inequality follows from the definition of v_2. Since the benefit function $b(\ )$ is increasing, we have $v_0(\ ) \quad v_2(\ , t_2)$. Combining this with the fact that no type $v \in \mathbf{U}$ switches after the cost reduction takes place, we see that $v_1(\ , t_2) \quad v_0(\ ) \quad v_2(\ , t_2)$. This double inequality implies that, in fact, almost no type v finds it optimal to engage in verified sharing, since this would require verified sharing to dominate both not sharing (i.e., $v \quad v_2$) and unverified sharing (i.e., $v \quad v_1$), which cannot happen generically as $v_1 \quad v_2$. We conclude that there cannot be a transfer of positive mass from $\mathbf{N}$ to $\mathbf{V}$ if no users in $\mathbf{U}$ switch first to $\mathbf{V}$.

A.2.2 Proof of Proposition 3

As in the proof of Proposition 2-(ii), since b/ℓ and ℓ are respectively increasing and decreasing in v, the set of users who engage in unverified sharing is $\mathbf{U} = [\max\{v_0, v_1\}, \bar{v}]$, with $v_0(\ )$ and $v_1(\ , t_2)$ defined in (6). We'll show that if $\mathbf{U}$ remains constant after a reduction in t, then a positive mass of users transfers from $\mathbf{N}$ to $\mathbf{V}$. To this end, consider t_1 and t_2 with $t_1 > t_2 > \quad \ell(v_0(\ ))$. This last inequality ensures that $\mathbf{U}$ remains constant after t falls from t_1 to t_2, as it implies that $v_0(\ ) > v_1(\ , t_2) > v_1(\ , t_1)$.

Now, consider types $v \in \mathbf{N}$, and $v_2(\ , t)$ defined in (A.1). Since b is decreasing, v_2 falls in t. Also, any type $v \in \mathbf{N}$ prefers not sharing to verified sharing (i.e., $v \quad v_2$) and prefers not sharing to unverified sharing (or, $v \quad v_0$). So if $t > \quad \ell(v_0)$ then $v_0 > v_2$, since benefits b are decreasing and type v_2 obtains a higher payoff from engaging in unverified sharing than type v_0. Consequently, $\mathbf{N} = [v_2; v_0]$ is a non-empty interval with positive mass.

Having characterized $\mathbf{N}$, notice that as t falls from t_1 to $t_2 < t_1$, every type v in the interval $[v_2(\ ;t_1); v_2(\ ;t_2))$ will exit $\mathbf{N}$. Clearly, this set has a positive mass. We'll show that all these types migrate to $\mathbf{V}$, namely, each turns to share verified news. First, because b is decreasing, any type $v < v_2(\ ;t_2)$ prefers verified sharing to not sharing. It remains to check that any type $v < v_2(\ ;t_2)$ also prefers verified sharing to unverified sharing (which would happen if $v_2(\ ;t_2) < v_1(\ ;t_2)$). But this is true because $v_0(\ ) > v_1(\ ;t_2)$ (as $\mathbf{U}$ remains constant) and so the following inequality holds:

$$(1 \quad)b(v_2) \quad t_2 = 0 = (1 \quad)b(v_0) \quad `(v_0) > (1 \quad )b(v_1) \quad `(v_1) = (1 \quad)b(v_1) \quad t_2:$$

Thus, $v_2(\ ;t_2) < v_1(\ ;t_2)$ since b is decreasing. Altogether, we have shown that every type $v\ 2\ [v_2(\ ;t_1); v_2(\ ;t_2))$ finds it optimal to share verified news, i.e., they switch from $\mathbf{N}$ to $\mathbf{V}$ as verification cost t falls from t_1 to t_2. 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: $v_0(\ )$ is convex for $\quad ^{-}{}_0$. Let $p(v) \quad b(v)=`(v)$ with $p(0) = \lim_{v\#0} b(v)=`(v)$. By (6), it follows that $v_0(\ )\ 2\ (0; \bar{v})$ is determined by $p(v_0(\ )) \qquad =(1 \qquad)$. Also, since $p : [0; \bar{v}] \ ! \ [p(0); p(\bar{v})]$ is strictly increasing and concave, its inverse $p\ ^{1} : [p(0); p(\bar{v})] \ ! \ [0; \bar{v}]$ is strictly increasing and convex. Thus, for $\_{}_0$ and $^{-}{}_0$ respectively solving $p(0) = \_{}_0=(1 \quad \_{}_0)$ and $p(\bar{v}) = {}^{-}{}_0=(1 \quad ^{-}{}_0)$, we have $v_0(\ ) = p\ ^{1} \frac{}{1}$ for $\quad 2\ [\_{}_0; ^{-}{}_0]$. For $\quad < \_{}_0$, we have $v_0(\ ) = 0$ by (6). Since the map $\quad 7! \quad =(1 \qquad)$ is increasing and convex, it follows that $v_0(\ )$ is increasing and convex for $\qquad ^{-}{}_0$.

