STAFF
REPORTS

NO. 1028
AUGUST 2022

REVISED
NOVEMBER 2024

Misinformation in Social Media: The Role of Verification Incentives

Gonzalo Cisternas | Jorge Vásquez

FEDERAL RESERVE BANK *of* NEW YORK

Misinformation in Social Media: The Role of Verification Incentives

Gonzalo Cisternas and Jorge Vásquez

Federal Reserve Bank of New York Staff Reports, no. 1028
August 2022; revised November 2024
JEL classification: D40, L10, L50

Abstract

We develop a model of a platform featuring producers of fake news as well as users who can share content and verify it at a cost. Since users supply news to other users, their actions affect fake news prevalence and strategic complementarities can arise: high levels of verification can lead to low prevalence of fake content, in turn inducing more unverified sharing that sustains high levels of verification. Equilibria in this market then arise as intersection points between a standard supply curve and a novel correspondence that generalizes a demand function to account for the users' strategic environment. Equilibria exhibiting more fake news production and diffusion can be consistent with higher user welfare due to the strong verification complementarities at play. We also quantify externalities associated with users affecting the average quality of news items in the platform and examine the effects on outcomes of (i) lowering verification costs, (ii) certifying verified content, and (iii) using algorithmic filters.

Key words: misinformation, news verification, social media

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

1 Introduction

The phenomenon of misinformation online has attracted substantial attention, increasingly threatening societies in areas as diverse as elections, markets, and disease spread.[1] Furthermore, it is argued that as artificial intelligence advances, the problem may become even worse.[2] Social media platforms have therefore responded by taking important steps in the area of *fact-checking*: they have partnered with independent professional entities specialized in the verification of content;[3] deployed algorithms designed to detect misinformation;[4] and begun labeling content as true or false.[5]

Nevertheless, underlying these responses is the key principle that users themselves must ultimately assess the veracity of news and decide how to act upon it. To empower users, therefore, suspicious news items are now accompanied either by reports that assess the content's trustworthiness, or by related material that provides context. What this means is that the success of fact-checking initiatives is inevitably linked to users' willingness to verify the truthfulness of the news items encountered. However, such a verification process is naturally costly, even if the evidence is readily available.

In this paper, we develop a model of fake news production, verification and sharing to understand how verification incentives—an understudied yet critical aspect of the fake news problem—determine the extent of misinformation diffusion, which is a natural first step towards evaluating the magnitude of this threat. We show how this topic can be analyzed with competitive analysis tools featuring a "strategic spin" linked to natural *social influence effects* at play, leading to richer predictions than in those traditional analyses.

Model and equilibrium We develop a stationary matching model in which a large number of small users encounter news in any period. Such items originate from a large set of small producers, and a fraction of them can be false. Upon encountering a news item, a user can uncover its veracity only after paying a cost; and after this decision is made, the user can decide whether to share the news. Thus, the pool of news in any period consists of fresh items recently introduced and those produced in the past that were shared by users.

[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. (2020) 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 termed fake news as a major global risk (Howell, 2013), with artificial intelligence deployed to "deepfake" videos 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] https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

[5] https://transparency.meta.com/en-gb/features/how-fact-checking-works/.

We assume that all news items live for two periods—e.g., the platform's algorithm deletes old content—so sharing decisions from only a period earlier matter. Further, all users enjoy sharing true content but dislike sharing fake news; thus, misinformation spreads only when it is *not verified*. Matches between users and news are random, a proxy for residual uncertainty in how the platform's algorithm decides to allocate news across users.

In practice, an inference problem familiar to all of us is at play when encountering news online: to which extent has the content been vetted by others in the past? This uncertainty simply is a reflection of all past choices along *sharing chains* not being readily observable to any individual user. To capture this situation, we assume that individual sharing and verification decisions by other users are not observable: say, our platform displays content via "news feeds" that collect news based on novelty (a fresh, "vintage one," item) and popularity (a "vintage two" item that was shared). The *prevalence* of fake news—the proportion of false items among those circulating in any period—is then not only determined by producers, but also directly affected by users' choices. Importantly, this happens in a way that users cannot discern between items that have definitely not, or may have been, verified in the past.

As users' verification and sharing choices affect the likelihood of encountering misinformation, they ultimately influence other users' same choices. A key finding is that this feedback loop—a "prevalence-driven" social influence effect—can manifest in *strategic complementarities*: high, intermediate and trivial levels of verification can arise at fixed levels of fake news production. In parallel, different degrees of unverified sharing—the mass of users who skip verification and always share, which determines the rate of diffusion of fake content—emerge, also ranked in the same order: for example, high diffusion rates of misinformation are supported by high levels of verification, because the latter induce low levels of prevalence that sustain high degrees of unverified sharing at the same time. This finding is non-trivial because it arises for intermediate—and endogenous—levels of fake news production. Otherwise, verification is uniquely pinned down either because not verifying news is a dominant strategy, or because verification choices are strategic substitutes: as more users verify and prevalence falls, more users enjoy sharing content without paying verification costs.

We make the natural assumption that fake news producers positively respond to misinformation diffusion rates, so an increasing supply curve emerges on the production side. Stationary equilibria then arise as intersection points between a standard supply curve and a *correspondence* capturing the possibility of multiple outcomes on the users' side. The novelty of this correspondence lies on its resemblance with traditional demand functions: as production grows and prevalence increases, unverified sharing is less attractive at all three possible levels of verification, which means that the "branches" of this correspondence are all weakly decreasing. The competitive-strategic duality of our setting is clear here. On the one

hand, any equilibrium is the outcome of forces akin to supply and demand locally balancing each other. On the other hand, the presence of multiple possibilities makes salient that an actual game among users is at play. Crucially, this dual aspect is because our demand side is not a traditional one: users act both as consumers and suppliers of news.

Welfare and externalities The complementarities uncovered have important implications. Concretely, when multiple stationary equilibria exist, the one displaying the largest fake news production maximizes user welfare. Indeed, this equilibrium exhibits such a relatively high level of verification that two phenomena occur simultaneously: (i) welfare per news item grows due to the prevalence of fake content falling sharply (quality effect) and (ii) total circulating news grows due to high induced rates of unverified sharing (volume effect). Outside this multiplicity region, or along each decreasing branch of the correspondence, the prevailing substitution effects lead to the opposite prediction: welfare falls with production (say, as the supply curve shift outwards) due to news falling in terms of quality and volume.

Altogether, users exert quality and volume externalities on others. We then turn to better understand the inefficiencies that arise through the more interesting quality channel, as volume can always be maximized by mandating unverified sharing for everyone (at the expense of quality, of course). Concretely, we show that for a planner interested in maximizing welfare per news, equilibrium verification is always too low when it takes place, and the diffusion rate of fake content too high (unless production is excessively high, in which verification becomes too costly). A revealed preference argument helps illustrate the benefits of more verification along this prevalence channel proxy for quality. First, as the likelihood of finding truthful news grows, the payoff for those originally sharing news necessarily increases—any switching between options reflects substitution effects from improved opportunities. Second, some users originally not sharing news at all will do it now—an extensive margin effect.

Policy We study three types of policies: lowering verification costs, say by facilitating access to fact-checking reports; using algorithmic filters, or technology that can assess news veracity without direct human aid; and news certification, labeling news as true or false depending on the users' verification outcomes. We focus on user welfare per news item to isolate how these policies affect users through the implied change in prevalence.

Verification costs affect outcomes directly through impacting users' payoffs and indirectly via behavior impacting prevalence. In equilibria with high verification, lowering such costs leads to more verification (at the expense of fake news diffusion) which reduces prevalence and raises user welfare. For equilibria displaying intermediate verification levels, the opposite occurs. Indeed, in close analogy with mixed strategy equilibria, only a fraction of individuals verify news in this case so that there is indifference between verifying news and not sharing

at all. As verification costs fall, equilibrium prevalence has to increase to maintain such indifference; thus, welfare falls, and in fact there is less verification in equilibrium.

Filters do not directly affect users' payoffs; instead, their effect on users is through the likelihood of encountering misinformation. With filters that (i) assess incoming fresh news items before these reach users and (ii) make errors only with fake content, prevalence falls ceteris paribus. As the chance of sharing misinformation falls, the pass-through correspondence weakly expands; meanwhile, the supply of fake content contracts, because fake content has to pass an extra layer of vetting. With a "demand expansion and a supply contraction" the "price" must necessarily increase: fake news diffuse more among users. Despite an ambiguous effect on production, prevalence must fall in the equilibrium with high verification to sustain higher diffusion rates—thus, welfare increases. After the imposition of a filter then, content will diffuse at higher rates conditional on reaching users; but this is just the reflection of a welfare-enhancing extra layer of protection.

Finally, the practical benefits of labeling fake content as such seem intuitive—but what are the effects of certifying truthful content? Our third policy exercise speaks to this issue. Concretely, as more content that has been verified to be truthful gets labeled as such, users can "remove" even more content from the pool of news with dubious quality. Thus, the composition of this residual pool worsens. In an equilibrium with high verification, fake news prevalence grows among such items, and hence there is less unverified sharing and lower user welfare among such items. In other words, introducing this policy lowers the rate at which uncertified false items diffuse, but this is a reflection of their relative abundance.

We also show that these policies have refining effects if sufficiently precise: they shrink the region where multiple equilibria can arise. When this occurs, the first two policies select the equilibrium with high verification, complementing each other towards this end.

Robustness We conclude the paper with a number of variations of our baseline model. First, we examine alternative options for our supply side. Second, we study network externalities, understood as *ex post payoffs*—i.e., benefits and losses—explicitly depending on aggregate behavior. Third, we perform an exhaustive list of variations for how (exogenous and independent) benefits and losses can vary across the population of users. Through this last exercise we uncover that changes in prevalence should have sufficiently strong "extensive verification effects" for strategic complementarities to arise: as prevalence falls, more users must enter the verified sharing world than those switching to unverified sharing because now misinformation is less prevalent. We discuss this topic in more detail at the end of the model section: irrespective of whether the complements or substitutes channel dominates, prevalence-driven social influence effects will be at play if verification takes place.

Roadmap We review the related literature next. Section 2 presents the model and its motivation, while the equilibrium analysis is in Section 3. Section 4 examines user welfare and Section 5 studies three policy exercises. Finally, Section 6 performs extensions and variations of the main model, and Section 7 concludes. All the proofs are in the Appendix.

Related literature There is a growing theoretical literature analyzing various aspects of the fake news problem. In Papanastasiou (2020) and Cheng and Hsiaw (2022), verification costs are present and the main focus is on long-run learning: the first focuses on informational cascades with an exogenous supply side, while the second features a sender-receiver setup where talk is cheap. In turn, Kranton and McAdams (2024) also follow a "supply and share" approach with two main differences: first, they feature an explicit network structure (number of neighbors), while in ours exposure to shared news is mediated by the platform; second, their verification technology is costless and imperfect. A common element, however, is that exposure to news is endogenous in both worlds: in their setup, users decide how many news to seek out from producers; instead, in our paper, users affect the user-to-user exposure margin through the implied value of prevalence—combining this latter feature with costly verification is at the heart of the complementarities that we uncover. Finally, other papers focus on the dispersion of beliefs when users share news: see Bowen et al. (2023) on belief polarization, or Acemoglu et al. (2023) on virality and eco chambers; in our setting, prior beliefs are homogeneous across users, but endogenously determined.

On empirics and experiments, our assumption that passing on fake content yields losses is consistent with Pennycook et al. (2021) where users find it important to share only accurate news, and with Altay et al. (2022) where users worry about their reputations when fake news is shared. In turn, Pennycook et al. (2020) show that labeling only a subset of false news articles leads users to believe that untagged articles are more accurate, which increases their sharing: this is consistent with the effect of algorithmic filters in our model, and is the mirror image of our false-news certification. On the supply side, Allcott and Gentzkow (2017) and Tucker et al. (2018) document that clicks are the main source of profitability for untrustworthy websites—but sharing rates are a key catalyst for clicks to happen.

The quality externality uncovered operates through the likelihood of encountering fake news, which is an endogenous belief in our model; thus, our mechanism resembles those in models with information externalities. In this line, Board and Meyer-ter Vehn (2021) examine how inspecting and adopting a product of unknown quality is affected by the network's structure when neighbors' adoption choices are observable, with ensuing implications about learning dynamics. Instead, our model is stationary, focusing on how unobserved sharing-verification choices affect users' perception of misinformation, with ensuing impli-

cations about the type of strategic interaction among users (verification choices acting as complements or substitutes). In turn, Che and Hörner (2018) examine how platforms can influence product adoption through recommendation systems that can depend on other users' past adoptions. Instead, we examine market outcomes absent this form of manipulation when users themselves have the ability to expose others to products of unknown quality.

Finally, our model contributes to the matching literature in settings in which individuals choose to protect themselves at a cost; see Quercioli and Smith (2015) and Vásquez (2022). However, we do this in the context of a novel externality stemming from protection choices by market participants non-trivially affecting the quality of matches that can take place in future rounds—see Chade et al. (2017) for a general survey of matching models, with and without transferable utility, and where other types of externalities are discussed.

2 Model

We develop a model of a platform over an infinite horizon in which a large number of infinitesimal users encounter fake content that originates from a large number of infinitesimal fake news producers. With agents who cannot affect aggregate variables, all players maximize flow payoffs at all times; and due to stationarity, these flows are identical across periods. We introduce the main elements of our model next, and subsequently justify our assumptions.

News viewers A unit mass of infinitesimal risk neutral *users* have access to an online platform where they encounter news of unknown veracity. Upon encountering a news item, each user can first decide to determine its truthfulness by paying a *verification cost* $t > 0$: the search costs incurred when consulting specialized websites for fact checks, or 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. While not sharing the item yields a payoff of zero, the payoff of sharing depends on the item's veracity: sharing truthful news yields a *benefit* $b > 0$ while sharing fake content entails a *loss* $\ell > 0$. Altogether, since users dislike passing on fake articles and the verification technology is perfect, misinformation is shared only when it is not verified.

Our baseline model features b and t as constant across users, with $b > t$ so verification can arise in equilibrium. In turn, the losses ℓ vary according to an atomless cumulative distribution function (CDF) $G(\cdot)$, with support $[0, +\infty)$ and differentiable density $g(\cdot)$. Alternative specifications are discussed at the end of this section.

Fake news producers In every period, a unit mass of news items, all different from one another, enters the platform, a fraction $\pi \in [0, 1]$ of them being false. Fake news production π

originates from a set of producers each facing the choice of producing a fake news article upon paying a *cost* $c \in [0,1]$: in this sense, our producers can be interpreted as "malicious" in that they specialize in fake content production and the costs that they bear reflect their forgone opportunities unrelated to news production. These costs vary across producers according to an atomless CDF $F(\cdot)$ with support $[0,1]$ and continuous density $f(\cdot)$. We assume that a producer receives a payoff of 1 every time that the item is viewed.

Random matching and prevalence The platform allocates all the news available randomly across users in every period. From a user's perspective then, the likelihood of an encountered item being false is determined by the proportion of fake news currently circulating—or *fake news prevalence*, which we will denote $\psi \in [0,1]$. This proportion not only depends on the current volume of fake news produced π, but also on users' past sharing and verification choices.[6] Letting $\sigma_U \in [0,1]$ denote the mass of users who share without verifying news, and $\sigma_V \in [0,1]$ that of those who verify (and hence who share only if truthful), we will have $\psi = \Psi(\pi, \sigma_U, \sigma_V)$ for some function Ψ.[7]

We assume that each news lives for two periods. In this case, $\Psi(\cdot)$ takes a simple form (see (1) in Section 3) and a producer's per unit expected payoff is $1 + \sigma_U$: fake content will reach a first user with certainty as it enters the platform, but subsequently this item will get a second view only if the first encounter was with a user doing unverified sharing, which happens with probability σ_U. Because the intercept is common across producers and the cost distribution F general, we normalize revenue to σ_U and have producers' costs distributed over $[0,1]$ as stated earlier. We refer to σ_U as the *misinformation pass-through rate*, as this is the rate at which fake content diffuses among users within the platform.[8]

Information and equilibrium The timing of moves within a period is as follows. First, producers simultaneously decide whether to produce or not. Second, the resulting cohort of fresh—vintage 1—news items is collected by the platform's algorithm and so are the (now) vintage 2 items that were "fresh" in the previous period and shared by users then. Third, given this pool of news, the algorithm allocates all the news to users randomly. Fourth, all users make their verification and sharing choices simultaneously.

We assume that the history of individual choices of all players are unobserved to their counterparties. In particular, users do not observe other users' individual sharing and verification choices from the previous period. This helps us focus on the following inference

[6] Thus, a user may receive more than one piece of news in any given period. We assume no attrition, and so this consideration is irrelevant because all benefits and costs are per news item.

[7] Since some users may never share news, $\sigma_U + \sigma_V < 1$, and carrying σ_U and σ_V separately is needed.

[8] We also use misinformation for users because they do not know an item's veracity when sharing it.

problem: when seeing a news item, to which extent has it been verified by others in the past? We elaborate on this modeling choice at the end of the section; the bottom line is, users can learn the quality of news only from their own verification choices.

Definition 1 (Equilibrium concept). *In a stationary equilibrium: (i) users' verification and sharing decisions, as well as producers' choices, are constant across time; and (ii) all players simultaneously best respond to one another.*

A stationary equilibrium will give rise to a triplet $(\pi, \sigma_U, \sigma_V)$ that is constant across time.[9] Moreover, as we will show, this triplet will end up encoding all the payoff-relevant information about the behavior of others: from any player's perspective, only $(\pi, \sigma_U, \sigma_V)$ is needed to find a best response—in the case of users, this occurs through the implied fake news prevalence value $\psi = \Psi(\pi, \sigma_U, \sigma_V)$. This is the topic of equilibrium analysis, which we examine in the next section. Before then, let us first justify some of our modeling choices.

Interpreting the model Our model is an approximation of a large network of individuals participating in an online platform that actively displays news to its users via so-called "news feeds:" a mix of recent news (our fresh items) and popular ones (i.e., items that have been shared by other users). In such a world, users' verification decisions, via their sharing choices, affect the average quality of the news observed by others, in turn influencing those same verification and sharing choices: such feedback loops then encode social influence effects. From this perspective, it is useful to break down a discussion of our assumptions as follows.

1. Information and news. While in practice users do get to see the sharing decisions of individuals in their own network, they need not see the sharing decisions of individuals they are not connected with. Importantly, platforms have the power to diffuse these decisions to other parts of the network through such news feeds, especially when articles are profusely shared—in this case, inferences must be made regarding the extent to which news items have been shared and verified in the past, as in our model.

   From this standpoint, allowing for news of only two vintages—proxy for a news feed algorithm that gives less relevance to older news—simplifies the inferences encoded in the prevalence function Ψ. But note that even if users were able to see when a news item is shared with them via another user—thus permitting discriminating between old and new content—letting news now live for three periods would make the inference problem reemerge: someone who receives a news item from another user has to evaluate

[9] With infinitesimal agents, the model's solution is independent of the observability assumptions on this triplet within and across periods (of course, producers do not observe $(\sigma_U, \sigma_V, \psi)$ in the same period, etc.)

whether the news is of vintage 2 (i.e., it is shared by a user for the first time) or of vintage 3. In other words, our model is the simplest laboratory for examining how people's perception of the pervasiveness of fake content is affected by others' past behavior when the actions along sharing chains are not perfectly observed.

A similar justification applies to our assumption that all news are different. If instead news cohorts carried non-trivial masses of identical news, users could differentiate news based on previous encounters, and so some uncertainty regarding vintages would be needed. By having different items, however, we eliminate mechanical "public goods" effects from the same item being consumed by many people (albeit at different points in time), which is important for differentiating the externalities that we study.

2. Preferences. Our assumptions on users' payoffs are natural for understanding how misinformation can diffuse despite everyone disliking to share fake content. The starting point is that verification is costly and people act in their own interest; having verification costs that are common to everyone simplifies our policy exercises. Equipped with this, our main point is that expectations regarding others' behavior can have non-trivial effects on outcomes, and the fact that benefits are concentrated while losses vary across users matters in this respect: strategic complementarities can emerge, manifested in large swings in the extent of verification that are supported by self-fulfilling expectations of what other users will do.

   While the mathematical details are in the next section, it is useful to anticipate the logic behind this finding. In order to verify news, users must consider two margins. First, is verified sharing—i.e., sharing only when content it truthful—profitable in *absolute* terms? Second, is it *relatively* more attractive than doing unverified sharing? On the first margin, note that verification requires an upfront payment (i.e., a cost with certainty) in exchange for a benefit that may not materialize (e.g., the verified item was false). When more users verify news and prevalence falls, the benefit accrues more often, and having these benefits concentrated can lead to an activation of verification incentives for many users, opening the way for complementarities to arise. The problem is that there are substitution effects too, which brings us to the second margin: if prevalence falls, more people will want to avoid paying the (certain) verification cost and simply take their chances with fake content. When losses vary smoothly, the strength of this substitution effect can be dominated for intermediate production levels.

   We chose our baseline model because it offers the richest set of predictions: it displays the dominance of the complements channel or the substitutes counterpart within the same framework. Away from this case, Section 6 presents an exhaustive list of varia-

tions regarding the distribution of losses (i.e., bounded support or even concentrated losses, in Section 6.3), and benefits (varying across users, with and without atoms, in Section 6.4). A general message is that even when the complementarity does not arise, social influence effects will remain at play—it is just that they have a unique resolution. In that section, we also explore more traditional network externalities: benefits and losses explicitly depending on the aggregate behavior of others (Section 6.2).

Regarding producers, we have assumed that their revenue is given by views; but it could also be driven by explicit "sharing clicks," in which case expected revenue (per item) would be non-linear in σ_U. Also, note that by assuming that the possible total inflow of news is fixed, increases in fake news production crowd out truthful content: this only strengthens our result that equilibria with more verification can deliver more welfare (Section 4). Importantly, what really matters in the end is that fake news production positively responds to σ_U, which is a natural property to have: in our baseline model, this manifests in a standard supply function emerging. Alternatively, in Section 6.1 we examine the case where producers face a non-trivial choice between producing costly truthful content versus costless fake news, thus resembling news outlets of dubious reputation than straight malicious actors. The treatment of this case ends up being qualitatively identical.

3. Random matching and stationarity. The random matching technology can be seen as stemming from producers' imperfect ability to reach specific users (i.e., those engaged in unverified sharing) when attempting to distribute news to populations of interest (i.e., based on observable characteristics). In practice, this can happen because users' preferences and actions can be private information (so achieving granular levels of targeting is difficult), but also because it is a platform's algorithm that ultimately decides who gets to see what in any news feed. Finally, the stationary aspect is for tractability: it is the simplest setting for incorporating realistic inter-temporal effects associated with an endogenous measure of news quality—namely, prevalence—being affected by sharing and verification choices in previous periods.

3 Equilibrium Analysis

Our model combines competitive and strategic elements. On the one hand, it features a large number of producers and consumers of news, none of whom can affect aggregate variables. In traditional competitive markets, however, only a single variable—the quantity traded or the price—suffice to characterize equilibrium outcomes. Instead, here we must augment π by

(σ_U, σ_V) due to the social influence effects at play: users' conjectures of different degrees of verification (here, of σ_V) can lead to different (π, σ_U) pairs that in turn sustain such levels of verification. This mechanism operates through people's perception of fake news prevalence, ψ, which introduces interdependence in expected payoffs—the "game" aspect of our model.

Fake news prevalence To determine this variable, consider the total number of news circulating in any period: there is an inflow 1 of new content; there is a fraction σ_U of the previous-period news cohort that gets passed to the current period without any verification (a mix of truthful and false content); and also a fraction $(1-\pi)\sigma_V$ of truthful news from the previous cohort gets passed on because of encounters with users who verify news.

Of this total circulating, only $1 + \sigma_U$ news items can be false, which happens with probability π. Thus, fake news prevalence reads

$$\psi = \frac{\pi + \pi\sigma_U}{1 + \sigma_U + (1-\pi)\sigma_V} =: \Psi(\pi, \sigma_U, \sigma_V). \tag{1}$$

The function $\Psi : [0,1]^3 \to [0,1]$ increases with σ_U (i.e., as more users engage in unverified sharing), and also with π (i.e., as the inflow of new fake content grows). Meanwhile, it falls with σ_V (i.e., as more users verify news before sharing). Also, $\Psi(0, \cdot) \equiv 0$ and $\Psi(1, \cdot) \equiv 1$.

Note that $\Psi(\pi, \sigma_U, \sigma_V) \leq \pi$, with strict inequality when $\sigma_V > 0$ due to the term $(1-\pi)\sigma_V$ in the denominator. This drop in prevalence captures the benefits from verification and is the channel through which the social influence effects will operate: verification by users affects others' perceptions of the severity of the fake news problem, thereby influencing their sharing and verification decisions—a non-trivial fixed point will emerge. Equipped with Ψ, we now characterize the possible values that (σ_U, σ_V) can take.

The "sharing game" Fix $\pi > 0$ in what follows. Given a perceived prevalence ψ, a necessary condition for users to be willing to verify news is

$$(1-\psi)b - t \geq 0. \tag{2}$$

This is the first margin discussed in our model section: in absolute terms, verified sharing is profitable when paying the verification cost t up front (i.e., with certainty) is compensated by sharing truthful news sufficiently often, i.e., when fake news prevalence is not too large.

Conjectures of others' behavior now matter for verification incentives, through the induced values that ψ can take via the function Ψ. We now examine equilibria of the resulting *sharing game among users* when taking as given the production level π—a form of partial equilibrium analysis regarding the set of stable predictions associated with user behavior.

