Misinformation in Social Media: The Role of Verification Incentives*

Gonzalo Cisternas and Jorge Vásquez

July 24, 2026

Abstract

We develop a model featuring producers of fake news and users who can take incorrect actions linked to fake news exposure, share news items with other users, and verify content at a cost. As users' sharing and verification choices affect fake news prevalence, they also impact the same choices of other users, and multiple equilibria can arise. Equilibria exhibiting the largest misinformation production and diffusion, and the lowest verification, can maximize user welfare: these conceal lower fake news prevalence, in which case users bear lower verification costs and are lured into wrong actions less frequently. At the core of these findings is that more unverified sharing—the channel through which falsehoods can diffuse—can in fact favor the long-run diffusion of truthful content. Policies targeting verification costs, the lifespan of content, the extent of news certification, and the quality of algorithmic filters are also studied.

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

1 Introduction

The frequency, speed, and scope of recent episodes of misinformation spreading online have raised important concerns worldwide, potentially becoming a threat to areas as diverse as political elections, financial markets, and health policy, among others.[1] Furthermore, as artificial intelligence technologies that lower the production costs of highly sophisticated content improve and become more widely used, a consensus has emerged that the fake news problem can only become worse.[2] Social media platforms have therefore taken important steps in a variety of areas. First, in human-based *fact-checking*, through "third-party" systems where independent professional entities verify content,[3] or through "community notes" systems where users themselves assess content.[4] Second, in the use of technology, as with algorithms designed to detect misinformation.[5] Third, in the area of certification, where stories that were verified receive a label indicating their 'true' or 'false' status.[6]

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 platform's initiatives to combat fake news 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.

Environment. In this paper we explore how verification incentives can affect metrics such as misinformation production, diffusion, and consumption. Specifically, we develop a model of a platform in which users encounter news items in every period, some of which can be false. Users can verify and share any item, and can also *follow it*, as a proxy for taking a real action in line with the information embedded in the news article—the consumption angle. In turn, fake news production positively responds to *conversions*: the revenue stream that unfolds as more users who blindly follow content are reached.

We aim to capture that the negative consequences of fake news exposure are rooted in

[1] Allcott and Gentzkow (2017) estimate that 760 million interactions with fake news occurred on the web around the 2016 U.S. presidential election, while Guess et al. (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 a major global risk (Howell, 2013), with artificial intelligence deployed to "deepfake" videos posing a major long-term threat (World Economic Forum, 2020).

[3] These organizations are typically affiliated with the International Fact-checking Code of Principles: https://www.ifcncodeofprinciples.poynter.org.

[4] https://transparency.meta.com/features/community-notes/

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

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

luring individuals into taking incorrect actions from which the producers of fake content benefit. In this line, the specification chosen is one where users suffer a loss when following false content, and accrue a benefit when the content is truthful; meanwhile, the outside option of 'not following' yields a payoff of zero. A natural example corresponds to a dubious advertisement about a new financial opportunity, where the following decision is to invest (and not doing it leaves utility unchanged). Indeed, recent reporting in the context of *prediction markets* documents the use of falsehoods in videos advertising binary bets: 'yes/no' outcomes related to events over a specific time period. Not only were false gains reported, but some portrayed evidence of the (ex post) incorrect bets winning. This occurred through footage that did display one of the two options materializing but were associated with events *before* the time frame specified in the bet; in particular, if someone had "followed the news item" in those cases, money would have been lost.[7]

To generate conversions, misinformation has to diffuse. We make three assumptions in this area. First, users can uncover the veracity of any item upon paying a cost. Second, users accrue 'sharing payoffs': if shared truthful content reaches another individual, a benefit is internalized; but if false, the user bears a loss—these payoffs can be interpreted as being driven by pro-social motives or by reputational concerns. Third, in any period the platform chooses to keep some old content that was shared and combines it with new incoming (truthful and false) news items to generate the pool of news to be randomly allocated to users.

Altogether, since users dislike sharing fake content, misinformation diffuses only when it is not verified. When such rates of unverified sharing rise, so do conversions, raising the supply of misinformation. With this, *fake news prevalence*—the likelihood of encountering fake news—increases, so some users now prefer to bear the costs of verifying content; in turn, this drives down unverified sharing, lowering conversions and so forth. In other words, in the steady state of this economy the market settles around a production-conversions pair, paralleling the balance of supply and demand forces in traditional competitive markets.

Findings. The novelty lies in the "demand side" of this market being non-standard: our users not only consume content through their following choices but also act as *suppliers* through their sharing decisions. But users are not mechanical suppliers: since they may or may not verify content, they can non-trivially affect the 'average quality of the good'

[7]See Long et al. (2026) and Ostroff et al. (2026). In principle, someone seeing the ad in real time may think that an arbitrage opportunity is available—and since the event has not happened in reality, the share price still trades below its incorrectly hypothesized fundamental value. Also, while content creators did not receive revenue based on users placing incorrect bets, it is alleged that they were paid based on views (National Association of Consumer Advocates, 2026)—our notion of conversions is more complex because it requires not only views to happen but also a specific action to be taken. For recent evidence on prediction markets see Akey et al. (2026), who document that a majority of participants lose money.

available. Social influence effects can then arise because the verification and sharing choices by users in the past affect current fake news prevalence, in turn influencing the same choices today. In a stationary environment, a feedback loop through the channel of fake news prevalence emerges: while users' verification and sharing choices depend on their perception of prevalence, those same choices must induce the level of prevalence initially conjectured.

The resolution to this feedback loop is in the form of the demand side being subsumed in a *correspondence*: for each production level, there can be up to three different values of conversions—and a fortiori up to three (stationary) market equilibria—all of them exhibiting (i) non-trivial levels of verification and (ii) every piece of unverified content being followed. Conversions and fake news production are highest, and verification lowest, in market equilibria supported by *full sharing*: everyone shares content, with users experiencing low verification costs acting as gatekeepers of misinformation, and their high-cost counterparts facilitating its diffusion through unverified sharing. At the other end, in an equilibrium supported by *no unverified sharing*, fake news production and conversions are the lowest, while verification is the highest: relative to the full-sharing case, more users join the task of verifying content, and the rest abstains from sharing and verifying.[8]

A central result is that the first type of equilibrium always maximizes user welfare; that is, the one featuring the largest (i) rates of unverified sharing, (ii) production of misinformation and (iii) number of users potentially taking incorrect actions (because it features the lowest verification). The reason is that such equilibria conceal lower levels of fake news prevalence. Indeed, while higher rates of unverified sharing do incentivize misinformation production and diffusion, they also help truthful items diffuse. In this line, a key insight uncovered is that if more indiscriminate sharing is complemented with non-trivial levels of verification, fake news prevalence can actually fall. This is because truthful news can diffuse through disproportionately more channels than fake content can, due to two types of users facilitating transmission—those who verify and those who do not. The role of verification incentives is key: if absent, the prevalence of fake content cannot be affected by the extent of unverified sharing, because the latter does not differentiate between true and false.

From this perspective, our message is not to downplay the perception of an increased frequency and scope of misinformation episodes such as the real-world example discussed earlier. Rather, the point is to offer a broader view, where verification incentives—or their absence, in the form of unverified sharing—can shape outcomes in subtle ways. Equilibria in which there is no unverified sharing illustrate this point well, as their relatively low levels of conversions and production may portray them as preferred outcomes. However, since

[8] In between is the third option, akin to a mix of the other two, which we describe in our analysis.

these outcomes emerge when most of the sharing is verified and verification is costly, it must be that misinformation is sufficiently prevalent—otherwise, bearing those costs would be suboptimal. In terms of welfare then, those taking correct real actions are incurring costs; meanwhile, those who save on those costs take incorrect actions frequently.

Policy. We examine how equilibria displaying full sharing respond to four types of interventions. First, a reduction in verification costs (say, because news articles are now accompanied with contextual information): these costs vary across the population of users and they fall as the new distribution becomes dominated in the first-order stochastic sense. Second, a change in the lifespan of content: our platform retains shared items for the next period only with some chance, so raising such survival probability increases the lifespan of content ceteris paribus. Third, the introduction of algorithmic filters, which can be imperfect in that they may miss identifying false content. Fourth, the certification of content that was verified, which enables users to segregate certified and uncertified items when deciding how to act.

The broad takeaway is that lowering verification costs and certifying verified content are the most effective policies at increasing user welfare, in that this happens over the entire parameter space under consideration. The commonality between these two policies is that each is directly targeted at lowering a barrier for accurate decision making: a cost in the first case, and the coarseness of information in the second. In fact, by always incentivizing more verification, both policies unequivocally lower conversions, fake news production, and misinformation prevalence, ultimately yielding more user welfare.

By contrast, the other two policies can have countervailing effects. For example, the filters studied always reduce equilibrium fake news prevalence, but they also weaken the incentives to verify content, thus leading to more unverified sharing. The latter channel is a force towards more conversions, and if it is strong enough, fake news production can grow. In other words, introducing a filter can attract more misinformation to a platform in this case, and in all cases any item that passes the platform's filter will diffuse more, or gain *virality*. Similarly, increasing the likelihood of retaining shared content for another period makes users more wary about passing on fake content, as their sharing losses gain sensitivity. While this is a force towards more verification, the direct effect of more retention is that it enhances misinformation diffusion, which is a force towards more conversions—if the latter effect generates a strong enough increase in production, fake news prevalence can increase.

To assess welfare in these cases, elasticity measures come into play: for conversions (our "demand" side) in the first case and for the supply of fake content in the second—we discuss their structure and implications in their corresponding sections. More generally, it is in this policy analysis that the tractability of our stationary dynamic random matching model is fully leveraged, as we can characterize several equilibrium variables of interest.

Roadmap. Next we review the related literature. After this, Section 2 lays out the model, and Section 3 develops the equilibrium analysis. Section 4 examines user welfare across possible equilibria, while Section 5 conducts the policy analysis. Finally, Section 6 discusses our assumptions and studies extensions of the main model. Section 7 concludes. All the proofs are relegated to the Appendix.

Related literature. The literature analyzing misinformation-related themes is growing—we focus on theoretical papers here and discuss other work as our analysis progresses.

Some recent papers feature verification costs but focus on learning dynamics: Papanastasiou (2020) imposes a sequential structure as in herding models, while Cheng and Hsiaw (2022) study a sender-receiver setup where talk is cheap. Relatedly, in the context of social learning more broadly, Board and Meyer-ter Vehn (2021) examine how inspecting and adopting a product of unknown quality are affected by the network's structure when neighbors' adoption choices are observable, with ensuing implications about learning dynamics.[9] Relative to these papers, our model is stationary and we take a large population approach, where all the social influence effects among users operate through an endogenous variable (prevalence); further, there is a supply side that responds to changes in user behavior.

On this latter point, in the rich framework of Kranton and McAdams (2024) not only an explicit network structure is examined, but they also allow for bona fide producers who can produce both true and false stories—the former at a cost—as well as an exogenous stock of fake news. Under some circumstances, an equilibrium with no production of truthful news can emerge in their setting. Our paper complements theirs not only because we examine a close counterpart to their supply side (strategic misinformation production and exogenously given truthful news producers), but also because our setting features a feedback loop among users' sharing choices—namely, a fixed-point within the demand side that is then embedded into the greater fixed-point of the two-sided economy that accounts for supply responses.

There is also a literature studying how platform design can shape market outcomes. Acemoglu et al. (2023) show that an engagement-maximizing algorithm may display untrustworthy content only to those inclined to believe it, thus creating an echo chamber that is shielded from other users who could deter fake news sharing. In turn, Germano et al. (2026) examine feedback loops in popularity-based recommendation algorithms: engagement metrics (e.g., sharing) shape the scores used for ranking content, but how content is displayed influences those same metrics—if the latter are given more weight, misinformation

[9] Also, in the context of learning dynamics, Bowen et al. (2023) develop a model featuring belief polarization; the friction there is agents' misperception of others' selective sharing rather than a cost of verification.

and polarization can increase.[10] By contrast, in our paper the engagement angle is captured in an algorithm that discards content that was not shared; meanwhile, we study platform parameters that can increase user welfare by reducing fake news prevalence.

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 a context where protection choices impose externalities on other market participants through the channel of affecting the quality of matches that can take place in the future—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 dynamic matching model of a social media platform over an infinite horizon in which: (i) platform users encounter fake content originating from a group of producers; (ii) users can verify and share news; and (iii) misinformation can prompt taking incorrect real actions that benefit producers. The model is stationary with the likelihood of encountering false content being determined by its proportion within the total volume circulating—such a *prevalence* of fake news is affected by users' sharing and verification choices.

News viewers. A unit mass of infinitesimal risk-neutral users have access to an online platform where they encounter news of unknown veracity. Upon receiving a news item in any period, users must make three decisions in the order described next.

First, there is a *verification choice*: any user can perfectly determine the veracity of the news item at hand upon choosing to pay a verification cost $t > 0$. The latter can represent the search costs incurred when consulting fact-checking websites, the time costs associated with reviewing articles presented as "contextual information," attention costs, etc.

Second, there is a *sharing decision*: any user decides whether to share the news item encountered. The physical act of sharing (e.g., a click) carries no cost. Rather, users accrue payoffs stemming from their sharing choices exposing others to news of varying veracity: if shared content reaches another individual in the next period, the user who shared it obtains a benefit $\hat{b} > 0$ if the item was truthful and a loss $\hat{\ell} > 0$ otherwise; in all other situations (not sharing, or the item shared not reaching anyone) the user gets zero. These 'sharing payoffs' can be seen as originating from pro-social or reputational considerations and are of

[10] See also Che and Hörner (2018) for behavior-based recommendation systems for product adoption, and Bonatti and Cisternas (2020) for score-based pricing recommendations that also use signals of past behavior.

a first-degree nature: users care about those directly affected by their sharing choices.[11] The important implication is that, since users dislike sharing fake content, misinformation can travel through users only when it has not been verified.

Third, there is a *following decision*: any user must decide whether to take a real action linked to the news item encountered. Concretely, when a user 'follows' the item at hand, she obtains a benefit $b > 0$ if the content is truthful, but bears a loss $\ell > 0$ if false; meanwhile, not following yields a payoff of zero.[12] These 'following payoffs' capture the tangible consequences of users themselves consuming content of unknown veracity. As an example, consider a news article advertising a novel but dubious financial investment: either a new type of financial product or a novel form of bet. The user follows when she is lured into investing: if the advertisement was true, she can obtain a positive return, but if false, she can even lose the entire principal with certainty due to the financial opportunity being fraudulent. If, instead, the user does not invest, there is no change in utility.

All told, the tuple $(b, \ell, \hat{b}, \hat{\ell}, t)$ completely identifies any user. While a fully homogeneous model is unrealistic, permitting heterogeneity in all dimensions is unnecessarily demanding. To make our points, we assume the following:

Assumption 1. *(i) Verification costs t vary across users according to an atomless distribution $G(\cdot)$, with support $[0, +\infty)$ and differentiable density $g(\cdot)$. (ii) The parameters $(b, \ell, \hat{b}, \hat{\ell})$ are identical across users, satisfying (ii.1) $\ell > b > \hat{b}$ and (ii.2) $b/(b + \ell) > \hat{b}/(\hat{b} + \hat{\ell})$.*

From (i), some users with low costs will verify news, but others will find it too costly: these users can pave the way for misinformation to diffuse if they decide to share content. In turn, (ii.1) has two implications. First, the (real) gains and losses from acting upon news (b and ℓ) matter more than any sharing benefit ($\hat{b}$). Second, following fake news items brings more pronounced downsides ($\ell > b$): in the financial investment example, the potential return, if truthful, never surpasses the size of the loss when the entire principal is wiped out.

Equipped with this, one can take $\hat{\ell}$ as a free parameter and choose its size depending on the situation of interest. We nevertheless impose that it cannot be too small, as dictated by (ii.2): as we will show, this implies that any user is willing to share unverified content only when she is also willing to follow it. This can be seen either as another form of prosocial behavior or as a form of reputational consistency: in the face of uncertainty, users would not expose others to unverified content if they are not willing to take a chance on it themselves.[13]

[11] In the experimental setting of Pennycook et al. (2021) users report that they dislike passing on fake content, while Altay et al. (2022) find that users worry about their reputations when fake news is shared.

[12] Thus, our use of "following news" differs from the colloquial one related to paying regular attention to current events reported in media outlets.

[13] Note that the implied lower bound is $\hat{\ell} > \ell\hat{b}/b$. Fixing ℓ and b then, as $\hat{b}$ falls, $\hat{\ell}$ can be chosen below ℓ,

Fake news producers. There is a unit mass of fake news producers. In every period, each of them faces the choice of producing a single fake news article upon paying a *cost* $c \in [0, \infty)$, or forgoing production and obtaining zero in that period: in this sense, our producers are "malicious" in that they specialize in fake content production. These costs are distributed according to a strictly increasing CDF H with $H(0) = 0$: this CDF may have finitely many atoms and is assumed differentiable away from them. The (endogenous) total volume of fake news produced in any period is denoted by $\pi \in (0, 1)$, and it reaches the platform alongside a corresponding cohort of news articles that are truthful. For simplicity, we assume that the total volume of "fresh" incoming news at any time is normalized to 1, so the volume of truthful items is $1 - \pi$; this is largely to ease the notation and, as we argue in Sections 4 and 6, not a main driver of the economics.[14]

We assume that a producer receives a flow revenue of 1 each time the item is followed and zero otherwise. Given our previous discussion of Assumption 1 (ii.2), if a news item reaches a user who shares content without verifying it, that user herself will generate a unit of revenue and will transmit the news to other users, opening the way for the same process to repeat. Thus, any news item generates a stream of cash flows as it is reshared and subsequently followed by other users. We refer to the resulting expected net present value as the item's *conversions*—we denote this endogenous variable by $\rho \in [0, +\infty)$.

Matching and prevalence. The platform discards items that users decline to share in the previous period, as they are deemed "unpopular"; it also randomly deletes a fraction $1 - \zeta \in (0, 1)$ among those that were shared. The remaining fraction ζ of old, shared, content is then joined by the set of incoming "fresh" news items entering the platform (of mass 1 as discussed) to constitute the pool of news to be distributed in the relevant period. The allocation—i.e., the matching between users and news—is random too: statistically, news items are equivalent to i.i.d. random variables taking two possible values from any user's perspective.[15] This matching is also independent across periods.

We denote the proportion of fake news items within the current pool of news—or *fake news prevalence*—by $\psi \in [0, 1]$. Given the matching described, this variable reflects the probability that each news item encountered within any period is fake.

i.e., sharing payoffs are below their real (following) counterparts. All our figures use this case.

[14] In Section 4 we show results along the lines of "more misinformation production, higher user welfare" despite this normalization lowering the inflow of truthful content as fake news production increases. Furthermore, in Section 6 we relax this normalization and show that results along the lines of "less misinformation production, higher user welfare" still go through when there is a fixed inflow of truthful news in any period.

[15] Thus, a user may receive more than one piece of news in any given period. We assume no attrition, so all the benefits and costs previously introduced constitute payoffs on a per-item basis.

Timing, information and equilibrium. In any period, producers first simultaneously decide whether to produce or not. The platform then collects all the news items to be distributed and allocates them to users. After this, all users simultaneously make their verification, sharing, and following choices. We note that, at the "following" stage, users can be informed about the veracity of the news item at hand only by engaging in verification, i.e., sharing payoffs do not convey information. While this isolates the role of verification incentives, it also has a rationale: for example, in the reputational interpretation, such consequences would be realized in the future—in this case, $(\hat{b}, \hat{\ell})$ are discounted payoff versions.

Users are long-lived but maximize their per-period expected flow utility because they are infinitesimal (hence do not affect aggregate market variables) and because the platform almost surely never allocates a news item to the same user more than once.[16] Similarly, producers are also long-lived, but their cost-benefit analysis is effectively a static one: we will not include a discount rate in ρ because ζ acts as one. Lastly, we assume that the history of individual choices and outcomes of all players is unobserved to their counterparties, which helps us focus on the following inference problem faced by users: when seeing content, to what extent has it been verified in the past? This problem, inherently of a backward nature, is complemented by a forward counterpart faced by producers: to what extent will misinformation remain unverified and shared with users who will follow it?

A stationary environment is the simplest setting for examining the economic implications of the intertemporal considerations just described.

Definition 1 (Equilibrium concept). *In a stationary equilibrium: (i) users' verification, sharing, and following decisions, as well as producers' choices, are constant across time; and (ii) all agents play a best response.*

In what follows, we drop the 'stationary' qualifier unless needed.

Interpretation. Our model is an approximation of a large network of individuals participating in an online platform that actively displays news items or stories via so-called "news feeds": a mix of recent, or "fresh," items and older, but popular, ones—in our model, those that have been shared. In this line, while in practice users could potentially see the sharing choices of individuals in their own network, through such news feeds users can be exposed to choices of other individuals they are not connected with. Users can then be uncertain whether the story being displayed is a new one or a popular, reshared one.

To maximize "time on site," the aforementioned feeds often take an "infinite scroll" form: a string of stories one after the other. Hence, in reality users do get to see more than one

[16] With users and news items coming from a continuum and the allocation being independent across time, the chance of two identical draws is zero.

story every time they log in, which is consistent with our users viewing more than one news item in any period (e.g., a day). In Section 6 we show how the model can easily accommodate congestion effects, where users consider only a fraction of the news that they are allocated.