Step 2: If $\qquad \Sigma_{\mathbf{N}}(\quad)$ then $jE\ (\Sigma)j \quad E\ (\Sigma_{\mathbf{N}})$. Let $\ _{\mathbf{V}} = \Sigma_{\mathbf{V}}(\quad)$ and $\ _{\mathbf{N}} = \Sigma_{\mathbf{N}}(\quad)$. Since $\Sigma(\ ) = 1 \quad (\Sigma_{\mathbf{V}}(\ ) + \Sigma_{\mathbf{N}}(\ ))$ for all $\qquad ^{-}$, it follows that for $\quad = \quad :$

$$jE\ (\Sigma)j = E\ (\Sigma_{\mathbf{V}})(\ _{\mathbf{V}}= \quad) + E\ (\Sigma_{\mathbf{N}})(\ _{\mathbf{N}}= \quad) \quad E\ (\Sigma_{\mathbf{N}}):$$

The inequality holds because both $\Sigma_{\mathbf{V}}(\ )$ and $\Sigma_{\mathbf{N}}(\ )$ are increasing in $\ $, since the thresholds $v_0(\ )$ and $v_1(\ )$ are, respectively, increasing and decreasing in $\ $; also, $\qquad _{\mathbf{N}}$.

Step 3: $E\ (\Sigma_{\mathbf{N}}) \quad 1$. First, notice that $\Sigma_{\mathbf{N}}(\ ) = G(v_0(\ ))$ is convex for $\qquad ^{-}{}_0$, since $g(v)$ is increasing in v, and $v_0(\ )$ is increasing and convex. Also, $\Sigma_{\mathbf{N}}(0) = 0$ because $v_0(0) = 0$, and so $\Sigma_{\mathbf{N}}(\ )$ rises from the origin at increasing rates. Thus, $\Sigma_{\mathbf{N}}$ must have an increasing secant: $(\Sigma_{\mathbf{N}}(\ )=\ )^{0}\quad 0$. But then, $E\ (\Sigma_{\mathbf{N}})\quad 1$. Altogether, $jE\ (\Sigma)j\quad E\ (\Sigma_{\mathbf{N}})\quad 1$.

Following similar logic, we now show that condition (b) also implies the desired result.
Step 1: If $1=\grave{}(v)$ is concave then $v_1(\ )$ is concave. By (6), $v_1(\ )\ 2\ (0;\bar{v})$ is characterized by $\tilde{}(v_1(\ ))\qquad =t$, where $\tilde{}(v)\quad 1=\grave{}(v)$ is monotone decreasing. Let $\_{}_1$ and $\bar{}_1$ solve $\tilde{}(0)=\bar{}_1=t$ and $\tilde{}(\bar{v})\quad \_{}_1=t$, respectively. Then, for $\ \ 2\ [\_{}_1;\bar{}_1]$, we have $v_1(\ )=\ \tilde{}\ ^1(\ \ =t)$. Note that for $\ <\ \_{}_1$, $v_1(\ )=\bar{v}$, given (6). Thus, for $\qquad \bar{}_1$, $v_1(\ )$ is decreasing and concave, since $\tilde{}\ ^1$ is the inverse of a monotone decreasing concave function.
Step 2: If $\qquad \Sigma_{\mathbf{V}}(\ \ )$ then $jE\ (\Sigma)j\quad E\ (\Sigma_{\mathbf{V}})$. Let $\ _{\mathbf{V}}=\Sigma_{\mathbf{V}}(\ \ )$ and $\ _{\mathbf{N}}=\Sigma_{\mathbf{N}}(\ \ )$. Since $\Sigma(\ )=1\quad(\Sigma_{\mathbf{V}}(\ )+\Sigma_{\mathbf{N}}(\ ))$ for all $\qquad \bar{}$, it follows that for $\quad=\qquad 2\ (0;\bar{})$:

$$jE\ (\Sigma)j=E\ (\Sigma_{\mathbf{V}})(\ _{\mathbf{V}}=\quad)+E\ (\Sigma_{\mathbf{N}})(\ _{\mathbf{N}}=\ )\qquad E\ (\Sigma_{\mathbf{V}});$$

where the inequality follows by the same reasons given in Step 2 above, but using $\qquad\qquad _{\mathbf{V}}$.
Step 3: $E\ (\Sigma_{\mathbf{V}})\quad 1$. First, notice that $\Sigma_{\mathbf{V}}$ is convex. Indeed, since $G(v)$ is concave, $1\quad G(v)$ is convex. Thus, $\Sigma_{\mathbf{V}}(\ )=1\quad G(v_1(\ ))$ is increasing and convex, since it is the composition of a decreasing concave function $v_1(\ )$, and a decreasing convex function $1\quad G(v)$. Moreover, $\Sigma_{\mathbf{V}}(0)=0$ for all $\qquad \_{}_1$, since $v_1(\ )=\bar{v}$. Altogether, $\Sigma_{\mathbf{V}}(\ )$ weakly rises from the origin at increasing rates, and thus $\Sigma_{\mathbf{V}}$ is superadditive and its secant must rise: $(\Sigma_{\mathbf{V}}=\ )^{0}\quad 0$. But then, $\Sigma_{\mathbf{V}}$ must be elastic, i.e., $E\ (\Sigma_{\mathbf{V}})\quad 1$. Therefore, $jE\ (\Sigma)j\quad E\ (\Sigma_{\mathbf{V}})\quad 1$.