Note that, if everyone expects others to skip verification, then $\psi = \pi$ by virtue of $\Psi \equiv \pi$ when setting $\sigma_V = 0$ in (1). Moreover, these expectations are self-fulling when (2) fails with $\psi = \pi$ in it. Or equivalently, when production is sufficiently large according to

$$\pi > \underline{\pi} := 1 - \frac{t}{b} \in (0, 1). \tag{3}$$

Above this value $\underline{\pi}$ then, there is always an equilibrium in which there is no verification, so $\sigma_V = 0$. In this case, σ_U is determined by the mass of individuals who find it profitable to do unverified sharing, namely, to gamble between b and ℓ when prevalence takes value π, or

$$(1 - \pi)b - \pi\ell \geq 0. \tag{4}$$

Since these users must experience (relatively low) losses—formally, $\ell \leq (1-\pi)b/\pi$—it follows that fake content will diffuse at rate $\sigma_U = \underline{\Sigma}(\pi)$ where

$$\underline{\Sigma}(\pi) := G\left(\frac{(1-\pi)b}{\pi}\right). \tag{5}$$

This "no-verification" equilibrium can arise on $[\underline{\pi}, 1]$, and users segment into two sets: those engaged in unverified sharing and those who do not share news at all.

Conversely, when $\pi < \underline{\pi}$, each user gets a positive payoff from verifying news even if nobody else does, because $\Psi(\cdot) \leq \pi < \underline{\pi}$. Thus, nobody abstains from sharing, and so $\sigma_V = 1 - \sigma_U$: users segment into two sets again, but now the split is between doing unverified and verified sharing. There are two important observations stemming from this finding. First, users doing unverified sharing must find this option more profitable than doing verified sharing, so these users must experience relatively low losses:

$$\underbrace{(1-\psi)b - \psi\ell}_{\text{unverified sharing}} > \underbrace{(1-\psi)b - t}_{\text{verified sharing}} \Leftrightarrow \ \ell < t/\psi. \tag{6}$$

This is the second margin discussed in the model section, encoding a tension at play when users expect more fellow users to verify news: as ψ falls, more users find it optimal to skip verification because (6) relaxes. In other words, expectations of others' behavior and actual best responses are strategic substitutes through this channel.

From (6), a total mass $G(t/\psi)$ of users engage in unverified sharing, and hence the following fixed-point condition must hold:

$$\sigma_U = G\left(\frac{t}{\Psi(\pi, \sigma_U, 1-\sigma_U)}\right), \tag{7}$$

where we have made explicit that $\psi = \Psi(\pi, \sigma_U, \sigma_V) = \Psi(\pi, \sigma_U, 1 - \sigma_U)$. This equation encodes the feedback loop already discussed, now using the pass-through rate σ_U as the main variable. Denote the (unique, as we will show) solution to (7) as $\pi \mapsto \overline{\Sigma}(\pi)$.

Our second observation is that (7) admits a solution over $[0, \bar{\pi})$ with $\bar{\pi} \in (\underline{\pi}, 1)$: that is verification can extend beyond the point at which the no-verification equilibrium emerges (while vanishing strictly before 1). The reason is the externalities that verification creates: as verification happens along $[0, \underline{\pi}]$, it follows that $\Psi < \underline{\pi}$ in this region, which helps sustain verification to the right of $\underline{\pi}$. We refer to this outcome as the "verification" equilibrium, which features $\psi < \pi$ everywhere on $[\underline{\pi}, \bar{\pi})$ and $1 - \overline{\Sigma}(\pi)$ users verifying news.

The existence of two equilibria over $\pi \in [\underline{\pi}, \bar{\pi}]$ opens the possibility for a third type of equilibrium, analogous to mixed-strategy equilibria in coordination games when two pure-strategy equilibria exist. In this equilibrium, only a fraction $\sigma_V \in [0, 1 - \overline{\Sigma}(\pi)]$ verify news so that prevalence remains constant at $\underline{\pi}$ and hence there is indifference between verified sharing and not sharing at all (i.e., $(1 - \underline{\pi})b - t = 0$ holds), which in turn sustains such partial levels of verification. Further, a constant mass $\sigma_U = \underline{\Sigma}(\underline{\pi})$ of users engages in unverified sharing on $[\underline{\pi}, \bar{\pi}]$. Users segment into three, the third set being users who never share news. The next result confirms that these three possibilities are exhaustive.

Proposition 1 (Sharing game equilibria). *There exists* $0 < \underline{\pi} < \bar{\pi} < 1$ *such that*

$$\sigma_U = \begin{cases} \overline{\Sigma}(\pi) & \text{for } \pi \in [0, \underline{\pi}) \\ (\overline{\Sigma}(\pi), \underline{\Sigma}(\underline{\pi}), \underline{\Sigma}(\pi)) & \text{for } \pi \in [\underline{\pi}, \bar{\pi}] \\ \underline{\Sigma}(\pi) & \text{for } \pi \in [\bar{\pi}, 1] \end{cases} \tag{8}$$

where $\underline{\Sigma} : [0,1] \to [0,1]$ *defined by* (5) *is continuous and decreasing, while* $\overline{\Sigma} : [0,1] \to [0,1]$ *is the unique solution to* (7)*, also continuous and decreasing.*[10] *Meanwhile,*

$$\sigma_V = \begin{cases} 1 - \overline{\Sigma}(\pi) & \text{for } \pi \in [0, \underline{\pi}) \\ (1 - \overline{\Sigma}(\pi), \alpha(\pi), 0) & \text{for } \pi \in [\underline{\pi}, \bar{\pi}] \\ 0 & \text{for } \pi \in [\bar{\pi}, 1] \end{cases} \tag{9}$$

where $\alpha(\pi) \in [0, 1 - \underline{\Sigma}(\pi)]$ *is continuous, increasing, and satisfies* $\Psi(\pi, \underline{\Sigma}(\underline{\pi}), \alpha(\pi)) = \underline{\pi}$ *for all* $\pi \in [\underline{\pi}, \bar{\pi}]$*, whereas the mass of individuals who never share news is* $1 - \sigma_U - \sigma_V$*. Finally,* $\overline{\Sigma}(\pi) > \underline{\Sigma}(\pi)$ *for* $\pi \in [\underline{\pi}, \bar{\pi}]$ *and* $\underline{\Sigma}(\underline{\pi}) = \overline{\Sigma}(\bar{\pi})$.

We refer to the right-hand side of (8) as the *pass-through correspondence*, because it

[10] Note that $\lim_{\pi \to 0} \underline{\Sigma}(\pi) = \lim_{\pi \to 0} \overline{\Sigma}(\pi) = 1$.

determines the possible rates at which fake content diffuses through the user population. By the last part of the proposition, the equilibrium with verification exhibits a pass-through rate that is higher than that of the no-verification equilibrium (i.e., $\overline{\Sigma} > \underline{\Sigma}$ on $[\underline{\pi}, \bar{\pi}]$): this is because verification reduces the prevalence of fake news—a result we establish in the next section—which in turn makes unverified sharing less risky. Three properties follow.

First, as π increases and fake news gains prevalence, misinformation necessarily diffuses less often along each of these two equilibria: $\overline{\Sigma}$ and $\underline{\Sigma}$ are decreasing, which in particular means that the extent of verification grows along the verification equilibrium as (9) shows. Second, the pass-through rate in the "mixed" equilibrium lies in between the other two, a consequence of its intermediate verification intensity: the latter is denoted by α, which increases with π because more verification is needed to keep ψ pegged at $\underline{\pi}$ as production grows. Third, the mixed equilibrium meets the verification one at $\bar{\pi}$ (the last equality in the proposition). This is because the mixed equilibrium exhausts the potential mass of users available to verify news at $\bar{\pi}$, after which the verification equilibrium vanishes.

The multiplicity discovered reflects endogenous strategic complementarities: high levels of verification induce low levels of prevalence that, in turn, sustain high levels of unverified sharing, and vice-versa. This happens for intermediate levels of production, because verification incentives can be activated or shut down for many users there. Concretely, as we enter $\pi > \underline{\pi}$, news verification can have a strong *extensive margin* (our first margin) effect through the implied drop in prevalence, which can make verified sharing profitable for many users who were not sharing news at all: these users are enticed to enter the "sharing world." This effect can overcome *inframarginal effects*—our second substitution margin—associated with some users switching to unverified sharing as prevalence falls. In other words, verification makes it possible to transition to both higher levels of verified *and* unverified sharing.

Above $\bar{\pi}$, activating the extensive margin would necessitate verification by so many users that it would require some of those "switchers" to join—things unravel and there is no verification. On the other hand, below $\underline{\pi}$, verification is active irrespective of what others do: substitution effects dominate and the outcome is unique. To see that social effects are still at play in this latter region, we compare our verification equilibrium with one where verification is prohibitively costly (a counterfactual case $t > b$): here, the pass-through rate σ_U is uniquely determined by $\underline{\Sigma}$, but now defined over the whole interval $[0, 1]$.

Proposition 2 (Verification effects). *Consider* $\underline{\Sigma}(\cdot)$ *as* (5) *and* $\overline{\Sigma}(\cdot)$ *solving* (7). *There is a unique* $\hat{\pi} \in (0, \underline{\pi})$ *such that* $\underline{\Sigma}(\pi) > \overline{\Sigma}(\pi)$ *if* $\pi \in (0, \hat{\pi})$ *and* $\overline{\Sigma}(\pi) > \underline{\Sigma}(\pi)$ *if* $\pi \in (\hat{\pi}, \bar{\pi}]$.

Consequently, $\underline{\Sigma}$ starts above $\overline{\Sigma}$ for π close to zero, and then falls under $\overline{\Sigma}$ permanently after a crossing point $\hat{\pi} < \underline{\pi}$. Indeed, as verification becomes feasible (i.e., as t falls below

b) the extensive margin effect is strong for low π: people begin verifying news, and the diffusion of fake content falls: $\underline{\Sigma} > \overline{\Sigma}$. But with more verification taking place as π increases, prevalence falls, and the substitution channel gets activated: $\overline{\Sigma} > \underline{\Sigma}$ eventually.

Characterization of stationary equilibria To find a stationary equilibrium—and in particular, to endogeneize π—we need only to incorporate producers' choices to our analysis. This side of the market is straightforward: given a conjectured pass-through rate $\sigma_U \in (0, 1)$, only producers with costs $c < \sigma_U$ will create fake content. This, in turn, leads to a total production of $F(\sigma_U)$, and a simple characterization of stationary equilibria follows.

Corollary 1. *In any stationary equilibrium, the induced pair (π, σ_U) is an intersection point between the pass-through correspondence (8) and the inverse supply function $\pi \mapsto F^{-1}(\pi)$. A stationary equilibrium always exists, but it may not be unique.*

The dual competitive-strategic nature of the setting studied is clear from Figure 1. On the one hand, any equilibrium is, locally, the outcome of two forces that balance each other: (i) an increasing supply and a (ii) weakly decreasing curve akin to an "aggregate demand" for fake content, as seen by producers. On the other hand, our setting differs from traditional competitive analyses in that there is more than one possibility for the demand side. This is because, this demand side also operates as a supply, albeit a non-trivial one: users ultimately affect the *average quality* of the good to be "consumed" by subsequent users.

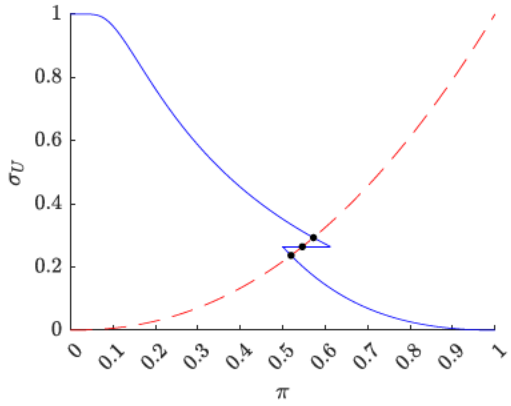

Figure 1: Stationary equilibria. $b = 1$, $t = 0.5$, $\ell \sim \text{Gamma}(2, 1)$, and $F(c) = c^{0.45}$.

4 Welfare Analysis

Welfare comparison across equilibria Towards comparing equilibria in terms of welfare, let us first characterize equilibrium prevalence.

Proposition 3 (Equilibrium prevalence). *Consider a stationary equilibrium with production π. Fake news prevalence can then take the following values:*

$$\psi = \begin{cases} \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) & \text{for } \pi \in [0, \underline{\pi}) \\ (\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)), \underline{\pi}, \pi) & \text{for } \pi \in [\underline{\pi}, \bar{\pi}] \\ \pi & \text{for } \pi \in [\bar{\pi}, 1] \end{cases} \tag{10}$$

Further, the mapping $\pi \mapsto \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))$ is (i) strictly increasing over $[0, \bar{\pi}]$; (ii) strictly below $\min\{\pi, \underline{\pi}\}$ for $\pi \in (0, \bar{\pi})$; and (iii) takes value $\underline{\pi}$ at $\pi = \bar{\pi}$.

From the result, increasing fake news production π leads to an increase in prevalence, except in the mixed equilibrium where it remains constant. In the no-verification equilibrium, the change is one-to-one (third row in (10)), while in the verification case (first row and first entry in the second row) the change is attenuated due to verification effects ($\pi \mapsto \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))$ is increasing, but always below π).[11] Along the verification equilibrium, therefore, prevalence ψ is not only below π, but in fact below $\underline{\pi}$ over $[\underline{\pi}, \bar{\pi})$; this is natural given that this equilibrium exhibits more verification than the mixed one, in which prevalence is fixed at $\underline{\pi}$. Figure 4 displays a typical prevalence correspondence (10) along with the identity function (synonym for prevalence absent any verification).

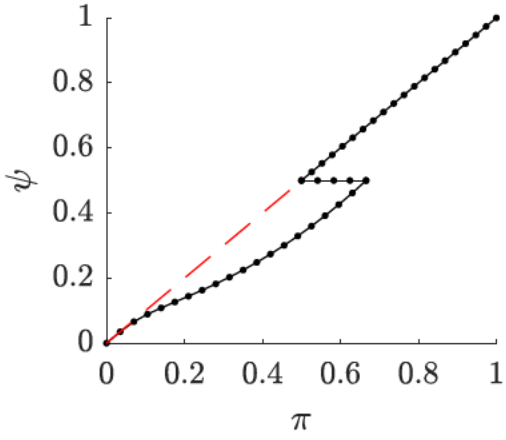

Figure 2: Prevalence correspondence. $b = 1$, $t = 0.5$, and $\ell \sim \text{Gamma}(5, 1)$.

Let $\mathcal{W}^{\text{ver}}(\pi), \mathcal{W}^{\text{mix}}(\pi)$, and $\mathcal{W}^{\neg\text{ver}}(\pi)$ denote aggregate *user welfare per news item* in the verification, mixed, and no-verification equilibria of the sharing game, for fixed π whenever they exist. Because of the random matching technology, total user welfare in each case is

[11] This can be seen from the fixed-point condition (7): increasing π raises prevalence all else being equal, in turn reducing σ_U on the left; but this puts downward pressure on Ψ through $1 - \sigma_U$ growing.

then given by per-news welfare times the corresponding volume of news in circulation, or

$$\mathcal{W}^x(\pi)[1 + \sigma_U^x(\pi) + (1-\pi)\sigma_V^x(\pi)],$$

where σ_U^x and σ_V^x denote the $x \in \{\text{ver}, \neg\text{ver}, \text{mix}\}$ entry of the correspondences (8) and (9).

Proposition 4 (Welfare comparison). *If multiple stationary equilibria exist, the one exhibiting the largest fake news production maximizes both per-news and total user welfare. Outside the multiplicity region, more fake news lead to lower (per-news and total) welfare.*

The proposition highlights the importance of strategic complementarities. When these arise, both components of user welfare—per-unit user welfare and circulating volume—satisfy a strong ranking across equilibria: the verification equilibrium generates more user welfare per news than the mixed uniformly on $[\underline{\pi}, \bar{\pi}]$, and so does the latter equilibrium relative to the no-verification one. And similarly for circulating volume: the mixed equilibrium reduces circulation by excluding some users from sharing news and by inducing less unverified sharing, while the no-verification equilibrium is even more extreme along both metrics. Thus, welfare increases as we go up along the supply curve reaching higher levels of verification that lower prevalence; but since more users do unverified sharing, too, fake news production increases.[12]

This does not happen away from the multiplicity region, where substitution effects dominate. There, higher fake news production increases prevalence despite the increase (if any) in verification intensity (Proposition 3)—but higher fake news prevalence implies lower per-unit welfare as we demonstrate in the proof. Thus, per-news user welfare decreases as we move down along each decreasing branch of the correspondence. Importantly, circulating volume falls too as π grows: in the verification equilibrium, this is because higher verification rates eliminate more news, while in the no-verification equilibrium, this is because unverified sharing decreases. This explains the last part of the proposition.[13]

Externalities In the verification equilibrium, $\sigma_U^{\text{ver}} + \sigma_V^{\text{ver}} \equiv 1$, and hence the total volume of news circulating is $2 - \pi\sigma_V^{\text{ver}}$. Thus, individual verification choices exert *externalities* on other users through two channels. First, by reducing the news volume circulating, which is a negative 'quantity' effect—this channel is rather straightforward in that total volume is maximized when everyone engages in unverified sharing. Second, by making encounters with truthful content more likely—a positive 'quality' effect driven by prevalence. This latter

[12] Kranton and McAdams (2024) find conditions under which an exogenous increase in misinformation leads to higher welfare in a model where welfare per news falls in response, but volume increases. Instead, in our model, welfare per news can grow across equilibria with more misinformation due to verification effects.

[13] These properties also hold in the multiplicity region. In the mixed case, per unit welfare is constant, while total welfare grows due to circulating volume growing as more people enter the (verified) sharing world.

channel has the potential to uncover important externalities: by reducing fake news prevalence, more verification increases the payoff of those originally sharing news (substitution effects reflecting better alternatives); and among those users originally not sharing news at all, some of them will find it profitable to do it now (the extensive margin effect).

Henceforth, we focus on the quality channel by considering a planner who maximizes *per-news* user welfare at each $\pi \in [0,1]$ using segmentations $(\sigma_U, \sigma_V) \in [0,1]^2$ with $\sigma_U + \sigma_V \leq 1$; this is an interim measure of user welfare that parallels the fact that our users act in a sequentially rational fashion (they make choices conditional on having encountered a news item). We use this measure as our efficiency criterion unless otherwise stated and return to the case of a planner who maximizes total surplus (i.e., volume included) in Remark 2.

For verification to deliver a positive payoff, however, recall that prevalence has to be lower than $\underline{\pi}$, in which case verified sharing provides more utility than not sharing at all (as in Section 3). Thus, we can restrict to $\sigma_V + \sigma_U = 1$ when searching for efficient outcomes featuring verification. Consider then

$$\bar{\pi}_P := \sup\{\pi \in [0,1] : \exists\ \sigma_U \in [0,1] \text{ s.t. } \Psi(\pi, \sigma_U, 1-\sigma_U) \leq \underline{\pi}\},$$

the maximum level of production that can feasibly sustain verification. Clearly, $\bar{\pi}_P$ is characterized by $\Psi(\bar{\pi}_P, 0, 1) = \underline{\pi}$, because prevalence is minimized when everyone is engaged in verification. Also, $\bar{\pi}_P > \bar{\pi}$, because in the verification equilibrium, a total mass of $\overline{\Sigma}(\bar{\pi}) > 0$ users engage in unverified sharing, so there is scope for a planner to further lower prevalence to the right of $\bar{\pi}$. And clearly, $\bar{\pi}_P < 1$ because $\underline{\pi} \in (0,1)$.[14]

Altogether, a planner can induce verification in the region $[0, \bar{\pi}_P]$ but not elsewhere. Since the planner can always opt to induce no verification at all (in which case the remaining mass $1 - \sigma_U$ of users would never share news), the planner's value function takes the form

$$\mathcal{W}_P(\pi) := \max\{\mathcal{W}_P^{\text{ver}}(\pi), \mathcal{W}_P^{\neg\text{ver}}(\pi)\}\mathbb{1}_{\pi \in [0, \bar{\pi}_P]} + \mathcal{W}_P^{\neg\text{ver}}(\pi)\mathbb{1}_{\pi \in [\bar{\pi}_P, 1]}$$

where

$$\begin{aligned} \mathcal{W}_P^{\text{ver}}(\pi) := \max_{0 \leq \sigma_U \leq 1} & (1 - \Psi(\pi, \sigma_U, 1-\sigma_U))b - t(1-\sigma_U) - \Psi(\pi, \sigma_U, 1-\sigma_U) \int_0^{G^{-1}(\sigma_U)} \ell dG \\ \text{s.t.} \quad & \Psi(\pi, \sigma_U, 1-\sigma_U) \leq \underline{\pi} \end{aligned} \tag{11}$$

and

$$\mathcal{W}_P^{\neg\text{ver}}(\pi) := \max_{0 \leq \sigma_U \leq 1} \int_0^{G^{-1}(\sigma_U)} [(1-\pi)b - \pi\ell] dG. \tag{12}$$

[14] By inspection of (1), $\bar{\pi}_P = 2\underline{\pi}/(1+\underline{\pi}) < 1$.

To understand (11), note that $\mathcal{W}_P^{\text{ver}}(\pi)$ requires a mass σ_U to engage in unverified sharing, and $1-\sigma_U$ to share only after a news item is verified to be truthful. This means that: since everyone shares news with some chance, all users bear the benefit b in expectation (first term); aggregate verification costs amount to $t(1-\sigma_U)$ (second term); and the aggregate (ex post) loss from unverified sharing equals $\int_0^{G^{-1}(\sigma_U)} \ell dG$ (third term), because it is always more efficient to induce high ℓ types to verify news. On the other hand, (12) captures user welfare when: no one verifies news (so $\Psi = \pi$); low types ℓ engage in unverified sharing; and higher types (i.e., $\ell > G^{-1}(\sigma_U)$) do not share news at all (thus getting a payoff of zero).

Let $\sigma_{U,P}^{\text{ver}}(\pi)$ and $\sigma_{U,P}^{\neg\text{ver}}(\pi)$ denote the solutions to (11) and (12), respectively.

Proposition 5 (Planner's solution and market efficiency). *The planner's value function $\mathcal{W}_P(\pi)$ has the following properties:*

(i) Region $[0, \bar{\pi}_P]$. There exists $\bar{\pi}_v \in (\bar{\pi}, \bar{\pi}_P)$ such that $\mathcal{W}_P(\pi) = \mathcal{W}_P^{ver}(\pi)$ for all $\pi \in (0, \bar{\pi}_v]$. Further, user verification satisfies $1-\sigma_{U,P}^{ver}(\pi) > 1-\overline{\Sigma}(\pi)$ in $[0, \bar{\pi}]$, while $1-\sigma_{U,P}^{ver}(\pi) > 0$ in $[\bar{\pi}, \bar{\pi}_v]$. Thus, equilibrium outcomes are inefficient in $[0, \bar{\pi}_v]$.

(ii) Region $[\bar{\pi}_P, 1]$. The maximizer $\sigma_{U,P}^{\neg ver}(\pi)$ that attains $\mathcal{W}_P^{\neg ver}(\pi)$ is such that $\sigma_{U,P}^{\neg ver} = \underline{\Sigma}(\pi)$. Thus, the no-verification equilibrium is always efficient in $[\bar{\pi}_P, 1]$.

Part (ii) states that when verification is not implementable from a social perspective, the market outcome is obviously efficient. Part (i) in turn speaks to the inefficiencies that can arise when verification is implementable. First, the planner prefers that verification happens strictly beyond $\bar{\pi}$ at which the verification equilibrium ceases to exist. Second, whenever the verification equilibrium exists, the planner would have preferred more verification to take place ($1-\sigma_{U,P}^{\text{ver}}(\pi) > 1-\overline{\Sigma}(\pi)$). Intuitively, requiring verification by some types to the left of $t/\Psi(\pi, \overline{\Sigma}(\pi), 1-\overline{\Sigma}(\pi))$—the lowest type that verifies news in the verification equilibrium (see (4))—can generate a large increase in welfare due to truthful news being more frequently encountered—this is the term $(1-\Psi)b$ in (11), which applies to all users (and thus ensures that those additional users who verify news continue having positive utility).

This additional verification is profitable for a planner only up to a level of production which we call $\bar{\pi}_v < \bar{\pi}_P$. This is because an exceedingly large volume of fake content created would require too many users to bear verification costs in order to obtain a meaningful reduction in prevalence: the planner would have to exchange too many mild losses ℓ for high costs t. Slightly to the right of $\bar{\pi}_v$ then, $\mathcal{W}_P^{\neg\text{ver}}(\pi) > \mathcal{W}_P^{\text{ver}}(\pi)$, as Figure 3 below shows—there, the market outcome is efficient despite verification being implementable.

Remark 1. The planner can implement the desired level of verification by subsidizing verification costs. Indeed, write the planner's solution as $\sigma_{U,P}^{\text{ver}}(\pi; t)$. Then, using (7), the subsidy

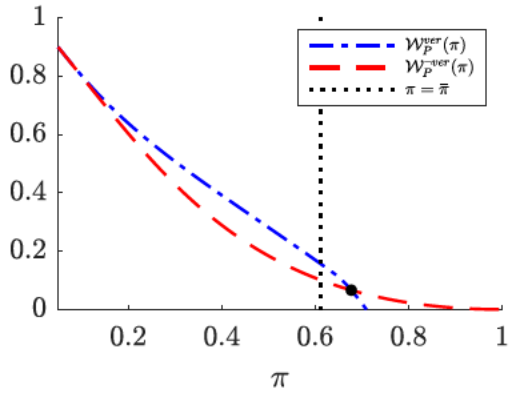

Figure 3: Planner's value function and $\bar{\pi}$ (vertical line). $b = 1$, $t = 0.45$, and $\ell \sim \text{Exp}(1)$.

is $t - t'$ where t' satisfies $\sigma_{U,P}^{\text{ver}}(\pi; t) = G\left(t'/\Psi(\pi, \sigma_{U,P}^{\text{ver}}(\pi; t), 1 - \sigma_{U,P}^{\text{ver}}(\pi; t))\right)$. Note that the discrepancy between t and t' in this expression implies that market outcomes (i.e., subsidies absent) would still be inefficient had verification costs been t' from the start.

