Fake News in Social Media: A Supply and Demand Approach*

Gonzalo Cisternas and Jorge Vásquez

November 25, 2020

Abstract

We introduce a model of a platform in which users encounter news of unknown veracity. Users vary in their propensity to share news and can learn the veracity of news at a cost. In turn, the production of fake news is both more sensitive to sharing rates and cheaper than its truthful counterpart. As in traditional markets, the equilibrium prevalence of fake news is determined by a *demand and supply of misinformation*. However, unlike in traditional markets, the exercise of market power is generally limited unless segmentation methods are employed. Combating fake news by lowering verification costs can be ineffective due to the demand for misinformation only being *weakly* reduced. Likewise, the use of algorithms that imperfectly filter news for users can lead to greater prevalence and *diffusion* of misinformation. Our findings highlight the important role of natural elasticity measures for policy evaluation.

1 Introduction

The spread of misinformation is a phenomenon that has gained substantial recent attention due to the rise of social media platforms that facilitate the rapid diffusion of news. In the 2016 U.S. presidential election, it is estimated that 760 million interactions with fake news occurred online (Allcott and Gentzkow, 2017); the top false election stories were shared more profusely on Facebook than the top counterparts from major news outlets (Silverman, 2016); and the same platform was a key gateway for landing on untrustworthy websites (Guess et al.,

*Cisternas: MIT Sloan School of Management, 100 Main St., Cambridge, MA 02142, gcistern@mit.edu. Vásquez: Smith College, Department of Economics, 10 Elm Street, Northampton, MA 01063, jvasquez@smith.edu. We thank Steven Durlauf, Jorge Lemus, Lones Smith, Marek Weretka, and Juanjuan Zhang for useful conversations, and Orrie Page for excellent research assistance.

2020). The overall scale of the "fake news" problem is still debated, with some arguing that exposure to fake news is low (Allen et al., 2020). However, the consequences can be real, potentially influencing elections, disease spread, and financial markets.[1] Indeed, this problem is considered on par with racism and income inequality by Americans (Mitchell et al., 2019) and is labeled as one of three major global risks by the World Economic Forum (Howell, 2013). The pace at which internet access, social media usage, and deception technologies evolve also suggest that this problem is far from disappearing.[2]

The response by key actors in the news world—most notably, social media platforms—can be synthesized based on three tenets. First, professional, independent fact-checking is required, resulting in the advent of a network of *third-party fact checkers* who verify the accuracy of the content linked to Facebook and YouTube.[3] Second, the incentives of fake news producers must be weakened. For example, Facebook now gives less relevance to news confirmed to be false and removes repeated offenders ("Fact Checking on Facebook", 2020), while initiatives such as the Global Disinformation Index rate untrustworthy websites in an attempt to curtail their advertising revenue. Third, critically, it is the *user* who must assess the veracity of news and ultimately choose how to act upon it. Flagged content is now accompanied by either fact checkers' reports or related material that provides the user with context. The rationale for this policy can have technological grounds—e.g., it is prohibitively costly for an algorithm to check all the content uploaded to a platform—but it also reflects a stance on the role of social media in society.[4]

In this paper, we introduce a flexible model of fake news that we leverage to better understand key economic forces driving the creation and diffusion of misinformation, as well as to determine how these forces are affected by policy interventions. In light of the aforementioned tenets, for instance, how impactful is the appearance of fact-checking that lowers the verification costs borne by users? How are equilibrium outcomes affected by changes in

[1]For instance, a famous recent video named "Plandemic" promoted falsehoods surrounding the COVID-19 pandemic (DiResta and Garcia-Camargo, 2020). In 2013, a temporary decrease of $130 billion in stock value occurred after a false tweet stated that an explosion had injured Barack Obama (Rapoza, 2017).

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

[3]Many of these organizations adhere to the International Fact Checking Code of Principles (`https://www.ifcncodeofprinciples.poynter.org`)—a set of guidelines intended to unify fact-checkers' practices as well as their transparency—that is promoted by the Poynter Institute for Media Sciences.