This brings us to the allocation technology. While the random matching is useful for tractability, it is not an unreasonable assumption given that news items are not horizontally differentiated in our model. Furthermore, relaxing this assumption can precisely open the way to congestion effects deserving a more careful consideration: for example, if the platform wants to maximize diffusion of news in general, it could channel news items to users more prone to engaging in unverified sharing; or if the goal is to minimize misinformation, those more likely to verify news (low-cost types) could be the targets. In other words, it is the joint assumption of random matching and no congestion that is sensible as a whole.

A further discussion on the user- and producer-side specifications can be better appreciated after seeing how the model works. Because of this, and to avoid breaking the flow between the equilibrium analysis and policy exercises, we relegate this discussion to Section 6. Concretely, we address the following topics: dropping Assumption 1 (ii.2), in which case the model is simpler; the advantages of the symmetry between sharing and following payoffs; a model with a flexible volume of incoming news in any period; and the case of conversions being driven by views rather than by inducing incorrect real actions.

3 Equilibrium Analysis

The economy studied is a two-sided market in which users take actions given the prevailing level of prevalence ψ, while fake news producers respond to the level of conversions ρ. These two sides are connected as follows. First, conversions determine producers' incentives to create fake content. Meanwhile, misinformation production affects its prevalence, which in turn affects user behavior. However, user behavior determines conversions: verification and sharing choices influence how far news items can travel and ultimately whether they will be followed. In this chain, the prevalence-behavior link—i.e., the backward aspect of the model—is non-trivial: while prevalence determines users' sharing and verification choices, those same choices themselves shape the prevalence of fake content. Once this feedback loop is solved, constructing conversions—the forward aspect—simply follows from accounting.

Consequently, our analysis proceeds in three steps. First, in Section 3.1 we characterize ψ and ρ, taking producer and user behavior as given. Second, in Section 3.2 we endogenize user behavior: users' individual choices must be optimal given their beliefs about prevalence, and the latter must be consistent with the same choices it generates. Third, in Section 3.3 we endogenize production: we show how stationary equilibria in this two-sided market emerge as "a balance between conversion and production forces"—in direct parallel with competitive markets, but with some important differences that we explain shortly.

3.1 Prevalence and Conversions

When it comes to sharing behavior, users can be segmented into three sets. First, users could verify news, in which case they share content only when truthful. Second, they could choose to engage in unverified sharing: skipping any verification and indiscriminately sharing what they see. Third, they could opt out of sharing: simply not sharing any content encountered. In a stationary equilibrium, the masses of users in each set are constant: we denote them by σ_V, σ_U and σ_N, respectively, all in $[0, 1]$ and adding up to 1.

With random matching, the way in which news items diffuse is fully determined by the aforementioned masses. To state our first result, let T and F denote the steady-state stocks of truthful and fake content in any period—a mix of fresh and old (shared) content of each type. By definition, fake news prevalence is $\psi = F/(T + F)$, and recall that with probability ζ the platform retains any item that was shared in the previous period.

Lemma 1. *In any equilibrium,*

$$T = \frac{1-\pi}{1-\zeta(\sigma_U+\sigma_V)} \quad \text{and} \quad F = \frac{\pi}{1-\zeta\sigma_U}, \tag{1}$$

and so prevalence takes the form

$$\psi = \pi\frac{1-\zeta(\sigma_U+\sigma_V)}{1-\zeta(\sigma_U+\sigma_V\pi)} := \Psi(\pi, \sigma_U, \sigma_V). \tag{2}$$

Meanwhile, conversions are given by

$$\rho = \frac{1-\sigma_V}{1-\zeta\sigma_U} := P(\sigma_U, \sigma_V) \tag{3}$$

if $\psi < b/(b + \ell)$, they take value $\rho \in [0, P(\sigma_U, \sigma_V)]$ if $\psi = b/(b + \ell)$, and are zero otherwise.

The steady-state stocks of truthful and false content given in (1) correspond to each inflow type amplified by the associated rate of transmission: conditional on being retained by the platform (ζ term), truthful content diffuses through both verified and unverified sharing ($\sigma_U + \sigma_V$ in the denominator of T), while misinformation does it only through the latter channel (σ_U in F). From these expressions, we obtain prevalence as given by (2).

The function Ψ reveals that, if $\pi \in (0,1)$, prevalence is strictly below π when $\sigma_V > 0$: the possibility of agents discriminating between false and truthful content at the moment of sharing reduces the prevalence of misinformation relative to its baseline abundance among new content. And naturally, prevalence increases with production, as implied by $\partial\Psi/\partial\pi > 0$.

To gain intuition, suppose that $\sigma_V > 0$, as will be the case in equilibrium. If $\sigma_U + \sigma_V < 1$, $\partial\Psi/\partial\sigma_V < 0$ as one would expect: increasing the extent of verification reduces fake news prevalence. What is interesting is that $\partial\Psi/\partial\sigma_U < 0$ holds too: if more users engage in unverified sharing and σ_V remains fixed, fake news prevalence falls. While this may seem surprising, it has a natural explanation: even though fake content can diffuse more, so can truthful items and, in fact, they can *diffuse through disproportionately more channels.* This is because two types of users facilitate the transmission of truthful content as opposed to one type for fake news, and these extra possibilities have a compounding effect over time.

In other words, a *complementarity* between verified and unverified sharing arises in this case, which will be central to our findings.[17] We note, however, that this property reverses when everyone engages in some type of sharing, or $\sigma_U + \sigma_V = 1$: marginally increasing σ_U carries an equal reduction in σ_V, resulting in higher prevalence as σ_U grows.

Turning to ρ, notice that conversions come from the users who do not verify news, a set that has mass $1 - \sigma_V$. Conditional on reaching such an agent, a news item will generate subsequent revenue provided that the user engages in unverified sharing (σ_U term) and the platform retains the item (with probability ζ). The result is an infinite sum with value (3), which is valid only when following news of unknown veracity is strictly profitable. As users' sharing payoffs have become irrelevant at the moment of deciding whether to follow content or not, the condition for following is

$$(1-\psi)b - \psi\ell > 0 \Leftrightarrow \psi < \frac{b}{b+\ell},$$

as stated in Lemma 1. Instead, if prevalence ψ equals $b/(b+\ell)$, there is indifference between following news and not doing it: those not verifying can split in any way between these options, so any $\rho \in [0, P(\sigma_U, \sigma_V)]$ is possible. If $\psi > b/(b+\ell)$, conversions drop to zero.

Towards finding equilibria, we need to determine the values that $(\sigma_U, \sigma_V, \sigma_N)$ can take as a function of π. Equipped with this, we can construct a "conversion locus"—a mapping between π and possible values of ρ via (3)—and find equilibria in the (π, ρ) space along with the supply of fake content, as it is routinely done in analyses of competitive markets.

The novelty is that $(\sigma_U, \sigma_V, \sigma_N)$ stems from an actual *game* played among users. Specifi-

[17]More formally, observe that $\frac{\partial^2(1-\Psi)}{\partial\sigma_U\partial\sigma_V} = \frac{\zeta^2\pi(1-\pi)}{(1-\zeta[\sigma_U+\sigma_V\pi])^3}[1 - \zeta\sigma_U + \zeta\sigma_V\pi] > 0$.

cally, users need to anticipate others' verification and sharing choices to correctly anticipate prevalence. In turn, prevalence guides individual verification and sharing choices. However, when aggregated, such choices must be consistent with the initial conjectures. This feedback loop can lead to different tuples $(\sigma_U, \sigma_V, \sigma_N)$ being sustained for a given π, reflecting that non-trivial social effects can be at play on the "demand" side of this market. This contrast with traditional competitive analyses is not only due to the good involved (news items) being of unknown quality, as occurs in many other settings. Rather, it is the combination of incomplete information with a demand side where agents play a dual role: not only do they consume goods but also act as suppliers via their sharing decisions. This latter channel can have non-trivial consequences when news verification is allowed: as the average quality of the good can be affected, different conjectures about others' behavior can be self-fulfilling.

3.2 The sharing game

The "sharing game" is the strategic environment faced by users, defined as follows: users choose optimal verification, sharing, and following decisions given (i) what other users are doing and (ii) that production is fixed at $\pi \in [0,1]$. Importantly, what others do matters to any individual user only through the implied level of fake news prevalence.

Consider first a user who verifies news: she pays the cost t upfront, and she gets to share and follow content only when truthful. If she conjectures a level of prevalence ψ, she will expect to accrue the (following) benefit b with probability $1-\psi$ and the sharing payoff $\hat{b}$ with probability $(1-\psi)\zeta$, as a fellow user will be exposed to the item in the next period only when the platform retains it. Thus, the (per unit) expected payoff from verifying is

$$(1-\psi)(b+\zeta\hat{b})-t.$$

The relevant outside option to verified sharing is either unverified sharing or not sharing at all. To determine which one dominates, observe first that a user who does not verify content will receive a payoff of $\max\{(1-\psi)b-\psi\ell, 0\}$ from the following choice, irrespective of whether she shares or not. The trade-off between doing unverified sharing and not sharing is thus shaped by the sharing payoffs exclusively, as the "following" counterpart just stated is common to both options. All told, unverified sharing dominates not sharing if and only if

$$(1-\psi)\zeta\hat{b}-\psi\zeta\hat{\ell} \geq 0 \Leftrightarrow \psi \leq \frac{\hat{b}}{\hat{b}+\hat{\ell}}, \tag{4}$$

that is, when prevalence is not too high. Accordingly, since $\hat{b}/(\hat{b}+\hat{\ell}) < b/(b+\ell)$, observe that any user who does unverified sharing will find it optimal to follow the content that she distributes. Conversely, if not sharing is the relevant outside option, users can still follow news as long as ψ is not too large: more precisely, when $\psi \in [\hat{b}/(\hat{b}+\hat{\ell}), b/(b+\ell)]$.

As users go about making their decisions, a tuple $(\sigma_U, \sigma_V, \sigma_N)$ emerges. The aforementioned feedback loop then manifests in an equation for prevalence encoding the equilibrium condition that users' beliefs must be correct: aggregate behavior $(\sigma_U, \sigma_V, \sigma_N)$ must, via (2), induce a level of prevalence that coincides with ψ initially conjectured. Before stating our main formal result, let us give an overview of the types of equilibria that can arise in this form of sub-game among users that takes $\pi \in [0,1]$ as given.

Full sharing. From the previous analysis, if ψ is below $\hat{b}/(\hat{b}+\hat{\ell})$, we must have $\sigma_U+\sigma_V = 1$: everyone must be doing some kind of sharing. In such a *full-sharing equilibrium*, a user verifies whenever her cost is low enough according to

$$\underbrace{(1-\psi)(b+\zeta\hat{b})-t}_{\text{payoff from verification}} \geq \underbrace{(1-\psi)\zeta\hat{b}-\psi\zeta\hat{\ell}}_{\text{unverified sharing}} + \underbrace{(1-\psi)b-\psi\ell}_{\text{following}} \Leftrightarrow \psi(\ell+\zeta\hat{\ell}) \geq t. \tag{5}$$

Thus, aggregate verification reads $\sigma_V = G(\psi(\ell+\zeta\hat{\ell}))$. For the users' beliefs to be correct, prevalence ψ must satisfy (2) under $\sigma_U + \sigma_V = 1$, where σ_V is as just stated, leading to

$$\psi = \pi \frac{1-\zeta}{1-\zeta(1-G(\psi(\ell+\zeta\hat{\ell}))+\pi G(\psi(\ell+\zeta\hat{\ell})))}. \tag{6}$$

As π goes to zero, so does ψ, and so this equilibrium is the only option for low levels of fake news production. As π rises, one would expect prevalence to rise too, in which case this equilibrium would cease to exist eventually—we confirm this shortly.

No unverified sharing. As π rises and users expect prevalence to exceed $\hat{b}/(\hat{b}+\hat{\ell})$, the relevant outside option becomes not sharing at all. Crucially, this can happen before the equilibrium with full sharing disappears: due to the complementarity discussed in Section 3.1, a full-sharing equilibrium exhibiting $\sigma_U > 0$ generates lower prevalence than one featuring $\sigma_U = 0$ for the same $\sigma_V > 0$ and $\pi \in (0,1)$. Thus, two types of equilibria (with $\sigma_U > 0$ and $\sigma_U = 0$) can coexist for intermediate values of production.

With this new outside option, and as long as $\psi \leq b/(b+\ell)$ so that following news items is still profitable, those who verify content must experience costs t such that

$$(1-\psi)(b+\zeta\hat{b})-t \geq \underbrace{(1-\psi)b-\psi\ell}_{\text{not sharing, but following}} \quad \Leftrightarrow \quad \zeta\hat{b}+\psi(\ell-\zeta\hat{b}) \geq t. \tag{7}$$

Thus, the mass of users who verify news is given by $\sigma_V = G(\zeta\hat{b}+\psi(\ell-\zeta\hat{b}))$. Inserting the latter into (2) along with $\sigma_U = 0$ then yields the following equation for ψ:

$$\psi = \pi\frac{1-\zeta G(\zeta\hat{b}+\psi(\ell-\zeta\hat{b}))}{1-\zeta G(\zeta\hat{b}+\psi(\ell-\zeta\hat{b}))\pi}. \tag{8}$$

As we will show, there is a production region wherein the solution to (8) is effectively less than $b/(b+\ell)$. In this case, the mass of users following news is $\sigma_N = 1-G(\zeta\hat{b}+\psi(\ell-\zeta\hat{b})) > 0$, which means that producers can still earn profits despite fake news never being reshared: news items can generate only one conversion when, upon entering the platform, they are matched with someone who follows news. As production grows and prevalence reaches $b/(b+\ell)$, however, everyone ceases to follow news and conversions drop to zero.[18]

Partial unverified sharing. A third equilibrium displaying indifference between the two outside options to verified sharing can arise in the region of coexistence of the previous two. The key is that the corresponding levels of prevalence and extent of verification are the same (i) at the lower production level at which the equilibrium without unverified sharing emerges and (ii) at the higher production level at which the full-sharing one vanishes. As a result, the users who do not verify news at either point can sort between unverified sharing and not sharing while remaining indifferent between these alternatives. In turn, this indifference requires prevalence ψ to be pegged at $\hat{b}/(\hat{b}+\hat{\ell})$.

Consequently, using (2) and the fact that at $\psi = \hat{b}/(\hat{b}+\hat{\ell})$ we have $\sigma_V = G(\psi(\ell+\zeta\hat{\ell})) = G(\zeta\hat{b}+\psi(\ell-\zeta\hat{b})) = G(\hat{b}(\ell+\zeta\hat{\ell})/(\hat{b}+\hat{\ell}))$, the sorting must feature σ_U satisfying

$$\frac{\hat{b}}{\hat{b}+\hat{\ell}} = \pi\frac{1-\zeta(\sigma_U+G(\hat{b}(\ell+\zeta\hat{\ell})/(\hat{b}+\hat{\ell})))}{1-\zeta(\sigma_U+\pi G(\hat{b}(\ell+\zeta\hat{\ell})/(\hat{b}+\hat{\ell})))}. \tag{9}$$

Importantly, as production increases from the lowest to the highest point in the coexistence region, σ_U must increase in this type of equilibrium to offset the tendency of prevalence to increase, in yet another demonstration of the complementarity between verified and unverified sharing. Throughout this region, however, the set of users who do not verify news has

[18] In this case, prevalence satisfies an equation analogous to (8) that uses the outside options of not sharing and not following, i.e., verification happens when $(1-\psi)(b+\zeta\hat{b})-t > 0$. See (A.6) in the Appendix.

a constant mass of $\sigma_N + \sigma_U = 1 - \sigma_V = 1 - G(\hat{b}(\ell + \zeta\hat{\ell})/(\hat{b} + \hat{\ell})))$—these users follow fake news and hence generate conversions. The next result formalizes the discussion thus far: in all cases, the ψ solutions referred to are continuous functions in π.

Proposition 1. *There are production levels $0 < \underline{\pi} < \bar{\pi} < 1$, and $\hat{\pi} \in (\underline{\pi}, 1)$, such that:*

(i) *Full-sharing. Equation (6) admits a solution $\psi(\pi)$ satisfying $\psi(\pi) \leq \hat{b}/(\hat{b} + \hat{\ell})$ if and only if $\pi \in [0, \hat{\pi}]$, with strict inequality if $\pi < \hat{\pi}$. This solution is unique and increasing in π. Hence, a full-sharing equilibrium always (and only) exists in this region.*

(ii) *No unverified sharing. An equilibrium with $\sigma_U = 0$—i.e., with prevalence greater than $\hat{b}/(\hat{b} + \hat{\ell})$—exists for all $\pi \in [\underline{\pi}, 1]$. Prevalence is also below $b/(b + \ell)$ if and only if $\pi \in [\underline{\pi}, \bar{\pi}]$, in which case the equilibrium is characterized as the unique solution $\psi(\pi)$ to (8), which is an increasing function. Hence, an equilibrium without unverified sharing, but with non-zero conversions, always (and only) exists in $[\underline{\pi}, \bar{\pi}]$.*

(iii) *Partial unverified sharing. An equilibrium with partial unverified sharing exists if and only if $\pi \in [\underline{\pi}, \hat{\pi}]$ due to (9) always admitting a unique solution $\sigma_U(\pi) \in [0, 1 - G(\hat{b}(\ell + \zeta\hat{\ell})/(\hat{b} + \hat{\ell})]$ in that region. Prevalence is constant at the value $\hat{b}/(\hat{b} + \hat{\ell})$ and so is the mass of users following fake news.*

The above constitute the full set of pure-strategy equilibria in the sharing game.[19]

Figure 1 below depicts prevalence (left panel) and the extent of verification (right panel) along the possible equilibria of the sharing game: for the chosen specification, $\hat{\pi}$ (where the full-sharing equilibrium ceases to exist) is below $\bar{\pi}$ (where users who do not verify news cease to follow items along the equilibrium with $\sigma_U = 0$). As $\bar{\pi}$ is the maximum level of production that can sustain conversions, we plot over the domain $[0, \bar{\pi}]$.

The lowest segment in each panel depicts the variable under consideration in a full-sharing equilibrium; in turn, the flat region pertains to the partial counterpart; and finally, the top relates to the one without unverified sharing. Thus, along each equilibrium, prevalence increases (or remains constant) as π grows; and pointwise in π, it ranks from lowest to highest as we transition across equilibria in the order just described. The extent of verification (right panel) follows an identical qualitative pattern for an obvious reason: as misinformation is more (less) prevalent, there are stronger (weaker) incentives to verify content. (The properties just mentioned are formally stated as part of the proof of Proposition 1.)

[19] The statement also extends to symmetric mixed-strategy equilibria under an application of the law of large numbers: in (iii), if those users with $t \geq \hat{b}(\ell + \zeta\hat{\ell})/(\hat{b} + \hat{\ell})$ do unverified sharing with probability $\sigma_U(\pi)$ and not sharing with $1 - \sigma_U(\pi)$, aggregate unverified sharing coincides with $\sigma_U(\pi)$ almost surely.

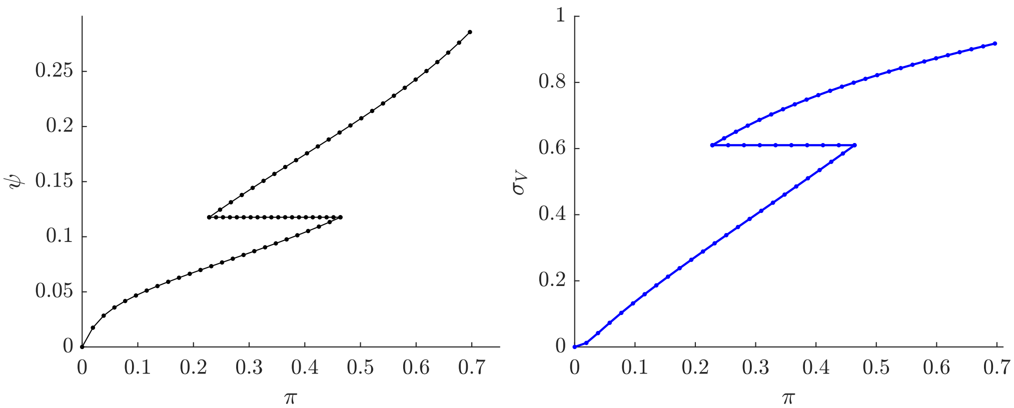

Figure 1: Parameter values: $\hat{b} = 0.2$; $\hat{\ell} = 1.5$; $b = 0.8$; $\ell = 2$; $\zeta = 0.9$; and $t \sim \text{Gamma}(3, 8)$.

In particular, the full-sharing equilibrium exhibits the lowest prevalence, despite its lower verification. In other words, when it comes to favoring truthful content, a full-sharing equilibrium's higher unverified sharing rate combined with a weaker but non-trivial verification is more effective than the stronger verification incentives acting alone in the no-unverified sharing counterpart. This better balance will be key in shaping user welfare (Section 4).

3.3 Market Equilibria: A "Supply and Demand" Formulation

The last step for finding equilibria in this economy corresponds to endogenizing π. As we anticipated, this is done by equating "supply and demand."