Remark 2. The problem of a planner interested in maximizing total user surplus—i.e., circulating volume included—is more complicated. Appendix B establishes conditions such that, for given π, the solution to this problem is in $(\sigma_{U,P}^{\text{ver}}(\pi), \sigma_U^{\text{ver}}(\pi))$: to promote more news circulation, the planner requires *less* verification than in the case just studied, but *more* than in the verification equilibrium, due to the existence of positive quality externalities.

5 Policy Interventions in Practice

We examine the impact of three policies that are currently used to combat fake news: lowering verification costs for users; using algorithmic filters; and certifying verified news.

Our focus is on two topics. First, towards assessing how stationary equilibria change, we examine how the sharing game's key outcomes—$(\sigma_U, \sigma_V, \psi)$ and welfare—vary for each of the three possible equilibria. We document changes in welfare per news to understand how these policies affect aggregate user utility through the prevalence channel. Appendix C contains the corresponding variations in the volume of news circulating as corollaries.

Second, towards the possibility of equilibria being refined, we study how the *multiplicity region* $[\underline{\pi}, \bar{\pi}]$ changes, where $\underline{\pi} := 1 - t/b$ (see (3)) while $\bar{\pi}$ is the unique solution to $\Psi(\bar{\pi}, \overline{\Sigma}(\bar{\pi}), 1 - \overline{\Sigma}(\bar{\pi})) = \underline{\pi}$ (see Proposition 1).

5.1 Verification Costs

Social media platforms have begun to offer more and better fact-checking services to users, while still allowing them to decide whether to share: in our model, this is a decrease in the verification cost t. Thus, we consider reductions in this variable while remaining strictly positive, as fully eliminating these costs is likely infeasible in practice.

Notice that $\underline{\pi}$ and $\overline{\pi}$ are decreasing in t: with more affordable verification, the verification equilibrium exists over a longer range $[0, \bar{\pi}]$; the no-verification equilibrium exists over a smaller range $[\underline{\pi}, 1]$; and any conclusion regarding the mixed equilibrium will depend on how the multiplicity region responds. But if a reduction in t decreases $\bar{\pi} - \underline{\pi}$, then the set of equilibria of this game is refined in favor of selecting the verification equilibrium.

The next result explores this latter topic while also examining how σ_U, ψ, and user welfare in the sharing game change. We study these changes in regions of π where these variables continue to be defined after the change in t—we refer to any such region as a "common domain," which will depend on the type of equilibrium of the sharing game at hand.[15] Recall also that in the verification equilibrium $\sigma_V = 1 - \sigma_U$.

Proposition 6 (Lowering verification costs). *Consider a small drop in verification costs from $t \in (0, b)$ to $t' < t$ such that all equilibria of the sharing game continue to exist.*

(i) For any fixed value of π in the corresponding common domain:

(i.1) Verification equilibrium: σ_U and ψ fall, while user welfare rises.

(i.2) Mixed equilibrium: σ_U and σ_V fall, ψ rises, and user welfare falls.

(i.3) No-verification equilibrium: σ_U, ψ, and user welfare do not change.

(ii) Refining effects. Both $\underline{\pi}$ and $\bar{\pi}$ are decreasing in t. Also, the wedge $\bar{\pi} - \underline{\pi}$:

(ii.1) rises when $0 < t' < t$ are sufficiently close to b;

(ii.2) falls when $0 < t' < t$ are sufficiently low.

For part (i), note that as t falls and more people verify news, misinformation diffuses more slowly in the verification equilibrium ($\overline{\Sigma}(\cdot)$ falls), so fake news prevalence falls; in turn, user welfare must increase. Now, recall that in the mixed equilibrium prevalence is pegged at $\underline{\pi}$ while the (constant) pass-through rate satisfies $\underline{\Sigma}(\underline{\pi}) = G(t/\psi) = G(t/\underline{\pi})$, which is increasing in t. Thus, lowering verification costs reduces unverified sharing, so to

[15] Clearly, given $t \neq t'$ and induced values $\{\underline{\pi}(t), \bar{\pi}(t), \underline{\pi}(t'), \bar{\pi}(t')\}$, these common domains are $[0, \min\{\bar{\pi}(t), \bar{\pi}(t')\}]$, $[\max\{\underline{\pi}(t), \underline{\pi}(t')\}, 1]$ and $[\max\{\underline{\pi}(t), \underline{\pi}(t')\}, \min\{\bar{\pi}(t), \bar{\pi}(t')\}]$.

induce indifference (between verifying news and not sharing at all) prevalence must increase—as a byproduct, *user welfare and verification both fall*. Finally, since the no-verification equilibrium does not explicitly depend on t, it does not change in the common domain induced by the policy change: it only emerges at higher levels of production, consistent with the lower pass-through rate of the mixed equilibrium. (Extrapolating to stationary equilibria—π endogenous—is straightforward given our upward sloping supply: the new crossing points on the verification and mixed branches exhibit less unverified sharing and less production, while nothing happens in the common domain of the no-verification branch.)

Part (ii) is also intuitive. Consider as a starting point the extreme case $t \geq b$ (referenced in Proposition 2) in which only the no-verification equilibrium can emerge. As we begin lowering verification costs, the other two equilibria necessarily emerge at some point. But as we further continue this process, things are reversed: verification becomes so attractive that it starts having refining effects, selecting the verification equilibrium.

Lowering verification costs can be a powerful tool because it affects user welfare directly by increasing the payoffs of those verifying news, and indirectly by lowering fake news prevalence for those doing unverified sharing. However, the fact that reductions in pass-through rates and in fake news inflows can be simultaneously accompanied by an increase in prevalence and drops in user welfare and verification rates—as the mixed equilibrium demonstrates—is an important warning sign to consider.

5.2 Algorithmic Filters

We now enrich the model to allow for detection algorithms. A key concern regarding such *filters* has been their potential use for removing content, which some studies document can be perceived as a form of censorship (e.g., Lazer et al., 2018). Our focus is different: we look at how these technologies affect users' inferences and their incentives to verify news.

We consider an algorithm that assesses news articles as they first enter the platform and before they reach consumers—a form of pre-screening that happens only once, a proxy for filtering for novel content in its early stages in a platform. Clearly, eliminating truthful news articles carries social costs. Thus, we focus on the more interesting case in which truthful news articles always pass the filter, but fake news articles are detected with probability $\phi \in [0, 1]$. Because of the public announcements that platforms have made on this topic, we assume that ϕ—which measures the filter's quality—is observable to both users and producers: in this way, changes in ϕ have the potential to affect all equilibrium variables.

In the sharing game, ϕ affects users only through prevalence (i.e., it does not affect payoffs

directly as the verification cost t did), which now takes the form

$$\Psi^{\phi}(\pi, \sigma_U, \sigma_V) := \frac{\overbrace{(1-\phi)\pi(1+\sigma_U)}^{\text{originally } \pi(1+\sigma_U)}}{\underbrace{(1-\phi\pi)(1+\sigma_U)}_{\text{originally } 1+\sigma_U} + (1-\pi)\sigma_V}. \tag{13}$$

To understand (13), note that only $1-\phi\pi$ news items are able to enter the platform in any period—the term highlighted in the denominator—of which only $(1-\phi)\pi$ news items are false, the term appearing in the numerator; both are multiplied by $(1+\sigma_U)$ due to the two possible vintages of news in any period. Finally, because the filter makes no mistakes with truthful content, the last term in the denominator is unchanged.

The introduction of a filter leads to a fall in prevalence all else equal. Indeed,

$$\Psi^{\phi}(\pi, \sigma_U, \sigma_V) = \Psi(\zeta(\pi;\phi), \sigma_U, \sigma_V), \quad \text{where} \quad \zeta(\pi;\phi) = \frac{(1-\phi)\pi}{1-\phi\pi} \tag{14}$$

is the fraction of fake content among those fresh items that passed the filter, and Ψ is our prevalence function (1). Since $\zeta < \pi$, prevalence falls (pointwise in $\pi > 0$): in other words, the presence of a filter naturally creates an "implied truth effect" (Pennycook et al., 2020).

Operationally, (14) implies that our entire analysis from Section 3—both towards obtaining the pass-through correspondence and characterizing stationary equilibria—admits a direct adaptation to this case, after two simple modifications are made. First, the thresholds $\underline{\pi}$ and $\bar{\pi}$ obviously change. More generally, we let $\underline{\pi}(\phi)$ denote the production level at which both the mixed and no-verification equilibria emerge in the presence of a filter with quality $\phi \in [0,1]$, while $\bar{\pi}(\phi)$ is the corresponding value at which the verification equilibrium ceases to exist—the values $\underline{\pi}(0)$ and $\bar{\pi}(0)$ are those in our baseline model (Proposition 1). Second, a producer's per-item expected payoff now reads $(1-\phi)\sigma_U$ because passing a filter is required for misinformation to diffuse—thus, a filter contracts the supply of fake content.[16]

Proposition 7 (Algorithmic filters). *Consider a filter as above, with quality $\phi \in (0,1)$.*

(i) In the sharing game, as ϕ increases over a common domain:

(i.1) Verification equilibrium: σ_U rises, ψ falls, and welfare rises.

(i.2) Mixed equilibrium: σ_V falls while σ_U, ψ and welfare all remain constant.

(i.3) No-verification equilibrium: σ_U rises, ψ falls, and welfare rises

[16] If we had not normalized the producers' payoffs, expected revenue per news item produced would read $(1-\phi)[1+\sigma_U]$, and the same contraction takes place.

(ii) Refining effects. The functions $\phi \mapsto \underline{\pi}(\phi)$ and $\phi \mapsto \bar{\pi}(\phi)$ are increasing. Also, if $t < b/2$, $\phi \mapsto \bar{\pi}(\phi) - \underline{\pi}(\phi)$ is decreasing at all $\phi \in (0,1)$.

(iii) Stationary equilibria. On the verification branch, equilibrium π rises as ϕ grows marginally if and only if $|\frac{\partial \overline{\Sigma}}{\partial \pi}|\frac{\zeta}{\overline{\Sigma}} > \frac{1}{1-\zeta}$ at that point; meanwhile, σ_U and welfare unequivocally rise, while ψ always falls. The same conclusions hold in the no-verification equilibrium (which uses $\underline{\Sigma}$). In the mixed case, π falls, and the rest is unchanged.

Regarding part (i.1), σ_U and ψ move in a different direction (unlike in Proposition 6): as ϕ increases and prevalence ψ falls all else equal, users find it more profitable to engage in unverified sharing because ϕ does not directly affect payoffs as t does. While verification rates drop, the direct effect of a better filter dominates, leading to lower fake news prevalence which enhances welfare. The same conclusions hold in the no-verification equilibrium, simply because the (countervailing) effect of reduced user verification is absent.

Consider the first part of (ii) now. Since prevalence falls pointwise in $\pi > 0$ due to $\zeta < \pi$, verification can be supported beyond $\bar{\pi}(0)$. Also, as fake news prevalence is given by $\zeta(\pi;\phi) < \pi$ in the absence of verification, $\underline{\pi}(\phi)$ satisfying $\zeta(\underline{\pi}(\phi);\phi) = \underline{\pi}(0)$ implies $\underline{\pi}(\phi) > \underline{\pi}(0)$—i.e., the no-verification and mixed equilibria must emerge at higher production levels. But since at $\underline{\pi}(\phi)$, prevalence continues to take value $1 - t/b$, the mixed equilibrium remains unchanged over a common domain, explaining (i.2). Finally, the last part of (ii) states that as long as verification costs are not too high, improving the filter's quality refines the set of equilibria: in other words, lowering verification costs and introducing algorithmic filters can complement each other in selecting the equilibrium with verification.

Lastly, part (iii) states what happens when we also incorporate supply effects to make π endogenous. With an expansion in the pass-through correspondence along the verification and no-verification branches, and a contraction in the supply of fake content, the impact on π depends on the elasticity of the relevant branch—the condition obtained is standard. But since σ_U necessarily increases as a byproduct, equilibrium prevalence must fall (implying that welfare must increase); this can be easily seen in the verification branch, wherein $\sigma_U = G(t/\psi)$.[17] In turn, in an equilibrium along the mixed branch, production necessarily falls after the supply contraction, while the rest of the variables remain fixed; this is possible because, as ϕ increases, $\underline{\pi}(\phi)$ adjusts to keep $\zeta(\underline{\pi}(\phi);\phi)$ pegged at $\underline{\pi}(0) = 1 - t/b$.

The relationship between user and algorithmic verification is subtle: they can behave as substitutes, reflected in user verification (weakly) falling as a filter is introduced; but they can also act as complements, in that a filter expands the region over which human

[17] Recall that in our baseline model, if there is no verification then $\sigma_U = G((1-\pi)b/\pi)$ (see (5)). The corresponding expression in the presence of a filter then is $\sigma_U = G((1-\zeta)b/\zeta)$, from where ζ must fall.

verification can arise. An advantage of filters is that they do not create unintended welfare effects through the mixed equilibria, while lower verification costs have the upside that, because users' payoffs are directly affected, any positive effects are more pronounced. Both policies, however, reinforce each other in selecting the equilibrium with verification if such costs are not too high, in which case prevalence falls and user welfare increases.

5.3 Certifying Verified News

Part of the identification problem faced by users is that they are uncertain whether a news item was encountered because someone verified it to be truthful in the past. News certification—whereby news items that users have verified receive a "stamp" by the platform that certifies their status—likely alleviates this problem. Naturally, news certification is not a frictionless endeavor, as it requires communication from users to the platform (reporting that content was verified to be truthful) and subsequent efforts by the platform itself (fact-checking sources, reviewing evidence etc.).

To account for these issues, we introduce a variable $\beta \in [0,1]$ capturing the fraction of verified news items that get certified. This means that from the original $(1-\pi)\sigma_V$ volume of truthful items that were shared after a successful verification, a fraction $(1-\pi)\beta\sigma_V$ will be known to be truthful by the users who encounter them. Hence, this volume can be removed from the pool of news of unknown quality, which now has prevalence given by

$$\Psi^{\beta}(\pi, \sigma_U, \sigma_V) = \frac{\pi(1+\sigma_U)}{1+\sigma_U+\underbrace{(1-\pi)(1-\beta)\sigma_V}_{\text{originally } (1-\pi)\sigma_V}}, \tag{15}$$

where the last term in the denominator captures the contribution of uncertified, but verified-to-be-truthful content. Two observations are in order. First, since all news items are different, and items verified to be false are discarded, tagging verified fake content has no impact on prevalence: otherwise, if multiple replica circulate, there are obvious benefits from tagging fake news. Second, this policy enables us to examine the one term that the algorithmic filters studied in Section 5.3 did not affect (*cf.* Ψ^{ϕ} in (13)).

If $\beta = 0$ in (15), no verified items are certified, and we recover the baseline model. As β grows, users exclude more and more truthful items from the pool of dubious news; the composition of the latter pool worsens—Ψ^{β} is increasing in β (pointwise in π)—in a form of "implied fake effect" associated with untagged content. Finally, if $\beta = 1$, all verified-truthful content is certified, so prevalence $\Psi^{\beta} \equiv \pi$, as fake items within young and old vintages come in equal proportions in all periods.

As in the previous sections, a pass-through correspondence also emerges here. The old threshold $\underline{\pi}$ continues to remain fixed at $1-t/b$ for all $\beta \in [0,1]$, as shutting down verification (i.e., setting $\sigma_V = 0$) eliminates the contribution of β to Ψ^β in (15). On the other hand, the threshold $\bar{\pi}(\beta)$—at which the verification equilibrium ceases to exist—is decreasing in β: as certification grows and more truthful content gets separated out, verifying news is less attractive because the pool's composition becomes worse. We state these findings, among others, in the next proposition—observe that the measures of welfare discussed below pertain to the pool of news of unknown quality (not news items that were certified to be true).

Proposition 8 (Certifying verified content). *(i) In the sharing game, as β grows and there is more certification, over a common domain:*

(i.1) Verification equilibrium: σ_U falls, ψ rises, and welfare falls.

(i.2) Mixed equilibrium: σ_V rises, while σ_U, ψ and welfare all remain constant.

(i.3) No-verification equilibrium: σ_U, ψ and welfare remain all constant.

(ii) Refining effects. The wedge $\bar{\pi}(\beta) - \underline{\pi}$ is decreasing in β. Also, $\bar{\pi}(\beta) - \underline{\pi} \to 0$ as $\beta \to 1$. In this case, $\Psi^\beta \equiv \pi$ and the pass-through correspondence is the continuous function

$$\sigma_U = \begin{cases} G(t/\pi) & \text{for } \pi \in [0, \underline{\pi}) \\ \underline{\Sigma}(\pi) & \text{for } \pi \in [\underline{\pi}, 1] \end{cases}$$

with $\underline{\Sigma}(\pi)$ given by (5); this equilibrium is efficient from the users' perspective.

The logic should be familiar. As fake news prevalence increases when certification β rises, doing verified sharing is more attractive than its unverified counterpart, and σ_U falls in the verification equilibrium of the sharing game (part (i.1)). Meanwhile, the mixed equilibrium is mostly unchanged because its implied prevalence $\underline{\pi}$ is independent of β; but to keep prevalence pegged at that level, σ_V must grow. In turn, the no-verification also remains the same because certification loses all its power in the absence of verification. And as before, user welfare (associated with uncertified news) and prevalence move in opposite directions.

The refining effects are strong here, as only one threshold can be affected. In the extreme case of complete certification ($\beta = 1$), the social-influence channel—news verification non-trivially affecting the prevalence of fake content among items of unknown quality—disappears: only one equilibrium emerges, which exhibits verification for moderately low levels of production ($\pi < \underline{\pi}$). Further, since certification eliminates all the externalities associated with dubious news, this equilibrium maximizes user welfare—but unlike with a perfect filter ($\phi = 1$), this efficient outcome features fake content circulating in the platform.

We conclude this section with two observations. First, since the pass-through correspondence (weakly) contracts when β increases, while the supply is unaffected, such a change (weakly) lowers π and σ_U in a stationary equilibrium; from here, prevalence increases and welfare falls. Second, related to this latter point, the fact that welfare can fall should not be taken as an argument against this policy, as (i) users only care about *sharing* content in our model and (ii) news items that have been certified are already in their last period of existence. Instead, if there were direct benefits from *seeing* truthful content, or news items lived for more periods, certification would bring additional tangible benefits to be considered.[18]

6 Extensions

6.1 Costly Truthful Content

As argued, our supply side is motivated by malicious suppliers for whom truthful content is simply not an option; hence their opportunity costs are linked to other alternatives. But one could also envision less-established news outlets who face a non-trivial trade off between producing truthful and fake content. In this case, truthful content can be more costly because it requires efforts devoted to accurate reporting. For simplicity then, suppose that producers vary in their costs of producing truthful content according to $c \sim F(\cdot)$ with support in $[0, 1]$, while producing fake content is costless. Note that the benefit of producing truthful content is that it diffuses at a rate $\sigma_U + \sigma_V$, which creates a non-trivial trade off.

Because σ_V is an additional variable of consideration for a producer, it turns out that it is easier to obtain a supply curve in terms of this variable instead of σ_U. Indeed, a producer will choose to create truthful content if and only if

$$\sigma_U + \sigma_V - c \geq \sigma_U \iff \sigma_V \geq c,$$

so the corresponding supply of trustworthy content is $1 - \pi = F(\sigma_V)$. Stationary equilibria are then found by intersecting the decreasing (in the π-axis) inverse supply $F^{-1}(1 - \pi)$ with the *verification correspondence* (9) for σ_V determined in Proposition 1, whose branches are weakly increasing—Figure 4 features an example where the multiplicity arises again.[19] Equipped with the intersection points, one can determine the extent of unverified sharing by turning to the pass-through correspondence (8).

[18] We can always interpret b as an intrinsic benefit from sharing news (i.e., accrued irrespective of the news vintage), in which case welfare would be positively affected by the news that were certified to be true.

[19] There is always an equilibrium with $\pi = 1$ and $\sigma_V = 0$, despite producers not being completely malicious.

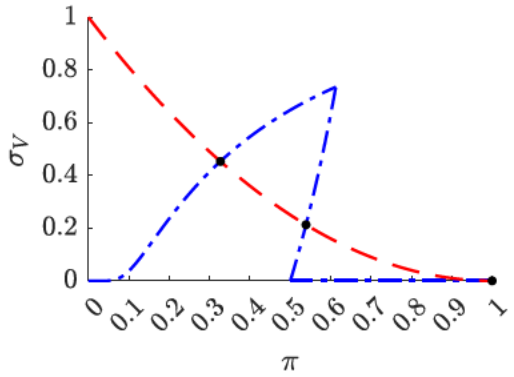

Figure 4: Parameter values: $b = 1$, $t = 0.5$, $\ell \sim \text{Gamma}(2, 1)$, and $F(c) = \sqrt{c}$. The supply curve is in red, while the verification correspondence in blue.

6.2 Network Externalities

The 'quality' externalities examined thus far are linked to a tangible measure such as news veracity. But since quality is unobserved, these externalities also have an informational flavor in that they operate through the (endogenous) belief about how pervasive fake content is. Thus, one message of the paper is that, in the context of misinformation, social influence effects can be at play despite not explicitly incorporating traditional *network effects*: situations when individual ex post payoffs explicitly depend on the behavior of others.

We can accommodate this type of phenomenon too. Concretely, consider losses as

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

where n is a differentiable function satisfying $n' > 0$ and $n(0) = 1$. That is, as a larger mass of users shares without verifying, the loss that each user suffers from sharing fake content decreases (e.g., because it is easier to justify such behavior if many others are doing it too). This dependence on an aggregate variable is reminiscent of Becker (1991).

To isolate how this payoff externality can generate complementarities, we assume two things. First, the benefit of sharing truthful news articles b scales in the same manner: in this way, substitution effects—verified versus unverified sharing—remain unchanged.[20] Second, we assume that all verified news items are certified ($\beta = 1$), enabling us to eliminate confounding effects from the aforementioned information externalities. Since $\psi \equiv \pi$ in this

[20]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. Thus, dividing b by n can be seen as a penalty associated with those long-term losses.

case, the following fixed point must hold:

$$\sigma_U = G\left(\min\left\{\frac{(1-\pi)b}{\pi}, \frac{tn(\sigma_U)}{\pi}\right\}\right).$$

Indeed, someone engaging in unverified sharing must find it more profitable than both not sharing at all (i.e., $\ell \leq (1-\pi)b/\pi$) and doing verified sharing (i.e., $\ell \leq tn(\sigma_U)/\pi$).[21]

The next figure depicts the solutions σ_U that can arise for each π (the discontinuity only a result of the grid used). The main difference is that the resulting pass-through correspondence can have *increasing* portions: higher prevalence can be consistent with larger unverified sharing rates because users' expectations of high unverified sharing behavior weakens verification incentives, thereby making those expectations self-fulfilling.

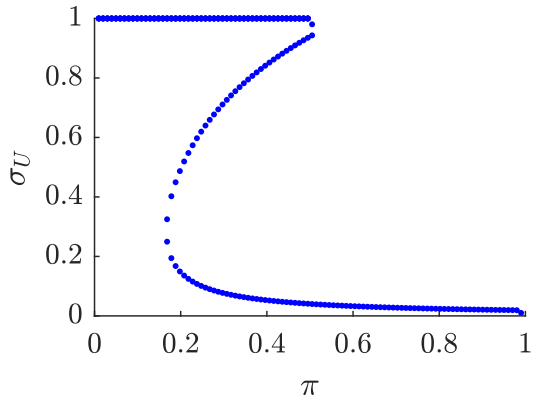

Figure 5: $t = 0.1$, $b = 5$, $n(\sigma_U) = \exp(3.5\sigma_U)$, and $\ell \sim U([0,5])$.

6.3 Other Distributions for Losses

Bounded support We have assumed that losses have unbounded support. Suppose instead that losses are distributed over $[0, \bar{\ell}]$, with $t < \bar{\ell} < +\infty$ (otherwise, no one will ever verify news). In this case, $\min\{\psi\bar{\ell}, t\} = \psi\bar{\ell}$ for sufficiently small ψ, hence everyone prefers to do unverified sharing. This means that a third threshold $\underset{\sim}{\pi} < \underline{\pi}$ may emerge such that the pass-through correspondence takes the value 1 over $[0, \underset{\sim}{\pi}]$; since no one verifies news at that threshold, it follows that $\psi = \underset{\sim}{\pi}$ there, so $\underset{\sim}{\pi} = t/\bar{\ell}$.

Proposition 9 (Pass-through correspondence with bounded losses). *Suppose that the users' losses ℓ take values in $[0, \bar{\ell}]$ and $\bar{\ell} > t$. Then, if $t < b/(1+b/\bar{\ell})$, $\sigma_U = 1$ for all $\pi \in [0, \underset{\sim}{\pi}]$, while for all $\pi > \underset{\sim}{\pi}$, it takes values according to the pass-through correspondence (8). Conversely, if $t \geq b/(1+b/\bar{\ell})$, $\sigma_U = 1$ for all $\pi \in [0, \underline{\pi}]$, while $\sigma_U = \underline{\Sigma}(\pi)$ otherwise.*

[21] The actual inequalities are $(1-\pi)b/n(\sigma_U) - \pi\ell/n(\sigma_U) \geq 0$ and $(1-\pi)b - \pi\ell/n(\sigma_U) \geq (1-\pi)b - t$.