Proposition A.1. *Suppose that b and $\grave{}$ are increasing and that assumption 1 holds. Consider an equilibrium $(\ \ ;\ \ )$, and let $v_0\quad v_0(\ \ )$ and $v_1\quad v_1(\ \ )$.*

- *(a) If $b(v)=\grave{}(v)$ is concave on $[0;v_0]$ and $g(v)$ non-decreasing on $[0;v_0]$, then $jE\ (\Sigma)j\quad 1$ in equilibrium, provided $\qquad \Sigma_{\mathbf{N}}(\ \ )$.*
- *(b) If $1=\grave{}(v)$ is concave on $[v_1;\bar{v}]$ and $g(v)$ non-increasing on $[v_1;\bar{v}]$, then $jE\ (\Sigma)j\quad 1$ in equilibrium, provided $\qquad \Sigma_{\mathbf{V}}(\ \ )$.*

Proof (a): First, as in proof of Proposition 4-(a), identity (9) allows to deduce that $E\ (\Sigma_{\mathbf{N}})$ is a lower bound for $jE\ (\Sigma)j$, provided $\Sigma\quad \Sigma_{\mathbf{N}}$. Second, the assumptions on $b=\grave{}$ and $g(v)$ imply that $\Sigma_{\mathbf{N}}(\ )=G(v_0(\ ))$ is of class C^1 and convex on $[0;\quad]$. Since, in addition, $\Sigma_{\mathbf{N}}(0)=0$, it follows that $\Sigma_{\mathbf{N}}(\ )$ is superadditive on $[0;\quad]$, and thus $(\Sigma_{\mathbf{N}}(\ )=\ )^{0}\quad 0$ for $\quad 2\ [0;\quad]$. Finally, it is then straightforward to see that: $(\Sigma_{\mathbf{N}}(\ )=\ )^{0}\quad 0\quad (\ )\quad E\ (\Sigma_{\mathbf{N}})\quad 1$.
Proof (b): The proof is analogous: (9) imply that $E\ (\Sigma_{\mathbf{V}})$ is a lower bound for $jE\ (\Sigma)j$, as long as $\Sigma\quad \Sigma_{\mathbf{V}}$. Next, the assumptions on $b;\grave{};g$ imply that $\Sigma_{\mathbf{V}}(\ )=1\quad G(v_1(\ ))$ is of class C^1 and convex on $[0;\quad]$. Moreover, since $v_1(0)=\bar{v}$, we have $\Sigma_{\mathbf{V}}(0)=0$; hence, $\Sigma_{\mathbf{V}}(\ )$ is superadditive on $[0;\quad]$, and so $E\ (\Sigma_{\mathbf{V}})\quad 1$ in equilibrium.

A.3.2 Proof of Proposition 5

We will show that each term in the right side of expression (9) is decreasing in t. First, notice that $\Sigma_{\mathbf{N}}(\ ) = G(v_0(\ ))$ is not affected by t, and so an increase in t reduces ratio $\Sigma_{\mathbf{N}}=\Sigma$, since t raises the amount of unverified sharing. Similarly, the ratio $(\Sigma_{\mathbf{V}}=\Sigma)$ falls as t rises, since t raises Σ but it lowers $\Sigma_{\mathbf{V}}$. Next, we verify that $E\ (\Sigma_{\mathbf{V}})$ in (10) falls in t.

First, we show that $v_1(\ ;t)$ is supermodular in $(\ ;t)$. Indeed, as in the proof of Proposition 4-(b), type $v_1(\ ;t)$ can be written as $v_1(\ ;t) = \tilde{\ }^{\ 1}(\ =t)$ with $\tilde{\ }(v)\quad 1=`(v)$. Moreover, the map $(\ ;t)\ 7!\quad =t$ is clearly submodular in $(\ ;t)$ (negative cross-partial), and increasing in and decreasing in t. Consequently, $v_1(\ ;t)$ is supermodular in $(\ ;t)$ (positive cross-partial) because it is the composition of a decreasing and concave function (i.e., $\tilde{\ }^{\ 1}$) and a submodular function that is increasing in and decreasing in t (i.e., $\ =t$).

Second, since $v_1(\ ;t)$ is increasing in t, it follows that if $g=(1\quad G(v))$ is non-increasing, then the hazard rate $g(v_1(\ ;t))=(1\quad G(v_1(\ ;t)))$ rises as t falls.

Altogether, $E\ (\Sigma_{\mathbf{V}})$ in (10) must rise as t falls, since it is the product of two positive and decreasing functions in t, namely, $g(v_1(\ ;t))=(1\quad G(v_1(\ ;t)))$ and $\quad @v_1(\ ;t)=@\ $.

A.4 Proofs of Section 6

In this section, we will show that condition (a) and (b), respectively, imply that the pass-through curve is sufficiently elastic, i.e., $E\ (\Sigma)\quad 1=(1\quad\ )$. This, in turn, will allow us to prove Proposition 6. The next auxiliary result exploits the properties of $\Sigma_{\mathbf{N}}$ and the conditions in (a) to derive the desired elasticity result.