On the demand side, from Proposition 1 and the expression for candidate equilibrium conversions—given by $P(\sigma_U, \sigma_V) = \frac{1-\sigma_V}{1-\zeta\sigma_U}$ from Lemma 1—one can construct a "conversion correspondence" yielding the possible levels of conversions induced by the equilibria of the sharing game at each π. The actual expression of this set-valued map and its properties can be found in the proof of the result that we present below (found in Appendix A.3).

As for the supply side, a producer with cost c has a strict incentive to enter the market when $c < \rho$. Thus, at any point of continuity of H, the quantity supplied is $H(\rho)$. Instead, if at cost level $\rho \in (0, \infty)$ the CDF H jumps, by virtue of indifference the quantity supplied can be any value in $[H(\rho^-), H(\rho)]$, where $\rho^- := \lim_{c \nearrow \rho} H(\rho)$. Clearly, the induced inverse supply function—defined on $\pi \in [0, 1]$ and with values in $[0, \infty)$—is continuous and non-decreasing.

Equipped with this, the complex task of finding equilibria for this economy can be reduced to a "market clearing condition" that equates both sides of this market. We refer to these as *market equilibria*, and starred variables denote the corresponding equilibrium outcomes.

Theorem 1. *A market equilibrium exists. The induced tuple (π^*, ρ^*) is an intersection point of the supply and the conversion correspondence, and it lies in $(0, \max\{\hat{\pi}, \bar{\pi}\}] \times (0, +\infty)$.*

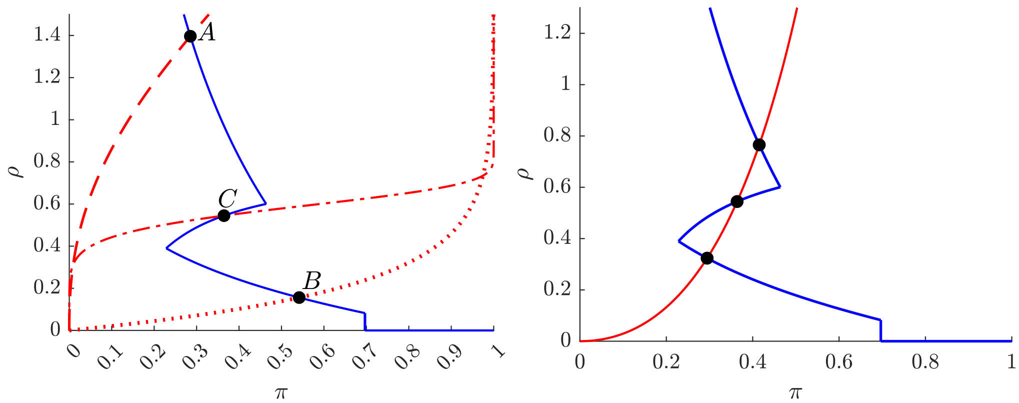

Figure 2: In both panels: $\hat{b} = 0.2$; $\hat{\ell} = 1.5$; $b = 0.8$; $\ell = 2$; $\zeta = 0.9$; and $t \sim \text{Gamma}(3, 8)$. Left panel: $c \sim$ Weibull with $(\text{scale}, \text{shape}) \in \{(2.2, 2.4), (0.61, 7), (0.2, 1)\}$. Right panel: $c \sim$ Weibull with $(\text{scale}, \text{shape}) = (2.66, 0.5)$.

Figure 2 illustrates, with both panels displaying the same conversions correspondence. The top segment pertains to possible outcomes exhibiting full sharing (i.e., $\sigma_U + \sigma_V = 1$). This segment is decreasing: as fake news production increases, so does prevalence, which induces more verification and thus less unverified sharing, implying that conversions drop.

The intermediate, increasing segment pertains to outcomes with partial sharing: as argued, σ_U needs to increase with π to maintain indifference between the outside options to unverified sharing. In this case, an increase in production leads to *more conversions* despite the total number of users following fake content remaining fixed (Proposition 1 (iii)): the key is that a larger fraction of these users engages in unverified sharing, enabling fake items to diffuse more—but since truthful items diffuse more too, prevalence can remain constant. Finally, the lowest decreasing portion features no unverified sharing ($\sigma_U = 0$): users either verify news or not share at all. As fake items are never shared, conversions take their lowest value—in fact, conversions can arise from first contacts only. Here, as in the top segment, increasing production also leads to a shift toward verification (σ_V grows as in the top segment of the right panel in Figure 1), but now by users who were not sharing at all.[20]

Turning to the supply side, if this one is very inelastic the market equilibrium will exhibit full sharing as at point A in the left panel, where low levels of production are associated with high conversions—this is sustained by low equilibrium prevalence, which justifies a large extent of unverified sharing. Conversely, in equilibrium B, a very elastic supply leads to a level of production high enough that no one engages in unverified sharing: here, fake news items are abundant so that $\psi^* \geq \hat{b}/(\hat{b} + \hat{\ell})$. Yet, conversions still happen in equilibrium (i.e., $\rho^* > 0$) precisely because market forces prevent production from becoming excessively high:

[20] The vertical drop to zero occurs around $\bar{\pi} = 0.7$, where following unverified news ceases to be profitable and ρ is multi-valued as stated in Lemma 1.

the equilibrium features $\psi^* \leq b/(b+\ell)$, which means that some—but not many—users who make first contact with fresh content will be willing to follow it; these modest conversions nevertheless suffice for generating substantial fake content given the elastic supply.

Lastly, an equilibrium with partial sharing is depicted by point C: here the set of users is fully segmented into the three possible options ($\sigma_X > 0$ for $X = U, V, N$).[21] Meanwhile, the right panel illustrates a situation where multiple equilibria can arise due to the social effects at play: users' different conjectures of what others are doing can be self-fulfilling.

4 User Welfare

To evaluate different types of equilibria, we examine user welfare in its utilitarian form: because the environment is stationary, we can focus on its per-period version. Further, we evaluate the latter criterion from an ex-ante perspective, before the matching in any period is realized, so all the parameters $(t, b, \hat{b}, \ell, \hat{\ell})$ can in principle affect utility depending on what the users do. To illustrate, suppose that the market equilibrium exhibits full sharing (i.e., $\sigma_U^* + \sigma_V^* = 1$). In such equilibrium, users with costs $t \leq \psi^*(\ell + \zeta\hat{\ell})$ verify news, while everyone else faces losses ℓ and $\hat{\ell}$ with chance ψ^*. Recalling that there is a unit mass of users, and letting T^* and F^* denote the equilibrium stocks of truthful and fake news in any period (the expressions of which are in Lemma 1), user welfare reads

$$\mathcal{W} = \left[(1-\psi^*)(b+\zeta\hat{b}) - \psi^*(\ell+\zeta\hat{\ell})[1-G(\psi^*(\ell+\zeta\hat{\ell}))] - \int_0^{\psi^*(\ell+\zeta\hat{\ell})} tdG(t)\right] \times [T^* + F^*]. \quad (10)$$

In expression (10), total welfare is decomposed into users' expected utility on a per-news basis—the first bracket—and how it scales with the number of news items seen in every period—the volume effect captured by $T^* + F^*$. We refer to the former as *welfare per news item*. To understand its structure, observe that conditional on an item being truthful, everyone earns $b + \zeta\hat{b}$ due to all the users following content and doing some type of sharing—the first term. In turn, the second and third terms are the losses associated with skipping and not skipping verification, respectively.

Analogous expressions pertaining to the other two equilibria can be found in the proof of the next result, which offers strong support for market equilibria featuring full sharing.

Proposition 2 (User welfare). *Fix a conversion correspondence. Then, any market equilibrium exhibiting full sharing generates more (per news and total) welfare than any possible*

[21] It is clear from the figure that there are conditions under which this equilibrium with indifference can be the only stable prediction in this economy.

market equilibrium of the other two types. Also, for equilibria along a fixed supply, the equilibrium with no unverified sharing has the lowest (per news and total) welfare if it exists.

The first part offers a uniform dominance result, in that it allows comparisons across economies that can have different production sides: once the 'demand side' is fixed, market equilibria with full sharing dominate, from a welfare standpoint, all market equilibria of the other two types, irrespective of the levels of production and conversions being compared. In turn, the last part says that if the comparison is *within* an economy, a complete ranking of possible market equilibria can always be established.

The ranking is driven by two properties that we establish in the proof. First, welfare per news, as an average notion of welfare, is negatively related to fake news prevalence; in particular, any market equilibrium supported by full sharing features lower prevalence than any counterpart supported by partial sharing, and so forth (the left panel in Figure 1). The second property is that equilibria featuring full sharing always feature a larger total volume of news circulating—hence, the two components of total welfare are larger. And in the particular case of equilibria along a fixed supply, the fact that their production levels are ranked facilitates a complete ranking of total volumes.

We conclude that equilibria with more fake news production and unverified sharing—hence with lower verification—can be more beneficial to users, as these equilibria can mask a lower prevalence of misinformation. Indeed, such equilibria can strike a better balance between two opposing forces: while more unverified sharing leads to more conversions and hence to more production, it also helps truthful content to diffuse more in relative terms. Our paper *uncovers* that the latter channel can be the dominant force: the increase in production does not outweigh the advantage of truthful content ultimately being transmitted in many more ways, resulting in a larger pool of news items that is tilted towards truthful content. Verification incentives are crucial because it is only when $\sigma_V > 0$ that the complementarity between unverified and verified sharing emerges.

Two final observations. First, the previous welfare ranking of equilibria is obtained in a setting where increases in fake news production reduce the supply of truthful content—this crowding-out effect strengthens our result further. Second, the fact that full-sharing equilibria feature everyone sharing news should not be taken literally, as one can always enrich the model with agents who do not internalize a benefit from sharing content. The point is that a larger fraction of users sharing news, with a non-trivial mass of individuals engaged in unverified sharing, need not be construed as a signal that misinformation is prevalent. Conversely, equilibria without unverified sharing reveal that low levels of conversions and production are likely associated with most news sharing being verified—but since verification is costly, it must be that misinformation is sufficiently prevalent. This is a signal of *low welfare*: those taking correct real actions bear verification costs, while those who save on the latter frequently take incorrect actions because of high prevalence.

Remark 1 (Supply shifts). If instead we compare different equilibria *within* the full-sharing class, more fake news production will lower welfare (established in the proof of Proposition 2). This happens when the supply of fake content shifts to the right, in which case the new market equilibria are traced along the correspondence's full-sharing branch.

5 Policies

This section studies four policy interventions aimed at combating misinformation: lowering verification costs, modulating the lifespan of content, introducing algorithmic filters, and certifying content that has been verified by users.

Given the welfare properties of market equilibria that feature full sharing, our focus will be on this type of market outcome. As a result, when we say that the "conversion locus shifts", we are always referring to its full-sharing branch; also, a reduction in unverified sharing is equivalent to an increase in verification (because $\sigma_U + \sigma_V = 1$). Unless otherwise stated, we assume that the (inverse) supply of fake news has a strictly positive slope in equilibrium (in which case it is also differentiable at (π^*, ρ^*), given our assumptions on H).

5.1 Verification Costs

The fact-checking initiatives that social media platforms have implemented in recent years (e.g., professional fact-checking through third parties, or user-based community notes) can be seen as reductions in the verification costs borne by users: without necessarily eliminating these costs, as users still may need to review material to decide for themselves.

We model this situation as an improvement in the distribution of verification costs according to (reversed) first-order stochastic dominance: say, there is a new cost distribution G' with the property that $G' \geq G$ pointwise.

Proposition 3 (Lower verification costs). *As the cost distribution improves, fake news production π^*, conversions ρ^*, extent of unverified sharing σ_U^*, fake news prevalence ψ^*, and the total volume of fake news F^* all fall; meanwhile, the total volume of truthful news T^* rises. Finally, both per news and total welfare rise.*

The direct effect of lower verification costs is that more users find it optimal to verify content, all else equal. Since this implies fewer users engaging in unverified sharing, the conversion locus shifts downwards, triggering a downward movement along the supply: fake news production and conversions fall in equilibrium. As the set of incoming news is now tilted towards truthful content, the steady-state stock of truthful (fake) news rises (falls).

Now, with lower production and more verification, prevalence falls: while this effect tends to weaken verification incentives (see (5)), the direct effect of lower verification costs dominates. Because welfare per news moves in the opposite direction of prevalence and lower verification costs directly increase users' utility, welfare per news necessarily rises.

Regarding total welfare, although it is affected by the total volume of news circulating, the proposition shows that it always increases. To see why, suppose that $T^* + F^*$ rises. Then, because welfare per news rises, total welfare must also rise. If instead $T^* + F^*$ falls, simple calculations show that total welfare (10) from Section 4 can be written as

$$\mathcal{W} = (b + \zeta\hat{b})T^* - [T^* + F^*] \int_0^{\psi^*(\ell+\zeta\hat{\ell})} [1 - G(t)]dt, \tag{11}$$

where $T^* = \frac{1-\pi^*}{1-\zeta}$ and $F^* = \frac{\pi^*}{1-\zeta\sigma_U^*}$ (from (1) and $\sigma_U + \sigma_V = 1$). In this expression, the integral term encompasses both verification costs and the real losses from sharing and following news, and it is increasing in prevalence. If $T^* + F^*$ falls then, the loss term in (11) becomes smaller, because ψ^* is falling too; meanwhile, the first term rises because T^* is increasing in $1 - \pi^*$.

Importantly, in Section 6 we show that these conclusions do not depend on the inflow of fresh content being fixed. Thus, the takeaway is that lowering verification costs is a powerful tool for improving welfare, partly because this policy *directly* targets the relevant margin for verification (the associated costs) without generating other side effects. Recent evidence on community-based fact checking supports the idea that lowering the barriers to verification reduces the resharing of false content: Chuai et al. (2026) show that the presence of this form of fact-checking substantially reduces the spread of misleading posts. Interestingly, the main driver of this finding is that those who become less willing to reshare are users who are not connected to the source—through the lens of our model, one possible interpretation is that not being connected creates uncertainty over whether the content was verified.

5.2 Lifespan of News

Platforms actively manage the lifespan of news to maximize user experience and engagement.[22] In our model, the engagement element is that any story that was not shared is

[22] Stoddard (2015) documents two formulas that have determined in practice how articles have been displayed on social media platforms: each puts a positive weight on a measure of user engagement and penalizes how old an article is.

immediately discarded. But to modulate the balance between new and old content in any period, shared items are kept for the next period only with probability ζ.

Thus, the latter parameter is a key determinant of news items' lifespans. Unlike the targeted approach of verification costs, however, ζ can affect equilibrium outcomes through three channels. First, it directly controls the relevance of sharing payoffs $(\hat{\ell}, \hat{b})$ for users: if a shared item is less likely to reach someone, this channel becomes weaker. Second, also on the user side, ζ affects user welfare indirectly through fake news prevalence, as it influences the stock of truthful and fake content circulating (Lemma 1). The third channel is linked to producers: ζ influences misinformation production through its direct impact on conversions $\rho = \frac{\sigma_U}{1-\zeta\sigma_U}$ (in which $\frac{1}{1-\zeta\sigma_U}$ is the expected lifespan of fake content).

The next result examines two extreme cases that offer substantial tractability. First, the supply is completely inelastic in equilibrium: locally at (π^*, ρ^*), the inverse supply is vertical, which fixes equilibrium production after small changes. Second, the supply is perfectly elastic in equilibrium: the inverse supply is horizontal over an interval around π^*, a consequence of the distribution of producers' costs exhibiting an atom at ρ^*.

Proposition 4 (Increasing the lifespan of news). *As $\zeta \in [0, 1)$ rises, in equilibrium,*

(i) if the supply is inelastic, this change lowers the equilibrium prevalence ψ^* and raises both the stock of truthful content T^* and welfare per news. A sufficient condition for equilibrium conversions ρ^* and the stock of fake news F^* to increase in ζ locally is

$$\left[1 - G(\psi^*(\ell + \zeta\hat{\ell}))\right]^2 \geq g(\psi^*(\ell + \zeta\hat{\ell}))\psi^*\hat{\ell}, \tag{12}$$

in which case total welfare increases;

(ii) if the supply is elastic, σ_U^* falls to keep ρ^* fixed. A sufficient condition for prevalence ψ^*, production π^*, and the stock of fake news F^* all rise locally is (12).

Part (i) illustrates the benefits of letting news live longer. First, as users are more sensitive to sharing losses, this pushes towards more verification (recall (5)). Second, truthful items gain more long-run diffusion, as captured by $\partial\Psi/\partial\zeta < 0$ from (2). Both effects map into a larger stock of truthful content circulating—and when production π^* is fixed, they end up reducing equilibrium prevalence. In addition, observe that users are also more sensitive to sharing benefits (via $\zeta\hat{b}$), which leads to the direct contribution of ζ to welfare per news—the first channel above—always being positive: only a fraction of users suffer from increased sharing losses $\zeta\hat{\ell}$ in (10), while everyone enjoys larger sharing benefits $\zeta\hat{b}$.[23] This effect, coupled with lower fake news prevalence, results in greater welfare per news.

[23]Formally, from (10) it is easy to see that welfare per news is increasing in ζ while keeping ψ fixed.

The predictions for conversions are ambiguous. The direct effect of ζ in $\rho = \frac{\sigma_U}{1-\zeta\sigma_U}$ is to increase this variable by fostering transmission. But there is a tendency towards more verification that we just mentioned, which limits the diffusion of fake content: $\sigma_U = 1 - G(\psi(\ell + \zeta\hat{\ell}))$ falls as a consequence of $\ell + \zeta\hat{\ell}$ growing, which lowers ρ. Condition (12) guarantees that the former direct effect dominates the latter after a marginal increase in ζ, which manifests itself in a local upward shift in the conversion locus.[24] As conversions grow in this case, the stock of fake news also rises locally due to $F^* = \pi^*(1+\zeta\rho^*)$. But this means that the total volume of news circulating, and a fortiori total welfare, increase too.

Instead, with an elastic supply the prospects for welfare to increase become dimmer. In fact, since (12) continues to trigger an upward shift in the conversion locus, π^* must increase now. The proposition then states that this effect is strong enough to raise both the stock of fake news and its prevalence (despite having more verification due to $\sigma_V^* = 1 - \sigma_U^*$ increasing). In particular, higher fake news prevalence is now a force pushing both welfare measures downward. To further explore total welfare, recall (11), i.e.,

$$\mathcal{W} = (b + \zeta\hat{b})T^* - [T^* + F^*] \int_0^{\psi^*(\ell+\zeta\hat{\ell})} [1 - G(t)]dt.$$

So, even if $T^* = (1-\pi^*)/(1-\zeta)$ were to rise after an increase in ζ—thus making the benefits component larger—the loss term would rise too because the increase in ψ^* would amplify (not offset, as it occurred with an inelastic supply) the negative contribution of a larger total volume $T^* + F^*$ circulating in the second term.

To interpret the findings in this section, observe that the advent of social media platforms can be seen as ζ taking off from an initial value $\zeta = 0$: a move towards a world in which it is considerably easier for users to share news with others.[25] At the same time, it is widely recognized that the emergence of online platforms has made it easier for content producers to reach users—from the lens of the model, an increase in ζ has led to *more conversions*. From this perspective, our findings suggest that increasing the lifespan of news items is a less effective policy when facing a more elastic supply.

Cost homogenization among fake news producers contributes to such an increase in elasticity. Further, the more producers share a common cost the more extensive these highly

[24] This condition can be weakened, as changes in ζ through ψ push towards a local expansion of ρ (see (B.6)-(B.8) in the Appendix): intuitively, lower prevalence induces more unverified sharing, which increases ρ. Note that a stronger (global) version for (12) is $[1 - G(\psi\lambda)]^2 \geq g(\psi\lambda)\psi\hat{\ell}$ for all $0 \leq \psi \leq \frac{\hat{b}}{\hat{b}+\hat{\ell}}$ and $\ell \leq \lambda \leq \ell + \hat{\ell}$.

[25] In such a "no platform benchmark," news sharing (and hence the effect of $\hat{b}$ and $\hat{\ell}$) is shut down. Thus, $\psi = \pi$, $\rho = \sigma_U = 1 - G(\pi\ell)$, $T = 1 - \pi$, and $F = \pi$. Further, as long as $\pi \leq b/(b + \ell)$, those who do not verify content do follow news items, so welfare per news reads $(1 - \pi)b - \pi\ell[1 - G(\pi\ell)] - \int_0^{\psi\ell} t dG(t)$.

elastic regions will be: the extreme case of all producers having the same cost leads to a perfectly elastic (i.e., horizontal everywhere) supply. The recent advances in artificial intelligence that make sophisticated content-production capabilities widely available at low cost are a strong force towards such a homogenization. From this perspective, our results suggest that increasing content turnover would be an appropriate policy.[26]

5.3 Algorithmic Filters

We now enrich the model to allow for detection algorithms, or *filters*. Specifically, we consider a filter that assesses news articles as they first enter the platform and before they are displayed to users—a proxy for screening content at an early stage on a platform. As eliminating truthful news articles clearly carries social costs, we focus on the more interesting case in which truthful news articles always pass the filter, but fake news articles are detected with probability $\phi \in [0, 1]$. This measure of filter quality is known by everyone.