[4]Facebook also argues that this contextual approach works better at reducing sharing than exclusively using labels (Lyons, 2017). On YouTube, this approach is arguably the only viable option due to the high verification costs associated with the nature of its content; thus, users are granted access to related information when they are about to view potentially fraudulent content. At the other extreme is Twitter's approach: fact-checking is performed mostly in-house, and only recently has content flagging begun.

the attractiveness of the market as measured by the degree of competition on the supply side? How effective is the implementation of internal filters that detect fake content before reaching the platform's users? A key asset of our model is that these and other questions can be answered via a framework akin to the traditional supply and demand paradigm.

Model and findings. A platform is simply a venue where users encounter news that can be true or false. When a user and a piece of news match, the user can choose to perfectly learn the veracity of the news encountered at a fixed cost and subsequently opt to share the news with an (unmodeled) group of followers. Importantly, users vary in their gains and losses when sharing true versus false news, respectively. In our baseline model, we capture such heterogeneity via a one-dimensional variable whereby high types enjoy more when sharing truthful news and suffer less from passing on fake news. Finally, news are homogeneous in their veracity from any user's perspective, so the probability of the news encountered being fake is simply the fraction of those types of news circulating—or the fake news *prevalence*. In line with recent studies, we assume that fake news are cheaper to produce than their truthful counterparts, and that their associated revenue is also more sensitive to users' sharing decisions.

The *demand for misinformation* maps, for any level of fake news prevalence, the mass of users who share news without verifying them. These are the users who ultimately pass on fake news, and their number declines as the prevalence of misinformation rises (i.e., demand is downward sloping). The *supply of misinformation* maps, for any mass of users who share without verifying, the fraction of firms that choose to produce fake news. Because fake news are more sensitive to sharing rates, the number of firms willing to forgo the cost of producing truthful news rises as the aforementioned mass of users grows larger (i.e., supply is upward sloping). The equilibrium prevalence is thus determined via the intersection of these curves.

The properties of the demand for misinformation are shaped by users' incentives to verify and hence are central to our analysis. Naturally, verification has the scope to arise only when its cost is below a threshold. Fixing any verification cost in the relevant range, however, fake news prevalence matters: for low prevalence, the loss incurred when sharing fake news happens so rarely that those who share skip verification; likewise, for sufficiently high prevalence, the lowest type willing to share absent any verification technology is so high that now her loss of sharing fake news is small. Thus, verification occurs only for intermediate levels of prevalence, with users self-selecting into three groups: high types who determine the demand for misinformation, intermediate types who verify and thus share only truthful news, and low types who neither share nor verify.

Our first finding then pertains to reductions in verification costs not necessarily acting as

standard "demand shifters" that reduce demand at all possible levels of prevalence. Rather, demand can remain unchanged at extreme values of prevalence because users are insensitive to such changes in those regions.[5] Depending on the distance between (i) the equilibrium level of prevalence when verification is not possible and (ii) a key level of prevalence around which the demand for misinformation "contracts", drops in verification costs may not lower the equilibrium prevalence of fake news. Consequently, better verification mechanisms, while affecting the demand, need not improve the veracity of information in the news market.

We then turn to the analysis of market power by studying the case of a single news producer. From a producer's perspective, the sharing rate in the demand for misinformation plays the role of the price in a traditional demand function, while the prevalence variable acts as the quantity. However, (i) the monopolist does not directly control users' sharing decisions, and (ii) fake news prevalence is not directly observed by the users. This informational problem implies that the monopolist is unable to "move along" a downward sloping demand, implying that the equilibrium prevalence is unchanged when the monopolist inherits the cost structure of our baseline model. That said, we uncover an important consequence that verification efforts can have for monopolistic strategies beyond "uniform" ones. Specifically, since the demand for misinformation declines only for intermediate levels of prevalence as verifying becomes cheaper, potential convexities that can be exploited with segmentation techniques arise. Indeed, the monopolist can ameliorate its losses even by trivially segmenting the market into identical sub-markets such that she operates on the *concave closure* of the demand for misinformation. As a result, more misinformation is created and diffused.