Proposition A.2. *Suppose that $b(v)=`(v)$ is concave, and the density $g(v)$ is increasing. If, in equilibrium, $\quad\ \Sigma_{\mathbf{N}}(\ )$, then Σ is sufficiently elastic, namely, $E\ (\Sigma)\quad 1=(1\quad\ )$.*

Proof: First, we characterize $v_0(\ )$ in (6). To this end, let $p(v) := b(v)=`(v)$ and $\&(\ ) :=\ =(1\quad\ )$. As in the proof of Proposition 4-(a), $v_0(\ ) = p^{\ 1}(\&(\ ))$ for $\ 2\ [\underline{\ }_0;\bar{\ }_0]$, where $\underline{\ }_0$ and $\bar{\ }_0$ respectively solve $p(0) = \underline{\ }_0=(1\quad \underline{\ }_0)$ and $p(\bar{v}) = \bar{\ }_0=(1\quad \bar{\ }_0)$. (For $\ < \underline{\ }_0$, $v_0(\ ) = 0$.)

Second, we notice that $\ 2\ (\underline{\ }_0;\bar{\ }_0)$. Indeed, since $\ 2\ (0;1)$, we must have $\ < \bar{\ }_0$; also, $\ > \underline{\ }_0$ as $0 <\quad\ \Sigma_{\mathbf{N}}(\ )$. Given this, $\Sigma_{\mathbf{N}}(\ )$ obeys $\Sigma_{\mathbf{N}}(\ ) = G(p^{\ 1}(\&(\ )))$.

Third, we observe that because $p(v)$ is increasing and concave, its inverse $p^{\ 1}(z)$ is increasing and convex and satisfies $p^{\ 1}(z) = 0$ for $z\quad \&(\underline{\ }_0)$. Thus, for $z\ 2\ (\&(\underline{\ }_0);\&(\bar{\ }_0))$, $p^{\ 1}(z)$ must have an increasing secant, i.e., $(p^{\ 1}(z)=z)^0\quad 0$, which implies $E_z(p^{\ 1})\quad 1$. Likewise, $G(v)$ is increasing and convex with $G(0) = 0$; thus, by the same logic, $E_v(G)\quad 1$.

Fourth, we use that the elasticity of the composition of functions equals the product of their elasticities. Equipped with this, we compute the elasticity of $\Sigma_{\mathbf{N}}(\ )$ at $\ =\ $, and then use the bounds previously found to get that $\Sigma_{\mathbf{N}}$ is sufficiently elastic at $\quad = \quad$:

$$E\ (\Sigma_{\mathbf{N}}) = E_v(G) \quad E_z(p\ ^{1}) \quad E\ (\&) \quad E\ (\&) = 1{=}(1 \qquad){:}$$

Finally, since $\qquad \Sigma_{\mathbf{N}}(\quad)$, identity (9) allows us to conclude that $jE\ (\Sigma)j \quad E\ (\Sigma_{\mathbf{N}})$, and so Σ is sufficiently elastic in equilibrium. This concludes the proof.

As in the result above, we now use condition (b) and properties of $\Sigma_{\mathbf{V}}$ to bound $E\ (\Sigma)$.

Proposition A.3. *Suppose that $1{=}`(v)$ is concave and that the hazard rate $g(v){=}(1 \quad G(v))$ is non-increasing. If $\qquad \Sigma_{\mathbf{V}}(\quad)$ when $\quad = 0$, and $g(0) > `^{0}(0){=}`(0)$, then there exists $\hat{t} > 0$ such that for all verification cost t 2 $(0;\hat{t})$, the pass-through curve Σ is sufficiently elastic in equilibrium: $E\ (\Sigma) \quad 1{=}(1 \qquad)$.*

To prove this proposition, we leverage a pair of claims. The first one examines how t impacts the equilibrium prevalence $\quad (t)$, and the mass of users who engage in unverified sharing $\Sigma(\quad(t);t)$ and verified sharing $\Sigma_{\mathbf{V}}(\quad(t);t)$, respectively. Since Proposition A.3 focuses on the case in which $\qquad \Sigma_{\mathbf{V}}(\quad)$, we next restrict attention to verification cost $t > 0$ for which the equilibrium $(\quad;\quad)$ satisfies $\Sigma_{\mathbf{V}}(\quad;t) > 0$, and thus $v_1(\quad;t) < \bar{v}$.

Claim A.1. *As t increases, $\quad(t)$ and $\Sigma(\quad(t);t)$ both increase, while $\Sigma_{\mathbf{V}}(\quad(t);t)$ decreases.*

Proof: Given $t > 0$, we know that $\quad(t)$ must satisfy $F(\Sigma(\quad(t);t)) = \quad(t)$, or:

$$\Sigma(\quad(t);t) := G(v_1(\quad(t);t)) \quad G(v_0(\quad(t))) = F\ ^{1}(\quad(t)){:}$$

By the Implicit Function Theorem, it is immediate to see that:

$$\frac{d}{dt} = \frac{g(v_1)@v_1{=}@t}{(F\ ^{1})^{0} + g(v_0)@v_0{=}@ \qquad g(v_1)@v_1{=}@} > 0{;}$$

since $@v_1{=}@t > 0 > @v_1{=}@$, $@v_0{=}@ \qquad 0$, and $F^{0} > 0$. Consequently, $d\Sigma(\quad(t);t){=}dt = (F\ ^{1})^{0}d\quad{=}dt > 0$, since $F\ ^{1}$ is strictly increasing.