As we show in the proof of our next result, under full sharing the expressions for fake news prevalence and conversions from Lemma 1 become

$$\psi = \Psi(\Delta(\pi; \phi), \sigma_U, 1 - \sigma_U), \text{ where } \Delta(\pi; \phi) = \frac{(1 - \phi)\pi}{1 - \phi\pi}, \text{ and } \rho = \frac{(1 - \phi)\sigma_U}{1 - \zeta\sigma_U}. \tag{13}$$

In other words, prevalence admits the same functional form as in the baseline model except for π being replaced by Δ, the fraction of incoming fake content that passes the filter. Thus, we can write $\psi = \psi(\Delta)$ to denote equilibrium prevalence in the sharing game (i.e., for fixed π), from where the extent of unverified sharing reads $\sigma_U = \sigma_U(\Delta) = 1 - G(\psi(\Delta)(\ell + \zeta\hat{\ell}))$ per (5). Further, through this latter channel, we can write $\rho = \rho(\Delta, \phi)$: the new expression for conversions in (13) simply reflects that the expected revenue that a fake item generates is now accrued only when the initial screening is passed.

The direct effect of increasing ϕ is to shift the conversion locus downwards—the reduced revenue effect. Its indirect effect is through the prevalence-unverified sharing channel. In particular, since $\partial\Delta/\partial\phi < 0$—i.e., the filter tilts the composition of incoming news towards truthful content—prevalence falls for each level of production. As this *weakens verification incentives*, the conversion locus shifts upwards. The strength of this effect is measured by

$$\epsilon_P(\Delta) := \frac{\partial\rho}{\partial\Delta}\frac{\Delta}{\rho}, \tag{14}$$

[26]See Graffius (2026) for estimates of the average time it takes a post to accumulate half of its lifetime engagement, which can be as short as less than an hour on some platforms. Such estimates correspond to the actual expected lifetime (in our model, $\sigma_U/(1 - \zeta\sigma_U)$ for fake content), rather than to ζ itself.

which is the elasticity of conversions with respect to Δ. Continue denoting by π^* the equilibrium production—where the dependence on ϕ is omitted—and use $\Delta^* := \Delta(\pi^*, \phi)$.

Proposition 5 (Effect of filters). *Marginally increasing ϕ leads both the equilibrium prevalence ψ^* and stock of false news F^* to fall; instead, the extent of unverified sharing σ_U^* and welfare per news rise. Further, $|\epsilon_P(\Delta^*)| < \frac{1}{1-\Delta^*}$ is a necessary and sufficient condition for equilibrium production π^* and conversions ρ^* to fall and for the stock of truthful content T^* to rise. The condition is also sufficient for total welfare to increase.*

As ϕ grows and more fresh items are denied entry to the platform, an improved filter always reduces the stock of fake content circulating. Ultimately, this leads to a lower equilibrium prevalence, despite having more unverified sharing (recall that $\partial\Psi/\partial\sigma_U > 0$ when $\sigma_U + \sigma_V = 1$) and potentially more equilibrium production (discussed below). Further, with a better average quality of news items, welfare per news necessarily increases. Thus, measured by per-news metrics, improving the quality of detection algorithms is a compelling tool whose direct effect dominates any countervailing indirect effects.

Turning to total welfare, the fact that a filter affects outcomes through only two channels enables us to obtain sharper results than in the previous policy exercise. The key is that under the elasticity condition in the proposition, the direct effect of a better filter is the dominant force: the conversion locus shifts downwards, which lowers both equilibrium production and conversions, while the volume of truthful news T^* rises—that total welfare increases in this case follows from the exact same argument used in Section 5.1 studying verification costs.[27]

However, because the extent of unverified sharing always increases in the presence of a filter, fake items' average lifespan conditional on not being detected ($1/(1 - \zeta\sigma_U)$) will be longer: put differently, fake items can gain *virality*, despite users potentially being better off at the same time. On the other hand, if the elasticity condition is reversed—i.e., $|\epsilon_P(\Delta^*)| > 1/(1 - \Delta^*)$—fake news production will rise.[28] This means that introducing or improving a filter can *attract more false content to a platform*. While these predictions may be seen as counterintuitive, they reflect a filter lowering the chance of encountering fake content, which effectively weakens verification incentives.[29]

To conclude, observe that sufficiently elastic conversions imply that the stock of truthful content, and a fortiori the total volume circulating, will fall. Because in this model news items

[27]Namely, if $T^* + F^*$ rises, the increase in welfare is automatic because welfare per news rises. But if it falls, we can show that the benefits component in (11) increases, while the loss falls.

[28]Such a reversal does not guarantee that total welfare will fall: while the first term in $\mathcal{W}$ falls, the second is ambiguous when $T^* + F^*$ increases (because the integral falls), or smaller when $T^* + F^*$ falls.

[29]See Pennycook et al. (2020) for a related "implied truth effect" that arises when the certification of fake content is incomplete.

inform users about the real actions to be taken, a third prediction is the following: if fake items have the ability to crowd-out truthful content, as it occurs here, filters may indirectly suppress information about possible welfare-enhancing opportunities—see Lazer et al. (2018) for a related concern regarding filters potentially constituting a form of censorship.

5.4 Certification of Content

Part of the identification problem faced by users is that they are uncertain whether a news item was encountered because someone shared it knowing it is true. Thus, we now study *news certification*, whereby a news item that was verified to be truthful can receive a "true stamp" (since all news items are different and false content is never shared, certifying false items does not change anything). Certification creates two pools: an uncertified pool whose items may be true or false, and a certified pool whose items are publicly known to be truthful (hence immediately shared and followed). We use the same notation thus far for variables pertaining to the uncertified pool and add a subscript C for certified content.

Importantly, news certification is not frictionless: in the case of third-party fact-checking, it requires users to communicate with the platform, followed by review and validation efforts performed by professional entities; in the case of user-based fact-checking, users must communicate among themselves, gather evidence, and reach consensus. To capture these frictions and preserve stationarity, we assume that if a currently uncertified truthful item is verified as true in any given period, its certification happens with probability $\beta \in (0, 1)$—in all other cases (e.g., reaching someone engaged in unverified sharing, or the certification not materializing) the item remains uncertified until another verification window opens in the future—in which case certification happens with the same chance.

Thus, with full sharing, the prevalence of fake content among uncertified items satisfies

$$\psi = \Psi(\pi, \sigma_U, (1-\beta)(1-\sigma_U)) \text{ where } \sigma_U = 1 - G(\psi(\ell + \zeta\hat{\ell})). \tag{15}$$

That is, the only change pertains to the third entry in Ψ: only a fraction $1-\beta$ of verified truthful items remains uncertified. Further, it is easy to see that raising β increases the right-hand side of the prevalence equation, all else fixed (in the baseline model, less verification increased prevalence). Thus, prevalence rises for each π: as more truthful items are certified as such, the *average quality of the residual uncertified pool worsens*. As a result, there is more verification; and since conversions continue to read $\rho = \sigma_U/(1-\zeta\sigma_U)$, while $\sigma_U = 1 - \sigma_V$, the conversion locus shifts downwards. Altogether, we have the following result.

Proposition 6 (Certification). *As certification β rises, both the equilibrium production of fake news π^* and conversions ρ^* fall, while the extent of verification σ_V^* rises. In addition:*

(i) Both the total stock of fake news F^* and the total volume of uncertified items $T^* + F^*$ fall, while the total stock of (certified plus uncertified) truthful content $T^* + T_C^*$ rises;

(ii) The prevalence of fake items within the uncertified pool ψ^* rises, but the unconditional likelihood of encountering fake news falls;

(iii) Total welfare—that is, volume effects included, accounting both for all certified and uncertified items—rises.

The beginning of the proposition follows from the shift just described. Further, with lower production and more verification, the total volume of circulating news (certified and uncertified) shifts towards truthful content (last part in (i)). Thus, the total probability of encountering a fake item falls, despite the uncertified pool's quality worsening (part (ii)).

Total welfare always increases per part (iii). To understand why, observe that the total volume of truthful news is now $T^* + T_C^*$ while prevalence within the uncertified pool is $\psi^* = F^*/(T^* + F^*)$. Thus, expression (11) for total welfare reads

$$\mathcal{W} = (b + \zeta\hat{b})[T^* + T_C^*] - \frac{F^*}{\psi^*}\int_0^{\psi^*(\ell+\zeta\hat{\ell})}[1 - G(t)]dt. \tag{16}$$

In this expression, the first term increases because $T^* + T_C^* = \frac{1-\pi^*}{1-\zeta}$ and production falls. Regarding the loss term, we show in the proof that it always falls with β: because F^* falls while $\psi \mapsto \frac{1}{\psi}\int_0^{\psi(\ell+\zeta\hat{\ell})}[1 - G(t)]dt$ is decreasing. This is intuitive: as more certification takes place, the costs that users bear originate from a shrinking set of uncertified items, with prevalence (which grows) only determining which types of costs different users will bear.[30]

These findings are strikingly similar to those when lowering verification costs: fake news production, conversions, and the overall probability of encountering fake content all fall; meanwhile, there is more verification and welfare. The reason is that both policies impact a relevant margin for decision making: in Section 5.1, this was a cost that prevented distinguishing true from false, while here it is the coarseness of information standing in between verified and unverified content. The key is that these policies affect these margins in a targeted manner: they increase the incentives to verify without creating any side effects.

[30]The proof also shows that the welfare associated with uncertified items—which carries the same loss term as in (16)—falls. This owes to the volume of uncertified truthful news falling and uncertified prevalence rising.

6 Discussion

Consistency between sharing and following unverified news. As discussed earlier, our users are willing to share unverified news items only when they are willing to follow them. Instead, if we revert (ii.2) in Assumption 1, as production increases along a full-sharing equilibrium, those who do not verify content eventually switch from following to not following news before any unverified sharing stops, due to $b/(b+\ell) < \hat{b}/(\hat{b}+\hat{\ell})$ now holding. Furthermore, as we show in Corollary A.1 in the Appendix, non-trivial conversions can only arise when there is full sharing: this is because any outcome displaying partial sharing or no unverified sharing necessitates prevalence to be above $\hat{b}/(\hat{b}+\hat{\ell})$. Thus, the conversion correspondence is dramatically simplified: it is initially decreasing as it occurred for low values of π in Figure 2, until it reaches the production level where the switch happens; after this, conversions vanish.[31] Our policy analysis continues to apply in the initial part.

Other 'real' payoffs. The symmetry between users' sharing and following choices—namely, both having an outside option of zero and the tuple $(b, \ell, \hat{b}, \hat{\ell})$ consisting of scalars—achieves two goals: real (i.e., following) payoffs dominate their virtual (i.e., sharing) counterparts (Assumption 1) while the fixed-point equations for prevalence (6), (8), and (9) all have a tractable structure.[32] The first requirement is important for aggregate utility in the model, constituting a suitable representation of the welfare consequences of exposing individuals to fake news. The second is useful for conducting a variety of comparative statics as our policy exercises show. But one could certainly study other more asymmetric environments, such as when both type I and type II errors matter (in particular, not following is costly when news items are truthful). In this case, the challenge in maintaining the dominance of real payoffs is solely that the fixed-point equations become somewhat more involved.

Congestion effects. We have assumed that users have enough capacity to view all the news they receive. Suppose instead that users are constrained, which we model as choosing to view only a fraction $\varrho \in (0,1)$ of news displayed to them in any period; meanwhile, any item that is not viewed cannot be verified, shared, or followed.

We maintain that the platform retains old content for the next period only if these items were shared. Letting T and F denote the (steady-state) volumes of truthful and false content that *are viewed* in any period, the following recursive relations hold:

$$T = \varrho[1 - \pi + \zeta(\sigma_U + \sigma_V)T] \quad \text{and} \quad F = \varrho[\pi + \zeta\sigma_U F].$$

[31] As in the baseline model, at the switching point there is a vertical line (*cf.* Figure 1).

[32] Essentially, all these equations boil down to a decreasing function intersecting an increasing counterpart.

To illustrate, consider the first equality. In the right-hand side, if T truthful items were seen by users in the previous period, a fraction $\zeta(\sigma_U + \sigma_V)T$ of them are shared and subsequently retained by the platform for the next period. To this set of news, the platform will add $1-\pi$ incoming truthful news, thus constituting the set of truthful news to be displayed. Within this set, only a fraction ϱ is viewed by the users.

Consequently, the steady-state prevalence is exactly as in the baseline model except that the survival rate ζ is replaced by $\zeta\varrho$: views are a necessary condition for sharing and ultimately for an item to be retained by the platform. Similarly, conversions under full sharing become $\rho = \varrho\sigma_U/(1 - \varrho\zeta\sigma_U)$: to get conversions (numerator) and transmission (denominator), fake items have to be seen by users. Qualitatively speaking then, changes in ϱ operate as changes in the survival rate while *holding the contribution of ζ to sharing payoffs fixed*. Importantly, recall from Section 5.2 that it was this latter channel that introduced ambiguity regarding changes in equilibrium production and conversions.

By symmetry with Proposition 4, we next study an increase in ϱ (less congestion) in a full-sharing equilibrium. Denote the "own-price" elasticity of supply as $\epsilon_S(\rho) := \rho H'(\rho)/H(\rho)$.

Proposition 7 (Congestion effects). *In equilibrium, as users become marginally less congested: fake news production π^*, conversions ρ^*, and the stock of fake content F^* always rise; meanwhile, prevalence ψ^* falls if and only if $\epsilon_S(\rho^*) < \frac{\varrho\zeta(1-\pi^*)(1-\sigma_U^*)}{1-\zeta\varrho}$. This condition is also sufficient for the stock of truthful news T^*, total volume $T^* + F^*$, and total welfare to rise.*

Holding production fixed, less congestion lowers prevalence: $\Psi_\varrho < 0$ just like $\Psi_\zeta < 0$, i.e., truthful content gains relative strength in terms of diffusion. But with lower prevalence, unverified sharing σ_U grows. Both channels imply that the conversion locus shifts upward—both π^* and ρ^* must therefore rise, and so does the stock of fake news F^*.[33]

Whether prevalence falls or rises depends on the strength of the change in production, which is linked to the elasticity of supply—we can obtain tighter results relative to Proposition 4 because ϱ does not affect users' sharing payoffs. However, the main qualitative findings remain: if the supply is sufficiently inelastic, prevalence falls, and welfare per news rises. Moreover, since a fall in prevalence necessarily requires an increase in the stock of truthful content (because F^* has grown), the total volume of news and total welfare rise.[34]

[33] This follows from $F^* = \frac{\varrho\pi^*}{1-\zeta\varrho\sigma_U^*}$ while $\varrho\sigma_U^* = \frac{\rho^*}{1+\zeta\rho^*}$.

[34] A world with congestion is not fundamentally different from the one studied in the baseline model. To see why, fix a model with congestion $\varrho < 1$ and survival rate ζ. It is easy to see that the induced equilibrium variables $(\pi^*, \rho^*, \sigma_U^*, \sigma_V^*, \psi^*)$, as well as welfare per news, are the same as those in a model without congestion and the following modifications: the survival rate is $\zeta' = \varrho\zeta$; users' sharing payoffs are $\hat{\ell}' = \hat{\ell}/\varrho$ and $\hat{b}' = \hat{b}/\varrho$; and a producer of cost c now bears $c\varrho$, $c \in [0, \infty)$.

That said, the business model of social media platforms depends heavily on "time on site": how much time users spend on them. As this variable increases, it is likely that more user attention is devoted to viewing stories—an increase in ϱ. This factor, coupled with a supply of fake content potentially having become more elastic (Section 5.2), highlights a downside risk present in the online world: more attention can increase misinformation prevalence and potentially lower welfare. This is important because while platform competition may lower the time that users spend on any single platform, the average time users spend across all platforms has been reported to grow over time (Saba, 2025).

Fixed supply of truthful news. By fixing the volume of fresh news entering the platform, we have implicitly assumed that an increase in fake news production crowds out truthful content. While this choice was made for convenience, it is not an unreasonable one: fake news can be engineered to achieve user engagement, which algorithms may prioritize when deciding what to display.[35] However, relaxing this assumption is instructive because of a dichotomy at play. On the one hand, the crowding-out effect strengthens any result predicting an increase in welfare per news accompanied by an increase in fake news production (e.g., welfare comparison across equilibria in Section 4), but if accompanied by a *decrease* in misinformation production, the converse "crowding-in" effect—mechanically introducing truthful news—can generate an upward bias if truthful content does not respond one-by-one.

Thus, we now consider the other extreme, in which there is an exogenously fixed volume $\kappa > 0$ of truthful news entering the platform in any period. If we continue to denote π its fake counterpart, we see that (1) becomes $T = \frac{\kappa}{1-\zeta(\sigma_U+\sigma_V)}$ and $F = \frac{\pi}{1-\zeta\sigma_U}$; that is, only the numerator in T changes. From here, if there is full sharing,

$$\psi = \tilde{\pi}\frac{1-\zeta}{1-\zeta(\sigma_U+(1-\sigma_U)\tilde{\pi})} \quad \text{where } \tilde{\pi} = \frac{\pi}{\pi+\kappa}.$$

In other words, prevalence is driven by the same function $\Psi(\tilde{\pi}, \sigma_U, 1-\sigma_U)$ but with a modified first argument $\tilde{\pi} \in (0,1)$ representing the fraction of incoming content that is false. Obviously, conversions are unchanged: $\rho = \sigma_U/(1-\zeta\sigma_U)$ when there is full sharing.

The next result revisits the exercise of lowering verification costs conducted in Section 5.1: recall that both welfare per news and total welfare increased there, and a key question is whether these findings were driven by the crowding-in effect just described. Here, instead, the stock of truthful news is constant at $T^* = \kappa/(1-\zeta)$ in an equilibrium with full sharing.

[35] Vosoughi et al. (2018) show that falsehoods have a better ability to diffuse than truthful stories, are also perceived as more novel, and can trigger emotions such as "fear, disgust and surprise." Meanwhile Brady et al. (2023) argue that, to maximize engagement, algorithms exploit human biases towards moral and emotional information, among others.

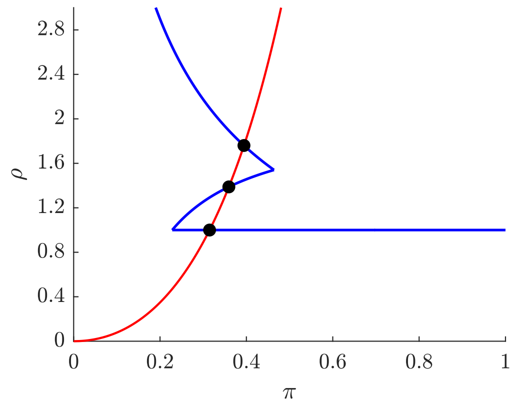

Figure 3: $\hat{b} = 0.2$; $\hat{\ell} = 1.5$; $b = 0.8$; $\ell = 2$; $\zeta = 0.9$; $t \sim \text{Gamma}(3, 8)$; and $c \sim$ Weibull with $(\text{scale}, \text{shape}) = (7, 0.5)$.

Proposition 8 (Fixed inflow of truthful content). *Fix an inflow κ of true news in any period, and suppose that verification costs fall. In a full-sharing equilibrium: production π^*, prevalence ψ^*, conversions ρ^*, the stock of fake news F^*, the extent of unverified sharing σ_U^*, and total news volume all fall. Meanwhile, both welfare per news and total welfare rise.*

Consequently, the major conclusions from Section 5.1 do not depend on our simplifying assumption. Indeed, while the total volume of news $T^* + F^*$ unequivocally falls in this case, total welfare continues to increase: the key is to revisit (11), where the only change is that the benefits component is now constant according to $(b + \zeta\hat{b})\kappa/(1 - \zeta)$.

Conversions from views. Suppose instead that producers receive a payoff of 1 whenever the item receives a view; for instance, malicious actors must reach out to specialized content producers and offer a view-based piece-rate contract.

On the demand side, the only change is the computation of conversions, as the set of equilibria in the sharing game is unchanged. In this line, if there is no unverified sharing, a fake news item generates a payoff of 1. Instead, if there is full sharing, conversions equal $1/(1 - \zeta\sigma_U)$: a payoff of 1 is accrued with probability one upon entering the platform, and subsequently with probability $\zeta\sigma_U$. Equilibria with partial unverified sharing can be obtained as in the baseline model. Overall, the model is simplified, as illustrated by Figure 3.

7 Conclusions

We have studied how verification incentives can affect the diffusion of fake news, showing that equilibria featuring lower verification, as well as more fake news production and conversions, can nevertheless deliver more welfare for users. This does not mean that verification incentives are not important, but the opposite: non-trivial levels of verification are necessary for truthful content to diffuse more relative to fake news, ultimately leading to lower levels of misinformation prevalence that justify less intensive scrutiny. At the center of this finding is a key dichotomy regarding unverified sharing: while the latter favors fake news production, it can also complement verified sharing to endow truthful content with more diffusion power.

There are at least two directions that our work could follow. The first is to study the interplay between *imperfect* public and private information about news: in our model verification is a perfect private signal of a news item's underlying type; meanwhile, certification, when it happens, is a perfect public counterpart. But information acquisition can be imperfect, and public signals could build on privately acquired information; for instance, in community notes systems users must reach a consensus before notes are shown (Kaplan, 2025). Questions about the extent of information aggregation and how this impacts welfare then arise, which can be particularly relevant if news items are allowed to exhibit some element of interdependence: knowing about one type of news can inform about related ones.