The possibility of $\underset{\sim}{\pi}$ playing an active role depends on verification costs: t can no longer be arbitrarily close to b to induce verification in an equilibrium of the sharing game. Indeed, it is easy to see that $\underset{\sim}{\pi} < \underline{\pi}$ is equivalent to $t < b/(1 + b/\bar{\ell})$—which is tighter than $t < b$—so if this condition is violated, it is not possible to induce verification precisely where it is most profitable: namely, for low levels of production, or $\pi \leq \underline{\pi}$. Since it is not possible to induce verification for low values of π, it is not possible to induce it beyond $\underline{\pi}$, so the equilibrium does not exhibit verification—as the last part of the proposition confirms.

Otherwise, verification can be an equilibrium outcome of the sharing game; but as π falls, it will vanish at $\underset{\sim}{\pi}$, i.e., strictly before reaching $\pi = 0$. Further, as $\bar{\ell}$ grows, verification is more likely to arise because the bound on t relaxes and $\underset{\sim}{\pi}$ falls: in the limit as $\bar{\ell} \to \infty$, we recover the original pass-through correspondence (8) when $t < b$.

Concentrated losses One may also wonder what happens when instead we allow for concentrated losses rather than benefits. We explore two cases. First, a mirror image of our baseline model where ℓ is a scalar such that $t < \ell$ (otherwise nobody would verify), while b is a random variable distributed over $[0, \infty)$ via a cdf G. Second, our baseline model modified to have both benefits and losses that are concentrated.

Proposition 10 (Pass-through *function* with concentrated losses). *Suppose that G is continuously differentiable and that $\sup_{z \geq 0} zG'(z) \leq 1$. Then, there are production cutoffs $\underset{\sim}{\pi} := t/\ell$ and $\bar{\pi} \in (0, 1)$ with $\underset{\sim}{\pi} < \bar{\pi}$ such that the pass-through and verification rates are determined by the following functions:*

$$(\sigma_U, \sigma_V) = \begin{cases} \left(1 - G\left(\frac{\pi\ell}{1-\pi}\right), 0\right) & \text{for } \pi \in [0, \underset{\sim}{\pi}) \\ \left(\alpha(\pi), 1 - G\left(\frac{\pi\ell}{1-\pi}\right) - \alpha(\pi)\right) & \text{for } \pi \in [\underset{\sim}{\pi}, \bar{\pi}] \\ (0, \nu(\pi)) & \text{for } \pi \in [\bar{\pi}, 1] \end{cases} \tag{16}$$

where $\alpha(\pi)$ uniquely solves $\Psi(\pi, \alpha, 1 - G\left(\underset{\sim}{\pi}\ell/(1 - \underset{\sim}{\pi})\right) - \alpha) = \underset{\sim}{\pi}$, and $\nu(\pi)$ uniquely solves $t/G^{-1}(1 - \nu) = (1 - \pi)\nu/(1 + (1 - \pi)\nu)$.[22] *On the other hand, if G is degenerate at $b > 0$ and $t < b/(1 + b/\ell)$, σ_U is a weakly decreasing continuous function as well.*

The proposition states that the pass-through correspondence can become a function when losses are concentrated: while the technical condition on G guarantees this property when benefits vary smoothly, this always happens when benefits are concentrated.[23] The logic is again linked to the two margins discussed: substitution effects gain strength when losses become concentrated, while extensive margin effects weaken if benefits vary too smoothly.

[22] $\bar{\pi}$ uniquely solves $\Psi(\bar{\pi}, 0, 1 - G(\underset{\sim}{\pi}\ell/(1 - \underset{\sim}{\pi}))) = \underset{\sim}{\pi}$.

[23] If $t > b/(b/\ell + 1)$, σ_U is a step function, so uniqueness of equilibria in the sharing game is generic.

Interestingly, the pass-through function that arises in the mirror image case features three equilibria akin to those we studied before, but each appearing in different regions of the production domain. Indeed, with losses that are fixed, a sufficiently low prevalence makes unverified sharing more attractive: this is the no-verification outcome in the first segment $[0, \underset{\sim}{\pi})$, where only high-benefit users do unverified sharing (i.e., those $b \geq \pi\ell/(1-\pi)$). After that, only a fraction $\alpha(\pi) \in [0, 1 - G(\underset{\sim}{\pi}\ell/(1-\tilde{\pi}))]$ do unverified sharing, while $1 - G(\underset{\sim}{\pi}\ell/(1-\tilde{\pi})) - \alpha(\pi)$ verify news (which means that the latter users come from the set $b \geq \underset{\sim}{\pi}\ell/(1-\underset{\sim}{\pi})$, as low-benefit users do not share news at all). This pegs prevalence at $\underset{\sim}{\pi}$ as in the equilibrium with indifference in the baseline model; in turn, this sustains verification with a mass of users that ranges from zero at $\tilde{\pi}$ to $1 - G(\underset{\sim}{\pi}\ell/(1-\underset{\sim}{\pi}))$ at $\bar{\pi}$, the point where the set of potential "news-checkers" that are available to maintain this peg is fully utilized.

Beyond $\bar{\pi}$, unverified sharing is strictly dominated and an equilibrium with higher levels of verification emerges (third line in (16)), analogous to the upper branch of our original pass-through correspondence. The last equation for the mass of users who verify news, $\nu(\pi)$, is another way of writing that high-benefit types b find this profitable (i.e., $(1-\psi)b - t > 0$) when nobody does unverified sharing (so $\psi = \Psi(\pi, 0, \nu)$).

6.4 Heterogeneous Benefits and Losses

A model incorporating both heterogeneous benefits and losses is considerably less tractable due to σ_U and σ_V satisfying integral expressions, which can also present discontinuities if the distribution of benefits concentrates around specific points as in the main model. To accommodate bidimensional heterogeneity, Appendix A.12 derives the resulting expressions for $(\sigma_U, \sigma_V, \psi)$ when b and ℓ are independent and the distribution of benefits is non-trivial, exhibiting a single atom of size $p \in [0, 1)$. Our main analytical result is next.

Proposition 11. *Suppose that $p > 0$. In the sharing game, there exist production levels $0 < \pi_1 < \pi_2 < 1$ such that there exists an equilibrium with constant prevalence over the interval $[\pi_1, \pi_2]$. Moreover, $\pi_2 - \pi_1 \to 0$ as $p \to 0$.*

The possibility of clusters around specific levels of benefits guarantees the existence of an equilibrium analogous to our mixed one, as a large mass of users can be split into verified sharing and not sharing at all. Whether this equilibrium—which is easier to establish due to the indifference conditions between the previous two options—ensures the existence two other equilibria (exhibiting more and less verification) is more complicated, due to the fixed-point equation becoming considerably more involved.

But we can explore this question numerically. Concretely, we refer the reader to Figure 6 in Appendix A.12, which plots the two sides of the fixed-point equation on

$$\psi = \frac{\pi(1 + \sigma_U(\psi))}{1 + \sigma_U(\psi) + (1 - \pi)\sigma_V(\psi)}$$

for the case of *exponential distributions* for benefits and losses.[24] (A fixed point on ψ is easier given all the constraints present with a two-dimensional domain.) In line with our intuition, there is a single crossing point if there is no atom, and three crossing points otherwise: because benefits have an unbounded domain, all these equilibria exhibit non-trivial verification. In the same section, we derive a sufficient condition (on endogenous variables, and hence that can always be verified ex-post) such that one can confirm analytically that multiple equilibria can exist in $[\pi_1, \pi_2]$ from Proposition 11 as long as this interval is non-trivial.

7 Concluding Remarks

A common theme in the response of social media platforms to misinformation has been empowering users: facilitating people's ability to determine the veracity of content without taking away their choice to share. As fake news prevalence is actively affected by users' choices, our model explains how prevalence-driven feedback loops can emerge: prevalence shapes fake news diffusion through verification and sharing choices; diffusion in turn affects producers' incentives; and production ultimately feeds back into prevalence jointly with users' choices. Further, as verification is costly, such feedback loops can lead to social-influence effects whereby users who verify news make it more attractive for other to join the "social cause," which can make high levels of fake news production compatible with high welfare. More generally, our model provides a broad picture regarding the set of possibilities for variables such as fake news production, diffusion, and prevalence, which is a first step towards assessing the consequences of misinformation. It also informs how different policies used nowadays may affect these variables.

We have examined these questions through a flexible framework that incorporates both competitive and strategic elements, and that makes it salient why the market for misinformation is different from other traditional markets. There are two natural extensions that can be explored. First, users derive utility from sharing news in our model, and the possibility of sharing when content is fake hurts them. While this is a reasonable first step—users may still fear sharing fake content despite not having full clarity about its consequences—it

[24] Benefits take value $\hat{b}$ with chance p, or are drawn from an exponential distribution with chance $1 - p$.

would be natural to augment the model to incorporate a final action, ideally in the context of a concrete threat to society. Second, it would be interesting to reevaluate the independence assumption between benefits and losses by examining how these variables correlate in practice, and to possibly include more layers of exposure. One could then obtain more precise predictions regarding the variables studied that can also account for the possibility of producers directly targeting populations based on specific observable characteristics.

A Proofs

A.1 Proof of Proposition 1: Sharing game equilibria

The next two lemmata respectively characterize $\overline{\Sigma}(\cdot)$, the solution to equation (7), and $\alpha(\cdot)$, the solution to $\Psi(\pi, \underline{\Sigma}(\underline{\pi}), \alpha) = \underline{\pi}$, where Ψ is given in (1), $\underline{\pi}$ in (3) and $\underline{\Sigma}$ in (5).

Lemma A.1. $\overline{\Sigma} : [0,1] \to [0,1]$ *is well-defined, continuous, and strictly decreasing with* $\overline{\Sigma}(0) = 1$ *and* $\overline{\Sigma}(1) = G(t)$.

Proof: Step 1: Equation (7) has a unique solution. Fix $\pi \in (0,1)$. Notice that function $\sigma_U \mapsto \Psi(\pi, \sigma_U, 1-\sigma_U)$ is strictly increasing, as differentiating (1) yields:

$$\frac{\partial \Psi(\pi, \sigma_U, 1-\sigma_U)}{\partial \sigma_U} = \frac{2(1-\pi)\pi}{(1+\sigma_U+(1-\pi)(1-\sigma_U))^2} > 0.$$

Thus, function $\sigma_U \mapsto G(t/\Psi(\pi, \sigma_U, 1-\sigma_U))$ is strictly decreasing. Moreover, $\Psi(\pi, 1, 0) = \pi$ by (1), and so $G(t/\pi) < 1$. Hence, by the Intermediate Value Theorem (IVT), there is a unique solution on $(0,1)$ to the equation, $G(t/\Psi(\pi, \sigma_U, 1-\sigma_U)) - \sigma_U = 0$. This proves that $\overline{\Sigma}(\pi)$ is well-defined for $\pi \in (0,1)$. Next, take a sequence $\pi_n \to 0$. Clearly, $G(t/\Psi(\pi_n, \sigma_U, 1-\sigma_U)) \to 1$, and thus $\overline{\Sigma}(\pi_n) \to 1$. So we can extend $\overline{\Sigma}$, defining $\overline{\Sigma}(0) := \lim_{\pi_n \to 0} \overline{\Sigma}(\pi_n)$. Conversely, if $\pi = 1$ then $\Psi(\pi, \sigma_U, 1-\sigma_U) \equiv 1$ and so $\overline{\Sigma}(1) = G(t) < 1$. We conclude that $\overline{\Sigma}(\pi)$ is well-defined for all $\pi \in [0,1]$. Finally, since $G(\cdot)$ and $\Psi(\cdot)$ are continuously differentiable in their respective domains, the Implicit Function Theorem ensures the continuity of $\overline{\Sigma}$.
Step 2: $\overline{\Sigma}(\pi)$ is strictly decreasing. Notice that $\Psi(\cdot, \sigma_U, 1-\sigma_U)$ is strictly increasing:

$$\frac{\partial \Psi(\pi, \sigma_U, 1-\sigma_U)}{\partial \pi} = \frac{2(1+\sigma_U)}{(1+\sigma_U+(1-\pi)(1-\sigma_U))^2} > 0$$

Thus, for any $\pi_2 > \pi_1$, we have $G(t/\Psi(\pi_1, \sigma_U, 1-\sigma_U)) - \sigma_U > G(t/\Psi(\pi_2, \sigma_U, 1-\sigma_U)) - \sigma_U$

for all $\sigma_U \in [0,1]$. Evaluating at $\sigma_U = \overline{\Sigma}(\pi_2)$ and using the definition of $\overline{\Sigma}(\cdot)$:

$$G\left(\frac{t}{\Psi(\pi_1, \overline{\Sigma}(\pi_2), 1-\overline{\Sigma}(\pi_2))}\right) - \overline{\Sigma}(\pi_2) > G\left(\frac{t}{\Psi(\pi_1, \overline{\Sigma}(\pi_1), 1-\overline{\Sigma}(\pi_1))}\right) - \overline{\Sigma}(\pi_1)$$

Since the map $\sigma_U \mapsto G(t/\Psi(\pi_1, \sigma_U, 1-\sigma_U)) - \sigma_U$ is strictly decreasing (Step 1), we must have $\overline{\Sigma}(\pi_1) > \overline{\Sigma}(\pi_2)$. This proves the lemma. □

Lemma A.2. *$\alpha : [\underline{\pi}, \bar{\pi}] \to [0, 1-\overline{\Sigma}(\bar{\pi})]$ is well-defined, continuous, and strictly increasing. Also, $\alpha(\underline{\pi}) = 0$, $\alpha(\bar{\pi}) = 1-\overline{\Sigma}(\bar{\pi})$ and $\alpha(\pi) \in (0, 1-\overline{\Sigma}(\pi))$ for $\pi \in (\underline{\pi}, \bar{\pi})$.*

Proof: First, since $\Psi(\pi, \sigma_U, \sigma_V)$ is decreasing in σ_V and, by the definition of $\bar{\pi}$ in Proposition 1—i.e., $\Psi(\bar{\pi}, \overline{\Sigma}(\bar{\pi}), 1-\overline{\Sigma}(\bar{\pi})) = \underline{\pi}$—for each $\pi \in [\underline{\pi}, \bar{\pi}]$ there is a unique $\alpha \in [0, 1-\overline{\Sigma}(\bar{\pi})]$ solving $\Psi(\pi, \overline{\Sigma}(\bar{\pi}), \alpha) = \underline{\pi}$. Also, for $\pi > \underline{\pi}$, we have $\alpha(\pi) > 0$, as $\Psi(\pi, \sigma_U, 0) = \pi$ given (1).

Second, we show that $\alpha(\pi)$ is increasing and continuous. The former holds because $\Psi(\pi, \sigma_U, \sigma_V)$ is increasing in π but decreasing in σ_V; thus, an increase in π must elicit an increase in α to keep prevalence Ψ fixed at $\underline{\pi}$. The latter follows from the continuity of $\Psi(\cdot)$.

Finally, take $\pi \in (\underline{\pi}, \bar{\pi})$ and recall that $\Psi(\pi, \overline{\Sigma}(\pi), 1-\overline{\Sigma}(\pi)) < \underline{\pi}$. Since $\overline{\Sigma}(\pi)$ is decreasing (Lemma A.1), and $\Psi(\pi, \sigma_U, \sigma_V)$ is increasing in σ_U, the next inequality holds:

$$\Psi(\pi, \overline{\Sigma}(\bar{\pi}), 1-\overline{\Sigma}(\pi)) < \Psi(\pi, \overline{\Sigma}(\pi), 1-\overline{\Sigma}(\pi)) < \underline{\pi} = \Psi(\pi, \overline{\Sigma}(\bar{\pi}), \alpha(\pi)).$$

Hence, $\Psi(\pi, \overline{\Sigma}(\bar{\pi}), 1-\overline{\Sigma}(\pi)) < \Psi(\pi, \overline{\Sigma}(\bar{\pi}), \alpha(\pi))$, and thus $\alpha(\pi) < 1-\overline{\Sigma}(\pi)$ since $\Psi(\pi, \sigma_U, \sigma_V)$ is decreasing in σ_V. □

Proof of Proposition 1: We breakdown the equilibrium analysis in three exhaustive cases: $\Psi(\pi, \sigma_U, \sigma_V) > \bar{\pi}$; $\Psi(\pi, \sigma_U, \sigma_V) < \bar{\pi}$; and $\Psi(\pi, \sigma_U, \sigma_V) = \bar{\pi}$.

(i) High prevalence: $\Psi(\pi, \sigma_U, \sigma_V) > \underline{\pi}$. If users anticipate high prevalence, then all strictly prefer not sharing to verified sharing. Thus, in any "high prevalence" equilibrium, $\sigma_V = 0$, and so $\Psi(\pi, \sigma_U, \sigma_V) = \pi$ by (1). Hence, a high prevalence equilibrium emerges, provided $\pi > \underline{\pi}$, with $\underline{\pi}$ defined in (3). The pass-through rate σ_U is given by (4) for prevalence $\Psi(\pi, \sigma_U, \sigma_V) = \pi$. Solving (4) for ℓ and integrating yields $\sigma_U = \underline{\Sigma}(\pi)$, where $\underline{\Sigma}(\cdot)$ is defined in (5). $\underline{\Sigma}(\cdot)$ is decreasing and continuous on $(0,1]$, and can be continuously extended by defining $\underline{\Sigma}(0) := 1 = \lim_{\pi \to 0} \underline{\Sigma}(\pi)$. Altogether, when $\pi > \bar{\pi}$, $\sigma_V = 0$, $\sigma_U = \underline{\Sigma}(\pi)$ is an equilibrium in the sharing game. We next show that this is not the only equilibrium that can arise for $\pi > \bar{\pi}$.

(ii) Low Prevalence: $\Psi(\pi, \sigma_U, \sigma_V) < \underline{\pi}$. If users anticipate low prevalence, then verified sharing is strictly preferred to not sharing. Thus, in any "low prevalence" equilibrium,

$\sigma_V = 1 - \sigma_U$. The pass-through rate σ_U is now determined by (6) for prevalence $\Psi(\pi, \sigma_U, 1 - \sigma_U)$. Integrating over types ℓ for which (6) holds implies that, in equilibrium, σ_U must solve equation (7). Lemma A.1 shows that, for each $\pi \in [0, 1]$, equation (7) has a unique solution, $\sigma_U = \overline{\Sigma}(\pi)$. This solution is strictly monotone.

Thus, the low prevalence equilibrium arises when $\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) < \underline{\pi}$. Since G is strictly increasing, we have $\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) = t/G^{-1}(\overline{\Sigma}(\pi))$ by (7). Therefore, $\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) < \underline{\pi}$ if and only if

$$t/G^{-1}(\overline{\Sigma}(\pi)) < \bar{\pi} \iff \overline{\Sigma}(\pi) > G(t/\underline{\pi}) = \underline{\Sigma}(\underline{\pi}).$$

Since $\overline{\Sigma}(\pi)$ is continuous and strictly decreasing (Lemma A.1) and $\overline{\Sigma}(0) = 1 > G(t/\underline{\pi})$ and $\overline{\Sigma}(1) = G(t) < G(t/\underline{\pi})$, the IVT implies that there exists a unique $\bar{\pi} \in (0, 1)$ such that $\Psi(\pi, \sigma_U, 1 - \sigma_U) < \bar{\pi}$ for all $\pi < \bar{\pi}$, where $\bar{\pi}$ solves:

$$\overline{\Sigma}_U(\bar{\pi}) = G(t/\underline{\pi}) = \underline{\Sigma}(\underline{\pi}).$$

Finally, $\bar{\pi} > \underline{\pi}$ since $\overline{\Sigma}(\underline{\pi}) < 1$ (as $\underline{\pi} > 0$), and so $\Psi(\underline{\pi}, \overline{\Sigma}(\underline{\pi}), 1 - \overline{\Sigma}(\underline{\pi})) < \underline{\pi}$ by (1).

Altogether, for $\pi \in [0, \bar{\pi})$, $\sigma_U = \overline{\Sigma}(\pi)$ and $\sigma_V = 1 - \sigma_U$ constitute an equilibrium in the sharing game. It is the only equilibrium for $\pi \in [0, \underline{\pi})$: if there were another equilibrium $(\tilde{\sigma}_U, \tilde{\sigma}_V)$, then $\tilde{\sigma}_V > 0$ as $\Psi(\pi, \tilde{\sigma}_U, \tilde{\sigma}_V) \leq \pi < \underline{\pi}$ by (1); thus, $\tilde{\sigma}_V = 1 - \tilde{\sigma}_U$ and the same construction would follow.

(iii) Constant prevalence: $\Psi(\pi, \sigma_U, \sigma_V) = \underline{\pi}$. If users conjecture prevalence to be equal to $\underline{\pi}$, then all types are indifferent between verified sharing and not sharing. The pass-through rate then obeys $\sigma_U = \underline{\Sigma}(\underline{\pi}) = \overline{\Sigma}(\bar{\pi})$. For this to be an equilibrium, a mass individuals $\sigma_V \in [0, 1 - \underline{\Sigma}(\underline{\pi})]$ must choose to verify news so that prevalence $\Psi(\pi, \sigma_U, \sigma_V)$ remains equal to $\underline{\pi}$. Notice that this cannot happen in the region $(0, \bar{\pi})$, because $\Psi < \bar{\pi}$ always in that region by (1). Likewise, prevalence cannot remain constant for $\pi \in (\bar{\pi}, 1]$ as σ_V cannot exceed $1 - \overline{\Sigma}(\bar{\pi})$. Hence, this equilibrium arises only if $\pi \in [\underline{\pi}, \bar{\pi}]$. Let $\alpha(\pi)$ be implicitly defined as $\Psi(\pi, \overline{\Sigma}(\bar{\pi}), \alpha(\pi)) \equiv \underline{\pi}$. Lemma A.2 characterizes $\alpha(\cdot)$, showing that it is continuous and strictly increasing. All in all, for $\pi \in [\underline{\pi}, \bar{\pi}]$, $\sigma_U = \underline{\Sigma}(\underline{\pi})$ and $\sigma_V = \alpha(\pi)$ is an equilibrium in the sharing game.

We have shown that there is a unique equilibrium when $\pi \in [0, \underline{\pi}) \cup (\bar{\pi}, 1]$. We now argue that our previous case analyses exhaust all equilibria that can arise when $\pi \in [\underline{\pi}, \bar{\pi}]$. Indeed, suppose that $\tilde{\pi} \in [\underline{\pi}, \bar{\pi}]$ and $(\tilde{\sigma}_U, \tilde{\sigma}_V)$ is an equilibrium in the sharing game, given $\tilde{\pi}$. If $\Psi(\tilde{\pi}, \tilde{\sigma}_U, \tilde{\sigma}_V) < \underline{\pi}$ then all users strictly prefer verified sharing to not sharing; hence, $\tilde{\sigma}_V = 1 -$

$\tilde{\sigma}_U$ and so we are back to case (ii) where we showed that $\tilde{\sigma}_U = \overline{\Sigma}(\tilde{\pi})$. If $\Psi(\tilde{\pi}, \tilde{\sigma}_U, \tilde{\sigma}_V) > \underline{\pi}$ then all users strictly prefer not sharing to verified sharing; thus, $\tilde{\sigma}_V = 0$ and $\Psi(\tilde{\pi}, \tilde{\sigma}_U, \tilde{\sigma}_V) = \tilde{\pi}$. This leads to case (i), where the unique possibility for $\tilde{\sigma}_U$ is $\underline{\Sigma}(\tilde{\pi})$. Finally, if $\Psi(\tilde{\pi}, \tilde{\sigma}_U, \tilde{\sigma}_V) = \underline{\pi}$, then the unique possibility for $\tilde{\sigma}_U$ is $\underline{\Sigma}(\underline{\pi})$, and thus $\tilde{\sigma}_V$ must equal $\alpha(\tilde{\pi})$ by case (iii).

Finally, we show that $\overline{\Sigma}(\pi) > \underline{\Sigma}(\pi)$ for $\pi \in [\underline{\pi}, \bar{\pi})$. To see this, recall that for $\pi < \bar{\pi}$, we have $\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) < \underline{\pi}$. Thus, using (3) and that $\underline{\Sigma}$ is strictly decreasing:

$$\overline{\Sigma}(\pi) = G\left(\frac{t}{\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))}\right) > G\left(\frac{t}{\underline{\pi}}\right) = G\left(\frac{(1-\underline{\pi})b}{\underline{\pi}}\right) = \underline{\Sigma}(\underline{\pi}) > \underline{\Sigma}(\pi).$$

This concludes the proof. □

A.2 Proof of Proposition 2: Verification effects

We will show that there exists a unique pair $(\hat{\sigma}_U, \hat{\pi}) \in (0,1) \times (0, \underline{\pi})$ such that $\hat{\sigma}_U = \overline{\Sigma}(\hat{\pi}) = \underline{\Sigma}(\hat{\pi})$. To this end, we first characterize all pairs (σ_U, π) such that

$$G\left(\frac{t}{\Psi(\pi, \sigma_U, 1 - \sigma_U)}\right) = G\left(\frac{(1-\pi)b}{\pi}\right).$$

Since G is strictly increasing, this reduces to $\Psi(\pi, \sigma_U, 1 - \sigma_U) = \pi t/((1-\pi)b)$. Using (1), we solve for σ_U to get

$$\sigma_U = \frac{b(1-\pi) - t(2-\pi)}{t - (b+t)(1-\pi)} =: K(\pi)$$

Note that $K(\pi)$ is continuous and strictly increasing: $K'(\pi) = 2t^2/(t - (b+t)(1-\pi))^2 > 0$. Moreover, using (3), $K(\underline{\pi}) = 1$ and so $K(\pi) < 1$ for all $\pi < \underline{\pi}$.