Finally, we evaluate the efficacy of detection algorithms or platform's *internal filters*. Specifically, we consider the case of a filter that makes only type II errors (i.e., mistakenly letting some false news enter the platform) whose accuracy is observed by both users and producers—if a piece of fake news is caught, it is removed from the platform before reaching users. In this context, we show that the prevalence and diffusion of misinformation can be *non-monotonic* functions of the filter quality, which means that a rise in filter precision could increase the creation and diffusion of fake news. Intuitively, by affecting the inference made by users at the moment of receiving news, such filters can induce more unverified sharing, an equilibrium effect that can outweigh the negative direct effect that they have on production for a large range of qualities.

Applied relevance. The market for online fake news features (i) minimal entry barriers (e.g., website creation), (ii) low content production costs (e.g., savings on accurate reporting),

[5] We establish conditions on primitives guaranteeing that, as verification costs fall, the demand formally becomes more *elastic* in the region where a reduction occurs.

and (iii) whether ideologically or profit-driven, clicks being the main source of profitability (Allcott and Gentzkow, 2017; Tucker et al., 2018), all of which our model captures. In turn, the greater sensitivity of fake news profits to sharing rates can be justified on three grounds. The first, related to (iii), is that traditional media outlets are driven by reputational long-term goals and their platform-related revenue is usually less critical than that for small websites. Second, fake news typically have a novelty element that can shield them from congestion losses related to duplication when reaching users. Third, there are behavioral considerations: it has been found that fake news spread faster, deeper and broader than true news, with fake news evoking more emotional responses (Vosoughi et al., 2018). While not modeling these elements explicitly, we believe that they make a strong case for having an upward-sloping supply of misinformation.

The number of fact-checking outlets has grown worldwide from 44 in 2014 to almost 300 as of June 2020 (Stencel and Luther, 2020). While this trend has lowered search costs for users, it has not eliminated verification costs: to make an informed decision, users have to review reports whether searching independently at specialized sites (e.g., Snopes.com or PolitiFact.com) or when costlessly accessing them as part of contextual information provided in platforms. Furthermore, the efficacy of fact checking is still in dispute: apart from a well-recognized capacity problem, fact checks may not reach their targets (Guess et al., 2020), and when they do, the effect on sharing can be short-lived (Friggeri et al., 2014). Our results indicate that even if information is readily available, the fact that individuals' verification incentives are inevitably linked to the pervasiveness of fake news can be an avenue that jeopardizes fact-checking attempts. This highlights the importance of analyzing users' and producers' decisions jointly.

Complementing this attempt to affect demand via empowering consumers with access to information are platforms' efforts to directly affect the stock of misinformation reaching users. For instance, Facebook has allowed fact-checkers to proactively identify fraudulent stories and modified its machine learning algorithms to improve detection rates and reduce the visibility of false news (Lyons, 2018). However, this platform has approached these measures with caution, with its CEO even stating that "Facebook should not be the arbiter of truth" (Halon, 2020). Our finding of increased prevalence after the introduction of a filter that attempts to remove false news before reaching users provides an economic rationale for the concerns associated with letting social media platforms dictate what is true or not. Equally important, our analysis is able to link the severity of the problem to key underlying primitives, which is crucial for designing effective policy interventions.[6]

[6]According to YouTube's policy, "Most of what we remove is first detected by machines, which means we actually review and remove prohibited content before you ever see it. But no system is perfect, so we make

The problem of fake news is obviously more complex in that other considerations, such as behavioral biases and political ideologies can be at play in specific instances. Our analysis nevertheless uncovers fundamental forces that will continue to prevail even after accounting for them. Moreover, it shows that elasticity measures traditionally used in applied work have natural counterparts in this world, and their estimation is necessary for policy evaluation.

The paper proceeds as follows. We first review the related literature below. Section 2 then introduces the model and in Section 3 we perform the equilibrium analysis. Section 4 is devoted to comparative statics, Section 5 examines the extent of monopoly power, and Section 6 studies the equilibrium effects of internal filters. All proofs are in the appendix.

Related literature. In addition to the empirical and institutional work referenced above, our paper connects to several strands of literature.