A second potential area for exploration is to examine a platform that can strategically direct news items to particular users. Platforms likely have information about users' characteristics beyond how these are distributed in the aggregate, and it could use that information to engage in third-degree discrimination by dividing users into segments that differ in their propensities to verify and/or share. The platform could then direct uncertain content toward users who are more likely to verify it, or limit the exposure of users who are especially susceptible to misinformation. Because behavior and diffusion can differ across segments, this targeting could generate dispersion in prevalence levels across clusters of users. This, in turn, can open the way for a broader set of distributional questions such as how information about user types can be used to influence the diffusion of misinformation, and whether protecting some segments shifts exposure or harmful propagation toward others. These and other questions are left for future research.

A Proofs of Sections 3 and 4: Equilibrium Analysis and User Welfare

A.1 Proof of Lemma 1: Prevalence and Conversions

Prevalence. Time is discrete with periods $\tau = 0, 1, 2, 3, \ldots$ Let $\pi_{\tau+1}$ denote and $1 - \pi_{\tau+1}$ denote the mass of fake and truthful news entering the platform between periods τ and $\tau+1$. Let $\sigma_{U,\tau}$ denote the mass of users who share without verifying, and let $\sigma_{V,\tau}$ denote that of users who verify before sharing. Let F_τ and T_τ denote, respectively, the stocks of fake and truthful news circulating at the *beginning* of period τ. Since fake news from period t will carry on to period $t+1$ only when they had been shared without being verified, the following law of motion holds

$$F_{\tau+1} = \zeta \sigma_{U,\tau} F_\tau + \pi_{\tau+1}. \tag{A.1}$$

Similarly, let T_t denote the stock of truthful news in period t. Since truthful news are shared as long as they are seen by any user who shares, then it follows that

$$T_{\tau+1} = \zeta(\sigma_{U,\tau} + \sigma_{V,\tau})T_\tau + 1 - \pi_{\tau+1}. \tag{A.2}$$

In any stationary equilibrium, tuple $(F_\tau, T_\tau, \pi_\tau, \sigma_{U,\tau}, \sigma_{V,\tau})$ is constant in τ. Thus, using (A.1)-(A.2), we obtain the steady-state levels of fake and truthful news, respectively, as a function of $(\pi, \sigma_U, \sigma_V)$:

$$F = \frac{\pi}{1 - \zeta\sigma_U} \qquad \text{and} \qquad T = \frac{1-\pi}{1 - \zeta(\sigma_U + \sigma_V)}. \tag{A.3}$$

The prevalence of fake news ψ captures the proportion of fake news circulating in the platform in steady state, i.e., $F/(T+F)$. Using (A.3), we obtain (2).

Conversions. Consider stationary sharing rates $(\sigma_U, \sigma_V, \sigma_N)$, adding up to 1, and a fake news item that enters the platform in some period. Since a user who verifies the item learns that it is fake and therefore does not follow it or share it, conversions can only come from users who do not verify. Moreover, if prevalence $\psi > b/(b+\ell)$, following news is suboptimal for users who do not verify since $(1-\psi)b - \psi\ell < 0$; thus, $\rho = 0$ for $\psi > b/(b+\ell)$.

Suppose that $\psi < b/(b+\ell)$. Then, every user who skips verification strictly prefers to follow the news item. Thus, if a fake news item enters into the platform in period τ, it will be followed with probability $1 - \sigma_V$. It could also be followed in subsequent periods if the item is shared and survives the platform's deletion. Since producers' payoff equals 1 when the item is followed (and zero otherwise), stationarity implies that the net present value of a fake news item obeys:

$$\rho = 1 - \sigma_V + \zeta\sigma_U\rho.$$

Solving for ρ yields (3), as desired.

Finally, let $\psi = b/(b+\ell)$. Then, users who do not verify are indifferent between following and not following. Thus, if a fraction $\alpha \in [0,1]$ of them breaks the indifference in favor of following, then $\rho = \alpha P(\sigma_U,\sigma_V) \in [0, P(\sigma_U,\sigma_V)]$. This completes the proof. $\square$

A.2 Proof of Proposition 1: Equilibria in the Sharing Game

We first examine the properties of the prevalence function $\Psi(\pi,\sigma_U,\sigma_V)$ defined in (2).

Lemma A.1. *The prevalence function $\Psi(\pi,\sigma_U,\sigma_V)$ in (2) obeys $\Psi(\pi,\sigma_U,\sigma_V) \leq \pi$ with strict inequality if $\sigma_V > 0$. Moreover, Ψ is strictly increasing in π for all $\sigma_U,\sigma_V \geq 0$ with $\sigma_U+\sigma_V \leq 1$, strictly decreasing in σ_V for all $\sigma_U \in [0,1)$ and $\pi \in (0,1)$, and strictly decreasing in σ_U for $\sigma_V \in (0,1)$ and $\pi \in (0,1)$. Additionally, if $\sigma_U+\sigma_V = 1$, then $\sigma_U \mapsto \Psi(\pi,\sigma_U,1-\sigma_U)$ is strictly increasing in σ_U for all $\pi \in (0,1)$.*

Proof: Notice that if $\sigma_V = 0$, then (2) yields $\Psi(\pi,\sigma_U,\sigma_V) \equiv \pi$. Conversely, if $\sigma_V > 0$, then (2) again yields $\Psi(\pi,\sigma_U,\sigma_V) < \pi$. The monotonicity properties follow from direct calculations of the partial derivatives of Ψ, which are given below:

1. Prevalence Ψ is strictly increasing in π for all $\sigma_U,\sigma_V \in [0,1]$ with $\sigma_U+\sigma_V \leq 1$:

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

2. Prevalence Ψ is strictly decreasing in σ_V for all $\sigma_U \in [0,1)$ and $\pi \in (0,1)$:

$$\frac{\partial\Psi}{\partial\sigma_V} = -\frac{\zeta(1-\zeta\sigma_U)(1-\pi)\pi}{(1-\zeta(\sigma_U+\sigma_V\pi))^2} < 0$$

3. Prevalence Ψ is strictly decreasing in σ_U for all $\sigma_V \in (0,1)$ and $\pi \in (0,1)$:

$$\frac{\partial\Psi}{\partial\sigma_U} = -\frac{\zeta^2\sigma_V(1-\pi)\pi}{(1-\zeta(\sigma_U+\sigma_V\pi))^2} < 0$$

4. Suppose $\sigma_U+\sigma_V = 1$. Then, using (2), function $\Psi(\pi,\sigma_U,1-\sigma_U)$ is given by:

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

This function is strictly increasing in σ_U for all $\pi \in (0,1)$:

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

This completes the proof. $\square$

We are now ready to prove Proposition 1. To this end, let $\Sigma_V : [0,1] \to [0,1]$ denote the mass of users who find it optimal to verify news, given prevalence ψ. This function takes a different form, depending on the type of equilibrium being analyzed, as depicted in Figure 1.

Proof of part (i): In a full sharing equilibrium, $\Sigma_V(\psi) \equiv G(\psi(\ell + \zeta\hat{\ell}))$. Then, (6) can be written as

$$\psi = \Psi(\pi, 1-\Sigma_V(\psi), \Sigma_V(\psi)). \tag{A.4}$$

Notice that Σ_V is continuously differentiable (as G is of class C^1) and strictly increasing, with $\Sigma_V(0) = 0$ and $\Sigma_V(1) < 1$. Hence, $\psi \mapsto \Psi(\pi, 1-\Sigma_V(\psi), \Sigma_V(\psi))$ is also C^1 on $[0,1]$. Also, by Lemma A.1, Ψ is bounded: $\Psi(\pi, \sigma_U, \sigma_V) \leq \pi \leq 1$ for all $\sigma_U, \sigma_V \in [0,1]$. Thus, by the Intermediate Value Theorem (IVT), there exists $\hat{\psi} \in [0,1]$ that solves (A.4).

Next, we show that, for each $\pi \in [0,1]$, $\hat{\psi}$ is unique. First, let $\pi \in \{0,1\}$. Then, $\hat{\psi} = \pi$, as $\Psi \equiv \pi$ for $\pi \in \{0,1\}$. Now, let $\pi \in (0,1)$. Then, by Lemma A.1, $\Psi(\pi, 1-\sigma_V, \sigma_V)$ is strictly decreasing in σ_V, so $\psi \mapsto \Psi(\pi, 1-\Sigma_V(\psi), \Sigma_V(\psi))$ is strictly decreasing, as $\Sigma_V(\cdot)$ is strictly increasing. Thus, for each $\pi \in (0,1)$, fixed-point $\hat{\psi}$ is unique by IVT.

Also, since Ψ and Σ_V are C^1, the Implicit Function Theorem ensures the continuity of $\pi \mapsto \hat{\psi}$. Also, since $\Psi(\pi, 1-\sigma_V, \sigma_V)$ is strictly increasing in π and strictly decreasing in σ_V for $\pi \in (0,1)$, the Implicit Function Theorem ensures that $\pi \mapsto \hat{\psi}$ is strictly increasing.

Finally, we show that there exists a unique cutoff $\hat{\pi} \in (0,1)$ such that a full sharing equilibrium exists if and only if $\pi \in [0, \hat{\pi}]$. To see this, consider $\pi \mapsto \hat{\psi}$. Since $\hat{\psi}(\cdot)$ is continuous and strictly increasing, with $\hat{\psi}(0) = 0$ and $\hat{\psi}(1) = 1$, IVT ensures the existence and uniqueness of $\hat{\pi} \in (0,1)$ such that $\hat{\psi}(\hat{\pi}) = \hat{b}/(\hat{b}+\hat{\ell}) \in (0,1)$. Since a full sharing equilibrium requires $\hat{\psi} \leq \hat{b}/(\hat{b}+\hat{\ell})$, such an equilibrium can arise if and only if $\pi \in [0, \hat{\pi}]$. $\square$

Proof of part (ii): An equilibrium with $\sigma_U = 0$ requires prevalence $\psi \geq \hat{b}/(\hat{b}+\hat{\ell})$, whereas an equilibrium with "unverified following" necessitates $\psi \leq b/(b+\ell)$. For prevalence ψ in that range (which is non-empty by Assumption 1), users choosing not to verify find it (weakly) optimal to follow news. In this case, $\Sigma_V(\psi) \equiv G((1-\psi)\zeta\hat{b} + \psi\ell)$, which is strictly increasing as $\ell > \hat{b}$ by Assumption 1. Then, equation (8) can be written as:

$$\psi = \Psi(\pi, 0, \Sigma_V(\psi)). \tag{A.5}$$

The argument for existence of a fixed point $\hat{\psi} \in [0,1]$ solving (A.5) is analogous to that given in the proof of part (i), and thus omitted. For uniqueness, note that $\Sigma_V(\psi)$ is strictly increasing in ψ (as $\ell > \hat{b}$, by Assumption 1); thus, $\Psi(\pi, 0, \Sigma_V(\psi))$ is strictly decreasing in ψ for $\pi \in (0,1)$, as $\Psi(\pi, 0, \sigma_V)$ is strictly decreasing in σ_V for $\pi \in (0,1)$ (Lemma A.1). Thus, for any $\pi \in (0,1)$, $\hat{\psi}$ is unique by IVT. For $\pi \in \{0,1\}$, $\hat{\psi} = \pi$, as $\Psi \equiv \pi$ for those π values.

For the monotonicity and continuity of $\pi \mapsto \hat{\psi}$, observe that since Ψ and Σ_V are C^1, and $\Psi(\pi, 0, \sigma_V)$ is strictly increasing in π and strictly decreasing in σ_V for $\pi \in (0,1)$, the Implicit Function Theorem implies that $\hat{\psi}(\pi)$ is continuous and strictly increasing. As a result, since $\hat{\psi}(0) = 0$ and $\hat{\psi}(1) = 1$, IVT ensures the existence of cutoffs $\underline{\pi}, \bar{\pi} \in (0,1)$, with $\underline{\pi} < \bar{\pi}$, such that $\hat{\psi}(\underline{\pi}) = \hat{b}/(\hat{b} + \hat{\ell})$ and $\hat{\psi}(\bar{\pi}) = b/(b+\ell)$. Hence, an equilibrium with $\sigma_U = 0$ and $\rho \geq 0$ (unverified following) exists if and only if $\pi \in [\underline{\pi}, \bar{\pi}]$. In particular, for $\hat{\psi} = b/(b+\ell)$, users not verifying are indifferent between following and not following news, as $(1-\hat{\psi})b - \hat{\psi}\ell = 0$.

Therefore, by construction, for all $\pi > \bar{\pi}$, the equilibrium cannot exhibit $\sigma_U > 0$, or $\sigma_U = 0$ and $\rho > 0$. The only possibility then is $\sigma_U = \rho = 0$. In such a case, when no user engages in unverified sharing or unverified following, the mass of users who find it optimal to verify given ψ is the map $\psi \mapsto G((1-\psi)(\zeta\hat{b} + b))$. Since this function is C^1 and Ψ is bounded, the fixed point equation

$$\psi = \pi \frac{1 - \zeta G((1-\psi)(\zeta\hat{b} + b))}{1 - \zeta G((1-\psi)(\zeta\hat{b} + b))\pi} \tag{A.6}$$

admits a solution $\hat{\psi} \in [0,1]$. Also, for $\pi > \bar{\pi}$, any solution to (A.6) involves $\hat{\psi} > b/(b+\ell)$. Indeed, note that for $\pi > \bar{\pi}$ and $z = b/(b+\ell)$,

$$\Psi(\pi, 0, G((1-z)(\zeta\hat{b} + b))) > \Psi(\bar{\pi}, 0, \Sigma_V(z)) = z,$$

where the inequality holds from Ψ being strictly increasing in π, and the equality from $G((1-z)(\zeta\hat{b} + b)) = G((1-z)\zeta\hat{b} + \psi\ell) = \Sigma_V(z)$, and the definition of $\bar{\pi}$. $\square$

Proof of part (iii): Suppose $\psi = \hat{b}/(\hat{b} + \hat{\ell})$ so that users (not verifying news) are indifferent between unverified sharing and not sharing. Consequently, the mass of users who find it optimal to verify is constant and equals $\hat{\sigma}_V := G(\hat{b}(\ell + \zeta\hat{\ell})/(\hat{b} + \hat{\ell}))$. Suppose ς_U of those indifferent users break the indifference in favor of unverified sharing (and thus $1 - \hat{\sigma}_V - \varsigma_U$ breaks it in favor of not sharing). Then, an equilibrium with constant prevalence can be sustained, provided (9) holds, namely:

$$\Psi(\pi, \varsigma_U, \hat{\sigma}_V) = \hat{b}/(\hat{b} + \hat{\ell}),$$

where $0 \leq \varsigma_U \leq 1 - \hat{\sigma}_V$. Now, by the definitions of $\underline{\pi}$ and $\hat{\pi}$:

$$\Psi(\underline{\pi}, 0, \hat{\sigma}_V) = \Psi(\hat{\pi}, 1 - \hat{\sigma}_V, \hat{\sigma}_V) = \hat{b}/(\hat{b} + \hat{\ell}),$$

implying that $\hat{\pi} > \underline{\pi}$, since $\Psi(\pi, \sigma_U, \sigma_V)$ in (2) is strictly increasing in π but decreasing in σ_U. Also, since $\Psi(\pi, \cdot, \hat{\sigma}_V)$ is continuous and strictly decreasing for $\pi \in (0, 1)$, IVT ensures that for each $\pi \in (\underline{\pi}, \bar{\pi})$ there exists a unique $\varsigma_U(\pi) \in (0, 1 - \hat{\sigma}_V)$ with $\Psi(\pi, \varsigma_U(\pi), \hat{\sigma}_V) = \hat{b}/(\hat{b} + \hat{\ell})$. Moreover, $\varsigma_U(\pi)$ is strictly increasing in π, by the properties of Ψ in Lemma A.1. □

A.3 Proof of Theorem 1: Stationary Equilibria

To prove Theorem 1, we first characterize the conversion correspondence induced by the equilibrium in the sharing game. To this end, let $\hat{\sigma}_U(\pi) := 1 - G(\hat{\psi}_f(\pi)(\ell + \zeta\hat{\ell}))$, where $\hat{\psi}_f(\pi)$ is the unique solution to (6). Let $\hat{\sigma}_V(\pi) := G((1 - \hat{\psi}_n(\pi))\zeta\hat{b} + \hat{\psi}_n(\pi)\ell)$, where $\hat{\psi}_n(\pi)$ is the unique solution to (8). Finally, let $\varsigma_U(\pi)$ denote the unique solution to (9), and recall that for any sharing rates (σ_U, σ_V) conversions are given by $\rho = P(\sigma_U, \sigma_V)$ in (3).

Lemma A.2 (Conversion correspondence). *Let $\underline{\pi}, \hat{\pi}, \bar{\pi} \in (0, 1)$ as in Proposition 1.*

1. *In any full sharing equilibrium, conversions read $\rho = \hat{\rho}_{full}(\pi) := P(\hat{\sigma}_U(\pi), 1 - \hat{\sigma}_U(\pi)) > 0$, where $\hat{\rho}_{full} : [0, \hat{\pi}] \to \mathbb{R}_+$ is strictly decreasing and continuous;*

2. *In any equilibrium with no unverified sharing, $\rho = \hat{\rho}_o(\pi) := P(0, \hat{\sigma}_V(\pi)) > 0$ if $\pi < \bar{\pi}$, $\rho \in [0, \hat{\rho}_o(\bar{\pi})]$ if $\pi = \bar{\pi}$, and $\rho = 0$ if $\pi > \bar{\pi}$, where $\hat{\rho}_o : [\underline{\pi}, \bar{\pi}] \to [0, 1]$ is strictly decreasing and continuous;*

3. *In any equilibrium with partial sharing, $\rho = \hat{\rho}_{partial}(\pi) := P(\varsigma_U(\pi), 1 - \hat{\sigma}_U(\hat{\pi})) > 0$, where $\hat{\rho}_{partial} : [\underline{\pi}, \hat{\pi}] \to \mathbb{R}_+$ is strictly increasing and continuous.*

Moreover, $\hat{\rho}_{full}(\hat{\pi}) = \hat{\rho}_{partial}(\hat{\pi})$ and $\hat{\rho}_{partial}(\underline{\pi}) = \hat{\rho}_o(\underline{\pi})$.

Proof:

1. Full sharing: $\sigma_U + \sigma_V = 1$. By Proposition 1, this equilibrium can be sustained if and only if $\pi \in [0, \hat{\pi}]$. In this equilibrium, prevalence $\pi \mapsto \hat{\psi}_f$ is continuous and strictly increasing in π. Thus, the resulting unverified sharing rate, $\hat{\sigma}_U(\pi)$, is thus continuous and strictly decreasing in π. Thus, using (3), the total conversion value of a fake item, as a function of π, is continuous and given by:

$$\hat{\rho}_{\text{full}}(\pi) := P(\hat{\sigma}_U(\pi), 1 - \hat{\sigma}_U(\pi)) = \frac{\hat{\sigma}_U(\pi)}{1 - \zeta\hat{\sigma}_U(\pi)}. \tag{A.7}$$

Since $\hat{\psi}_f(0) = 0$, it follows that $\hat{\sigma}_U(0) = 1$, and so $\hat{\rho}(0) = 1/(1-\zeta) > 1$. Moreover, $\hat{\rho}_{\text{full}}(\pi)$ is strictly decreasing in π because $\hat{\sigma}_U$ is strictly decreasing in π.

2. No unverified sharing: $\sigma_U = 0$. By Proposition 1, this equilibrium can be sustained—along with non-zero conversions— if and only if $\pi \in [\underline{\pi}, \bar{\pi})$. In this equilibrium, prevalence $\hat{\psi}_n$ is continuous and strictly increasing in π, and the resulting verified sharing rate, $\hat{\sigma}_V(\pi)$, is continuous and strictly increasing in π. So, given (3), the total conversion value of a fake item is continuous and obeys

$$\hat{\rho}_o(\pi) := P(0, \hat{\sigma}_V(\pi)) = 1 - \hat{\sigma}_V(\pi), \tag{A.8}$$

which is strictly decreasing in π. Also, since $\hat{\psi}_n(\bar{\pi}) = b/(b+\ell)$, it follows that all users (not verifying) are indifferent between following and not following news; hence, $\rho \in [0, \hat{\rho}_o(\bar{\pi})]$. For $\pi > \bar{\pi}$, $\rho = 0$, as users not verifying prefer not following news.