Next, we examine $\Sigma_{\mathbf{V}}(\quad(t);t) = 1 \quad G(v_1(\quad(t);t))$. Differentiating this in t yields:

$$\frac{d\Sigma_{\mathbf{V}}(\quad(t);t)}{dt} = \quad g(v_1(\quad(t);t)) \quad \frac{dv_1(\quad(t);t)}{dt}{:}$$

Therefore, the sign of $d\Sigma_{\mathbf{V}}(\quad(t);t){=}dt$ depends on that of $dv_1(\quad(t);t){=}dt$. We'll show that $dv_1(\quad(t);t){=}dt > 0$. To see this, consider the expression for $d\quad(t){=}dt$ above. Notice that this expression can be bounded from above by:

$$\frac{d\ }{dt} = \frac{g(v_1)\partial v_1/\partial t}{(F^{-1})' + g(v_0)\partial v_0/\partial\ \quad g(v_1)\partial v_1/\partial\ } < \quad \frac{g(v_1)\partial v_1/\partial t}{g(v_1)\partial v_1/\partial\ } = \quad \frac{\partial v_1/\partial t}{\partial v_1/\partial\ }.$$

This last inequality effectively implies that $v_1(\ \ (t); t)$ must increase in t, since:

$$\frac{d\ }{dt} < \quad \frac{\partial v_1/\partial t}{\partial v_1/\partial\ } \quad (\ ) \quad \frac{dv_1(\ \ (t);t)}{dt} = \frac{\partial v_1}{\partial\ } \quad \frac{d\ }{dt} + \frac{\partial v_1}{\partial t} > 0.$$

As a result, $d\Sigma_{\mathbf{V}}(\ \ (t);t)/dt < 0$.

The next claim shows that the equilibrium prevalence vanishes as verification costs vanish.

Claim A.2. $\ (t) \to 0$ *as* $t \to 0$.

Proof: By contradiction, suppose that $\lim_{t\downarrow 0}\ \ (t) = \tilde{\ } > 0$. Since $\ (t)$ is an equilibrium:

$$G(v_1(\ \ (t);t)) \quad G(v_0(\ \ (t))) = F^{-1}(\ \ (t)).$$

Taking the limit as $t \to 0$, we see that

$$G\ \ \lim_{t\downarrow 0} v_1(\ \ (t);t) \quad\quad G\ \ \lim_{t\downarrow 0} v_0(\ \ (t)) \quad = F^{-1}(\tilde{\ }) > 0,$$

where we have used the continuity of G and F, as well as the fact that F is strictly increasing. Furthermore, because G is also strictly increasing, we must have:

$$\lim_{t\downarrow 0} v_1(\ \ (t);t) > \lim_{t\downarrow 0} v_0(\ \ (t)).$$

But $v_0(\ )$ in (6) is continuous, since it is the generalized inverse of a continuous function. Thus, $\lim_{t\downarrow 0} v_0(\ \ (t)) = v_0(\tilde{\ })$. Next, we notice that $v_1(\ \ (t);t) \in (0,\bar{v})$, since $\Sigma, \Sigma_{\mathbf{V}} > 0$. Thus, $v_1(\ \ (t);t)$ in (6) satisfies:

$$\ (t)\ell(v_1(\ \ (t);t)) = t.$$

Taking the limit as $t \to 0$, and using that ℓ is continuous, we get

$$\ell\ \ \lim_{t\downarrow 0} v_1(\ \ (t);t)\ \ = 0.$$

Since $\ell(0)\ \ 0$, it follows that $\lim_{t\downarrow 0} v_1(\ \ (t);t)\ \ 0$. But this is a contradiction because it implies that $v_0(\tilde{\ }) < 0$. We conclude that $\lim_{t\downarrow 0}\ \ (t) = 0$.

Proof of Proposition A.3: First, by Claim A.2, $\lim_{t\#0} v_1(\ \ (t);t) = \lim_{t\#0} v_0(\ \ (t)) = 0$; thus, $\lim_{t\#0} \Sigma(\ \ (t);t) = 0$ and $\lim_{t\#0} \Sigma_{\mathbf{V}}(\ \ (t);t) = 1$. Hence, by continuity, $\Sigma(\ \ (t);t) < \Sigma_{\mathbf{V}}(\ \ (t);t)$ for small enough $t > 0$. On the other hand, no one verifies when t is too high, i.e.: $\Sigma_{\mathbf{V}}(\ \ (t);t) = 0 < \Sigma(\ \ (t);t))$.[26] Moreover, by Claim A.1, $\Sigma_{\mathbf{V}}(\ \ (t);t)$ falls in t while $\Sigma(\ \ (t);t)$ rises in t. Since these functions are continuous and monotone, the Intermediate Value Theorem ensures the existence and uniqueness of $\hat{t}_0 > 0$ such that for all $0 < t \quad \hat{t}_0$,

$$\Sigma(\ \ (t);t) \quad \Sigma_{\mathbf{V}}(\ \ (t);t):$$

Second, we prove that the -elasticity of $\Sigma_{\mathbf{V}}(\ ;t)$ evaluated at $\ = \ (t)$, i.e., the mapping

$$t\ 7\!\ \ \Phi(t) := \frac{@\Sigma_{\mathbf{V}}(\ ;t)=@}{\Sigma_{\mathbf{V}}(\ ;t)}_{\ =\ \ (t)}$$

is decreasing in t. Since $\Sigma_{\mathbf{V}}(\ ;t) = 1 \quad G(v_1(\ ;t))$, with $v_1(\ ;t)$ in (6), it follows that $\Phi(t)$ can be written as:

$$\Phi(t) = \frac{g(v_1(\ \ (t);t))}{1 \quad G(v_1(\ \ (t);t))} \quad \frac{`(v_1(\ \ (t);t))}{`^0(v_1(\ \ (t);t))}:$$