First, there is a recent body of literature on fake news social media characterized by two features: users dislike sharing fake news, and they have access to technologies that enable them to inspect the news at hand. We share with Kranton and McAdams (2019) that fake news is an equilibrium phenomenon, but our modeling and spirit are different: our focus is on the users' incentives to engage in *costly verification*, while theirs is on how the network structures affects outcomes. Henry et al. (2020) in turn develop a model like our demand side with linear costs and imperfect verification to explain properties of their experimental data showing that fact-checking reduces fake news sharing; by allowing more general propensities for sharing, as well as a production side, our model illuminates on the conditions for such reductions to effectively take place. Finally, in a dynamic model with an exogenous supply of fake news, Papanastasiou (2020) shows that "news virality" can be understood as a traditional rational cascade also facilitated by a costly acquisition of signals.

Regarding our analysis of competition, there is work that explores the extent to which market forces discipline outlets' incentives to distort the truth or *slant* their reports (e.g., Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2008); our invariance result points to an information friction—namely, users not directly observing the prevalence of fake news—as a likely source limiting a monopolist's ability to exercise market power in the traditional way. Relatedly, the optimal segmentation we obtain via the concave closure of the demand function relates to the techniques in Kamenica and Gentzkow (2011) for Bayesian persuasion problems (see Bergemann et al., 2015, for the connections with price discrimination). Yet, a key novelty is that fake news prevalence, acting as the prior, is endogenous in our model.

Our non-monotonicity results relate to the phenomenon whereby policies aimed at pro-

sure if you see something that doesn't belong on YouTube, you can flag it for us and we'll quickly review it." For more information, see https://blog.youtube/news-and-events/safer-internet-day.

tecting individuals backfire due to a change in people's behavior, automobile safety being a classic example (Peltzman, 1975). In this line, the experimental design by Pennycook et al. (2020) shows that labeling only a subset of false news leads users to believe that those untagged are more accurate, which positively influences their sharing. While users in our model are certain that news have passed a filter, the underlying logic is similar—thus, we see their evidence as compelling when applied to our exercise of evaluating the use of imperfect filters. In addition, we show that this negative effect can result from individuals' *verification* incentives being undermined and that the problem can be exacerbated by more fake news reaching platforms due to suppliers' response.

To conclude, this paper contributes to the literature exploiting the tractability of matching models in settings where individuals choose to protect themselves from harm with endogenous intensity: Quercioli and Smith (2015) examines the economics of counterfeiting, while Vásquez (2019) develops an equilibrium theory of crime and vigilance.

2 The Model

We develop a static model of a platform intended to capture a situation in which a large number of heterogeneous users interact with a large number of heterogeneous news producers without established reputations for veracity.

A unit mass of infinitesimal risk neutral *users* participating in an online platform encounter "uncertified" news, i.e., news for which truthfulness cannot be determined upon first contact (e.g., by reading the headline, the originating website, or even the whole article). These encounters are one-to-one and random from the perspective of all platform participants.[7] A fraction $\nu \in [0,1]$ of those news items are false: we refer to this variable as fake news *prevalence*.

Upon encountering a piece of news, each user can decide to uncover its veracity at a fixed cost $t \geq 0$—for instance, the time costs reviewing a series of related articles presented as part of "contextual information" or the search costs when consulting specialized websites for fact checks, etc. After this decision is made, the user can subsequently share the news depending on her sharing preferences, which vary across the population. We model this heterogeneity via a one-dimensional variable $z \in [0, z_{max}]$, with $z_{max} \in \mathbb{R}_+ \cup \{\infty\}$, reflecting that the *benefit* of sharing truthful news is z, while the absolute *loss* from sharing fake news

[7] Of course, this randomness assumption is not intended to reflect that encounters are truly accidental. Instead, it captures a limited reach by news outlets over specific individuals when targeting a subpopulation of interest. This could be because the platform's targeting service offers an imperfect degree of granularity at the user level, or because the algorithm mediates matches using information not possessed by news producers.

is $\ell(z)$, where $\ell : [0, z_{\max}] \mapsto \mathbb{R}_+ \cup \{\infty\}$. We assume that types z are distributed according to a cumulative distribution function (CDF) $G(\cdot)$ which is continuously differentiable with density $g(\cdot) \equiv G'(\cdot)$ and support $[0, z_{max}]$. The payoff of not sharing news is normalized to zero (so all payoffs can be seen as net of a consumption benefit).[8]

In this baseline specification, the loss function $\ell(\cdot)$ is strictly decreasing. Economically, this implies that higher types have (i) a higher *propensity to share news* (the ratio $z/\ell(z)$ is increasing in z) and (ii) a *weaker incentive to verify news* (the ratio $t/\ell(z)$ is increasing in z). We believe that this relationship between propensity to share and incentives to verify is the most plausible in light of the studies documenting the demographics of those who share more fake stories.[9] We further make the following assumption.