Next, we show that curve $\sigma_U = K(\pi)$ intersects curve $\sigma_U = \underline{\Sigma}(\pi)$ uniquely. This is straightforward because $K(0) < 1 = \underline{\Sigma}(0)$ and $K(\underline{\pi}) = 1 > \underline{\Sigma}(\underline{\pi})$. Hence, by the Intermediate Value Theorem, there exists a unique value $\hat{\pi} \in (0, \underline{\pi})$ such that $\underline{\Sigma}(\hat{\pi}) = K(\hat{\pi})$. Now, let $\hat{\sigma}_U = \underline{\Sigma}(\hat{\pi})$. By definition, $\hat{\sigma}_U = K(\hat{\pi})$; thus, $\Psi(\hat{\pi}, \hat{\sigma}_U, 1 - \hat{\sigma}_U) = \hat{\pi}t/((1-\hat{\pi})b)$, and so

$$G\left(\frac{t}{\Psi(\hat{\pi}, \hat{\sigma}_U, 1 - \hat{\sigma}_U)}\right) = G\left(\frac{(1-\hat{\pi})b}{\hat{\pi}}\right) = \hat{\sigma}_U.$$

Hence, $\hat{\sigma}_U = \overline{\Sigma}(\hat{\pi}) = \underline{\Sigma}(\hat{\pi})$.

To conclude, take $\pi < \hat{\pi}$ and $\sigma_U = \underline{\Sigma}(\pi)$. Then, $\sigma_U > K(\pi)$, which is equivalent to $\Psi(\pi, \sigma_U, 1 - \sigma_U) > \pi t/((1-\pi)b)$, as $z \mapsto \Psi(\pi, z, 1-z)$ is strictly increasing. Thus,

$$G\left(\frac{t}{\Psi(\pi, \sigma_U, 1 - \sigma_U)}\right) < \sigma_U = G\left(\frac{(1-\pi)b}{\pi}\right).$$

Therefore, $\overline{\Sigma}(\pi) < \sigma_U = \underline{\Sigma}(\pi)$ for $\pi < \hat{\pi}$. The case $\pi \in (\hat{\pi}, \underline{\pi})$ is analogous. □

A.3 Proof of Corollary 1: Stationary equilibria

Any stationary equilibrium gives rise to a triplet $(\pi, \sigma_U, \sigma_V)$, reflecting all players best responses. That is, $\pi = F(\sigma_U)$ (producers optimize given σ_U) and, by Proposition 1, $(\sigma_U, \sigma_V) \in \{(\underline{\Sigma}(\pi), 0), (\underline{\Sigma}(\underline{\pi}), \alpha(\pi)), (\overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))\}$ (users optimize given π). As for existence, let $\underline{\pi}$ and $\bar{\pi}$ as in Proposition 1, such that $\underline{\Sigma}(\underline{\pi}) = \overline{\Sigma}(\bar{\pi})$. If inverse supply $F^{-1}(\bar{\pi}) \geq \overline{\Sigma}(\bar{\pi})$, then existence is ensured by IVT applied to $F^{-1}(\cdot)$ and $\overline{\Sigma}(\cdot)$ on $[0, \bar{\pi}]$. If $F^{-1}(\bar{\pi}) < \overline{\Sigma}(\bar{\pi})$, then $F^{-1}(\underline{\pi}) < \underline{\Sigma}(\underline{\pi})$, existence follows from IVT applied to $F^{-1}(\cdot)$ and $\underline{\Sigma}(\cdot)$ on $[\underline{\pi}, 1]$. Since $\underline{\pi} < \bar{\pi}$, we have shown that an equilibrium always exists. □

A.4 Proof of Proposition 3: Equilibrium prevalence

The prevalence ψ characterization follows directly from Proposition 1, as $\psi = \Psi(\pi, \sigma_U, \sigma_V)$ where (σ_U, σ_V) is an equilibrium of the sharing game, given π. Let $\overline{\Sigma} : [0, \bar{\pi}] \to [0, 1]$ as defined in Proposition 1. To see why $\pi \mapsto \psi = \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))$ is strictly increasing, recall that, by definition, $\overline{\Sigma} \equiv G(t/\psi)$. Since G is strictly increasing, $\psi = t/G^{-1}(\overline{\Sigma})$. Thus, ψ must be strictly increasing in π since $\overline{\Sigma}$ strictly decreasing in π (Lemma A.1).

Next, by the proof of Proposition 1-(ii), $\psi = \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) < \underline{\pi}$ for all $\pi \in [0, \bar{\pi})$. Meanwhile, by the properties of Ψ (equation (1)), $\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) \leq \pi$ with strict inequality for $\pi \in (0, \bar{\pi})$. Consequently, $\psi \leq \min\{\pi, \underline{\pi}\}$ with strict inequality for $\pi \in (0, \bar{\pi})$.

Finally, the proof of Proposition 1 shows that $\bar{\pi}$ is the unique value of π that solves $\overline{\Sigma}_U(\bar{\pi}) = \underline{\Sigma}(\underline{\pi})$. Therefore, for $\pi = \bar{\pi}$, we have $G((1 - \underline{\pi})b/\underline{\pi}) = G(t/\psi)$ and thus $\psi = \underline{\pi}$ since, by definition, $(1 - \underline{\pi})b = t$, and G is strictly increasing. □

A.5 Proof of Proposition 4: Welfare comparison

In Lemma A.3, we characterize the welfare per news functions in terms of prevalence only. This allows us to show in Lemma A.4 a strong ranking of equilibria in terms of welfare per news. We next show in Lemma A.5 that a similar ranking applies to the amount of circulating news across equilibria. Finally, we use these results to prove Proposition 4.

Lemma A.3. *Welfare per news functions $\mathcal{W}^{ver}(\pi), \mathcal{W}^{mix}(\pi)$, and $\mathcal{W}^{\neg ver}(\pi)$ are given by:*

(i) $\mathcal{W}^{ver}(\pi) = (1 - \psi)b - t + \psi \int_0^{t/\psi} G(\ell)d\ell$, *with* $\psi = \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))$;

(ii) $\mathcal{W}^{mix}(\pi) = \underline{\pi} \int_0^{(1-\underline{\pi})b/\underline{\pi}} G(\ell)d\ell$;

(iii) $\mathcal{W}^{\neg ver}(\pi) = \pi \int_0^{(1-\pi)b/\pi} G(\ell)d\ell$.

Proof-(i): Let $\psi = \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))$. Recall that, in this equilibrium, users with type $\ell \leq t/\psi$ find it optimal to engage in unverified sharing, while the rest verifies before sharing. Hence, the total user welfare is given by:

$$\begin{aligned}\mathcal{W}^{\text{ver}}(\pi) &= \int_0^{t/\psi} [(1-\psi)b - \psi\ell]dG + \int_{t/\psi}^{\infty} [(1-\psi)b - t]dG \\ &= (1-\psi)b - t - \psi \int_0^{t/\psi} \ell dG + tG(t/\psi)\end{aligned}$$

Integrating the middle term by parts yields the desired results.

Proof-(ii): In this equilibrium, prevalence is constant and equal to $\psi = \underline{\pi}$. Hence, $(1-\psi)b - t = 0$ and so total user welfare reflects the welfare of those who engage in unverified sharing, namely, types ℓ for which $(1-\psi)b - \psi\ell \geq 0$. Thus,

$$\mathcal{W}^{\text{mix}}(\pi) = \int_0^{(1-\psi)b/\psi} [(1-\psi)b - \psi\ell]dG = (1-\psi)bG((1-\psi)b/\psi) - \psi \int_0^{(1-\psi)b/\psi} \ell dG.$$

Integrating by parts the last term, using that $\psi = \underline{\pi}$ yields the result.

Proof-(iii): Finally, in the no verification equilibrium, prevalence $\psi = \pi$ and users' welfare coincides with the welfare of those who engage in unverified sharing:

$$\mathcal{W}^{\neg \text{ver}}(\pi) = \int_0^{(1-\pi)b/\pi} [(1-\pi)b - \pi\ell]dG$$

Integrating by parts yields the desired result. □

Lemma A.4. *$\mathcal{W}^{ver}(\pi)$ and $\mathcal{W}^{\neg ver}(\pi)$ are strictly decreasing, while $\mathcal{W}^{mix}(\pi)$ is constant. Moreover,* $\min_{\pi \in [0,\bar{\pi}]} \mathcal{W}^{ver}(\pi) = \mathcal{W}^{mix}(\pi) = \max_{\pi \in [\underline{\pi},1]} \mathcal{W}^{\neg ver}(\pi)$.

Proof: First, $\mathcal{W}^{\text{ver}}(\pi)$ is strictly decreasing. To see this, observe that $0 < z \mapsto (1-z)b - t + z\int_0^{t/z} G(\ell)d\ell$ is strictly decreasing in z, since its derivative $-b + \int_0^{t/z}[G(\ell) - G(t/z)]d\ell < 0$, as G is strictly increasing. Thus, by Lemma A.3-(i), $\mathcal{W}^{\text{ver}}(\pi) = (1-\psi)b - t + \psi\int_0^{t/\psi} G(\ell)d\ell$ must be strictly decreasing in π, since $\psi = \Psi(\pi, \overline{\Sigma}(\pi), 1-\overline{\Sigma}(\pi))$ is strictly increasing (Proposition 3). Next, by Lemma A.3-(iii), for $\pi < 1$:

$$\frac{d\mathcal{W}^{\neg \text{ver}}(\pi)}{d\pi} = \int_0^{\bar{\ell}} G(\ell)d\ell - G(\bar{\ell})\frac{b}{\pi} < \int_0^{\bar{\ell}} G(\ell)d\ell - G(\bar{\ell})\bar{\ell} = \int_0^{\bar{\ell}} [G(\ell) - G(\bar{\ell})]d\ell < 0,$$

where $\bar{\ell} = (1-\pi)b/\pi < b/\pi$. Also, $\mathcal{W}^{\text{mix}}(\pi)$ is clearly constant by Lemma A.3-(ii).

Finally, notice that $\mathcal{W}^{\text{ver}}(\bar{\pi}) = \mathcal{W}^{\neg\text{ver}}(\underline{\pi})$, since $\psi = \underline{\pi}$ and so $t/\psi = t/\underline{\pi} = (1-\underline{\pi})b/\underline{\pi}$. Moreover, $\mathcal{W}^{\text{mix}}(\pi) = \mathcal{W}^{\neg\text{ver}}(\underline{\pi})$. By the monotonicity properties of $\mathcal{W}^{\text{ver}}, \mathcal{W}^{\text{mix}}$, and $\mathcal{W}^{\neg\text{ver}}$, it easily follows that $\min_{\pi\in[0,\bar{\pi}]} \mathcal{W}^{\text{ver}}(\pi) = \mathcal{W}^{\text{mix}}(\pi) = \max_{\pi\in[\underline{\pi},1]} \mathcal{W}^{\neg\text{ver}}(\pi)$. □

Let $\tau^{\text{ver}}(\pi), \tau^{\text{mix}}(\pi)$, and $\tau^{\neg\text{ver}}(\pi)$ denote the total news volume in the verification, mixed, and no-verification equilibria of the sharing game, for fixed π whenever they exist.

Lemma A.5. *$\tau^{ver}(\pi)$ and $\tau^{\neg ver}(\pi)$ are strictly decreasing, while $\tau^{mix}(\pi)$ is strictly increasing. Moreover, $\min_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{ver}(\pi) = \max_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{mix}(\pi)$ and $\min_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{mix} = \max_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{\neg ver}(\pi)$.*

Proof: First, consider the verification equilibrium. Then, $\tau^{\text{ver}}(\pi) = 1-\pi(1-\overline{\Sigma}(\pi))$. Since $1-\overline{\Sigma}(\cdot)$ is strictly increasing (Proposition 1), $\tau^{ver}(\pi)$ must be strictly decreasing in π. Similarly, $\tau^{\neg\text{ver}}(\pi) = 1+\underline{\Sigma}(\pi)$ and $\underline{\Sigma}(\cdot)$ is strictly decreasing (Proposition 1). In the mixed equilibrium, $\tau^{\text{mix}}(\pi) = 1 + \underline{\Sigma}(\underline{\pi}) + (1-\pi)\alpha(\pi)$, where α is the unique solution to $\Psi(\pi, \underline{\Sigma}(\underline{\pi}), \alpha) = \underline{\pi}$ (see Proposition 1). Let us show that $(1-\pi)\alpha(\pi)$ is strictly increasing. To this end, use (1) to solve $\alpha(\pi)$ in closed form:

$$\alpha(\pi) = \frac{(\pi-\underline{\pi})(1+\underline{\Sigma}(\underline{\pi}))}{\underline{\pi}(1-\pi)} \implies (1-\pi)\alpha(\pi) = \frac{(\pi-\underline{\pi})(1+\underline{\Sigma}(\underline{\pi}))}{\underline{\pi}}$$

Notice that $(1-\pi)\alpha(\pi)$ increases linearly in π, implying the same holds for $\tau^{\text{mix}}(\pi)$.

Next, by the aforementioned monotonicity properties, $\min_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{\text{ver}}(\pi) = \tau^{\text{ver}}(\bar{\pi}) = 1 + \overline{\Sigma}(\bar{\pi}) + (1-\bar{\pi})(1-\overline{\Sigma}(\bar{\pi}))$. Also, $\max_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{\text{mix}}(\pi) = \tau^{\text{mix}}(\bar{\pi}) = 1 + \underline{\Sigma}(\underline{\pi}) + (1-\bar{\pi})\alpha(\bar{\pi})$. By Proposition 1 , $\underline{\Sigma}(\underline{\pi}) = \overline{\Sigma}(\bar{\pi})$ and $\alpha(\bar{\pi}) = 1-\overline{\Sigma}(\bar{\pi})$, and thus $\tau^{\text{mix}}(\bar{\pi}) = \tau^{\text{ver}}(\bar{\pi})$. Finally, $\min_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{\text{mix}}(\pi) = \tau^{\text{mix}}(\underline{\pi}) = 1 + \underline{\Sigma}(\underline{\pi})$, while $\max_{\pi\in[\underline{\pi},\bar{\pi}]} \tau^{\neg\text{ver}}(\pi) = \tau^{\neg\text{ver}}(\underline{\pi}) = 1 + \underline{\Sigma}(\underline{\pi})$. □

Proof of Proposition 4: Suppose multiple stationary equilibria exist, and let $\pi^*_{\text{ver}}, \pi^*_{\text{mix}}$, and $\pi^*_{\neg\text{ver}}$ denote the equilibrium production of fake news in the verification, mixed, and no-verification stationary equilibria. By Proposition 1, we know that all these levels must belong to $[\underline{\pi}, \bar{\pi}]$. Moreover, since the supply curve is strictly increasing, and $\overline{\Sigma}(\cdot) > \underline{\Sigma}(\cdot)$ on $[\underline{\pi}, \bar{\pi}]$, it follows that $\pi^*_{\text{ver}} \geq \pi^*_{\text{mix}} \geq \pi^*_{\neg\text{ver}}$ (with at least one strict inequality). Hence, Lemma A.4 implies that $\mathcal{W}^{\text{ver}}(\pi^*_{\text{ver}}) \geq \mathcal{W}^{\text{mix}}(\pi^*_{\text{mix}}) \geq \mathcal{W}^{\neg\text{ver}}(\pi^*_{\neg\text{ver}})$ (with at least one strict inequality). Moreover, Lemma A.5 implies $\tau^{\text{ver}}(\pi^*_{\text{ver}}) \geq \tau^{\text{mix}}(\pi^*_{\text{mix}}) \geq \tau^{\neg\text{ver}}(\pi^*_{\neg\text{ver}})$ (with at least one strict inequality). Thus, $\mathcal{W}^{\text{ver}}(\pi^*_{\text{ver}})\tau^{\text{ver}}(\pi^*_{\text{ver}}) \geq \mathcal{W}^{\text{mix}}(\pi^*_{\text{mix}})\tau^{\text{mix}}(\pi^*_{\text{mix}}) \geq \mathcal{W}^{\neg\text{ver}}(\pi^*_{\neg\text{ver}})\tau^{\neg\text{ver}}(\pi^*_{\neg\text{ver}})$.

Finally, the last claim in the proposition holds, since Lemma A.4 and Lemma A.5 imply that $\mathcal{W}^{\text{ver}}(\pi)\tau^{\text{ver}}(\pi)$ and $\mathcal{W}^{\neg\text{ver}}(\pi)\tau^{\neg\text{ver}}(\pi)$ are each strictly decreasing in π. □

A.6 Proof of Proposition 5: Planner's solution and efficiency

We begin with some preliminary observations. First, as explained in the main text, it is straightforward to see that $\sigma_U = \underline{\Sigma}(\pi)$ solves problem (12); therefore, $\mathcal{W}_P^{\neg\text{ver}}(\pi) = \mathcal{W}^{\neg\text{ver}}(\pi)$. Next, we also know that, for any $\pi \leq \bar{\pi}$, the equilibrium sharing $\sigma_U = \overline{\Sigma}(\pi)$ is feasible in the planner's problem (see Proposition 3). Thus, $\mathcal{W}_P^{\text{ver}}(\pi) \geq \mathcal{W}^{\text{ver}}(\pi)$ for all $\pi \leq \bar{\pi}$. Finally, by Proposition 4, we obtain $\mathcal{W}_P^{\text{ver}}(\pi) \geq \mathcal{W}^{\text{ver}}(\pi) \geq \mathcal{W}^{\neg\text{ver}}(\pi) = \mathcal{W}_P^{\neg\text{ver}}$ for all $\pi \in [\underline{\pi}, \bar{\pi}]$. In the next lemma, we show that this ranking extends to $\pi \in (0, \underline{\pi})$.

Lemma A.6. *$\mathcal{W}^{ver}(\pi) > \mathcal{W}^{\neg ver}(\pi)$ for all $\pi \in (0, \underline{\pi})$.*

Proof: Fix $\pi \in (0, \underline{\pi})$ and let $\psi = \Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi))$. Then,

$$\begin{aligned}
\mathcal{W}^{\neg\text{ver}}(\pi) &= \int_0^{(1-\pi)b/\pi} [(1-\pi)b - \pi\ell] dG + \int_{(1-\pi)b/\pi}^{\infty} 0 dG \\
&< \int_0^{(1-\pi)b/\pi} [(1-\pi)b - \pi\ell] dG + \int_{(1-\pi)b/\pi}^{\infty} [(1-\pi)b - t] dG \\
&< \int_0^{(1-\psi)b/\pi} [(1-\psi)b - \psi\ell] dG + \int_{(1-\pi)b/\pi}^{\infty} [(1-\psi)b - t] dG \\
&< \int_0^{t/\psi} [(1-\psi)b - \psi\ell] dG + \int_{t/\psi}^{\infty} [(1-\psi)b - t] dG = \mathcal{W}^{\text{ver}}(\pi).
\end{aligned}$$

The first inequality holds since $\pi < \bar{\pi}$, and thus $(1-\pi)b - t > 0$, given (3). The second one holds because $\psi < \pi$ by Proposition 3. The third inequality obtains since $(1-\psi)b - \psi\ell \geq (1-\psi)b - t$ for all $\ell \leq t/\psi$. □

Proof of Proposition 5: First, using the observations above and Lemma A.6, we conclude that $\mathcal{W}_P^{\text{ver}}(\pi) \geq \mathcal{W}^{\text{ver}}(\pi) > \mathcal{W}^{\neg\text{ver}}(\pi) = \mathcal{W}_P^{\neg\text{ver}}$ for all $\pi \in (0, \bar{\pi}]$.

Second, we show that $\mathcal{W}_P^{\neg\text{ver}}(\pi) > \mathcal{W}_P^{\text{ver}}(\pi)$ at $\pi = \bar{\pi}_P$. To see this, recall that, by definition of $\bar{\pi}_P$, equation $\Psi(\bar{\pi}_P, \sigma_U, 1 - \sigma_U) = \underline{\pi}$ is uniquely solved by $\sigma_U = 0$. Thus, $\mathcal{W}_P^{\text{ver}}(\bar{\pi}_P) = 0$. On the other hand, $\mathcal{W}_P^{\neg\text{ver}}(\bar{\pi}_P) = \mathcal{W}^{\neg\text{ver}}(\bar{\pi}_P) > 0$, since $\bar{\pi}_P < 1$. Let us define $\bar{\pi}_v := \inf\{\pi > 0 : \mathcal{W}_P^{\text{ver}}(\pi) = \mathcal{W}_P^{\neg\text{ver}}(\pi)\}$. By continuity of $\mathcal{W}_P^{\text{ver}}$ and $\mathcal{W}_P^{\neg\text{ver}}$, this cutoff exists by the Intermediate Value Theorem. Thus, $\mathcal{W}_P^{\text{ver}}(\pi) \geq \mathcal{W}_P^{\neg\text{ver}}(\pi)$ for all $\pi \leq \bar{\pi}_v$.

Third, we'll show that $1 - \sigma_{U,P}^{\text{ver}}(\pi) > 1 - \overline{\Sigma}(\pi)$ in $[0, \bar{\pi}]$. This trivially holds if $\sigma_{U,P}^{\text{ver}}(\pi) = 0$. If $\sigma_{U,P}^{\text{ver}}(\pi) > 0$ then it must satisfy the first order condition:

$$t - \Psi(\pi, \sigma_{U,P}^{\text{ver}}, 1 - \sigma_{U,P}^{\text{ver}}) G^{-1}(\sigma_{U,P}^{\text{ver}}) \geq \frac{\partial \Psi}{\partial \sigma_U}\Big|_{\sigma_U = \sigma_{U,P}^{\text{ver}}} \times \left(b + \int_0^{G^{-1}(\sigma_{U,P}^{\text{ver}})} \ell dG \right),$$

with equality if $\Psi(\pi, \sigma^{\text{ver}}_{U,P}, 1-\sigma^{\text{ver}}_{U,P}) < \underline{\pi}$. Since $\Psi(\pi, \sigma_U, 1-\sigma_U)$ is strictly increasing in σ_U, we must have $t - \Psi(\pi, \sigma^{\text{ver}}_{U,P}, 1-\sigma^{\text{ver}}_{U,P})G^{-1}(\sigma^{\text{ver}}_{U,P}) > 0$. Also, since $\overline{\Sigma}(\pi)$ solves (7) (Proposition 1), we have $t = \Psi(\pi, \overline{\Sigma}, 1-\overline{\Sigma})G^{-1}(\overline{\Sigma})$. Thus,

$$\Psi(\pi, \sigma^{\text{ver}}_{U,P}, 1-\sigma^{\text{ver}}_{U,P})G^{-1}(\sigma^{\text{ver}}_{U,P}) < \Psi(\pi, \overline{\Sigma}, 1-\overline{\Sigma})G^{-1}(\overline{\Sigma}).$$

But then $\sigma^{\text{ver}}_{U,P} < \overline{\Sigma}$, since the mapping $\sigma_U \mapsto \Psi(\pi, \sigma_U, 1-\sigma_U)G^{-1}(\sigma_U)$ is strictly increasing. Altogether, $1 - \sigma^{\text{ver}}_{U,P}(\pi) > 1 - \overline{\Sigma}(\pi)$.

Finally, $1 - \sigma^{\text{ver}}_{U,P}(\pi) > 0$ in $[\bar{\pi}, \bar{\pi}_v]$. Otherwise, $\mathcal{W}^{\text{ver}}_P(\pi) < \mathcal{W}^{\neg\text{ver}}_P(\pi)$, which contradicts the definition of $\bar{\pi}_v$. Part (ii) of the proposition is proved in the main text. □

A.7 Proof of Proposition 6: Lowering verification costs

Proof-(i.1): Consider the verification equilibrium. Given Proposition 1, $\sigma_U = \overline{\Sigma}(\pi; t)$ is the unique solution to $\sigma_U = G(t/\Psi(\pi, \sigma_U, 1-\sigma_U))$. Since $G(\cdot)$ is strictly increasing, we have $G(t/\Psi(\pi, \sigma_U, 1-\sigma_U)) - \sigma_U > G(t'/\Psi(\pi, \sigma_U, 1-\sigma_U)) - \sigma_U$ for all $\sigma_U \in [0, 1]$. Evaluating at $\sigma_U = \overline{\Sigma}(\pi; t')$ and using the definition of $\overline{\Sigma}(\pi; t)$:

$$G\left(\frac{t}{\Psi(\pi, \overline{\Sigma}(\pi;t'), 1-\overline{\Sigma}(\pi;t'))}\right) - \overline{\Sigma}(\pi;t') > G\left(\frac{t}{\Psi(\pi, \overline{\Sigma}(\pi;t), 1-\overline{\Sigma}(\pi;t))}\right) - \overline{\Sigma}(\pi;t).$$

As shown in the proof of Proposition 1 (Step 1), the mapping $\sigma_U \mapsto G(t/\Psi(\pi_1, \sigma_U, 1-\sigma_U)) - \sigma_U$ is strictly decreasing; thus, $\overline{\Sigma}(\pi; t') < \overline{\Sigma}(\pi; t)$. This proves that, in the verification equilibrium, the unverified sharing rate σ_U falls if verification cost t falls.