Thus,

$$\Phi^0(t) = \quad \frac{g(v_1)}{1 \quad G(v_1)}{}^0 \frac{dv_1}{dt} \quad \frac{`(v_1)}{`^0(v_1)} + \frac{g(v_1)}{1 \quad G(v_1)} \quad \frac{`(v_1)}{`^0(v_1)}{}^0 \quad \frac{dv_1}{dt}:$$

Since $g=(1 \quad G)$ is decreasing, $`^0=`$ is increasing (as $1=`$ is concave[27]), and $dv_1=dt > 0$, we have that $\Phi^0(t) < 0$. Moreover, because $\lim_{t\#0} v_1(\ \ (t);t) = 0$, it follows that

$$\lim_{t\#0} \Phi^0(t) = g(0) \quad \frac{`(0)}{`^0(0)}:$$

But this latter expression is strictly greater than one, provided $g(0) > `^0(0)=`(0)$. On the other hand, the mapping $t\ 7\!\ \ 1=(1 \qquad (t))$ is increasing in t, as the equilibrium prevalence rises as t rises, and obeys $\lim_{t\#0} 1=(1 \qquad (t)) = 1$, since $\lim_{t\#0} \ \ (t) = 0$ by Claim A.1–A.2.

Altogether, by continuity, $\Phi(t) > 1=(1 \qquad (t))$ for small $t > 0$. We can then define

$$\hat{t}_1 := \inf f t \quad 0 : \Phi(t) \quad 1=(1 \qquad (t))g;$$

where, without loss, $\hat{t}_1 = \hat{t}_0$ if the above set is empty.

[26] Easily, if t is such that $t=`(\bar{v}) = b(\bar{v})=`(\bar{v})=(1 + b(\bar{v})=`(\bar{v}))$ then no type finds it optimal to verify news.

[27] Indeed, if $1=`$ is concave, then $\ `^0=`^2$ is decreasing in $v$, i.e., $(`^0=`) \ (1=`)$ must be increasing. But since $1=`$ is decreasing, $`^0=`$ must be increasing.

All in all, we have shown that for all $t\ 2\ (0;\hat{t})$ with $\hat{t} := \min f\hat{t}_0;\hat{t}_1 g$,

$$E\ (\Sigma) > E\ (\Sigma_{\mathbf{V}})\qquad\frac{1}{1\qquad};$$

where we have used (9) to bound $E\ (\Sigma)$. Thus, Σ is sufficiently elastic in equilibrium.

Equipped with Proposition A.2 and A.3, we are ready to prove Proposition 6.

Proof of Proposition 6: Let $(\ )$ denote the equilibrium prevalence given a filter of quality $.$ As argued in the main text of Section 6, this value is the unique solution to the equation

$$\underbrace{(1\qquad)\Sigma(\ \ (\quad(\ );\ \ ))}_{e(\ (\ \ (\ );\ ))} = \Pi\ ^{1}(\quad(\ ));$$

where $\Pi\ ^{1}$ is the the upward sloping inverse supply function. Totally differentiating the above equality with respect to $\ $, we get:

$$\Sigma + (1\qquad)\Sigma^0\ \ \frac{@}{@}[\quad]^0(\ ) + \frac{@}{@}\quad = [\Pi\ ^1]^0[\quad]^0(\ ):$$

Solving for $[\quad]^0(\ )$ yields:

$$[\quad]^0(\ ) = \frac{\Sigma + (1\qquad)\Sigma^0\frac{@}{@}}{[\Pi\ ^1]^0\quad(1\qquad)\Sigma^0\frac{@}{@}}$$

Since $[\Pi\ ^1]^0 > 0 > \Sigma^0$ and $@\ \ =@\ \ > 0$, it follows that the sign of $[\quad]^0$ is fully determined by the sign of the numerator of the above expression:

$$\Upsilon(\ )\qquad\Sigma(\ (\quad(\ );\ \ )) + (1\qquad)\Sigma^0(\ (\quad(\ );\ \ ))\frac{@}{@}(\quad(\ );\ \ ):$$

Taking the limit of the above expression as $\ !\ 0$,

$$\Upsilon(0) = \quad\Sigma(\quad(0))\quad\Sigma^0(\quad(0))\quad(0)(1\qquad(0));$$

where we have used that $(\ ;0) =\quad$ and $\frac{@}{@}(\ ;0) = \quad(1\qquad)$. If the conditions given in (a) hold, then Proposition A.2 ensures that $jE\ (\Sigma)j\quad 1=(1\qquad(0))$, and so $\Upsilon(0) > 0$. Likewise, Proposition A.3 would ensure the same result, provided the conditions in part (b) are satisfied. So if either (a) or (b) holds, then a small increase in the filter's quality leads to an increase in the equilibrium prevalence $(\ )$. Because a change in the filter has no *direct* effect on the supply function $\Pi(\ ^e)$, the new equilibrium is the result of an *upward* movement along the supply curve, and hence the new equilibrium displays a higher effective sharing ^{e} and, thus, a higher diffusion of fake news, $\Delta(\ )\quad(\ )\ ^e(\ )$. Intuitively, the small increase in shifts the effective sharing $(1\quad)\Sigma(\ )$ up in the $(\ ;\ ^e)$-space around the point studied. Conversely, as the filter becomes perfect, i.e., $!\ 1$, Σ^e shifts left towards the origin, with $(\ (\ );\ ^{e;}(\ ))\ !\ (0;0)$ and $\Delta(\ ) = \ ^e(\ )\quad(\ )\ !\ 0$.