3. Partial sharing: By Proposition 1, this equilibrium can be sustained if and only if $\pi \in [\underline{\pi}, \hat{\pi}]$. In this equilibrium, prevalence is pegged at $\hat{b}/(\hat{b}+\hat{\ell})$ for all $\pi \in [\underline{\pi}, \hat{\pi}]$. Thus, the sharing rate of verified news is also constant in π and given by $1-\hat{\sigma}_U(\hat{\pi})$. As shown in the proof of Proposition 1-(iii), the unverified sharing $\varsigma_U(\pi)$ uniquely solves $\Psi(\pi, \varsigma_U(\pi), 1-\hat{\sigma}_U(\hat{\pi})) = \hat{b}/(\hat{b}+\hat{\ell})$. This solution is strictly increasing and continuous in π. Thus, the total conversion value (3) is continuous and strictly increasing in π, given by

$$\hat{\rho}_{\text{partial}}(\pi) := P(\varsigma_U(\pi), 1-\hat{\sigma}_U(\hat{\pi})) = \frac{\hat{\sigma}_U(\hat{\pi})}{1-\zeta\varsigma_U(\pi)}. \tag{A.9}$$

Finally, since $\varsigma_U(\underline{\pi}) = 0$ and $\varsigma_U(\hat{\pi}) = \hat{\sigma}_U(\hat{\pi})$, it follows from (A.7)–(A.9) that $\hat{\rho}_{\text{full}}(\hat{\pi}) = \hat{\rho}_{\text{partial}}(\hat{\pi})$ and $\hat{\rho}_{\text{partial}}(\underline{\pi}) = \hat{\rho}_{\text{o}}(\underline{\pi})$. This completes the proof. $\square$

Before proving Theorem 1, let us derive the inverse supply function, which is used in the proof. To this end, recall that producer costs are distributed according to a strictly increasing CDF H with full support on $[0, +\infty)$. Since the distribution may contain atoms, given conversions $\rho \geq 0$, producers with costs $c < \rho$ find it optimal to produce, producers with $c > \rho$ strictly prefer not to produce, and producers $c = \rho$ are indifferent. Thus, the supply correspondence is given by:

$$\pi = \begin{cases} H(\rho), & \text{if } H(\rho^-) = H(\rho), \\ [H(\rho^-), H(\rho)], & \text{if } H(\rho^-) < H(\rho). \end{cases}$$

where $H(\rho-) = \lim_{c\uparrow\rho} H(c)$. At continuity points of H, this reduces to $\pi = H(\rho)$, and so the *inverse supply* obeys $\rho_S(\pi) \equiv H^{-1}(\pi)$. At an atom ρ_a, there exists a production interval $[H(\rho_a-), H(\rho_a)]$ such that $\rho_S(\pi) \equiv \rho_a$ for all $\pi \in [H(\rho_a-), H(\rho_a)]$. It follows that $\rho_S(\pi)$ is increasing and continuous, exhibiting flat regions if atoms exist. In fact, $\rho_S(\pi)$ coincides with the generalized inverse of H, namely, $\inf\{\rho \geq 0 : H(\rho) \geq \pi\}$.

Proof of Theorem 1: Consider the inverse supply function $\rho_S(\pi)$ described above, which is continuous and increasing, with $\rho_S(0) = 0$. WLOG assume that $\hat{\pi} \leq \bar{\pi}$. (The proof for the case $\hat{\pi} > \bar{\pi}$ is analogous, and thus omitted.) Thus, if an equilibrium exists, it cannot have a production of fake content above $\bar{\pi}$ because the conversion correspondence vanishes for $\pi > \bar{\pi}$ (Proposition A.2), whereas the inverse supply $\rho_S(\pi) > 0$ for all $\pi > \bar{\pi}$.

We will show that an equilibrium with production in $[0, \bar{\pi}]$ always exists. To this end, consider $\hat{\rho}_{\text{full}}$, $\hat{\rho}_{\text{o}}$, and $\hat{\rho}_{\text{partial}}$ as defined in Proposition A.2. We will proceed by cases.

1. Case 1: Suppose $\rho_S(\hat{\pi}) \geq \hat{\rho}_{\text{full}}(\hat{\pi})$. Then existence follows from the Intermediate Value Theorem (IVT) applied to $\hat{\rho}_{\text{full}}$ and ρ_S on $[0, \hat{\pi}]$, since $\hat{\rho}_{\text{full}}(0) = 1/(1-\zeta) > 0 = \rho_S(0)$.

2. Case 2: Suppose $\rho_S(\hat{\pi}) \in (\hat{\rho}_{\text{o}}(\hat{\pi}), \hat{\rho}_{\text{full}}(\hat{\pi}))$. Then, if $\rho_S(\underline{\pi}) \geq \hat{\rho}_{\text{partial}}(\underline{\pi})$, then existence follows from IVT applied to $\hat{\rho}_{\text{partial}}$ and ρ_S on $[\underline{\pi}, \hat{\pi}]$, since $\rho_S(\hat{\pi}) < \hat{\rho}_{\text{full}}(\hat{\pi}) = \hat{\rho}_{\text{partial}}(\hat{\pi})$ (Proposition A.2). Conversely, if $\rho_S(\underline{\pi}) < \hat{\rho}_{\text{partial}}(\underline{\pi})$ then existence follows from IVT applied to $\hat{\rho}_{\text{o}}$ and ρ_S on $[\underline{\pi}, \hat{\pi}]$, as $\hat{\rho}_{\text{partial}}(\underline{\pi}) = \hat{\rho}_o(\underline{\pi})$ by Proposition A.2.

3. Case 3: Suppose that $\rho_S(\hat{\pi}) \leq \hat{\rho}_o(\hat{\pi})$. If $\rho_S(\bar{\pi}) > \hat{\rho}_o(\bar{\pi})$, existence follows from applying IVT to $\hat{\rho}_o$ and ρ_S on $[\hat{\pi}, \bar{\pi}]$. Conversely, $\rho_S(\bar{\pi}) \in [0, \hat{\rho}_o(\bar{\pi})]$, and thus $\rho_S(\bar{\pi})$ belongs the conversion correspondence at $\pi = \bar{\pi}$ (Proposition A.2). $\square$

Corollary A.1. *Suppose Assumption 1-(ii.2) is replaced by $b/(b+\ell) < \hat{b}/(\hat{b}+\hat{\ell})$, while maintaining $\ell > b > \hat{b} > 0$. Let $\hat{\psi}(\pi)$ denote the solution to (6), and let $\hat{\sigma}_U(\pi) = 1 - G(\hat{\psi}(\pi)(\ell + \zeta\hat{\ell}))$. Then there exists a unique production level $\pi_F \in (0,1)$, defined by $\hat{\psi}(\pi_F) = b/(b+\ell)$, such that the conversion correspondence obeys $\rho = \hat{\rho}_{full}(\pi)$ (A.7) for all $\pi < \pi_F$. At $\pi = \pi_F$, the conversion correspondence contains the vertical segment $\rho \in \left[0, \frac{\hat{\sigma}_U(\pi_F)}{1-\zeta\hat{\sigma}_U(\pi_F)}\right]$. Finally, for every $\pi > \pi_F$, all conversion values are zero, $\rho = 0$.*

Proof: Since $\hat{b}/(\hat{b}+\hat{\ell}) > b/(b+\ell)$, users who do not verify stop following news before they stop sharing it. Moreover, strictly positive conversions require that some users follow news without verifying it, which can occur only if prevalence $\psi \leq b/(b+\ell)$. In this case, these same users also strictly prefer unverified sharing to not sharing. In other words, whenever $\psi < b/(b+\ell)$, users who do not verify both share and follow, and the verification cutoff is exactly the same as in the full-sharing branch of the baseline model, namely, $t \leq \psi(\ell + \zeta\hat{\ell})$.

Thus, for any π supporting positive conversions, the sharing-game equilibrium must satisfy

$$\sigma_U = 1 - G(\psi(\ell + \zeta\hat{\ell})), \qquad \psi = \Psi(\pi, \sigma_U, 1 - \sigma_U),$$

which jointly lead to (6). Let $\hat{\psi}(\pi)$ and $\hat{\sigma}_U(\pi)$ denote the corresponding solution. By the proof of Proposition 1-(i), $\hat{\psi}(\pi)$ is continuous and strictly increasing in π, with $\hat{\psi}(0) = 0$ and $\hat{\psi}(1) = 1$. Hence, there is a unique $\pi_F \in (0,1)$ such that $\hat{\psi}(\pi_F) = b/(b+\ell)$. Thus, $\hat{\psi}(\pi) < b/(b+\ell)$ if and only if $\pi < \pi_F$. For all such π, the conversion correspondence has the same form as in the full-sharing branch of the baseline model: $\rho = \frac{\hat{\sigma}_U(\pi)}{1-\zeta\hat{\sigma}_U(\pi)}$. At $\pi = \pi_F$, users who do not verify are indifferent between following and not following news. Since they still strictly prefer sharing, the conversion correspondence obeys $\rho \in \left[0, \frac{\hat{\sigma}_U(\pi_F)}{1-\zeta\hat{\sigma}_U(\pi_F)}\right]$.

Finally, no sharing-game equilibrium with $\pi > \pi_F$ can generate strictly positive conversions. To see this, suppose otherwise. If conversions are strictly positive, some non-verifying users must follow unverified news, so $\psi \leq b/(b+\ell)$. If $\psi = b/(b+\ell)$, then the preceding characterization implies $\pi = \pi_F$, contradicting $\pi > \pi_F$. If instead $\psi < b/(b+\ell)$, then the equilibrium must lie on the full-sharing branch described above, so $\pi < \pi_F$, again a contradiction. Thus, for every $\pi > \pi_F$, the conversion correspondence vanishes, $\rho \equiv 0$. □

A.4 Proof of Proposition 2: Welfare Effects

Lemma A.3. *Consider the characterization of the sharing game in Proposition 1.*

(i) In any equilibrium with full sharing and prevalence ψ, welfare per news equals $\mathcal{W}_{full}(\psi) = (1-\psi)(b+\zeta\hat{b}) - \psi(\ell+\zeta\hat{\ell}) + \int_0^{\psi(\ell+\zeta\hat{\ell})} G(t)dt$.

(ii) In any equilibrium with no unverified sharing—but with non-zero conversions—and prevalence ψ, welfare per news equals $\mathcal{W}_o(\psi) = (1-\psi)b - \psi\ell + \int_0^{\zeta\hat{b}+\psi(\ell-\zeta\hat{b})} G(t)dt$.

(iii) In any equilibrium with partial sharing, welfare per news $\mathcal{W}_{partial} = \mathcal{W}_{full}(\hat{b}/(\hat{b}+\hat{\ell}))$.

Proof:

(i) As discussed in the main text, in a full sharing equilibrium ($\sigma_U + \sigma_V = 1$) with prevalence ψ, users find it optimal to verify when their verification cost $t \leq \psi(\ell+\zeta\hat{\ell})$. Moreover, those users who do not verify find it optimal to both share and follow news.

Thus, given prevalence ψ, total welfare per news equals

$$\begin{aligned}\mathcal{W}_{\text{full}}(\psi) &= (1-\psi)(b+\zeta\hat{b}) - \psi(\ell+\zeta\hat{\ell})[1-G(\psi(\ell+\zeta\hat{\ell}))] - \int_0^{\psi(\ell+\zeta\hat{\ell})} tdG(t) \\ &= (1-\psi)(b+\zeta\hat{b}) - \psi(\ell+\zeta\hat{\ell}) + \int_0^{\psi(\ell+\zeta\hat{\ell})} G(t)dt,\end{aligned}$$

where the second equality follows from integration by parts.

(ii) Consider an equilibrium with $\sigma_U = 0$ but with non-zero conversions. As outlined in the main text, in this equilibrium, users who choose not to verify find it optimal to follow news. Thus, given prevalence ψ, users with verification cost $t \leq (1-\psi)\zeta\hat{b} + \psi\ell =: \hat{t}$ find it optimal to verify. Thus, total welfare per news equals

$$\begin{aligned}\mathcal{W}_{\text{o}}(\psi) &= (1-\psi)b + (1-\psi)\zeta\hat{b}G(\hat{t}) - \psi\ell[1-G(\hat{t})] - \int_0^{\hat{t}} tdG(t) \\ &= (1-\psi)b - \psi\ell + \hat{t}G(\hat{t}) - \int_0^{\hat{t}} tdG(t) = (1-\psi)b - \psi\ell + \int_0^{\hat{t}} G(t)dt,\end{aligned}$$

where the last equality follows from integration by parts.

(iii) In a partial sharing equilibrium, prevalence is fixed at $\psi = \hat{b}/(\hat{b}+\hat{\ell})$. Thus, the mass of users verifying is fixed too, and given by $\hat{\sigma}_V = G(\psi(\ell+\zeta\hat{\ell}))$, while the remaining mass $1-\hat{\sigma}_V$ splits between unverified sharing and not sharing at all, namely, $\sigma_U+\sigma_N = 1-\hat{\sigma}_V$. All users not verifying find it optimal to follow news. So, welfare per news equals:

$$\begin{aligned}\mathcal{W}_{\text{partial}} &= (1-\psi)b + (1-\psi)\zeta\hat{b}(\sigma_U+\hat{\sigma}_V) - \psi(\ell+\zeta\hat{\ell})\sigma_U - \psi\ell\sigma_N - \int_0^{\psi(\ell+\zeta\hat{\ell})} tdG(t) \\ &= (1-\psi)b + [(1-\psi)\zeta\hat{b} - \psi(\ell+\zeta\hat{\ell})](\sigma_U+\hat{\sigma}_V) - \psi\ell\sigma_N + \int_0^{\psi(\ell+\zeta\hat{\ell})} G(t)dt \\ &= (1-\psi)b - \psi\ell + \int_0^{\psi(\ell+\zeta\hat{\ell})} G(t)dt,\end{aligned}$$

where the second equality follows from integration by parts, while the third one from prevalence satisfying $(1-\psi)\hat{b} - \psi\hat{\ell} = 0$. Thus, $\mathcal{W}_{\text{partial}} = \mathcal{W}_{\text{full}}(\hat{b}/(\hat{b}+\hat{\ell}))$. □

Proof of Proposition 2: The first part of the proof consists of ranking full sharing equilibria with any other type of equilibria in terms of welfare per news and total welfare. We first establish the ranking in terms of welfare per news and then in terms of total volume. The conclusion on total welfare then follows immediately. The second part of the proof ranks equilibria along a fixed supply curve.

Welfare per news. Using the welfare per news expressions from Lemma A.3, we see that $\mathcal{W}_{\text{full}}(\cdot)$ and $\mathcal{W}_{\text{o}}(\cdot)$ are both strictly decreasing. Indeed,

$$\begin{aligned}\frac{\partial \mathcal{W}_{\text{full}}}{\partial \psi} &= -(b+\zeta\hat{b}) - (\ell+\zeta\hat{\ell})[1-G(\psi(\ell+\zeta\hat{\ell}))] < 0 \\ \frac{\partial \mathcal{W}_{\text{o}}}{\partial \psi} &= -b - \zeta\hat{b}G(\hat{t}) - \ell[1-G(\hat{t})] < 0, \quad \text{where} \quad \hat{t} = \zeta\hat{b} + \psi(\ell - \zeta\hat{b}).\end{aligned}$$

Next, by Proposition 1, a full sharing equilibrium can be sustained iff $\pi \in [0, \hat{\pi}]$, whereas an equilibrium with conversions but no unverified sharing can be sustained iff $\pi \in [\underline{\pi}, \bar{\pi}]$. Since the equilibrium prevalence is strictly increasing in π, for each type of equilibria, it follows that $\mathcal{W}_{\text{full}}$ is *minimized* when $\pi = \hat{\pi}$, in which case, prevalence equals $\hat{b}/(\hat{b}+\hat{\ell})$. By contrast, $\mathcal{W}_{\text{o}}$ is *maximized* when $\pi = \underline{\pi}$, in which case prevalence again takes the value $\hat{b}/(\hat{b}+\hat{\ell})$. Finally, using the expressions $\mathcal{W}_{\text{full}}(\cdot)$, $\mathcal{W}_{\text{partial}}$, and $\mathcal{W}_{\text{o}}(\cdot)$ from Lemma A.3, it follows that

$$\mathcal{W}_{\text{full}}(\hat{b}/(\hat{b}+\hat{\ell})) = \mathcal{W}_{\text{o}}(\hat{b}/(\hat{b}+\hat{\ell})) = \mathcal{W}_{\text{partial}}.$$

Thus, in the sharing game, any equilibrium exhibiting full sharing delivers higher welfare per news than any equilibrium with partial sharing, which, in turn, generates more welfare per news than any equilibrium featuring unverified sharing (and non-zero conversions).

Volume. We now establish that any equilibrium with full sharing generates more news volume than any equilibrium with partial sharing or with no unverified sharing. To this end, use (2) and (A.3) to express the *total volume* of news circulating in the platform in terms of the equilibrium triplet $(\sigma_U, \sigma_V, \psi)$ arising from the sharing game:

$$TV(\sigma_U, \sigma_V, \psi) := T + F = \frac{1}{1-\zeta(\sigma_U + (1-\psi)\sigma_V)}. \tag{A.10}$$

In a full sharing equilibrium, $\sigma_U + \sigma_V = 1$, and so total volume (A.10) can be written as

$$TV_{full}(\sigma_V, \psi) := TV(1-\sigma_V, \sigma_V, \psi) = \frac{1}{1-\zeta+\zeta\psi\sigma_V}.$$

Since a full sharing equilibrium is sustained iff $\pi \in [0, \hat{\pi}]$, and both the equilibrium prevalence ψ and σ_V are strictly increasing in π, it follows that TV_{full} is strictly decreasing in π, achieving its *minimum* when $\pi = \hat{\pi}$ at which point prevalence $\psi = \hat{b}/(\hat{b}+\hat{\ell})$. Since in a full-sharing equilibrium both welfare per news and total volume fall as π rises, it follows that total welfare falls in π, as documented in Remark 1.

In contrast, in a partial sharing equilibrium, prevalence is fixed at $\psi = \hat{b}/(\hat{b}+\hat{\ell})$, while σ_V is fixed at $\hat{\sigma}_V = G(\psi(\ell+\hat{\ell}))$. The unverified sharing rate $\sigma_U \in [0, 1-\hat{\sigma}_V]$ is strictly increasing in π (to keep prevalence constant). Hence, by (A.10), the total volume of news, $TV_{partial}(\sigma_U) := TV(\sigma_U, \hat{\sigma}_V, \hat{b}/(\hat{b}+\hat{\ell}))$, where σ_U solves (9) (see Proposition 1), is strictly increasing in π, reaching its *maximum* when $\pi = \hat{\pi}$ and so $\sigma_U = 1-\hat{\sigma}_V$. Moreover, notice that $TV_{partial}(1-\hat{\sigma}_V) = TV_{full}(\hat{\sigma}_V, \hat{b}/(\hat{b}+\hat{\ell}))$. Thus, any equilibrium with full sharing generates more news volume than any equilibrium with partial sharing.

Finally, for equilibrium with no unverified sharing (and non-zero conversions) (A.10) yields $TV_o(\sigma_V, \psi) := 1/(1-\zeta(1-\psi)\sigma_V)$. Let $\hat{x} := \hat{b}/(\hat{b}+\hat{\ell})$ and $\sigma_{V,\hat{x}} := G(\hat{x}(\ell+\zeta\hat{\ell}))$. Since in this equilibrium, $\psi \geq \hat{x}$, we have

$$1-\zeta(1-\psi)\sigma_V \geq 1-\zeta+\zeta\hat{x} \geq 1-\zeta+\zeta\hat{x}\sigma_{V,\hat{x}} \implies TV_o \leq \frac{1}{1-\zeta+\zeta\hat{x}\sigma_{V,\hat{x}}}.$$

Thus, $TV_o(\sigma_V, \psi)$ is less than the lowest volume that can arise in a full sharing equilibrium.

We conclude that any equilibrium with full sharing generates more (per news and total) welfare than any possible equilibria of the other two types.

Supply. Suppose multiple stationary equilibria exist, and let $\pi^*_{\text{full}}, \pi^*_{\text{partial}},$ and π^*_{o} denote the equilibrium production of fake news in the full, partial, and no-unverified sharing stationary equilibria. By Proposition 1, these levels must belong to interval $I = [\underline{\pi}, \min\{\hat{\pi}, \bar{\pi}\}]$. By Proposition A.2, the conversion correspondence obeys $\min_{\pi\in I}\hat{\rho}_{\text{full}}(\pi) \geq \max_{\pi\in I}\hat{\rho}_{\text{partial}}(\pi)$ and $\min_{\pi\in I}\hat{\rho}_{\text{partial}}(\pi) \geq \max_{\pi\in I}\hat{\rho}_{\text{o}}(\pi)$. Moreover, since the supply curve is strictly increasing, $\pi^*_{\text{full}} \geq \pi^*_{\text{partial}} \geq \pi^*_{\text{o}}$ (with at least one strict inequality), and the conversion values obey $\hat{\rho}_{\text{full}}(\pi^*_{\text{full}}) \geq \hat{\rho}_{\text{partial}}(\pi^*_{\text{partial}}) \geq \hat{\rho}_{\text{o}}(\pi^*_{\text{o}})$ (with at least one strict inequality). Part (i) allows us to conclude that total welfare is higher for $\pi = \pi^*_{\text{full}}$ than for $\pi = \pi^*_{\text{partial}}$, as the former equilibrium features full sharing, whereas the latter exhibits partial sharing only.