The effect on prevalence is direct as $\sigma_U \mapsto \Psi(\pi, \sigma_U, 1-\sigma_U)$ is strictly increasing, and $\psi(t) = \Psi(\pi, \overline{\Sigma}(\pi;t), 1-\overline{\Sigma}(\pi;t))$. Since $\overline{\Sigma}(\pi;t)$ falls in t, prevalence ψ must also fall in t. Finally, by Lemma A.3-(i), welfare can be written as $\mathcal{W}^{\text{ver}}(z;t) = (1-z)b - t + z\int_0^{t/z} G(\ell)d\ell$ for $z > 0$. Since $\mathcal{W}^{\text{ver}}(z;t)$ is decreasing in z (Proposition 4) and is also clearly decreasing in t, it follows that $\mathcal{W}^{\text{ver}}(\psi(t);t)$ must be decreasing in t. □

Proof-(i.2): In the mixed equilibrium, $\psi = \underline{\pi}(t) = 1 - t/b$ (Proposition 3). Hence, equilibrium prevalence ψ rises as verification cost t falls. The effect on σ_U is direct since $\sigma_U = \underline{\Sigma}(\underline{\pi}(t))$, and $\underline{\Sigma}(\cdot)$ is decreasing (Proposition 1). As for σ_V, note that since σ_U falls and $\underline{\pi}$ rises, σ_V must fall to keep $\Psi(\pi, \sigma_U, \sigma_V) = \underline{\pi}$, given (1). Finally, by Lemma A.3, welfare is given by $\mathcal{W}^{\text{mix}}(\pi) = \mathcal{W}^{\neg\text{ver}}(\underline{\pi})$. Since $\mathcal{W}^{\neg\text{ver}}(\pi)$ is decreasing in π (Proposition 4), $\mathcal{W}^{\neg\text{ver}}(\underline{\pi}(t))$ must rise in t. Hence, welfare in the mixed equilibrium must also fall as t falls. □

Proof-(i.3): This is direct as $\sigma_U = \underline{\Sigma}(\pi)$, where $\underline{\Sigma}(\cdot)$ is given by (5) and is independent of t. Also, by Proposition 3, in a no-verification equilibrium, $\psi = \pi$. Since prevalence ψ is

unaffected by t in this equilibrium, so is welfare $\mathcal{W}^{\neg\text{ver}}(\pi) = \pi \int_0^{(1-\pi)b/\pi} G(\ell)d\ell$ (Lemma A.3-(iii)). This concludes part-(i) of the lemma. □

Proof-(ii.1): Let us define $\Delta : [0,1]^2 \to [0,1]$ as

$$\Delta(\pi, \sigma_U) := \frac{\pi(1-\pi)(1-\sigma_U)}{1+\pi+\sigma_U(1-\pi)},$$

and let $\gamma(t) := \bar{\pi}(t) - \underline{\pi}(t)$, where $\underline{\pi}, \bar{\pi}$ are characterized in Proposition 1. We'll show that there exists $t^* \in (0, b)$ such that for all $t \geq t^*$, we have $\partial\gamma/\partial t < 0$. First, recall that $\bar{\pi}$ is the unique solution to $\Psi(\bar{\pi}, \overline{\Sigma}(\bar{\pi}), 1 - \overline{\Sigma}(\bar{\pi})) = \underline{\pi}$ (Proposition 1). Using (1) and that $\overline{\Sigma}(\bar{\pi}) = \underline{\Sigma}(\underline{\pi})$, we obtain

$$\bar{\pi} = \frac{2\underline{\pi}}{1+\underline{\pi}+(1-\underline{\pi})\underline{\Sigma}(\underline{\pi})}. \tag{17}$$

Next, using (3), (5), and (17), we can express γ as $\gamma(t) = \Delta(\underline{\pi}(t), \underline{\Sigma}(\underline{\pi}(t)))$, where $\underline{\Sigma}(\underline{\pi}(t)) = G(t/\underline{\pi}(t))$. Thus,

$$\frac{\partial\gamma}{\partial t} = \frac{\partial\Delta}{\partial\pi} \times \frac{\partial\underline{\pi}}{\partial t} + \frac{\partial\Delta}{\partial\sigma_U} \times \frac{\partial\underline{\Sigma}}{\partial t}.$$

Notice that $\partial\underline{\pi}/\partial t < 0$, as $\underline{\pi} = 1 - t/b$. Similarly, $\partial\underline{\Sigma}(\underline{\pi})/\partial t > 0$, since $\underline{\Sigma}(\cdot)$ is strictly decreasing. Also, $\Delta(\pi, \sigma_U)$ is clearly decreasing in σ_U for each π. The effect of π on Δ is non-monotone: $\Delta(\pi, \sigma_U)$ is strictly concave in π for each σ_U, since:

$$\frac{\partial\Delta}{\partial\pi} = \frac{(1-\sigma_U)(1+\sigma_U(1-\pi)^2-\pi^2-2\pi)}{(1+\pi+\sigma_U(1-\pi))^2} \qquad \text{and} \qquad \frac{\partial^2\Delta}{\partial\pi^2} = \frac{-4(1-\sigma_U^2)}{(1+\pi+\sigma_U(1-\pi))^3} < 0$$

Moreover, $\partial\Delta/\partial\pi = 0$ if and only if (σ_U, π) solves the first-order condition:

$$\sigma_U = \frac{\pi(2+\pi)-1}{(1-\pi)^2}.$$

Define $\pi^*(t) \in (0,1)$ as the value of π that uniquely solves:

$$\underbrace{G(t/\pi)}_{LHS} = \underbrace{\frac{\pi(2+\pi)-1}{(1-\pi)^2}}_{RHS}$$

Notice that $\pi^*(t)$ is well-defined by the Intermediate Value Theorem: LHS is strictly decreasing in π, while RHS is strictly increasing in it.[25] Moreover, $\pi^*(t)$ is increasing in t, since t raises LHS but leaves RHS unaffected. Now, define t^* as the unique solution to $\underline{\pi}(t) = \pi^*(t)$. Again, this expression is well-defined by IVT since $\underline{\pi}(t)$ is strictly decreasing

[25] Moreover, for $RHS < LHS$ for $\pi = 0$ and $RHS > LHS$ for π close enough to 1.

in t, while $\pi^*(t)$ is increasing in it. Further, $\underline{\pi}(0) = 1 > \pi^*(0)$ and $\underline{\pi}(b) = 0 < \pi^*(b)$.

Finally, take $t \geq t^*$. Then, $\underline{\pi}(t) \leq \pi^*(t)$ and thus

$$\underline{\Sigma}(\underline{\pi}(t)) = G(t/\underline{\pi}(t)) \geq \frac{\underline{\pi}(t)(2 + \underline{\pi}(t)) - 1}{(1 - \underline{\pi}(t))^2}$$

This implies that $\partial\Delta(\underline{\pi}(t), \underline{\Sigma}(\underline{\pi}(t)))/\partial\pi \geq 0$. Consequently,

$$\frac{\partial\gamma}{\partial t} = \underbrace{\frac{\partial\Delta}{\partial\bar{\pi}}}_{\geq 0} \times \underbrace{\frac{\partial\bar{\pi}}{\partial t}}_{<0} + \underbrace{\frac{\partial\Delta}{\partial\bar{\sigma}_U}}_{<0} \times \underbrace{\frac{\partial\bar{\sigma}_U}{\partial t}}_{>0} < 0$$

This completes the proof. □

Proof-(ii.2): Fix $t < b$. First, note that $\Delta(\pi, \sigma_U)$ is continuously differentiable, as its partial derivatives $\partial\Delta/\partial\pi$ and $\partial\Delta/\partial\sigma_U$ both exist for all $(\pi, \sigma_U) \in [0,1]^2$ and are continuous. Next, consider $\bar{\pi}(t) = 1 - t/b$ and $\underline{\Sigma}(\underline{(t)}) = G(t/\bar{\pi}(t))$, and observe that both functions are continuously differentiable in t. Thus, the composition $t \mapsto \gamma(t) \equiv \Delta(\underline{\pi}(t), \underline{\Sigma}_U(\underline{(t)}))$ is continuously differentiable. Let us compute $\gamma'(t)$ and evaluate the resulting expression at $t = 0$, leveraging that $\underline{\pi}(0) = 1$ and $\underline{\Sigma}(1) = 0$:

$$\gamma'(0) = \frac{\partial\Delta(1,0)}{\partial\pi} \times \bar{\pi}'(0) + \frac{\partial\Delta(1,0)}{\partial\sigma_U} \times \underline{\Sigma}'(1) = \frac{1}{2b} > 0$$

Since γ' is continuous, there exists $\underline{t} > 0$ such that $\gamma'(t) > 0$ for all $t \leq \underline{t}$. □

A.8 Proof of Proposition 7: Algorithmic filters

Prevalence computation: To find the misinformation prevalence $\Psi^\phi(\pi, \sigma_U, \sigma_V)$, consider production π and sharing rates (σ_U, σ_V). Notice that the mass of newly produced news items that pass the filter is $1 - \phi\pi \leq 1$. This mass of news is allocated to a unit mass of users at random. This means that a fraction σ_U of these news items is shared without verification, while a fraction σ_V is first verified and then shared if truthful. By Bayes' rule, the chance of the latter event is:

$$\frac{1 - \pi}{1 - \phi\pi}.$$

Thus, the total number of shared news is

$$(1 - \phi\pi) \times \sigma_U + (1 - \phi\pi) \times \frac{1 - \pi}{1 - \phi\pi} \times \sigma_V = (1 - \phi\pi)\sigma_U + (1 - \pi)\sigma_V,$$

while total number of fake news is:

$$(1-\phi)\pi + (1-\phi\pi) \times \frac{(1-\phi)\pi}{1-\phi\pi} \times \sigma_U = (1-\phi)\pi(1+\sigma_U).$$

Thus, the misinformation prevalence $\Psi^{\phi}(\pi, \sigma_U, \sigma_V)$ (ratio of fake news to total) is (13).

Proof-(i.1): Consider the verification equilibrium and $\pi \in (0,1)$. First, given (14) and Proposition 1, $\sigma_U = \overline{\Sigma}(\zeta(\pi;\phi))$. Since $\zeta(\pi;\cdot)$ is strictly decreasing, an increase in ϕ raises σ_U, as $\overline{\Sigma}(\cdot)$ is strictly decreasing. Next, equilibrium prevalence $\psi(\phi)$ must decrease in ϕ, since by (7): $\psi(\phi) = t/G^{-1}(\overline{\Sigma}(\zeta(\pi;\phi)))$ and $\overline{\Sigma}(\zeta(\pi;\phi))$ rises in ϕ. Finally, welfare increases in ϕ because prevalence ψ falls in ϕ and $\mathcal{W}^{\text{ver}}$ falls in ψ (Proposition 4). □

Proof-(i.2): To sustain the mixed equilibrium, prevalence $\psi = 1 - t/b$ (Proposition 3) so that users are indifferent between verified sharing and not sharing. Hence, equilibrium prevalence ψ is constant in ϕ, and so are σ_U and welfare $\mathcal{W}^{\text{mix}}$. As for σ_V, observe that since σ_U is unaffected by ϕ, σ_V must fall in ϕ to keep $\Psi(\zeta(\pi;\phi), \sigma_U, \sigma_V) = 1 - t/b$. □

Proof-(i.3): Consider the no-verification equilibrium and $\pi \in (0,1)$. By (14), prevalence ψ must equal to $\zeta(\pi;\phi)$; thus, $\sigma_U = \underline{\Sigma}(\zeta(\pi;\phi))$. As in the proof of (i.1), σ_U rises in ϕ because ζ falls in ϕ and $\underline{\Sigma}$ is decreasing. Prevalence $\psi = \zeta(\pi;\phi)$ clearly falls, and thus welfare rises as $\mathcal{W}^{\neg\text{ver}}$ is decreasing in prevalence (Proposition 4). □

Proof-(ii): Let $\phi < 1$. Recall that $\bar{\pi}(\phi) \in (0,1)$ is determined by $\zeta(\bar{\pi}(\phi);\phi) = \bar{\pi}(0)$, where $\zeta(\pi;\phi)$ is given by (13) and $\bar{\pi}(0) \in (0,1)$ is the π value that solves $\overline{\Sigma}(\pi) = \underline{\Sigma}(1 - t/b)$. Hence,

$$\frac{\partial \bar{\pi}(\phi)}{\partial \phi} = -\frac{\zeta_\phi(\bar{\pi}(\phi);\phi)}{\zeta_\pi(\bar{\pi}(\phi);\phi)} = \frac{(1-\bar{\pi}(\phi))\bar{\pi}(\phi)}{1-\phi} > 0,$$

where $\zeta_\phi \equiv \partial\zeta/\partial\phi$ and $\zeta_\pi \equiv \partial\zeta/\partial\pi$. Similarly, $\underline{\pi}(\phi) \in (0,1)$ is given by $\zeta(\underline{\pi}(\phi);\phi) = \underline{\pi}(0)$, where $\underline{\pi}(0) = 1 - t/b \in (0,1)$. Thus,

$$\frac{\partial \underline{\pi}(\phi)}{\partial \phi} = -\frac{\zeta_\phi(\underline{\pi}(\phi);\phi)}{\zeta_\pi(\underline{\pi}(\phi);\phi)} = \frac{(1-\underline{\pi}(\phi))\underline{\pi}(\phi)}{1-\phi} > 0.$$

Now, suppose that $t < b/2$. Since $\underline{\pi}(\phi)$ and $\bar{\pi}(\phi)$ are strictly increasing, with $\bar{\pi}(\phi) > \underline{\pi}(\phi)$, it follows that $\bar{\pi}(\phi) > \underline{\pi}(\phi) > \underline{\pi}(0) \geq 1/2$ for all $\phi \in (0,1)$. Thus, wedge $\bar{\pi}(\phi) - \underline{\pi}(\phi)$ must be strictly decreasing, since:

$$\frac{\partial \hat{\pi}(\phi)}{\partial \phi} - \frac{\partial \bar{\pi}(\phi)}{\partial \phi} = \frac{(1-\bar{\pi}(\phi))\bar{\pi}(\phi) - (1-\underline{\pi}(\phi))\underline{\pi}(\phi)}{1-\phi} < 0,$$

where the inequality holds because the mapping $z \mapsto (1-z)z$ is decreasing for $z \geq 1/2$. □

Proof-(iii): As shown in (i.1) and (i.3), an increase in ϕ shifts up the respective curves $\sigma_U = \overline{\Sigma}(\zeta(\pi;\phi))$ and $\sigma_U = \underline{\Sigma}(\zeta(\pi;\phi))$, while the curve $\pi = F((1-\phi)\sigma_U)$ shifts left. Thus, the new stationary equilibrium in the respective verification and no verification branches unambiguously exhibits a higher pass-through rate. The effect on prevalence is immediate because prevalence and pass-through are negatively related. In the verification branch, $\sigma_U = G(t/\psi)$, while in the no verification one, $\sigma_U = G((1-\psi)b/\psi)$. Hence, welfare $\mathcal{W}^{\text{ver}}$ and $\mathcal{W}^{\neg\text{ver}}$ must increase, as these functions are decreasing in prevalence.

The effect on equilibrium production π is, in general, ambiguous. If the pass-through curve shifts more than supply does, then an increase in ϕ leads to an increase in production of fake content. Otherwise, equilibrium production falls. Ultimately, this depends on the elasticities of the pass-through curves $\sigma_U = \overline{\Sigma}(\zeta(\pi;\phi))$ and $\sigma_U = \underline{\Sigma}(\zeta(\pi;\phi))$. To fix ideas, consider the no-verification branch (the analysis for the verification branch is analogous). In a stationary equilibrium, (π, σ_U) must solve:

$$\sigma_U = \underline{\Sigma}(\zeta(\pi;\phi)) \qquad \text{and} \qquad \pi = F((1-\phi)\sigma_U).$$

Since F is strictly increasing, this system can be rewritten as:

$$(1-\phi)\underline{\Sigma}(\zeta(\pi;\phi)) = F^{-1}(\pi).$$

Let $\pi' := d\pi/d\phi$. Implicitly differentiating the above equality with respect to ϕ, we get:

$$\pi' = \frac{-\underline{\Sigma} + (1-\phi)\underline{\Sigma}'\partial\zeta/\partial\phi}{[F^{-1}]' - (1-\phi)\underline{\Sigma}'\partial\zeta/\partial\pi}$$

Since $[F^{-1}]' > 0 > \underline{\Sigma}'$ and $\partial\zeta/\partial\pi > 0$, the sign of π' is fully determined by the sign of the numerator. Therefore, using (14), easy algebra shows that $\pi' > 0$ if and only

$$\left|\frac{\zeta\underline{\Sigma}'(\zeta)}{\underline{\Sigma}(\zeta)}\right| \times (1-\zeta) > 1.$$

Finally, in the mixed equilibrium, the pass-through curve is inelastic at $\sigma_U = \underline{\Sigma}(1-t/b)$. Thus, the production of fake content $\pi = F[(1-\phi)\underline{\Sigma}(1-t/b)]$ falls as ϕ rises. However, prevalence must remain constant at $1-t/b$ to sustain the equilibrium, leaving welfare $\mathcal{W}^{\text{mix}}$ unaffected. This concludes the proof. □

A.9 Proof of Proposition 8: Certifying verified content

Proof-(i.1): First, notice that $\Psi^{\beta}(\pi, \sigma_U, \sigma_V) \equiv \Psi(\pi, \sigma_U, (1-\beta)\sigma_V)$. Thus, an increase in β increases Ψ^{β} (ceteris paribus), given (1). Hence, $\sigma_U = \overline{\Sigma}(\pi; \beta)$ must decrease in β because σ_U solves $\sigma_U = G(t/\Psi(\pi, \sigma_U, (1-\beta)(1-\sigma_U)))$ and $(\sigma_U, \beta) \mapsto G(t/\Psi(\pi, \sigma_U, (1-\beta)(1-\sigma_U)))$ is decreasing in σ_U and decreasing in β. Next, prevalence ψ must increase since $\psi = t/G^{-1}(\sigma_U)$ and σ_U decreases with β. Finally, welfare must decrease as $\mathcal{W}^{\text{ver}}$ is decreasing in ψ. □

Proof-(i.2): By the same reasons given in the proof of Proposition 7-(i.2), equilibrium prevalence ψ, pass-through σ_U, and welfare are all unaffected by β. That said, σ_V must rise in β to keep prevalence $\Psi(\pi, \sigma_U, (1-\beta)\sigma_V) = 1 - t/b$. □

Proof-(i.3): This is straightforward because in the no-verification equilibrium, $\sigma_V = 0$ and thus prevalence $\Psi^{\beta} \equiv \pi$ for all $\beta \in [0, 1]$. Since prevalence is unaffected by β, it follows that σ_U and welfare are unaffected by it. □

Proof-(ii): First, the no-verification equilibrium emerges when prevalence $\Psi^{\beta} = \pi \geq 1 - t/b$; hence $\underline{\pi} = 1 - t/b$. Second, the verification-equilibrium can be sustained, provided $\Psi^{\beta}(\pi, \sigma_U, 1-\sigma_U) \leq \underline{\pi}$ with $\sigma_U = \overline{\Sigma}(\pi; \beta)$. As in the proof of Proposition 1, this condition reduces to $\overline{\Sigma}(\pi; \beta) \geq \underline{\Sigma}(\underline{\pi})$. Since $\overline{\Sigma}(\pi; \beta)$ is decreasing in π, this condition is satisfied when $\pi \leq \bar{\pi}(\beta)$, where $\bar{\pi}(\beta)$ is the π value that solves $\overline{\Sigma}(\pi; \beta) = \underline{\Sigma}(\underline{\pi})$. Since $\overline{\Sigma}(\pi; \beta)$ decreases in β, cutoff $\bar{\pi}(\beta)$ must also decrease in β to keep $\overline{\Sigma}(\bar{\pi}(\beta); \beta)$ constant. As a result, wedge $\bar{\pi}(\beta) - \underline{\pi}$ is decreasing in β.

Now, as $\beta \to 1$, $\Psi^{\beta} \to \pi$ and thus by continuity of G and Ψ, $\overline{\Sigma}(\pi; \beta) \to G(t/\pi)$. Let $\bar{\pi}_1 = \lim_{\beta \to 1} \bar{\pi}(\beta)$. We'll show that $\bar{\pi}_1 = \underline{\pi}$. To see this, notice that by continuity of $\overline{\Sigma}$, we have $\overline{\Sigma}(\bar{\pi}_1; 0) = \underline{\Sigma}(\underline{\pi})$. But, $\overline{\Sigma}(\bar{\pi}_1; 0) = G(t/\bar{\pi}_1)$ and $\underline{\Sigma}(\underline{\pi}) = G((1-\underline{\pi})b/\underline{\pi})$. Moreover, by definition of $\underline{\pi}$, $(1-\underline{\pi})b = t$, and thus $\bar{\pi}_1 = \underline{\pi}$, as G is strictly increasing.

Finally, for $\beta = 1$, it is easy to see that for $\pi \leq \underline{\pi}$, $\mathcal{W}_P^{\text{ver}}$ in (11) obeys $\mathcal{W}_P^{\text{ver}}(\pi) = (1-\pi)b - t(1 - G(t/\pi)) - \pi \int_0^{t/\pi} \ell dG = \mathcal{W}^{\text{ver}}(\pi)$. On the other hand, $\mathcal{W}_P^{\neg\text{ver}}$ in (12) is given by $\mathcal{W}_P^{\neg\text{ver}}(\pi) = (1-\pi)bG((1-\pi)b/\pi) - \pi \int_0^{(1-\pi)b/\pi} \ell dG = \mathcal{W}^{\neg\text{ver}}(\pi)$. Thus, by Lemma A.6, $\mathcal{W}_P^{\text{ver}}(\pi) > \mathcal{W}_P^{\neg\text{ver}}(\pi)$ for $\pi \in (0, \underline{\pi})$. Hence, the equilibrium solves the planner's problem. □

A.10 Proof of Proposition 9: Pass-through correspondence with bounded losses

Case 1: Let $t < b/(1 + b/\bar{\ell})$ so that $\underset{\sim}{\pi} < \underline{\pi}$. We identify four cases to analyze.

(i) Let $\pi \in [0, \underset{\sim}{\pi}]$. Then, by (1), it follows that $\Psi(\pi, \sigma_U, \sigma_V) \leq \underset{\sim}{\pi} < \underline{\pi}$. Thus, all types prefer verified sharing to not sharing; hence $\sigma_U + \sigma_V = 1$. Likewise, since $\Psi(\pi, \sigma_U, \sigma_V) \leq \underset{\sim}{\pi}$,

all types $\ell < \bar{\ell}$ strictly prefer unverified sharing to verified sharing. Since G is atomless, we have $\sigma_V = 0$. Consequently, $\Psi(\pi, \sigma_U, \sigma_V) = \pi$ and $\sigma_U = 1$.

(ii) Let $\pi \in (\underset{\sim}{\pi}, \underline{\pi})$. By (1), we have $\Psi(\pi, \sigma_U, \sigma_V) < \underline{\pi}$, and so again all types prefer verified sharing to not sharing: $\sigma_V = 1 - \sigma_U$. If $\Psi(\pi, \sigma_U, \sigma_V) \leq \underset{\sim}{\pi}$, then $\sigma_V = 0$ and $\sigma_U = 1$, but then $\Psi(\pi, \sigma_U, \sigma_V) = \Psi(\pi, 1, 0) = \pi > \underset{\sim}{\pi}$, a contradiction. Thus, $\Psi(\pi, \sigma_U, \sigma_V) \in (\underset{\sim}{\pi}, \underline{\pi})$ and so σ_U solving fixed-point equation (7), i.e., $\sigma_U = \overline{\Sigma}(\pi) \in (0, 1)$. For this to be an equilibrium, we must verify that $\Psi > \underset{\sim}{\pi}$. Notice that $\overline{\Sigma}_U(\pi) < G(t/\underset{\sim}{\pi}) = 1$, implying that prevalence $\Psi(\pi, \overline{\Sigma}(\pi), 1 - \overline{\Sigma}(\pi)) = t/G^{-1}(\overline{\Sigma}(\pi)) > \underset{\sim}{\pi}$. Moreover, $\Psi(\pi, \overline{\Sigma}, 1 - \overline{\Sigma}) = \underset{\sim}{\pi}$ if and only if $\overline{\Sigma}(\pi) = 1$, which happens only at $\pi = \underset{\sim}{\pi}$, by (1).

(iii) Let $\pi = \underline{\pi}$. Then, in any equilibrium, $\Psi(\underline{\pi}, \sigma_U, \sigma_V) > \underset{\sim}{\pi}$ (same logic as in case (ii)). Next, we have two possible cases to consider. If $\Psi(\underline{\pi}, \sigma_U, \sigma_V) = \underline{\pi}$, then $\sigma_V = 0$ by (1). Hence, $\sigma_U = \underline{\Sigma}(\pi)$, where $\underline{\Sigma}$ is given in (5). Conversely, if $\Psi(\underline{\pi}, \sigma_U, \sigma_V) < \underline{\pi}$, then all types prefer verified sharing to not sharing. Hence, $\sigma_U = \overline{\Sigma}(\pi)$ solves (7).

(iv) Finally, let $\pi > \underline{\pi}$. Again, in any equilibrium, $\Psi(\underline{\pi}, \sigma_U, \sigma_V) > \underset{\sim}{\pi}$ (same logic as in case (ii)). As in case (iii), we have two possible subcases: $\Psi(\pi, \sigma_U, \sigma_V) = \underline{\pi}$ and $\Psi(\pi, \sigma_U, \sigma_V) < \underline{\pi}$. The analysis of these subcases is identical to the case in which the support of ℓ is unbounded. There is multiplicity of equilibria for $\pi \in (\underline{\pi}, \bar{\pi}]$ where $\overline{\Sigma}(\bar{\pi}) = \underline{\Sigma}(\underline{\pi})$. For $\pi > \bar{\pi}$, there is a unique equilibrium: $\sigma_U = \underline{\Sigma}(\pi)$ and $\sigma_V = 0$.

Case 2: Let $t \geq b/(1 + b/\bar{\ell})$. In this case, $\underset{\sim}{\pi} \geq \underline{\pi}$. We identify three relevant cases to analyze.