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

A.5 Proofs of Section 7

A.5.1 Proof of Proposition 7

Take $0 < \ _{\tilde{}} < \ \tilde{} < 1;\ \ 2\ (_{\tilde{}};\tilde{})$ and suppose that $\Sigma(\ ) < \Sigma^{co}_{\tilde{};\tilde{}}(\ )$. By Carathéodory's Theorem (e.g., Theorem 17.1 in Rockafellar, 1970), there exists $\tilde{}\ \ 2\ (0;1)$ and prevalence $_A;\ _B\ 2\ [0;1]$ such that $_A + (1\quad)\ _B = $ and $\Sigma^{co}_{\tilde{};\tilde{}}(\ ) = \ \Sigma(\ _A) + (1\quad)\Sigma(\ _B)$ with $_A;\ _B\ 2\ [_{\tilde{}};\tilde{}]$. Without loss of generality, assume $_A < \ < \ _B$. As explained in the main text, the monopolist can induce $_A$ and $_B$ by choosing segment sizes $!$ and $1\quad!$ for A and B, respectively; and also, by sending an extra amount of truthful news to segment A, with $\quad 0$. Following the steps outlined in the main text, since the variables $(\ ;\ _A;\ _B)$ are already fixed by the concavification (given), it follows that the pair $(!;\ )$ adjusts to induce the desired prevalence $(\ _A;\ _B)$:

$$_A = \frac{}{\ +(1\quad)!+\ } \qquad \text{and} \qquad _B = \frac{(1\quad)}{(1\quad)\ +(1\quad)(1\quad!)}:$$

Using that $= (\ _B\quad)=(\ _B\quad _A)$, it is simple to see that the above system of equations has a unique solution given by:

$$! \ = \ \frac{(\ _B\quad)(\ _B\quad _A +\ (1\quad _B))}{_B(1\quad)(\ _B\quad _A)}\ 2\ (0;1) \tag{A.2}$$

$$= \ \frac{(\quad _A)(\ _B\quad)}{_A\ _B}\ 2\ (0;\ 1\ ): \tag{A.3}$$

Moreover, any prevalence for which $\Sigma(\ ) = \Sigma^{co}_{\tilde{};\tilde{}}(\ )$ can be trivially implemented by setting segment sizes $! = $ and $= 0$ so that $_A = \ _B = $.

Finally, if the equilibrium under uniform policies $\ 2\ (_{\tilde{}};\tilde{})$ and satisfies $\Sigma(\ ) <$

$\Sigma^{co}_{\tilde{};\sim}(\quad)$, then the above segmentation strategy yields strictly higher revenues, since

$$\Sigma(\ _{A}) + (1 \qquad)\quad \Sigma(\ _{B}) = \quad \Sigma^{co}_{\tilde{};\sim}(\quad) > \quad \Sigma(\quad):$$

This can make the segmentation strategy more profitable than a uniform policy.

A.6 Additional Figures

Increasing benefits and decreasing losses (Section 4.1). The figure below depicts the case in which both sharing benefits and losses are decreasing in v. Observe that rigidities in the pass-through curve to decreases in verification costs emerge not only at low prevalence levels, but also at high ones.

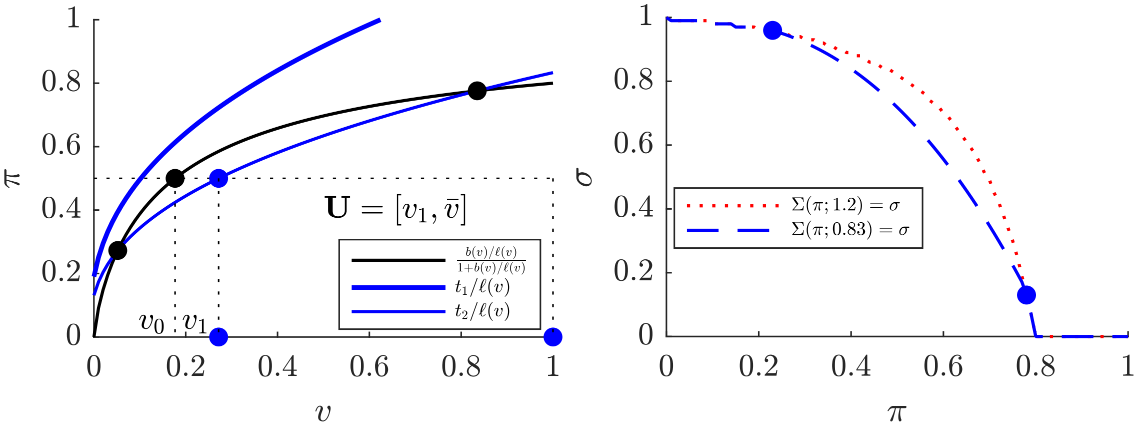

Figure 9: In both panels, we use the same parameters as in Figure 3 (right panel). With decreasing losses, the propensity to skip verification *t*=` is an increasing function. Thus, high types are more prone to prefer unverified to verified sharing. As verification cost t falls (from $t = 1{:}2$ to $t = 0{:}83$), the pass-through curve Σ contracts for mid prevalence levels.