To complete the proof, we will show that total welfare is the lowest for $\pi = \pi^*_{\text{o}}$. To see this, note that, given (A.3), the equilibrium with no-unverified sharing ($\sigma_U = 0$) has the lowest total volume of fake news, $F = \pi^*_{\text{o}}$. At the same time, Proposition 1 establishes that this equilibrium has the highest prevalence level $\psi \geq \hat{b}/(\hat{b}+\hat{\ell})$. Since prevalence equals the ratio of fake to total news, this implies that this equilibrium must have the lowest amount of total news. As previously shown, any equilibrium with partial sharing delivers more welfare per news than any equilibrium with no-unverified sharing; it follows that both welfare per news and total volume are the lowest in the no-unverified sharing equilibrium. $\square$

B Proofs of Section 5: Policies

Preliminaries. In this section, we focus on full sharing equilibria (i.e., $\sigma_U + \sigma_V = 1$), and so the expressions for the stock of truthful and fake content are given by:

$$F = \frac{\pi}{1 - \zeta\sigma_U} \qquad \text{and} \qquad T = \frac{1-\pi}{1-\zeta}. \tag{B.1}$$

The prevalence function $\Psi(\pi, \sigma_U, \sigma_V)$ takes the form:

$$\tilde{\Psi}(\pi, \sigma_U) := \Psi(\pi, \sigma_U, 1 - \sigma_U) = \frac{(1-\zeta)\pi}{1 - \zeta(\pi + (1-\pi)\sigma_U)}, \tag{B.2}$$

which is increasing in both π and σ_U (Lemma A.1). Welfare per news is given by:

$$w(\psi) := \mathcal{W}_{\text{full}}(\psi) = (1-\psi)(b + \zeta\hat{b}) - \int_0^{\psi(\ell + \zeta\hat{\ell})} [1 - G(t)]dt. \tag{B.3}$$

Total welfare is $W := w(\psi)(T+F)$, and since $1 - \psi = T/(T+F)$, total welfare can be recast as:

$$W = (b + \zeta b)T - L(\psi)(T+F), \tag{B.4}$$

where $L(\psi) := \int_0^{\psi(\ell+\zeta\hat{\ell})}[1 - G(t)]dt$ is increasing in ψ.

B.1 Proof of Proposition 3: Lowering Verification Costs

Consider a family $\{G(\cdot;\eta)|\eta \in \mathbb{R}_+\}$ capturing possible distributions for verification costs. We assume that higher values of η reduce such costs in the sense of (reversed) first-order stochastic dominance, i.e., $G(\cdot;\eta) \geq G(\cdot;\eta')$, for all $\eta > \eta'$.

Fix η and a production level π for which a full-sharing equilibrium exists. For each π in this range, we denote by $\hat{\sigma}_U(\pi;\eta)$ and $\hat{\psi}(\pi;\eta)$ the rate of unverified sharing and prevalence level that arises in the sharing game (where π is fixed). Using (B.2), pair $(\hat{\sigma}_U, \hat{\psi})$ solves:

$$\psi = \tilde{\Psi}(\pi, \sigma_U) \qquad \text{and} \qquad \sigma_U = 1 - G\left(\psi(\ell + \hat{\ell});\eta\right).$$

The former map $\sigma_U \mapsto \Psi(\pi, \sigma_U, 1 - \sigma_U)$ is strictly increasing (Lemma A.1), while the latter one $\psi \mapsto 1 - G(\psi(\ell + \hat{\ell});\eta)$ is strictly decreasing. Hence, in the (σ_U, ψ)-space, an increase in η shifts down the unverified sharing curve (pointwise in ψ), without directly affecting the increasing prevalence curve. Thus, for each π, both $\hat{\sigma}_U(\pi;\eta)$ and $\hat{\psi}(\pi;\eta)$ fall in η.

Next, consider the full-sharing conversion locus, namely,

$$\rho = \frac{\hat{\sigma}_U(\pi;\eta)}{1-\zeta\hat{\sigma}_U(\pi;\eta)}.$$

Since for each π, $\hat{\sigma}_U(\pi;\eta)$ falls in η, it follows that $\hat{\rho}_{\text{full}}(\pi;\eta)$ falls in η. Thus, in the (π,ρ)-space, lower verification costs shift the full-sharing conversion locus downward.

The supply curve, i.e., $\pi = H(\rho)$, is unchanged. As a result, the new stationary equilibrium features less production of fake news π^* and lower conversion value ρ^*. Since, in equilibrium, $\rho^* = \sigma_U^*/(1-\zeta\sigma_U^*)$, it follows that the unverified sharing rate σ_U^* falls. Prevalence ψ^* falls, since both π^* and σ_U^* fall and $\psi^* = \Psi(\pi^*, \sigma_U^*, 1-\sigma_U^*)$.

Now, consider welfare per news under full sharing:

$$w(\psi^*;\eta) = (1-\psi^*)(b+\zeta\hat{b}) - \int_0^{\psi^*(\ell+\zeta\hat{\ell})}[1-G(t;\eta)]dt.$$

For fixed ψ, increasing η weakly lowers the integrand by definition of the family $\eta \mapsto G(\cdot;\eta)$. In addition, welfare per news is strictly decreasing in ψ. Since equilibrium prevalence ψ^* falls as η rises, welfare per news $w(\psi^*;\eta)$ rises.

Finally, consider total welfare. If total volume T^*+F^* rises, then total welfare clearly rises, as $W^* = w(\psi^*;\eta)(T^*+F^*)$. Otherwise, if T^*+F^* falls, then W^* rises too, by using decomposition (B.4). Indeed, if T^* rises but (T^*+F^*) falls, then the benefit term $(b+\zeta b)T^*$ rises, while the loss terms $L(\psi^*;\eta)(T^*+F^*)$ falls. Altogether, W^* rises. □

B.2 Proof of Proposition 4: News Longevity

Preliminaries. Fix a production level π for which a full-sharing equilibrium exists. This equilibrium is characterized by the pair $(\hat{\sigma}_U(\pi;\zeta), \hat{\psi}(\pi;\zeta))$ that solves:

$$\psi = \Psi(\pi, \sigma_U, 1-\sigma_U;\zeta) \qquad \text{and} \qquad \sigma_U = 1 - G\big(\psi(\ell+\zeta\hat{\ell})\big). \tag{B.5}$$

Recall from Lemma A.1 that $\Psi_\pi > 0$ and $\Psi_{\sigma_U} > 0$. Also, direct computations show that

$$\Psi_\zeta = -\frac{\pi(1-\pi)(1-\sigma_U)}{\big[1-\zeta(\pi+(1-\pi)\sigma_U)\big]^2} < 0.$$

Proof of Proposition 4-(i): We examine the effects of increasing ζ for a fixed production level π, reflecting a perfectly inelastic supply curve. This reduces to study the effects of ζ on the sharing game equilibrium characterized by (B.5). First, we examine prevalence effects.

Differentiating (B.5) in ζ yields:

$$\frac{\partial\hat{\psi}}{\partial\zeta} = \Psi_\zeta + \Psi_{\sigma_U}\left[-g\frac{\partial\hat{\psi}}{\partial\zeta}(\ell+\zeta\hat{\ell}) - g\hat{\psi}\hat{\ell}\right],$$

which gives:

$$\frac{\partial\hat{\psi}(\pi;\zeta)}{\partial\zeta} = \frac{\Psi_\zeta - g\hat{\psi}\hat{\ell}\Psi_{\sigma_U}}{1+g(\ell+\zeta\hat{\ell})\Psi_{\sigma_U}} < 0, \tag{B.6}$$

where the sign follows from $\Psi_{\sigma_U} > 0 > \Psi_\zeta$. So, an increase in ζ lowers prevalence.

Next, because π is fixed, the stock of truthful news $T = (1-\pi)/(1-\zeta)$ (B.1) clearly rises. As for welfare per news (B.3), notice that $w_\zeta(\psi;\zeta) > 0$. Indeed,

$$\begin{aligned} w_\zeta(\psi;\zeta) &= (1-\psi)\hat{b} - [1-G(\psi(\ell+\zeta\hat{\ell}))]\psi\hat{\ell} \\ &= [(1-\psi)\hat{b} - \psi\hat{\ell}] + G(\psi(\ell+\zeta\hat{\ell}))\psi\hat{\ell} > 0. \end{aligned}$$

This expression is positive because in a full-sharing equilibrium, $\psi \leq \hat{b}/(\hat{b}+\hat{\ell})$. Thus, since an increase in ζ lowers prevalence $\hat{\psi}$, it follows that welfare per news $w(\hat{\psi};\zeta)$ rises.

We now examine the effects on conversions. To this end, assume (12). Conversions are given by the conversion "demand" locus:

$$\rho^D(\pi;\zeta) = \frac{\hat{\sigma}_U(\pi;\zeta)}{1-\zeta\hat{\sigma}_U(\pi;\zeta)}$$

Differentiating the conversion demand in ζ yields:

$$\frac{\partial\rho^D(\pi;\zeta)}{\partial\zeta} = \frac{1}{(1-\zeta\hat{\sigma}_U)^2}\left(\hat{\sigma}_U^2 + \frac{\partial\hat{\sigma}_U}{\partial\zeta}\right) = \frac{1}{(1-\zeta\hat{\sigma}_U)^2}\left(\hat{\sigma}_U^2 - g\frac{\partial\hat{\psi}}{\partial\zeta}(\ell+\zeta\hat{\ell}) - g\hat{\psi}\hat{\ell}\right). \tag{B.7}$$

Since $\partial\hat{\psi}/\partial\zeta < 0$ we have that:

$$\frac{\partial\hat{\rho}(\pi;\zeta)}{\partial\zeta} > \frac{1}{(1-\zeta\hat{\sigma}_U)^2}\left(\hat{\sigma}_U^2 - g\hat{\psi}\hat{\ell}\right) > 0, \tag{B.8}$$

where the last inequality follows from (12).

The stock of fake news (B.1) can be recast in terms of conversions as

$$F = \frac{\pi}{1-\zeta\sigma_U} = \pi(1+\zeta\rho). \tag{B.9}$$

Because conversions increase in ζ (under (12)) and π is fixed, the stock of fake news rises.

Finally, total welfare rises since T and F rise in ζ, and also welfare per news rises. $\square$

Proof of Proposition 4-(ii): We now examine the effects of ζ under a perfectly elastic supply curve, where conversions are effectively fixed at some value $\rho^* > 0$.

Because ρ^* is fixed, the equilibrium unverified sharing rate σ_U^* falls as ζ rises, since:

$$\rho^* = \frac{\sigma_U^*}{1 - \zeta\sigma_U^*} \qquad \Longrightarrow \qquad \sigma_U^* = \frac{\rho^*}{1 + \zeta\rho^*} \quad \text{and} \quad \frac{d\sigma_U^*}{d\zeta} = \frac{-\rho^{*2}}{(1 + \zeta\rho^*)^2} < 0.$$

Next, we examine prevalence effects. To this end, recall that in a full-sharing equilibrium, $\sigma_U^* \equiv 1 - G(\psi^*(\ell + \zeta\hat{\ell}))$. Thus, we can differentiate this in ζ, using (B.2), to get:

$$\frac{d\psi^*}{d\zeta} = \frac{\sigma_U^{*2} - g\psi^*\hat{\ell}}{g(\ell + \zeta\hat{\ell})}. \tag{B.10}$$

So, under (12), the numerator of (B.10) is positive, implying that prevalence ψ^* rises in ζ.

The effects on production π^* follow from the relation $\psi^* \equiv \Psi(\pi^*, \sigma_U^*, 1 - \sigma_U^*; \zeta)$. Since $\Psi_\pi, \Psi_{\sigma_U} > 0 > \Psi_\zeta$, we know that if ψ^* rises in ζ, then π^* must also rise in ζ, as unverified sharing σ_U^* falls in ζ. Thus, under (12), production π^* rises.

Finally, the stock of fake news F^* clearly rises, given (B.9), as π^* rises under (12). $\square$

B.3 Proof of Proposition 5: Algorithmic Filters

Preliminaries. Suppose that truthful news always passes the filter, while a fake news item is detected and removed with probability $\phi \in [0, 1]$. Thus, conditional on production π, the mass of fake news that reaches the platform after filtering is $(1 - \phi)\pi$, while the mass of truthful news remains $1 - \pi$. Let

$$\Delta(\pi, \phi) := \frac{(1 - \phi)\pi}{1 - \phi\pi}, \qquad 1 - \Delta(\pi, \phi) = \frac{1 - \pi}{1 - \phi\pi}, \tag{B.11}$$

denote the fraction of fake and truthful news among fresh items that pass the filter. Since

$$\Delta_\phi(\pi, \phi) = -\frac{\pi(1 - \pi)}{(1 - \phi\pi)^2} < 0, \tag{B.12}$$

a better filter lowers/raises the share of fake/truthful news among items that pass the filter.

To compute the misinformation prevalence, let F_τ denote the stock of fake news in period τ. Then, the following law of motion holds:

$$F_{\tau+1} = \zeta\sigma_U F_\tau + (1 - \phi)\pi.$$

Similarly, let T_τ represent the stock of truthful news in period τ. Then:

$$T_{\tau+1} = \zeta(\sigma_U + \sigma_V)T_\tau + 1 - \pi.$$

In steady state:

$$F = \frac{(1-\phi)\pi}{1-\zeta\sigma_U} \qquad \text{and} \qquad T = \frac{1-\pi}{1-\zeta(\sigma_U+\sigma_V)}.$$

Therefore, prevalence $\psi = F/(T+F)$ is given by:

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

Using Δ in (B.11), it follows that $\psi = \Psi(\Delta(\pi;\phi), \sigma_U, \sigma_V)$, where Ψ is in (2).

Henceforth, we focus on full-sharing equilibria, so $\sigma_U + \sigma_V = 1$, which is uniquely determined by:

$$\psi = \tilde{\Psi}(\Delta, \sigma_U), \qquad \sigma_U = 1 - G(\psi(\ell + \zeta\hat{\ell})), \tag{B.13}$$

where $\tilde{\Psi}(\Delta, \sigma_U) \equiv \Psi(\Delta, \sigma_U, 1-\sigma_U)$; see (B.2). We denote the solution to (B.13) by $\hat{\psi}(\Delta)$ and $\hat{\sigma}_U(\Delta)$, which reflects that, on the full-sharing branch, the filter affects user behavior only through Δ. The functions $\hat{\psi}$ and $\hat{\sigma}_U$ obey natural monotonicity properties: $\hat{\psi}(\Delta)$ and $\hat{\sigma}_U(\Delta)$ are increasing and decreasing in Δ, respectively. Indeed, in the (σ_U, ψ)-space, an increase in Δ shifts up the upward sloping curve $\sigma_U \mapsto \tilde{\Psi}$ along the fixed downward sloping locus $\sigma_U \mapsto 1 - G(\psi(\ell + \zeta\hat{\ell}))$. Hence, the new intersection exhibits more $\hat{\psi}$ and less $\hat{\sigma}_U$.

The conversion demand locus can be written in terms of Δ and ϕ as follows:

$$\rho^D(\Delta, \phi) := P(\hat{\sigma}_U(\Delta), 1 - \hat{\sigma}_U(\Delta); \phi) = \frac{(1-\phi)\hat{\sigma}_U(\Delta)}{1-\zeta\hat{\sigma}_U(\Delta)}. \tag{B.14}$$

Lemma B.1. *Fix $\pi \in (0,1)$. Then, an increase in filter quality ϕ lowers prevalence $\hat{\psi}$ and raises unverified sharing $\hat{\sigma}_U$. Moreover, the conversion locus shifts down if and only if it is not too elastic with respect to Δ, namely, $|\epsilon_P(\Delta)| < 1/(1-\Delta)$, where $\epsilon_P(\Delta)$ is the elasticity of $\rho^D(\Delta, \phi)$ with respect to Δ.*

Proof: Since Δ is decreasing in ϕ, it follows that an increase in ϕ lowers $\hat{\psi}(\Delta)$ and raises $\hat{\sigma}_U(\Delta)$. Next, let $\rho^D_\Delta := \partial\rho^D/\partial\Delta$ and $\rho^D_\phi := \partial\rho^D/\partial\phi$. Then, log-differentiating (B.14) with respect to ϕ— holding π fixed—and using (B.11) and (B.12) gives:

$$\begin{aligned}
\frac{d\rho^D/d\phi}{\rho^D} &= \frac{\rho^D_\Delta}{\rho^D}\Delta_\phi + \frac{\rho^D_\phi}{\rho^D} = \frac{\Delta\rho^D_\Delta}{\rho^D}\frac{\Delta_\phi}{\Delta} + \frac{\rho^D_\phi}{\rho^D} \\
&= \epsilon_P(\Delta)\frac{\Delta_\phi}{\Delta} - \frac{1}{1-\phi} = -\epsilon_P(\Delta)\frac{(1-\pi)}{(1-\phi\pi)(1-\phi)} - \frac{1}{1-\phi} \\
&= -\epsilon_P(\Delta)\frac{(1-\Delta)}{(1-\phi)} - \frac{1}{1-\phi}
\end{aligned}$$

Therefore, the conversion locus contracts if and only if $|\epsilon_P(\Delta)|(1-\Delta) < 1$. □

Define

$$R(\Delta) := \frac{\hat{\sigma}_U(\Delta)}{1-\zeta\hat{\sigma}_U(\Delta)}.$$

Then $\rho^D(\Delta,\phi) = (1-\phi)R(\Delta)$. Since $\hat{\sigma}'_U(\Delta) < 0$, we have $R'(\Delta) < 0$. Also, write the inverse supply as $\rho = S(\pi) := H^{-1}(\pi)$, and define the production level π that generates an incoming mass of fake news items equal to Δ as:

$$\Pi(\Delta,\phi) := \frac{\Delta}{1-\phi+\phi\Delta}.$$

Naturally, Π is increasing in Δ and in ϕ:

$$\frac{\partial\Pi(\Delta,\phi)}{\partial\Delta} = \frac{1-\phi}{(1-\phi+\phi\Delta)^2} > 0 \qquad \text{and} \qquad \frac{\partial\Pi(\Delta,\phi)}{\partial\phi} = \frac{\Delta(1-\Delta)}{(1-\phi+\phi\Delta)^2} > 0$$

A full-sharing market equilibrium can be determined directly in terms of Δ as

$$S(\Pi(\Delta^*,\phi)) = (1-\phi)R(\Delta^*).$$

Proof of Proposition 5: First, let $M(\Delta,\phi) := S(\Pi(\Delta,\phi)) - (1-\phi)R(\Delta)$. Direct computations show that $M_\Delta > 0$ and $M_\phi > 0$. Since in equilibrium, $M(\Delta^*,\phi) = 0$, the implicit function theorem yields: $d\Delta^*/d\phi = -M_\phi/M_\Delta < 0$. Since $\psi^* = \hat{\psi}(\Delta^*)$ and $\sigma^*_U = \hat{\sigma}_U(\Delta^*)$, Lemma B.1 implies that an increase in ϕ lowers ψ^* and raises σ^*_U. Welfare per news (B.3) rises, since ψ^* falls.

Next, consider production effects. Suppose the conversion is not too elastic. Then, by Lemma B.1, an increase in ϕ contracts the conversion demand locus while leaving the supply curve unaffected. Thus, π^* and ρ^* both fall. Hence, the stock of truthful news $T^* = \frac{1-\pi^*}{1-\zeta}$

rises since π^* falls in ϕ. The stock of fake news falls too, since it can be cast as

$$F^* = \frac{(1-\phi)\pi^*}{1-\zeta\sigma_U^*} = \frac{\pi^*\rho^*}{\sigma_U^*}.$$

As both π^* and ρ^* fall while σ_U^* rises, it follows that F^* falls. Next, we show that total welfare W^* rises. Indeed, if volume $T^* + F^*$ rises, then W^* rises too, as $W^* = w(\psi^*)(T^* + F^*)$. Conversely, if $T^* + F^*$ falls, then we can use decomposition (B.4): the benefit term $(b+\zeta\hat{b})T^*$ rises, while the loss term $L(\psi^*)(T^* + F^*)$ falls; hence, W^* rises.

Finally, suppose that the elasticity condition is non-trivially reversed. Then, by Lemma B.1, the conversion demand expands after an increase in ϕ, leading to more π^* and ρ^*. As a result, the stock of truthful news T^* falls. Moreover, since F^* can be written as $F^* = \frac{\psi^*}{1-\psi^*}T^*$, and both $\psi^*/(1-\psi^*)$ and T^* fall, it follows that F^* falls. □

B.4 Proof of Proposition 6: News Certification

Preliminaries. Let $\beta \in [0,1]$ denote the chance that a truthful item becomes certified after being verified. Let $T_{\neg C,\tau}$ denote the stock of uncertified truthful news at the beginning of period τ. Since truthful news enters the platform uncertified, if an uncertified truthful item is shared by a user who does not verify, then it remains uncertified. If it is shared by a user who verifies, it remains uncertified with chance $1-\beta$ and becomes certified with chance β. Hence,

$$T_{\neg C,\tau+1} = \zeta\left(\sigma_U + (1-\beta)\sigma_V\right)T_{\neg C,\tau} + 1 - \pi.$$

The stock of fake news is unchanged as these items are never certified. So, in steady state:

$$T_{\neg C} = \frac{1-\pi}{1-\zeta\left(\sigma_U + (1-\beta)\sigma_V\right)} \qquad \text{and} \qquad F = \frac{\pi}{1-\zeta\sigma_U}. \tag{B.15}$$

Therefore, prevalence among uncertified items is $\psi_{\neg C} = F/(T_{\neg C} + F)$:

$$\psi_{\neg C} = \frac{\left[1-\zeta\left(\sigma_U + (1-\beta)\sigma_V\right)\right]\pi}{1-\zeta\left(\sigma_U + (1-\beta)\sigma_V\pi\right)} \equiv \Psi(\pi, \sigma_U, (1-\beta)\sigma_V).$$

Now let $T_{C,\tau}$ denote the stock of certified truthful news. Certified items are known to be truthful and are shared without verification. In addition, a fraction β of the uncertified truthful items that are verified becomes certified. Thus, $T_{C,\tau}$ obeys:

$$T_{C,\tau+1} = \zeta T_{C,\tau} + \zeta\beta\sigma_V T_{\neg C,\tau}.$$

In steady state,

$$T_C = \frac{\zeta\beta\sigma_V T_{\neg C}}{1-\zeta} = \frac{\zeta\beta\sigma_V(1-\pi)}{(1-\zeta)\left[1-\zeta\left(\sigma_U+(1-\beta)\sigma_V\right)\right]}. \tag{B.16}$$

Henceforth, we focus on a full-sharing equilibrium, $\sigma_U+\sigma_V=1$. Using (B.15) and (B.16) then gives

$$T_{\neg C}+T_C = \frac{1-\pi}{1-\zeta}. \tag{B.17}$$

As a result, the overall probability of encountering fake news is

$$\psi = \frac{F}{T_{\neg C}+T_C+F} = \frac{(1-\zeta)\pi}{1-\zeta\left[\sigma_U+\pi(1-\sigma_U)\right]} \equiv \Psi(\pi,\sigma_U,1-\sigma_U), \tag{B.18}$$

which does not depend directly on β.