(i) Let $\pi \in [0, \underline{\pi})$. Then, by (1), $\Psi(\pi, \sigma_U, \sigma_V) \leq \pi < \underline{\pi}$, and so all types prefer verified sharing to not sharing. Also, since $\underline{\pi} \leq \underset{\sim}{\pi}$, $\Psi(\pi, \sigma_U, \sigma_V) \leq \pi < \underset{\sim}{\pi}$; thus, all types prefer unverified sharing to verified sharing. Altogether, $\sigma_U = 1$, $\sigma_V = 0$, and $\Psi = \pi$.

(ii) Let $\pi \in [\underline{\pi}, \underline{\pi})$. Then, by (1), $\Psi(\pi, \sigma_U, \sigma_V) \leq \pi < \underset{\sim}{\pi}$, implying that all types prefer unverified sharing to verified sharing. Hence, $\Psi(\pi, \sigma_U, \sigma_V) = \pi \geq \underline{\pi}$, implying that all types prefer not sharing to verified sharing. Thus, the relevant trade off is between unverified sharing and not sharing, resulting in $\sigma_U = \underline{\Sigma}(\pi)$ with Σ given by (5).

(iii) Let $\pi \geq \underset{\sim}{\pi}$. We analyze two possibilities. 1) If $\sigma_V = 0$ then $\Psi(\pi, \sigma_U, \sigma_V) = \pi > \underset{\sim}{\pi} \geq \underline{\pi}$, and so all types prefer not sharing to verified sharing, leading to $\sigma_V = 0$. In this equilibrium, $\sigma_U = \underline{\Sigma}(\pi)$. 2) If $\sigma_V > 0$ then $\Psi(\pi, \sigma_U, \sigma_V) < \pi$. If $\Psi(\pi, \sigma_U, \sigma_V) \leq \underset{\sim}{\pi}$, then almost all types prefer unverified sharing to verified sharing, hence $\sigma_V = 0$ which contradicts $\sigma_V > 0$. Conversely, if $\Psi(\pi, \sigma_U, \sigma_V) > \underset{\sim}{\pi} \geq \bar{\pi}$, then all types prefer not sharing to verified sharing, hence $\sigma_V = 0$, which again contradicts $\sigma_V > 0$. Altogether, when $\pi \geq \underset{\sim}{\pi}$, there is a unique equilibrium: $\sigma_V = 0$ and $\sigma_U = \underline{\Sigma}(\pi)$.

This concludes the proof. □

A.11 Proof of Proposition 10: Pass-through function with concentrated losses

Case 1: G is continuously differentiable with $\sup_{z\geq 0} zg(z) \leq 1$.

(i) HIGH PREVALENCE: $\Psi > t/\ell$. In this case, no user finds it optimal to engage in unverified sharing, since $(1-\Psi)b-\Psi\ell < (1-\Psi)b-t$. Thus, $\sigma_U = 0$ and so prevalence (1) turns to:

$$\Psi(\pi, \sigma_U, \sigma_V) = \frac{\pi}{1+(1-\pi)\sigma_V}.$$

The user population splits between those who verify and those who choose not to share. Given prevalence $\psi < 1$, the mass of users engaged in verified sharing equals $\sigma_V = \Sigma_V(\psi)$, where

$$\Sigma_V(\psi) := 1 - G\left(\frac{t}{1-\psi}\right).$$

Let us define $\Sigma_V(1) := \lim_{\psi\to 1}\Sigma_V(\psi) = 1$ so that Σ_V is well-defined on $[0,1]$.

A high prevalence equilibrium can be sustained iff $\Psi(\pi, 0, \Sigma_V(\psi)) = \psi > t/\ell$, or:

$$\psi = \frac{\pi}{1+(1-\pi)\Sigma_V(\psi)} \qquad \text{and} \qquad \psi > t/\ell$$

Notice that $\Sigma_V : [0,1] \to [0,1)$ is continuous and strictly decreasing. Thus, the right hand side of the fixed-point equation above $Q(\psi,\pi) := \Psi(\pi, 0, \Sigma_V(\psi))$ is continuous on $[0,1]^2$ and strictly increasing in ψ and in π. Moreover, $Q(\psi,\pi) \leq \pi$ and $Q(0,\pi) \geq 0$. Thus, by the Intermediate Value Theorem, there exists $\psi^* \in [0,1]$ such that $Q(\psi^*,\pi) = \psi^*$. Next, we show that this fixed point is unique. To avoid trivialities, let $\pi \in (0,1)$ so that $\psi^* \in (0,1)$. We'll show that if ψ^* solves the fixed-point equation, then $\partial Q(\psi^*,\pi)/\partial\psi < 1$. Using the expression for Σ_V and that ψ^* is a fixed-point:

$$\frac{\partial Q(\psi^*,\pi)}{\partial\psi} = \left(\frac{\psi^*}{1-\psi^*}\right)^2 \times \frac{1-\pi}{\pi} t \times g\left(\frac{t}{1-\psi^*}\right)$$

Since $\psi^* < \pi$ as $\Sigma_V(\psi^*) > 0$, it follows that,

$$\frac{\partial Q(\psi^*,\pi)}{\partial\psi} < \left(\frac{\psi^*}{1-\psi^*}\right) tg\left(\frac{t}{1-\psi^*}\right) < \left(\frac{t}{1-\psi^*}\right) g\left(\frac{t}{1-\psi^*}\right) \leq \sup_{z\geq 0} zg(z) \leq 1$$

Thus, $\partial Q(\psi^*,\pi)/\partial\psi < 1$, as desired.

Finally, we show that $\psi^* > t/\ell$ provided π is high enough. To see this, note that $Q : (0,1)^2 \to (0,1)$ is continuously differentiable, and $\partial Q(\psi^*, \pi)/\partial \psi < 1$ for (ψ^*, π) satisfying $Q(\psi^*, \pi) = \psi^*$. Thus, by the Implicit Function Theorem, we can express ψ^* as continuous function of $\pi \in (0,1) \mapsto \psi^* \in (0,1)$ (slightly abusing notation). That said, since $\partial Q(\psi^*, \pi)/\partial \psi < 1$, fixed-point $\psi^*(\pi)$ must strictly increase in π. Moreover, $\lim_{\pi \to 1} \psi^*(\pi) = 1 > t/\ell$ and $\lim_{\pi \to 0} \psi^*(\pi) = 0$. Hence, by the Intermediate Value Theorem, there exists a unique value $\bar{\pi} \in (0,1)$ such that $\psi^*(\bar{\pi}) = t/\ell$. In other words, $\bar{\pi}$ satisfies $\Psi(\bar{\pi}, 0, \Sigma_V(t/\ell)) = t/\ell$. Thus, $\psi^* > t/\ell$ if and only if $\pi > \bar{\pi}$. To conclude, notice that $\bar{\pi} > t/\ell$, since $\Psi(\bar{\pi}, 0, \Sigma_V(t/\ell)) < \bar{\pi}$ as $\Sigma_V(t/\ell) > 0$.

We have found that for $\pi \in (\bar{\pi}, 1)$, there exists a unique pair $(\sigma_V, \psi) \in (0,1)^2$ that solves $\psi = \Psi(\pi, 0, \sigma_V) > t/\ell$ and $\sigma_V = \Sigma_V(\psi)$. Letting $\psi = \Sigma_V^{-1}(\sigma_V)$, we must have that σ_V is the unique value ν that solves $\Sigma_V^{-1}(\sigma_V) = \Psi(\pi, 0, \sigma_V)$, namely:

$$1 - \frac{t}{G^{-1}(1-\nu)} = \frac{\pi}{1 + (1-\pi)\nu}.$$

Moreover, by continuity of G, $\nu \to 0$ as $\pi \to 1$.

(ii) Low prevalence: $\Psi < t/\ell$. In this case, nobody engages in verified sharing, as $(1-\Psi)b - \Psi\ell > (1-\Psi)b - t$. Thus, $\sigma_V = 0$ and so prevalence $\Psi(\pi, \sigma_U, \sigma_V) = \pi$. The population splits between those who engage in unverified sharing and those who choose not to share. Thus, $\sigma_U = \widetilde{\Sigma}_U(\pi)$, where:

$$\widetilde{\Sigma}_U(\pi) = 1 - G\left(\frac{\pi\ell}{1-\pi}\right)$$

Notice that $\widetilde{\Sigma}_U : [0,1) \to (0,1]$ is continuous and decreasing in π. Also, this equilibrium can be sustained if $\pi < t/\ell =: \underset{\sim}{\pi}$.

(iii) Constant prevalence: $\Psi = t/\ell$. In this case, users are indifferent between verified sharing and unverified sharing. Since prevalence is constant, the mass of users who find it optimal to share equals $\tilde{\sigma}_U := \widetilde{\Sigma}_U(t/\ell)$. Suppose α_U of those users break the indifference in favor of unverified (and thus $\bar{\sigma}_U - \alpha_U$ breaks it in favor of verified sharing). Then, an equilibrium with constant prevalence can be sustained as long as $\Psi(\pi, \alpha_U, \tilde{\sigma}_U - \alpha_U) = t/\ell$, or:

$$\frac{(1+\alpha_U)\pi}{1 + \alpha_U + (1-\pi)(\tilde{\sigma}_U - \alpha_U)} = t/\ell,$$

where $0 \leq \alpha_U \leq \tilde{\sigma}_U$.

Since $\Psi(\pi, \alpha_U, \tilde{\sigma}_U - \alpha_U)$ is continuous strictly increasing in α_U, and $\alpha_U = \Sigma_U(t/\ell)$ when $\pi = t/\ell$; and $\alpha_U = 0$ when $\pi = \bar{\pi}$, it follows by IVT that for each $\pi \in (t/\ell, \bar{\pi})$ there exists a unique α_U that solves $\Psi(\pi, \alpha_U, \tilde{\sigma}_U - \alpha_U) = t/\ell$. □

Case 2: G is degenerate at $b > 0$. In this case, notice that a user prefers unverified sharing to not sharing if $(1 - \Psi)b - \Psi\ell \geq 0$, or $\Psi \leq b/(b + \ell) \in (0, 1)$. That said, we examine two cases, depending on verification cost t.

1. Low verification cost: $t/\ell < b/(b + \ell)$. Here, the parameters satisfy:

$$t/\ell < b/(b + \ell) < 1 - t/b.$$

We identify five types of equilibria:

(i) No sharing equilibrium: $\Psi > 1 - t/b$. Here, both verified and unverified sharing are strictly dominated by not sharing at all; hence, $\sigma_U = \sigma_V = 0$ and $\Psi(\pi, \sigma_U, \sigma_V) = \pi$. This equilibrium can be sustained, provided $\pi > 1 - t/b$.

(ii) No verification equilibrium: $\Psi < t/\ell$. Verified sharing is dominated by unverified sharing; thus, $\sigma_V = 0$ and $\Psi \equiv \pi$ by (1). Also, $\Psi < t/\ell < b/(b+\ell)$ and so unverified sharing strictly dominates not sharing $\sigma_U = 1$. This equilibrium can be sustained if $\pi < t/\ell$.

(iii) Full verification equilibrium: $\Psi \in (t/\ell, 1 - t/b)$. Here, verified sharing strictly dominates both unverified sharing and not sharing. Thus, $\sigma_U = 0$ and $\sigma_V = 1$. Hence, $\Psi(\pi, \sigma_U, \sigma_V) \equiv \pi/(2 - \pi)$ by (1). Let $\underline{\pi} := 2t/(t + \ell)$ and $\bar{\pi} := 2(b-t)/(2b-t)$ so that $\Psi(\underline{\pi}, 0, 1) = t/\ell$ and $\Psi(\bar{\pi}, 0, 1) = 1-t/b$. As $\Psi(\cdot, 0, 1)$ is strictly increasing, this equilibrium can be sustained provided $\pi \in (\underline{\pi}, \bar{\pi})$.

(iv) Mixed equilibrium with full sharing: $\Psi = t/\ell$. Any type of sharing strictly dominates not sharing; thus, $\sigma_U + \sigma_V = 1$. Also, all users are indifferent between verified and unverified sharing. This equilibrium can be sustained provided $\pi \in [t/\ell, \underline{\pi}]$ and $\Psi(\pi, \sigma_U, 1 - \sigma_U) = t/\ell$, where $\underline{\pi}$ is defined in case (iii).

(v) Mixed equilibrium with verified sharing only: $\Psi = 1 - t/b$. Here, unverified sharing is strictly dominated, $\sigma_U = 0$, while all users are indifferent between verified sharing and not sharing. Thus, by (1), prevalence $\Psi(\pi, \sigma_U, \sigma_V) \equiv \pi/(1 + (1 - \pi)\sigma_V)$, with $\sigma_V \in [0, 1]$. This equilibrium can be sustained provided $\pi \in [1 - t/b, \bar{\pi}]$ and $\Psi(\pi, 0, \sigma_V) = 1 - t/b$, where $\bar{\pi}$ is defined in case (iii). □

2. HIGH VERIFICATION COST: $t/\ell \geq b/(b+\ell)$. Suppose the inequality is strict. Then, the parameters satisfy:

$$1 - t/b < b/(b+\ell) < t/\ell.$$

Thus, an equilibrium with verification cannot arise because it would require $\Psi \leq 1-t/b$ (i.e., verified sharing is preferred to not sharing) and $\Psi \geq t/\ell$ (i.e., verified sharing is preferred to unverified sharing), which cannot happen given the ranking above. Thus, in any equilibrium, $\sigma_V = 0$ and thus $\Psi(\pi, \sigma_U, \sigma_V) \equiv \pi$, given (1). Thus, $\sigma_U = 1$ for $\pi < b/(b+\ell)$; $\sigma_U \in [0,1]$ for $\pi = b/(b+\ell)$; and $\sigma_U = 0$ otherwise.

Finally, if $t/\ell = b/(b+\ell)$ then $1-t/b = b/(b+\ell)$. Hence, an equilibrium with verification needs $\Psi = t/\ell$ which, by our previous logic, can be sustained for $\pi \in [t/\ell, \underline{\pi}]$ where $\Psi(\underline{\pi}, 0, 1) = t/\ell$. Otherwise, $\sigma_V = 0$, and $\sigma_U = 1$ if $\pi < t/\ell$, and $\sigma_U = 0$ if $\pi > t/\ell$. □

A.12 Heterogeneous Benefits and Losses

Outline. To find equilibria in the sharing game, it is convenient to work on the "prevalence space," treating prevalence ψ as the main equilibrium variable. That is, we characterize optimal sharing behavior, given ψ, and then we require that, in equilibrium, prevalence ψ is consistent with production π and sharing rates σ_U, σ_V. To this end, we first introduce the heterogeneity in benefits and losses, allowing for a mass point in the benefit distribution. Second, we derive the optimal sharing rates (unverified and verified, respectively) given prevalence $\psi \in (0,1)$. Third, we introduce the fixed-point equation that prevalence must solve. We then show in Proposition A.1 that an equilibrium with constant prevalence can be sustained, provided the benefit distribution has a non-trivial mass point. Next, in Proposition A.2, we provide conditions under which multiplicity of equilibria can emerge in this general setting. Finally, we illustrate our findings using exponential distributions.

Suppose that with probability $p \in (0,1)$, benefit $b = \hat{b} > t$; and with probability $(1-p)$, benefit b is drawn from an atomless, and continuously differentiable cdf H supported on $[0, \infty)$. Also, suppose loss ℓ is drawn from an atomless, and continuously differentiable cdf G supported on $[0, \infty)$. Finally, assume that b and ℓ are independent.

Given prevalence $\psi \in (0,1)$, user with type (b, ℓ) finds it optimal to engage in unverified sharing if and only if the following conditions hold:

$$(1-\psi)b - \psi\ell \geq 0 \qquad \text{and} \qquad (1-\psi)b - \psi\ell \geq (1-\psi)b - t,$$

i.e., $\ell \leq \min\{(1-\psi)b/\psi, t/\psi\}$. So, the mass of users engaged in unverified sharing is

$\sigma_U = \Sigma_U(\psi)$, where:

$$\Sigma_U(\psi) = (1-p)\int_0^\infty G\left(\min\left\{\frac{t}{\psi}, \frac{(1-\psi)b}{\psi}\right\}\right) dH(b) + pG\left(\min\left\{\frac{t}{\psi}, \frac{(1-\psi)\hat{b}}{\psi}\right\}\right) \quad (18)$$

Likewise, user (b, ℓ) finds it optimal to engage in verified sharing if and only if:

$$(1-\psi)b - t \geq 0 \qquad \text{and} \qquad (1-\psi)b - t \geq (1-\psi)b - \Psi\ell,$$

namely $b \geq t/(1-\psi)$ and $\ell \geq t/\psi$. Let $\hat{\psi} := 1 - t/\hat{b}$ and define functions

$$\underline{\Sigma}_V(\psi) := (1-p)\left(1 - H\left(\frac{t}{1-\psi}\right)\right)\left[1 - G\left(\frac{t}{\psi}\right)\right] \quad (19)$$

$$\overline{\Sigma}_V(\psi) := \left[(1-p)\left(1 - H\left(\frac{t}{1-\psi}\right)\right) + p\right]\left[1 - G\left(\frac{t}{\psi}\right)\right] \quad (20)$$

The mass of users engaged in verified sharing is given by a correspondence $\sigma_V \in \Sigma_V(\psi)$:

$$\Sigma_V(\psi) = \begin{cases} \overline{\Sigma}_V(\psi) & \text{for } \psi \in (0, \hat{\psi}) \\ [\underline{\sigma}_V, \overline{\sigma}_V] & \text{for } \psi = \hat{\psi} \\ \underline{\Sigma}_V(\psi) & \text{for } \psi \in (\hat{\psi}, 1) \end{cases}$$

where $\underline{\sigma}_V := \underline{\Sigma}_V(\hat{\psi})$ and $\overline{\sigma}_V := \overline{\Sigma}_V(\hat{\psi})$.

A triplet $(\sigma_U, \sigma_V, \psi)$ is an equilibrium in the sharing game, given π, if $\sigma_U = \Sigma_U(\psi)$, $\sigma_V \in \Sigma_V(\psi)$, and ψ satisfies:

$$\Psi(\pi, \sigma_U, \sigma_V) = \psi,$$

where $\Psi(\pi, \sigma_U, \sigma_V)$ is given by (1). Proposition 11 can be restated as follows.

Proposition A.1. *There exists $0 < \pi_1 < \pi_2 < 1$ solving $\Psi(\pi_1, \Sigma_U(\hat{\psi}), \underline{\Sigma}_V(\hat{\psi})) = \hat{\psi}$ and $\Psi(\pi_2, \Sigma_U(\hat{\psi}), \overline{\Sigma}_V(\hat{\psi})) = \hat{\psi}$, respectively, such that for each $\pi \in [\pi_1, \pi_2]$, there exists an equilibrium in the sharing game in which prevalence equals $\hat{\psi} = 1 - t/\hat{b}$. Moreover, $\pi_2 - \pi_1 \to 0$ as the size of the atom $p \to 0$.*

Proof: Consider $\epsilon_1, \epsilon_2 > 0$ small. Since $\Psi(\cdot, \sigma_U, \sigma_V)$ is strictly increasing and continuous, with $\Psi(0, \sigma_U, \sigma_V) = 0$ and $\Psi(1, \sigma_U, \sigma_V) = 1$, the Intermediate Value Theorem ensures the existence and uniqueness of production thresholds $\pi_1^{\epsilon_1}, \pi_2^{\epsilon_2} \in (0, 1)$ such that

$$\begin{aligned} \Psi(\pi_1^{\epsilon_1}, \Sigma_U(\hat{\psi} + \epsilon_1), \Sigma_V(\hat{\psi} + \epsilon_1)) &= \hat{\psi} + \epsilon_1 \\ \Psi(\pi_2^{\epsilon_2}, \Sigma_U(\hat{\psi} - \epsilon_2), \Sigma_V(\hat{\psi} - \epsilon_2)) &= \hat{\psi} - \epsilon_2 \end{aligned}$$

Since $\Sigma_V(\hat{\psi}+\epsilon_1) = \underline{\Sigma}_V(\hat{\psi}+\epsilon_1)$ and $\Sigma_V(\hat{\psi}-\epsilon_2) = \overline{\Sigma}_V(\hat{\psi}-\epsilon_2)$, the above system turns to:

$$\begin{aligned}\Psi(\pi_1^{\epsilon_1}, \Sigma_U(\hat{\psi}+\epsilon_1), \underline{\Sigma}_V(\hat{\psi}+\epsilon_1)) &= \hat{\psi}+\epsilon_1 \\ \Psi(\pi_2^{\epsilon_2}, \Sigma_U(\hat{\psi}-\epsilon_2), \overline{\Sigma}_V(\hat{\psi}-\epsilon_2)) &= \hat{\psi}-\epsilon_2\end{aligned}$$

Take $\epsilon_1, \epsilon_2 \to 0$. Since $\Psi, \Sigma_U, \underline{\Sigma}_V, \overline{\Sigma}_V$ are continuous, we have $\pi_2^{\epsilon_2} \to \pi_2$ and $\pi_1^{\epsilon_1} \to \pi_1$, where

$$\begin{aligned}\Psi(\pi_1, \Sigma_U(\hat{\psi}), \underline{\Sigma}_V(\hat{\psi})) &= \hat{\psi} \\ \Psi(\pi_2, \Sigma_U(\hat{\psi}), \overline{\Sigma}_V(\hat{\psi})) &= \hat{\psi}\end{aligned}$$

Since $\Psi(\pi, \sigma_U, \cdot)$ is strictly decreasing and $\overline{\Sigma}_V(\hat{\psi}) > \underline{\Sigma}_V(\hat{\psi})$, we obtain $\pi_2 > \pi_1$. Moreover, $(\Sigma_U(\hat{\psi}), \underline{\Sigma}_V(\hat{\psi}), \hat{\psi})$ is an equilibrium of the sharing game, given π_1, while $(\Sigma_U(\hat{\psi}), \overline{\Sigma}_V(\hat{\psi}), \hat{\psi})$ is an equilibrium, given π_2.

Now, let $\alpha_V \in [\underline{\Sigma}_V(\hat{\psi}), \overline{\Sigma}_V(\hat{\psi})]$. Because $\Psi(\cdot, \sigma_U, \sigma_V)$ is strictly increasing, we can find a unique $\tilde{\pi}(\alpha_V) \in (\pi_1, \pi_2)$ such that $\Psi(\tilde{\pi}(\alpha_V), \Sigma_U(\hat{\psi}), \alpha_V) = \hat{\psi}$. Moreover, $\tilde{\pi}(\alpha_V)$ must be strictly increasing in α_V in order to keep prevalence at $\hat{\psi}$. Thus, we can define $\tilde{\alpha}_V := \tilde{\pi}^{-1} : [\pi_1, \pi_2] \to [\underline{\Sigma}_V(\hat{\psi}), \overline{\Sigma}_V(\hat{\psi})]$ such that, for each $\pi \in [\pi_1, \pi_2]$ there exists a unique value $\alpha_V = \tilde{\alpha}_V(\pi)$ such that

$$\Psi(\pi, \Sigma_U(\hat{\psi}), \alpha_V) = \hat{\psi}.$$

That is, $(\Sigma_U(\hat{\psi}), \tilde{\alpha}_V(\pi), \hat{\psi})$ is an equilibrium of the sharing game, given π. This proves that an equilibrium with prevalence equal to $\hat{\psi}$ exists for each $\pi \in [\pi_1, \pi_2]$. Intuitively, in this equilibrium, types $b = \hat{b}$ are indifferent between verified sharing and not sharing, with some breaking the indifference in favor of verified sharing as π rises from π_1.