Non-monotone losses (Section 4.2). In this 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.

Specifically, consider Figure 10 below: Without loss of generality, we preserves the monotonicity of the propensity-to-share function but introduce a non-monotonic *t*=`. In the top row, two crossing points for low prevalence levels (left panel) generate a single region of contraction (right panel); in the middle row, with a second crossing point, a second disconnected region of sensitivity appears for high ; and in the bottom panel, as high types exhibit larger losses and *t*=` decreases more, both regions connect and a subset of the original one exhibits more verification.

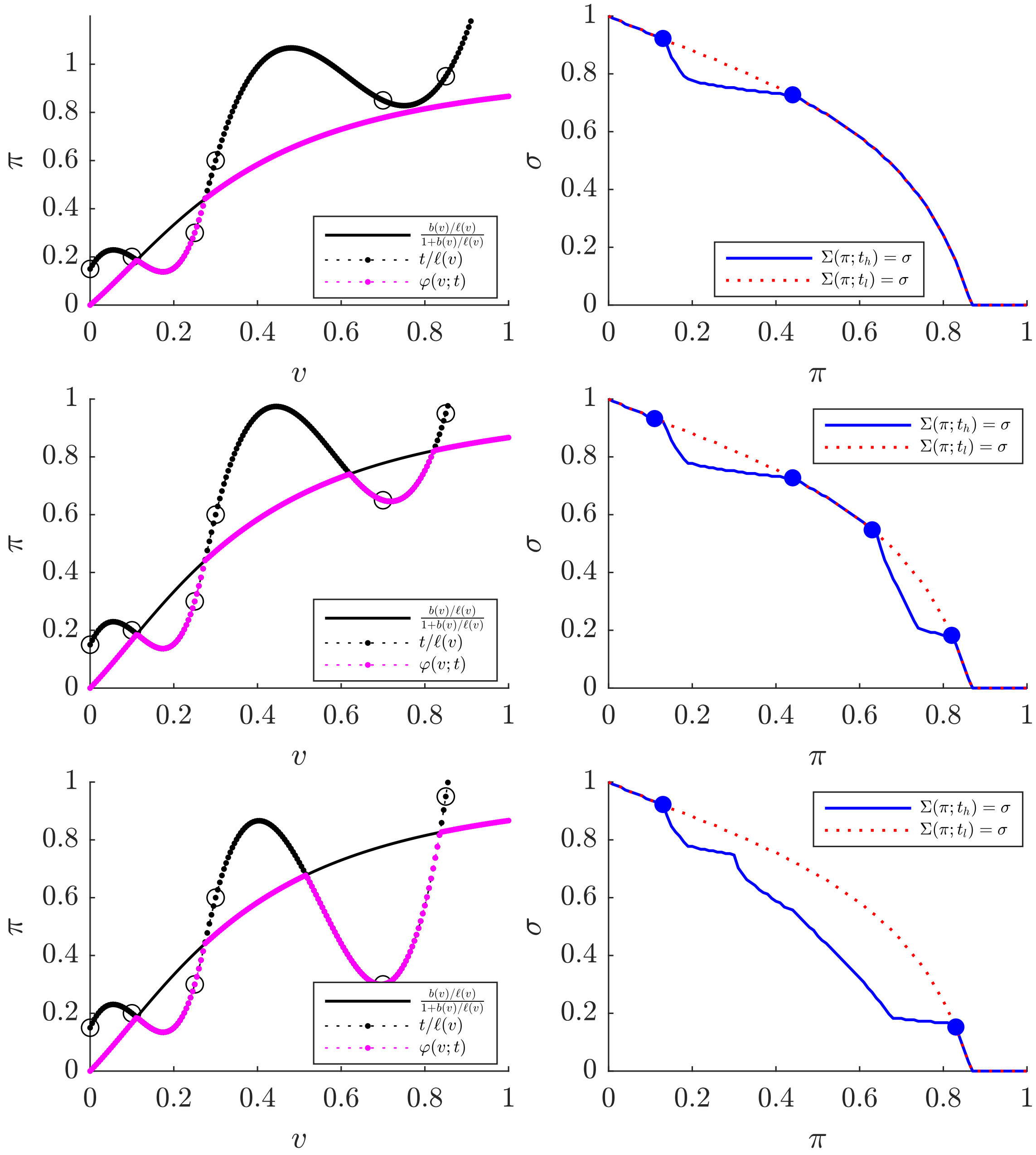

Figure 10: Non-monotone losses. Multiple intersections could lead to multiple disconnected verification regions (middle panel), and existing ones could exhibit more verification and become connected (bottom panel).

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*.

Chade, H., J. Eeckhout, and L. Smith (2017): "Sorting through search and matching models in economics," *Journal of Economic Literature*, 55, 493–544.

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.

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer (2019): "Fake news on Twitter during the 2016 US presidential election," *Science*, 363, 374–378.

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

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.

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.

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.

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*, 14, 255–303.

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