For fixed (π,β), the full-sharing equilibrium solves

$$\psi_{\neg C} = \Psi(\pi,\sigma_U,(1-\beta)(1-\sigma_U)), \qquad \sigma_U = 1-G(\psi_{\neg C}(\ell+\zeta\hat{\ell})). \tag{B.19}$$

Let $\hat{\psi}_{\neg C}(\pi,\beta)$ and $\hat{\sigma}_U(\pi,\beta)$ denote the solution to (B.19). Notice that increasing β shifts up the prevalence locus in the $(\sigma_U,\psi_{\neg C})$ space, while leaving the optimal sharing locus $\sigma_U = 1-G(\psi_{\neg C}(\ell+\zeta\hat{\ell}))$ unchanged. Therefore, $\hat{\psi}_{\neg C}$ rises and $\hat{\sigma}_U$ falls as β rises, given π.

The conversion demand locus is

$$\rho^D(\pi,\beta) := P(\hat{\sigma}_U(\pi,\beta),1-\hat{\sigma}_U(\pi,\beta)) = \frac{\hat{\sigma}_U(\pi,\beta)}{1-\zeta\hat{\sigma}_U(\pi,\beta)}. \tag{B.20}$$

Since $\hat{\sigma}_U$ falls in β for a fixed π, it follows that certification shifts conversion demand downward on the (π,ρ)-space. That is, given π, certification makes the uncertified category more suspicious, users verify more, unverified sharing falls, and conversions decline.

Proof of Proposition 6: Consider an increase in β. Because certification shifts conversion demand (B.20) downward while leaving supply (i.e., $\pi = H(\rho)$) unchanged, both equilibrium conversions ρ^* and equilibrium fake-news production π^* fall. Also, since $\rho^* = \frac{\sigma_U^*}{1-\zeta\sigma_U^*}$, the decline in equilibrium conversions ρ^* implies that σ_U^* necessarily falls. Therefore, by the sharing condition $\sigma_U^* = 1-G(\psi_{\neg C}^*(\ell+\zeta\hat{\ell}))$, equilibrium prevalence $\psi_{\neg C}^*$ rises.

(i) *Stocks:* The stock of fake news F^* (B.15) falls, as both π^* and σ_U^* fall, so F^* falls. Next, since $1-\psi_{\neg C}^* = T_{\neg C}^*/(T_{\neg C}^*+F^*)$, it follows that $T_{\neg C}^* = F^*(1-\psi_{\neg C}^*)/\psi_{\neg C}^*$. Since F^* falls and $\psi_{\neg C}^*$ rises, $T_{\neg C}^*$ falls. Total truthful news $T_{\neg C}^*+T_C^*$ (B.17) rises since π^* falls. Because T_U^* falls while $T_U^*+T_C^*$ rises, certified truthful news T_C^* must rise.

(ii) *Prevalence:* Overall prevalence is $\psi^* = \Psi(\pi^*,\sigma_U^*,1-\sigma_U^*)$ by (B.18). Prevalence function

Ψ is increasing in both π^* and σ_U^* (Lemma A.1). Since both π^* and σ_U^* fall with β, overall prevalence ψ^* falls.

(iii) *Welfare:* Certification creates two observable categories of news. Total welfare is the sum of total welfare in each category. For a certified item, there is no uncertainty (item is certified to be truthful) and no verification cost, so welfare per certified item is $W_C^* = (b + \zeta\hat{b})T_C^*$. For an uncertified item, the user problem is the same as in the baseline model except that the relevant prevalence is $\psi_{\neg C}$. Therefore, welfare per uncertified item is $w(\psi_{\neg C}^*)$, where w is in (B.3). Total welfare of uncertified items is $W_{\neg C}^* = w(\psi_{\neg C}^*)(T_{\neg C}^* + F^*)$. Since $1 - \psi_{\neg C} = T_{\neg C}/(T_{\neg C} + F)$, and $\psi_{\neg C} = F/(T_{\neg C} + F)$, using (B.4) and (B.17), total welfare can be written as

$$\begin{aligned} W_C^* + W_{\neg C}^* &= (b + \zeta\hat{b})(T_C^* + T_{\neg C}^*) - L(\psi_{\neg C}^*)(T_{\neg C}^* + F^*) \\ &= (b + \zeta\hat{b})\frac{1 - \pi^*}{1 - \zeta} - \frac{L(\psi_{\neg C}^*)}{\psi_{\neg C}^*}F^* \end{aligned}$$

Since π^* falls as β rises, the first term above rises. Also, $\psi \mapsto L(\psi_U)/\psi_U$ is decreasing,[36] and since F^* falls and $\psi_{\neg C}^*$ rises, the second term of the welfare expression, which enters with a minus sign, falls. Hence, total welfare rises.

Finally, we show that aggregate welfare from uncertified items falls. Since $T_{\neg C}^* + F^* = F^*/\psi_{\neg C}^*$ and F^* falls while $\psi_{\neg C}^*$ rises, the stock of uncertified items falls. Since, $w(\psi_{\neg C}^*)$ falls, total welfare from uncertified items falls. □

C Proofs of Section 6: Extensions

C.1 Proof of Proposition 7: Congestion Effects

Preliminaries. Suppose users process only a fraction $\varrho \in (0, 1]$ of the news allocated to them. Items that are not viewed are not followed and are not reshared, thus generating no welfare. Let T_τ and F_τ denote the stocks of truthful and fake news that are *seen* in period τ. Then, their laws of motion are

$$T_{\tau+1} = \varrho\left[1 - \pi + \zeta(\sigma_U + \sigma_V)T_\tau\right], \qquad \text{and} \qquad F_{\tau+1} = \varrho\left[\pi + \zeta\sigma_U F_\tau\right]. \tag{C.1}$$

Henceforth, we focus on a full-sharing equilibrium, so $\sigma_U + \sigma_V = 1$. In steady state, (C.1)

[36]Indeed, let $a := \ell + \zeta\hat{\ell}$. Direct differentiation yields: $[L(\psi)/\psi]' = \int_0^{\psi a}[G(t) - G(\psi a)]dt/\psi^2 < 0$.

yields:

$$T = \frac{\varrho(1-\pi)}{1-\zeta\varrho}, \qquad F = \frac{\varrho\pi}{1-\zeta\varrho\sigma_U}. \tag{C.2}$$

Thus, prevalence $\psi = F/(T+F)$ is

$$\psi = \frac{\pi(1-\zeta\varrho)}{1-\zeta\varrho\left[\sigma_U + \pi(1-\sigma_U)\right]} \equiv \tilde{\Psi}(\pi, \sigma_U; \varrho), \tag{C.3}$$

where $\tilde{\Psi}$ is in (B.2). Direct differentiation of (C.3) gives $\tilde{\Psi}_\varrho(\pi, \sigma_U; \varrho) < 0$.

A full-sharing equilibrium is determined by:

$$\psi = \tilde{\Psi}(\pi, \sigma_U; \varrho) \qquad \text{and} \qquad \sigma_U = 1 - G(\psi(\ell + \zeta\hat{\ell})). \tag{C.4}$$

Let us call $\hat{\sigma}_U(\pi, \varrho)$ and $\hat{\psi}(\pi, \varrho)$ the unique solution to (C.4). Notice that, given π, increasing ϱ shifts down the (increasing) prevalence locus in the (σ_U, ψ)-space, while leaving the (decreasing) sharing locus $\sigma_U = 1 - G(\psi(\ell + \zeta\hat{\ell}))$ unaffected. Thus, $\hat{\psi}$ falls and $\hat{\sigma}_U$ rises.

The conversion demand locus is then given by:

$$\rho^D(\pi, \varrho) := \frac{\varrho\hat{\sigma}_U(\pi, \varrho)}{1-\zeta\varrho\hat{\sigma}_U(\pi, \varrho)}. \tag{C.5}$$

The conversion demand rises in ϱ because an increase in ϱ raises conversions (holding fixed σ_U) and also raises σ_U. So the direct and indirect effects move in the same directions. Precisely,

$$\frac{\partial \rho^D}{\partial \varrho} = \frac{\hat{\sigma}_U + \varrho\frac{\partial \hat{\sigma}_U}{\partial \varrho}}{(1-\zeta\varrho\hat{\sigma}_U)^2} > 0. \tag{C.6}$$

Proof of Proposition 7: Consider an increase in ϱ. By our previous analysis, this expands the conversion demand while the supply curve $\pi = H(\rho)$ remains unchanged. Therefore, equilibrium conversions ρ^* and the equilibrium production π^* both rise. Next, conversions obey $\rho^* = \frac{\varrho\sigma_U^*}{1-\zeta\varrho\sigma_U^*}$, it follows that the "effective" unverified sharing rate $\varrho\sigma_U^* = \frac{\rho^*}{1+\zeta\rho^*}$ rises. The stock of fake news $F^* = \frac{\varrho\pi^*}{1-\zeta\varrho\sigma_U^*}$ rises because the numerator $\varrho\pi^*$ rises since both ϱ and π^* rise, while the denominator falls because $\varrho\sigma_U^*$ rises.

Next, we examine prevalence effects. Notice that, given equilibrium prevalence ψ^*, the unverified sharing rate is $\tilde{\sigma}(\psi^*) := 1 - G(\psi^*(\ell + \zeta\hat{\ell}))$, while conversions $\tilde{\rho}(\psi^*, \varrho) := \frac{\varrho\tilde{\sigma}(\psi^*)}{1-\zeta\varrho\tilde{\sigma}(\psi^*)}$. Since production $\pi^* = H(\rho^*)$, in equilibrium, ψ^* solves the following fixed point equation:

$$\psi^* = \Phi(\psi^*, \varrho), \qquad \text{where} \qquad \Phi(\psi, \varrho) := \tilde{\Psi}\left(H\left(\tilde{\rho}(\psi, \varrho)\right), \tilde{\sigma}(\psi); \varrho\right). \tag{C.7}$$

Notice that $\Phi_\psi(\psi, \varrho) < 0$. Indeed, holding ϱ fixed, higher prevalence ψ lowers $\tilde{\sigma}(\psi)$, which lowers conversions and hence production; it also lowers unverified sharing directly. Since $\tilde{\Psi}(\pi, \sigma_U)$ is strictly increasing in both π and σ_U, it follows that $\Phi_\psi(\psi, \varrho) < 0$.

Therefore, we can implicitly differentiate $\psi^* = \Phi(\psi^*, \varrho)$ to obtain:

$$\frac{d\psi^*}{d\varrho} = \frac{\Phi_\varrho}{1 - \Phi_\psi}. \tag{C.8}$$

So, the sign of $d\psi^*/d\varrho$ is determined by the sign of Φ_ϱ, which we derive next. To this end, we compute some useful elasticities of $\tilde{\Psi}$. Let

$$A := \sigma_U + \pi(1 - \sigma_U) \qquad \text{and} \qquad B := 1 - \zeta\varrho A.$$

Then, log-differentiate (C.3) to get:

$$\frac{\varrho\tilde{\Psi}_\varrho}{\tilde{\Psi}} = \frac{-\zeta\varrho A}{(1-\zeta\varrho)B} \qquad \text{and} \qquad \frac{\pi\tilde{\Psi}_\pi}{\tilde{\Psi}} = \frac{1 - \zeta\varrho\sigma_U}{B}. \tag{C.9}$$

Define the supply elasticity in conversions as:

$$\epsilon_\rho(H) := \frac{\rho h(\rho)}{H(\rho)}.$$

Now, we log-differentiate Φ in ϱ (C.7), and use: (C.9), the definition of $\tilde{\rho}$, and $\pi = H(\tilde{\rho})$:

$$\begin{aligned}
\frac{\Phi_\varrho}{\Phi} &= \frac{\tilde{\Psi}_\pi}{\tilde{\Psi}} H'(\tilde{\rho}) \frac{\tilde{\sigma}}{(1-\zeta\varrho\tilde{\sigma})^2} + \frac{\tilde{\Psi}_\varrho}{\tilde{\Psi}} = \frac{\pi\tilde{\Psi}_\pi}{\tilde{\Psi}} \frac{\tilde{\rho}H'(\tilde{\rho})}{H(\tilde{\rho})} \frac{1}{\varrho(1-\zeta\varrho\tilde{\sigma})} + \frac{\tilde{\Psi}_\varrho}{\tilde{\Psi}} \\
&= \frac{1-\zeta\varrho\tilde{\sigma}}{B} \frac{\epsilon_\rho(H)}{\varrho(1-\zeta\varrho\tilde{\sigma})} - \frac{\zeta(1-\pi)(1-\tilde{\sigma})}{(1-\zeta\varrho)B} = \frac{\epsilon_\rho(H)}{B\varrho} - \frac{\zeta(1-\pi)(1-\tilde{\sigma})}{(1-\zeta\varrho)B} \\
&= \frac{1}{B}\left[\frac{\epsilon_\rho(H)}{\varrho} - \frac{\zeta(1-\pi)(1-\tilde{\sigma})}{(1-\zeta\varrho)}\right]
\end{aligned}$$

Since $B > 0$, we see that $\Phi_\varrho < 0$ if and only if the conversion supply is sufficiently inelastic:

$$\Phi_\varrho < 0 \iff \epsilon_\rho(H) < \frac{\varrho\zeta(1-\pi)(1-\tilde{\sigma})}{(1-\zeta\varrho)} \tag{C.10}$$

Therefore, given (C.10), in equilibrium, a marginal increase in ϱ lowers equilibrium prevalence ψ^*, provided the supply is sufficiently inelastic. In such a case: (i) the stock of truthful news T^* rises, as $T^* = F^*(1-\psi^*)/\psi^*$; (ii) welfare per news $w(\psi^*)$ (B.3) rises; and (iii) total welfare W^* rises as welfare per news and total volume both rise. $\square$

C.2 Proof of Proposition 8: Fixed Inflow of Truthful Content

Preliminaries. Note that the only change in law of motion for truthful news (A.2) is that $1-\pi$ now becomes $\kappa > 0$. Meanwhile, the one for misinformation (A.1) remains unchanged. From here, it is straightforward to conclude that the steady state stocks of truthful and fake news are

$$T = \frac{\kappa}{1-\zeta(\sigma_U+\sigma_V)} \quad \text{and} \quad F = \frac{\pi}{1-\zeta\sigma_U}.$$

As a result, prevalence $\psi = F/(T+F)$ is given by

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

Now, divide the numerator and denominator of the right hand side by $\pi+\kappa$. Letting $\tilde{\pi} := \pi/(\pi+\kappa)$ and $\tilde{\kappa} := \kappa/(\pi+\kappa)$, and using that $\tilde{\pi}+\tilde{\kappa} = 1$,

$$\begin{aligned}
\psi &= \frac{(1-\zeta(\sigma_U+\sigma_V))\tilde{\pi}}{\tilde{\kappa}(1-\zeta\sigma_U)+\tilde{\pi}(1-\zeta(\sigma_U+\sigma_V))} \\
&= \frac{(1-\zeta(\sigma_U+\sigma_V))\tilde{\pi}}{(1-\tilde{\pi})(1-\zeta\sigma_U)+\tilde{\pi}(1-\zeta(\sigma_U+\sigma_V))} \\
&= \frac{(1-\zeta(\sigma_U+\sigma_V))\tilde{\pi}}{1-\zeta(\sigma_U+\sigma_V\tilde{\pi})} \equiv \Psi(\tilde{\pi},\sigma_U,\sigma_V).
\end{aligned}$$

We now revisit the comparative static of lowering verification costs.

Proof of Proposition 8: As in Section B.1, let the distribution of verification costs be indexed by a parameter η, so that an increase in η lowers verification costs in the sense of first-order stochastic dominance.

Equipped with this, write $\tilde{\pi}(\pi,\kappa)$ to make the dependence explicit and suppose that verification costs are distributed according to $G(\cdot;\eta)$. Given this, choose π such that an equilibrium with full sharing exists. As in the proof of Proposition 3, the extent of unverified sharing $\hat{\sigma}_U(\tilde{\pi};\eta)$ and prevalence $\hat{\psi}(\tilde{\pi};\eta)$ that arise from the sharing game solve

$$\psi = \Psi(\tilde{\pi},\sigma_U,1-\sigma_U) \qquad \text{and} \qquad \sigma_U = 1-G(\psi(\ell+\hat{\ell});\eta),$$

and the conversion demand is $\rho^D(\tilde{\pi},\eta) := \frac{\hat{\sigma}_U(\tilde{\pi};\eta)}{1-\zeta\hat{\sigma}_U(\tilde{\pi};\eta)}$. Crucially, since $\tilde{\pi}$ is increasing in π, it is easy to verify that, by the same arguments given in the proof of Proposition 3, an increase in η lowers equilibrium production π^* and conversions ρ^*; from here, both unverified sharing σ_U^* and prevalence ψ^* fall, and thus welfare per news $w(\psi^*)$ in (B.3) rises.

Finally, as for the stock of each type of news, note that $T^* = \kappa/(1-\zeta)$ in a full-sharing equilibrium, so this variable is constant in η. In turn, $F = \pi^*/(1 - \zeta\sigma_U^*)$ must fall because both π^* and σ_U^* fall. As a result, total volume $T^* + F^*$ falls.

As for total welfare W^*, consider decomposition (B.4). Since T^* is constant, the benefit term $(b + \zeta\hat{b})T^*$ is unchanged; however, the loss term $L(\psi^*; \eta)(T^* + F^*)$ falls, since $L(\psi^*; \eta)$ falls (same argument as in the proof of Proposition 3) and $T^* + F^*$ falls. Altogether, W^* rises. This concludes the proof. $\square$

References

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

Akey, P., V. Grégoire, N. Harvie, and C. Martineau (2026): "Who Wins and Who Loses in Prediction Markets? Evidence from Polymarket," Working paper. https://ssrn.com/abstract=6443103, accessed July 19, 2026.

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.

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

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

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.

Brady, W. J., J. C. Jackson, B. Lindström, and M. J. Crockett (2023): "Algorithm-mediated Social Learning in Online Social Networks," *Trends in Cognitive Sciences*, 27, 947–960.

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.*

Chuai, Y., M. Pilarski, T. Renault, D. Restrepo-Amariles, A. Troussel-Clément, G. Lenzini, and N. Pröllochs (2026): "Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter)," *Nature Communications*, 17, 4070.

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.

Germano, F., V. Gómez, and F. Sobbrio (2026): "Ranking for engagement: How social media algorithms fuel misinformation and polarization," *Journal of Public Economics*, 255, 105589.

Graffius, S. M. (2026): "Lifespan (Half-Life) of Social Media Posts: Update for 2026," https://www.scottgraffius.com/blog/files/lifespan-halflife-of-social-media-posts-update-2026.html, accessed July 19, 2026.

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.

Kaplan, J. (2025): "More Speech and Fewer Mistakes," Meta Newsroom, January 7, 2025 (updated February 6, 2026). https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/, accessed June 27, 2026.

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.

Long, K., C. Ostroff, N. Mehta, and B. T. Smith (2026): "They Looked Like They Were Getting Rich on Polymarket—but None of It Was Real," *The Wall Street Journal*, june 20, 2026. https://www.wsj.com/business/media/polymarket-social-media-bets-prediction-market-441cdeb5, accessed July 13, 2026.

National Association of Consumer Advocates (2026): "Complaint, *National Association of Consumer Advocates v. Blockratize, Inc. d/b/a Polymarket, et al.*" Superior Court of the District of Columbia, filed June 26, 2026. https://www.vacadaffanlaw.com/post/national-association-of-consumer-advocates-v-polymarket-et-al, accessed July 13, 2026.

Ostroff, C., N. Mehta, K. Long, and B. T. Smith (2026): "We Investigated Polymarket's Deceptive Marketing Campaign. Here's What We Found," *The Wall Street Journal*, june 22, 2026. https://www.wsj.com/business/media/we-investigated-polymarkets-deceptive-marketing-campaign-heres-what-we-found-51169858, accessed July 13, 2026.

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.

Saba, J. (2025): "Tech Titans Clash over Scraps of Limited Time," Reuters Breakingviews, May 29, 2025. https://www.reuters.com/breakingviews/tech-titans-clash-over-scraps-limited-time-2025-05-27/, accessed June 27, 2026.

Stoddard, G. (2015): "Popularity dynamics and intrinsic quality in reddit and hacker news," 9, 416–425.

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

Vásquez, J. (2022): "A theory of crime and vigilance," *American Economic Journal: Microeconomics*, 14, 255–303.

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