To conclude the proof, let $p \to 0$. Then, by continuity, $\overline{\Sigma}_V(\hat{\psi}) \to \underline{\Sigma}_V(\hat{\psi})$, and so

$$\hat{\psi} = \lim_{p \to 0} \Psi(\pi_2, \Sigma_U(\hat{\psi}), \overline{\Sigma}_V(\hat{\psi})) = \Psi(\pi_2, \Sigma_U(\hat{\psi}), \underline{\Sigma}_V(\hat{\psi}))$$

Thus, $\Psi(\pi_2, \Sigma_U(\hat{\psi}), \underline{\Sigma}_V(\hat{\psi})) = \Psi(\pi_1, \Sigma_U(\hat{\psi}), \underline{\Sigma}_V(\hat{\psi}))$ and so $\pi_2 = \pi_1$, since prevalence function $\Psi(\cdot, \sigma_U, \sigma_V)$ is strictly increasing. □

Proposition A.2. *Consider π_1 and π_2 from Proposition A.1. Suppose that function $\psi \mapsto \Psi(\pi_2, \Sigma_U(\psi), \overline{\Sigma}_V(\psi)) - \psi$ or $\psi \mapsto \Psi(\pi_1, \Sigma_U(\psi), \underline{\Sigma}_V(\psi)) - \psi$ is strictly decreasing near $\hat{\psi}$. Then, there exists a production region $\mathcal{R}$ such that for each $\pi \in \mathcal{R}$, the sharing game has multiple equilibria.*

Proof: Let $\overline{\Upsilon}(\pi, \psi) := \Psi(\pi, \Sigma_U(\psi), \overline{\Sigma}_V(\psi)) - \psi$. Suppose that $\overline{\Upsilon}(\pi_2, \cdot)$ is strictly decreasing near $\hat{\psi}$ (the other case is analogous and thus omitted). By Proposition A.1, $(\pi_2, \hat{\psi})$ satisfies

$\overline{\Upsilon}(\pi_2, \hat{\psi}) = 0$. Since $\overline{\Upsilon}(\pi, \psi)$ is continuously differentiable, as Ψ, G, H are continuously differentiable functions on their respective domains, the Implicit Function Theorem ensures the existence of open sets $U \ni \pi_2$ and $V \ni \hat{\psi}$ and a continuously differentiable function $\varsigma : U \to V$ such that $\overline{\Upsilon}(\pi, \varsigma(\pi)) = 0$ for $\pi \in U$. Moreover,

$$\varsigma'(\pi_2) = \frac{-\partial\overline{\Upsilon}(\pi_2, \hat{\psi})/\partial\pi}{\partial\overline{\Upsilon}(\pi_2, \hat{\psi})/\partial\psi} > 0,$$

where the inequality holds since $\partial\overline{\Upsilon}(\pi_2, \hat{\psi})/\partial\pi > 0 > \partial\overline{\Upsilon}(\pi_2, \hat{\psi})/\partial\psi$. Thus, ς is strictly increasing near π_2. Therefore, for each $\pi^* \in (\pi_1, \pi_2) \cap U$, $\psi^* = \varsigma(\pi^*) < \hat{\psi}$ and $\Upsilon(\pi^*, \psi^*) = 0$, i.e., $\Psi(\pi^*, \Sigma_U(\psi^*), \overline{\Sigma}_V(\psi^*)) = \psi^*$. Moreover, since $\psi^* < \hat{\psi}$, $\overline{\Sigma}_V(\psi^*) = \Sigma_V(\psi^*)$, and thus ψ^* solves $\Psi(\pi^*, \Sigma_U(\psi^*), \Sigma_V(\psi^*)) = \psi^*$, namely, $(\Sigma_U(\psi^*), \Sigma_V(\psi^*)\psi^*)$ is an equilibrium of the sharing game, given $\pi^* \in (\pi_1, \pi_2) \cap U$. Thus, we have found another equilibrium in addition to the one described in Proposition A.1. □

EXAMPLE: Suppose $H(b) = 1 - e^{-b}$ and $G(\ell) = 1 - e^{-\ell}$. Then, (18), (19), (20) turn to:

$$\begin{aligned}
\Sigma_U(\psi) &= (1-p)(1-\psi)\left(1 - e^{-\frac{t}{\psi(1-\psi)}}\right) + p\left(1 - e^{-\min\left\{\frac{t}{\psi}, \frac{(1-\psi)\hat{b}}{\psi}\right\}}\right) \\
\overline{\Sigma}_V(\psi) &= (1-p)e^{-\frac{t}{\psi(1-\psi)}} + pe^{-\frac{t}{\psi}} \\
\underline{\Sigma}_V(\psi) &= (1-p)e^{-\frac{t}{\psi(1-\psi)}}
\end{aligned}$$

In Figure 6, the increasing dashed line depicts the 45^o line, while the non-monotone solid locus plots $\Psi(\pi, \sigma_U, \sigma_V)$ with $\sigma_U = \Sigma_U(\psi)$ and $\sigma_V \in \Sigma_V(\psi)$. In the left panel, the benefit distribution is atomless, as opposed to the right one.

B Total User Welfare Maximization

Let us define the following functions

$$\begin{aligned}
W^{\text{ver}}(\sigma_U; \pi) &:= (1 - \Psi(\pi, \sigma_U, 1-\sigma_U))b - t(1-\sigma_U) - \Psi(\pi, \sigma_U, 1-\sigma_U)\int_0^{G^{-1}(\sigma_U)} \ell dG \\
T(\sigma_U; \pi) &:= 2 - \pi(1-\sigma_U) \\
\widetilde{W}^{\text{ver}}(\sigma_U; \pi) &:= W^{\text{ver}}(\sigma_U; \pi) \times T(\sigma_U; \pi)
\end{aligned}$$

Given fake news production π, the first function determines the total user welfare per news, provided types with low losses $\ell \leq G^{-1}(\sigma_U)$ engage in unverified sharing, while the rest verifies before sharing. The second function determines the total news volume circulating in

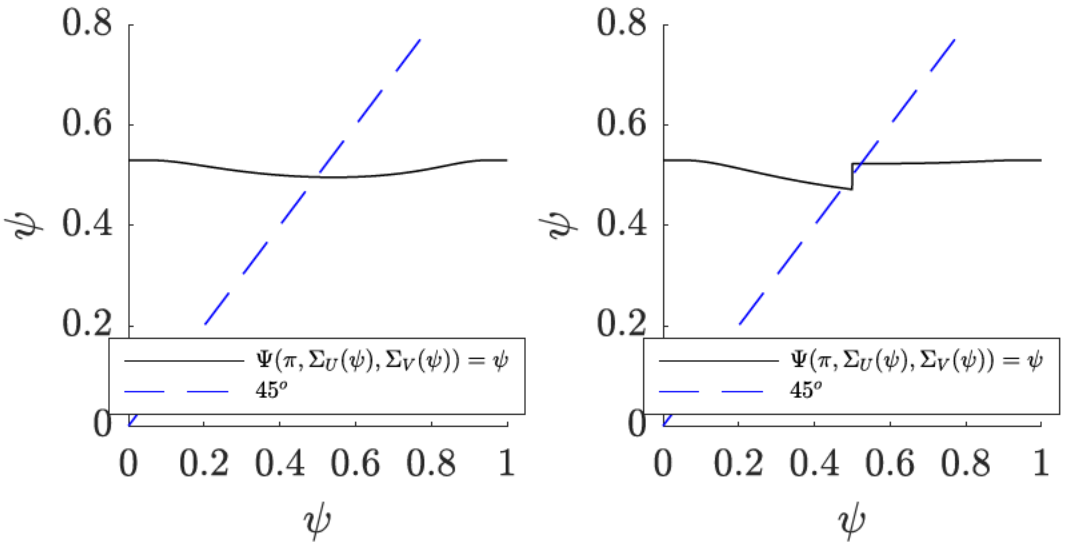

Figure 6: Parameter values: $\hat{b} = 0.8$, $t = 0.4$, $p = 0$ (left) and $p = 0.8$ (right).

the platform in a setting in which everyone finds it optimal to share, i.e., $1 + \sigma_U + (1 - \pi)\sigma_V$ with $\sigma_V = 1 - \sigma_U$. The last function determines total user welfare (per-news times volume).

Let $\sigma_U^{\text{ver}}(\pi) = \overline{\Sigma}(\pi)$ denote the level of unverified sharing in the verification equilibrium. Next, let $\sigma_{U,P}^{\text{ver}}(\pi)$ denote the solution to the planner's problem that maximizes per news welfare. By Proposition 5, $\sigma_{U,P}^{\text{ver}}(\pi)$ solves:

$$\max_{\sigma_U \in [0,1]} W^{\text{ver}}(\sigma_U; \pi) \qquad \text{such that} \qquad \Psi(\pi, \sigma_U, 1 - \sigma_U) \leq \underline{\pi} \tag{21}$$

Finally, let $\tilde{\sigma}_{U,P}^{\text{ver}}(\pi)$ denote the solution to the analogous problem (total welfare maximization):

$$\max_{\sigma_U \in [0,1]} \widetilde{W}^{\text{ver}}(\sigma_U; \pi) \qquad \text{such that} \qquad \Psi(\pi, \sigma_U, 1 - \sigma_U) \leq \underline{\pi} \tag{22}$$

Before introducing our main result, recall that the market exhibits no verification when fake news production $\pi > \bar{\pi}$ (see Proposition 1); thus, we consider the complementary case. In what follows, to minimize notation, we omit the dependence on π for the recently introduced objects (i.e., W^{ver}, $\widetilde{W}^{\text{ver}}$, T, $\sigma_U^{\text{ver}}, \tilde{\sigma}_{U,P}^{\text{ver}}, \sigma_{U,P}^{\text{ver}}$). All derivatives are in σ_U.

Proposition B.1. *Suppose that the planner is interested in maximizing total welfare and that verification is mandated at $\pi \in (0, \bar{\pi})$, where $\bar{\pi}$ is given in Proposition 1. Then, if $\widetilde{W}^{ver}$ is strictly quasi-concave, $\sigma_U^{ver} < \tilde{\sigma}_{U,P}^{ver} \leq \sigma_{U,P}^{ver}$, with strict inequality if $\pi < \underline{\pi}$. Finally, a sufficient condition that ensures that $\widetilde{W}^{ver}$ is strictly quasi-concave is that $\sup_{z \geq 0} g(z) \leq 1/b$.*

Proof: We prove the result in three steps.

STEP 1: $\tilde{\sigma}_{U,P}^{\text{VER}} \geq \sigma_{U,P}^{\text{VER}}$, WITH STRICT INEQUALITY IF $\Psi(\pi, \sigma_{U,P}^{\text{VER}}, 1-\sigma_{U,P}^{\text{VER}}) < \underline{\pi}$. To see this, suppose $\sigma_{U,P}^{\text{ver}} > 0$ (the result is trivial otherwise). If $\Psi(\pi, \sigma_{U,P}^{\text{ver}}, 1-\sigma_{U,P}^{\text{ver}}) = \underline{\pi}$, then the first-order condition for (21) implies that $\partial W^{\text{ver}}(\sigma_{U,P}^{\text{ver}})/\partial\sigma_U \geq 0$. Consequently,

$$\frac{\partial \widetilde{W}^{\text{ver}}(\sigma_{U,P}^{\text{ver}})}{\partial \sigma_U} = \frac{\partial W^{\text{ver}}(\sigma_{U,P}^{\text{ver}})}{\partial \sigma_U} \times T(\sigma_{U,P}^{\text{ver}}) + W^{\text{ver}}(\sigma_{U,P}^{\text{ver}}) \times T'(\sigma_{U,P}^{\text{ver}}) \geq 0,$$

where the inequality holds as $T, T' > 0$. So $\tilde{\sigma}_{U,P}^{\text{ver}} = \sigma_{U,P}^{\text{ver}}$ since $\widetilde{W}^{\text{ver}}$ is strictly quasi-concave.

Now, suppose $\Psi(\pi, \sigma_{U,P}^{\text{ver}}, 1-\sigma_{U,P}^{\text{ver}}) < \underline{\pi}$. The FOC implies, $\partial W^{\text{ver}}(\sigma_{U,P}^{\text{ver}})/\partial\sigma_U = 0$. So,

$$\frac{\partial \widetilde{W}^{\text{ver}}(\sigma_{U,P}^{\text{ver}})}{\partial \sigma_U} = \frac{\partial W^{\text{ver}}(\sigma_{U,P}^{\text{ver}})}{\partial \sigma_U} \times T(\sigma_{U,P}^{\text{ver}}) + W^{\text{ver}}(\sigma_{U,P}^{\text{ver}}) \times T'(\sigma_{U,P}^{\text{ver}}) = W^{\text{ver}}(\sigma_{U,P}^{\text{ver}}) \times T'(\sigma_{U,P}^{\text{ver}}) > 0$$

Again, since $\widetilde{W}^{\text{ver}}$ is strictly quasi-concave in σ_U, we must have $\tilde{\sigma}_{U,P}^{\text{ver}} > \sigma_{U,P}^{\text{ver}}$. Finally, observe that for any $\pi < \underline{\pi}$, we have $\Psi(\pi, \sigma_U, 1-\sigma_U) \leq \pi$; hence, the same conclusion applies. □

STEP 2: $\tilde{\sigma}_{U,P}^{\text{VER}} < \sigma_U^{\text{VER}}$. To see this, first use $\widetilde{W}^{\text{ver}}$ and T to write $\widetilde{W}^{\text{ver}}$ as:

$$\widetilde{W}^{\text{ver}}(\sigma_U) = 2(1-\pi)b - t(1-\sigma_U)T(\sigma_U) - \pi(1+\sigma_U)\int_0^{G^{-1}(\sigma_U)} \ell dG$$

Next, differentiate the above expression, using that $T' = \pi$, to get:

$$\frac{\partial \widetilde{W}^{\text{ver}}(\sigma_U)}{\partial \sigma_U} = tT(\sigma_U) - t(1-\sigma_U)\pi - \pi\int_0^{G^{-1}(\sigma_U)} \ell dG - (1+\sigma_U)\pi G^{-1}(\sigma_U)$$

Now divide both sides by volume $T(\sigma_U)$ and recall that $\Psi(\pi, \sigma_U, 1-\sigma_U) = (1+\sigma_U)\pi/T(\sigma_U)$:

$$\frac{1}{T}\frac{\partial \widetilde{W}^{\text{ver}}(\sigma_U;\pi)}{\partial \sigma_U} = t - \Psi(\pi, \sigma_U, 1-\sigma_U)G^{-1}(\sigma_U) - \frac{\pi}{T}\left(t(1-\sigma_U) + \int_0^{G^{-1}(\sigma_U)} \ell dG\right) \quad (23)$$

Finally, observe that if $\tilde{\sigma}_{U,P}^{\text{ver}} > 0$ (the result is trivial otherwise), the FOC for (22) implies $\partial \widetilde{W}^{\text{ver}}(\tilde{\sigma}_{U,P}^{\text{ver}})/\partial\sigma_U \geq 0$ (with equality if $\Psi(\pi, \tilde{\sigma}_{U,P}^{\text{ver}}, 1-\tilde{\sigma}_{U,P}^{\text{ver}}) < \underline{\pi}$). Thus, by expression (23):

$$t - \Psi(\pi, \tilde{\sigma}_{U,P}^{\text{ver}}, 1-\tilde{\sigma}_{U,P}^{\text{ver}})G^{-1}(\tilde{\sigma}_{U,P}^{\text{ver}}) > 0.$$

Meanwhile, in the verification equilibrium, $t - \Psi(\pi, \sigma_U^{\text{ver}}, 1-\sigma_U^{\text{ver}})G^{-1}(\sigma_U^{\text{ver}}) = 0$ (Proposition 1). Since $\sigma_U \mapsto \Psi(\pi, \sigma_U, 1-\sigma_U)G^{-1}(\sigma_U)$ is strictly increasing, $\tilde{\sigma}_{U,P}^{\text{ver}} < \sigma_U^{\text{ver}}$. □

STEP 3: IF $\sup_{z\geq 0} g(z) \leq 1/b$ THEN $\widetilde{W}^{\text{VER}}$ IS STRICTLY QUASI-CONCAVE. To prove this claim, we will show that the condition on primitives ensures that W^{ver} is strictly concave

(Lemma B.1), which, in turn, ensures that $\widetilde{W}^{\text{ver}}$ is strictly quasi-concave (Lemma B.2).

Lemma B.1. *If* $\sup_{z\geq 0} g(z) \leq 1/b$ *and* $\pi > 0$ *then* W^{ver} *is strictly concave in* σ_U.

Proof: Let us begin with some preliminary steps. First, rewrite W^{ver} as follows:

$$\begin{aligned} W^{\text{ver}}(\sigma_U) &= b - t(1-\sigma_U) - \Psi(\pi, \sigma_U, 1-\sigma_U)\left(\int_0^{G^{-1}(\sigma_U)} \ell dG + b\right) \\ &= b - t(1-\sigma_U) + \Psi(\pi, \sigma_U, 1-\sigma_U)\left[\int_0^{G^{-1}(\sigma_U)} G(\ell)d\ell - \sigma_U G^{-1}(\sigma_U) - b\right], \end{aligned}$$

where the second equality holds by integration by parts. Define

$$\xi(\sigma_U) := \int_0^{G^{-1}(\sigma_U)} G(\ell)d\ell - \sigma_U G^{-1}(\sigma_U) - b.$$

Notice that $\xi < 0$, $\xi' = -G^{-1}(\sigma_U) < 0$ and $\xi'' = -1/g(G^{-1}(\sigma_U)) < 0$.

Therefore, $W^{\text{ver}}(\sigma_U)$ is strictly concave if and only if $\Psi \times \xi$ is strictly concave. The latter is trivially true if $\pi = 1$ as in such a case $\Psi \equiv 1$, and so $\Psi \times \xi$ is strictly concave as $\xi'' < 0$.

In what follows, let $\pi \in (0,1)$ and define $\Xi := \Psi \times \xi/\pi$ (we divide by π because is treated as a constant in the remaining of the analysis, and it also leads to cleaner algebra). We will show that Ξ is strictly concave in σ_U. To this end, first recall that $\Psi(\pi, \sigma_U, 1-\sigma_U) = (1+\sigma_U)\pi/T(\sigma_U)$. Hence,

$$\Xi(\sigma_U) = \frac{(1+\sigma_U)\xi}{T} \implies \Xi' = \frac{2(1-\pi)\xi + (1+\sigma_U)\xi' T}{T^2},$$

where we used that $T' = \pi$ and $T = 2 - \pi + \pi\sigma_U$ to compute Ξ'. Next,

$$\Xi'' = \frac{[2(1-\pi)\xi' + (\xi' + (1+\sigma_U)\xi'')T + (1+\sigma_U)\xi'\pi)]T - [2(1-\pi)\xi + (1+\sigma_U)\xi' T]2\pi}{T^3},$$

where we used again that $T' = \pi$. Since $T > 0$, we want to show that the numerator of Ξ'' is strictly negative. In other words, we want to show that

$$2(1-\pi)\xi' T + (\xi' + (1+\sigma_U)\xi'')T^2 + \pi(1+\sigma_U)\xi' T - 2\pi(1+\sigma_U)\xi' T < 4\pi(1-\pi)\xi$$

$$\iff 2(1-\pi)\xi' T + (\xi' + (1+\sigma_U)\xi'')T^2 - \pi(1+\sigma_U)\xi' T < 4\pi(1-\pi)\xi \tag{24}$$

Now use that $T = 2 - \pi + \pi\sigma_U$ to see that

$$2(1-\pi)\xi' T + \xi' T^2 - \pi(1+\sigma_U)\xi' T = 4(1-\pi)\xi' T.$$

Using this observation, inequality (24) can be written as:

$$4(1-\pi)\xi' T + (1+\sigma_U)\xi'' T^2 < 4\pi(1-\pi)\xi$$

$$\iff \underbrace{4(1-\pi)(\xi' T - \pi\xi) + (1+\sigma_U)\xi'' T^2}_{\Omega(\sigma_U):=} < 0.$$

Now use the expressions for T and for ξ, recalling that $\xi' = -G^{-1}(\sigma_U)$. Then,

$$\begin{aligned} 4(1-\pi)(\xi' T - \pi\xi) &= -4(1-\pi)\left(G^{-1}(\sigma_U)(2-\pi) + \pi\int_0^{G^{-1}(\sigma_U)} G(\ell)d\ell\right) + 4(1-\pi)\pi b \\ &< 4(1-\pi)\pi b \quad \leq \quad b, \end{aligned} \tag{25}$$

where the last inequality holds, since $\pi(1-\pi) \leq \max_{z\in[0,1]} z(1-z) = 1/4$. On the other hand, $(1+\sigma_U)T^2 > 1$ since $T > 1$. Thus, since $\xi'' = -1/g(G^{-1}(\sigma_U))$, it follows that

$$(1+\sigma_U)\xi'' T^2 < \frac{-1}{g(G^{-1}(\sigma_U))} \leq \frac{-1}{\sup_{z\geq 0} g(z)} \tag{26}$$

Therefore, using inequalities (25) and (26):

$$\Omega(\sigma_U) < b - \frac{1}{\sup_{z\geq 0} g(z)} \leq 0,$$

as desired. We conclude that the numerator of Ξ'' is strictly negative, and thus Ξ is strictly concave. Hence, $W^{\text{ver}}(\sigma_U) = b - t(1-\sigma_U) + \pi\Xi(\sigma_U)$ is strictly concave in σ_U. □

Lemma B.2. *If $W^{ver}(\sigma_U)$ is strictly concave, then $\widetilde{W}^{ver}(\sigma_U)$ is strictly quasi-concave.*

Proof: We will show that, $(\widetilde{W}^{\text{ver}}(\sigma_U))' = 0$ implies $(\widetilde{W}^{\text{ver}}(\sigma_U))'' < 0$. Suppose σ_U satisfies $(\widetilde{W}^{\text{ver}}(\sigma_U))' = 0$. Then, since $\widetilde{W}^{\text{ver}} = W^{\text{ver}} \times T$, we obtain:

$$(W^{\text{ver}}(\sigma_U))' \underbrace{T(\sigma_U)}_{>0} + \underbrace{W^{\text{ver}}(\sigma_U)T'(\sigma_U)}_{\geq 0} = 0 \implies W^{\text{ver}'}(\sigma_U) < 0$$

$$\implies (W^{\text{ver}}(\sigma_U)T(\sigma_U))'' = (W^{\text{ver}}(\sigma_U))''T(\sigma_U) + 2(W^{\text{ver}}(\sigma_U))'T'(\sigma_U) < 0.$$

The equality holds as $T'' = 0$, while the inequality follows from: $(W^{\text{ver}})'' < 0 < T'$; and $(W^{\text{ver}})' < 0$ whenever $(\widetilde{W}^{\text{ver}})' = 0$. This concludes the proof. □

C Total Volume of News: Changes Across Policies

We will compute total news volume for each of the policy exercises from Section 5.

Lowering verification cost (t) Given $(\pi, \sigma_U, \sigma_V)$, the total news volume is

$$\tau = 1 + \sigma_U + (1 - \pi)\sigma_V$$

- *Verification equilibrium:* Here, $\sigma_U + \sigma_V = 1$ and so $\tau = 2 - \pi\sigma_V$. Proposition 6-(i.1) shows that σ_U falls, so σ_V must rise; hence, τ must fall.
- *Mixed equilibrium:* Here, τ cannot simplify further since, generically, $\sigma_U + \sigma_V < 1$. Still, Proposition 6-(i.2) shows that both σ_U, σ_V fall; hence, τ must fall too.
- *No-verification equilibrium:* Here, $\sigma_V = 0$ and $\tau = 1 + \sigma_U$. Proposition 6-(i.3) shows that σ_U is unchanged, and thus τ must be unchanged too.

Better algorithmic filters (ϕ) Given $(\pi, \sigma_U, \sigma_V, \phi)$, the total news volume is

$$\tau = (1 - \phi\pi)(1 + \sigma_U) + (1 - \pi)\sigma_V$$

Notice that news volume is directly affected by the policy variable ϕ.

- *Verification equilibrium:* Here, $\sigma_U + \sigma_V = 1$ and so $\tau = 2 - \pi[\phi + (1 - \phi)\sigma_V]$. Proposition 7-(i.1) shows that σ_U rises, so σ_V must fall; hence, the effect on τ^{ver} is, in general, ambiguous. Using that $\sigma_V = 1 - \overline{\Sigma}(\zeta)$, where ζ is given in (14), it is easy to see that:

$$\frac{\partial \tau}{\partial \phi} > 0 \iff \left|\frac{\partial \overline{\Sigma}}{\partial \zeta}\right| \times \frac{\zeta}{\overline{\Sigma}} > \frac{1}{1 - \zeta}.$$

  Thus, volume increases if the pass-through curve is "sufficiently" elastic (in absolute terms), given the current amount of newly produced fake news in circulation (i.e., ζ).
- *Mixed equilibrium:* Here, $\tau = (1 - \phi\pi)(1 + \sigma_U) + (1 - \pi)\sigma_V$. Proposition 7-(i.2) shows that σ_U is unchanged but σ_V falls; hence, τ must fall too.
- *No-verification equilibrium:* Here, $\sigma_V = 0$ and so $\tau = (1 - \phi\pi)(1 + \sigma_U)$. Proposition 7-(i.3) shows that σ_U rises; thus, the effect on τ is ambiguous. In this equilibrium, $\sigma_U = \underline{\Sigma}(\zeta)$, and so using (14) it is straightforward to see that:

$$\frac{\partial \tau}{\partial \phi} > 0 \iff \left|\frac{\partial (1 + \underline{\Sigma})}{\partial \zeta}\right| \times \frac{\zeta}{1 + \underline{\Sigma}} > \frac{\zeta}{1 - \zeta}.$$

In other words, $1 + \underline{\Sigma}$ must be "sufficiently" elastic (in absolute terms), given the current amount of newly produced fake news in circulation (i.e., ζ).

Increasing news certification (β) Given $(\pi, \sigma_U, \sigma_V, \beta)$, the total news volume is

$$\tau = 1 + \sigma_U + (1-\pi)(1-\beta)\sigma_V$$

Notice that news volume is directly affected by the policy variable β.

- *Verification equilibrium:* Here, $\sigma_U + \sigma_V = 1$ and so $\tau = 2 - [1 - (1-\pi)(1-\beta)]\sigma_V$. Proposition 8-(i.1) shows that σ_U falls, so σ_V must rise; thus, τ must fall.
- *Mixed equilibrium:* Here, $\tau = 1 + \sigma_U + (1-\pi)(1-\beta)\sigma_V$. Proposition 8-(i.2) shows that σ_U is unchanged but σ_V rises. Since in this equilibrium σ_V solves $\Psi(\pi, \sigma_U, \sigma_V) = \underline{\pi}$ (Proposition 1), we can use (1) to solve for σ_V in closed form to get:

  $$(1-\beta)\sigma_V = \frac{(\pi - \underline{\pi})(1 + \sigma_U)}{\underline{\pi}(1-\pi)}.$$

  Observe that the right hand side is unaffected by β. So, τ is unaffected by the policy.
- *No-verification equilibrium:* Here, $\sigma_V = 0$ and $\tau = 1 + \sigma_U$. Proposition 8-(i.3) shows that σ_U is unchanged; thus, τ is unchanged too.

References

Acemoglu, D., A. Ozdaglar, and J. Siderius (2023): "A model of online misinformation," *Review of Economic Studies*, rdad111.

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.

Board, S. and M. Meyer-ter Vehn (2021): "Learning dynamics in social networks," *Econometrica*, 89, 2601–2635.

Bowen, T. R., D. Dmitriev, and S. Galperti (2023): "Learning from shared news: When abundant information leads to belief polarization," *The Quarterly Journal of Economics*, 138, 955–1000.

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

Che, Y.-K. and J. Hörner (2018): "Recommender systems as mechanisms for social learning," *The Quarterly Journal of Economics*, 133, 871–925.

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.

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

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

Kranton, R. and D. McAdams (2024): "Social connectedness and information markets," *American Economic Journal: Microeconomics*, 16, 33–62.

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.

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.

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.