Assumption 1. *The loss function $\ell(\cdot)$ is twice continuously differentiable and satisfies:*

(i) No loss at the top: $\ell(z_{max}) = 0$.

(ii) Non-increasing elasticity of losses: $(z\ell'(z)/\ell(z))' \leq 0$ for all $z < z_{max}$.

Part (i) ensures that for any level of prevalence, there are always types who share news items without verifying them—this assumption is realistic, but more critically, it is innocuous in that our statements require minor qualifications after relaxing it. Part (ii) is just technically convenient—we explain its role in the next section. Examples of functions with these properties are $\ell(z) = z^{-\gamma}$ (with $z_{max} = \infty$), or $\ell(z) = (z_{max} - z)^{\gamma}$, $\gamma \geq 0$, and $z_{max} < \infty$.

We now turn to the supply side. There is a unit mass of infinitesimal risk-neutral producers, each of them facing the choice of producing a single unit of truthful versus fake content. Producing fake news is costless, while producing truthful content is costly (e.g., expenses associated with editorial norms such as accessing primary sources or cross-validation of information). Firms differ in their technologies: the unit cost of producing truthful news, r, varies according to a CDF $F(\cdot)$ that is continuously differentiable with density $f \equiv F'$ and support in $[0, 1]$.[10]

The benefit of producing a piece of fake news is given by the fraction of users who end up sharing this type of news; we denote this fraction by $\sigma \in [0, 1]$ and refer to it as the *sharing*

[8] Assuming linear sharing benefits in z is without loss: if benefits were $v(z) \in [0, z_{max}]$ and strictly increasing, then we could have redefined types $\tilde{z} \equiv v(z)$, CDF $\tilde{G}(\tilde{z}) \equiv G(v^{-1}(\tilde{z}))$, and losses $\tilde{\ell}(\tilde{z}) \equiv \ell(v^{-1}(\tilde{z}))$.

[9] Around the 2016 presidential election, sharing fake content was concentrated among older individuals (Grinberg et al., 2019), and age has been documented as the main predictor even after controlling for partisanship and ideology (Guess et al., 2019). Thus, propensity to share is likely negatively correlated with verification efforts, especially if the latter become more costly with age. In this line, a model with heterogeneous verification costs t is conceptually identical to ours provided the properties on t/ℓ are preserved—our choice intends to depict cleaner comparative statics with respect to verification costs. Section 7 discusses this inverse relation of sharing propensities and verification incentives as well as other modeling assumptions.

[10] We discuss some institutional details of the supply side of this market in the beginning of Section 5.

rate of unverified news, as fake news items are shared only if they are not verified. On the other hand, we will study the benchmark case in which the revenue of each unit of true content is completely insensitive to its sharing rate, with a value set to 1. Consider the case of advertising revenue from users visiting the host website. While truthful news may trigger fewer platform-originated visits per share relative to fake news (e.g., due to a perceived lack of novelty, as argued in Section 1), producers of trustworthy content can reach out to major news outlets in parallel, pass their stringent screening tests, and channel more visitors via references in those outlets. Thus, the revenue of truthful outlets is less sensitive to sharing rates on platforms. Our model examines the perfectly unresponsive benchmark case, which gives rise to an everywhere increasing supply of fake news.[11]

Definition 1 (Equilibrium). *A competitive equilibrium consists of a prevalence ν^* and a sharing rate of unverified news σ^* such that (i) users' sharing decisions are optimal given the prevalence level ν^* and (ii) producers' choices of news veracity are optimal given σ^*.*

In a competitive equilibrium, users make their sharing decisions taking as given the (correctly anticipated) fake news prevalence, while producers decide between fake and truthful content taking the mass of individuals who share without verifying as fixed.

This model can be seen as short for the steady state of a platform with a *rapidly evolving* inflow of news. Specifically, each period can be divided into two stages. First, a new cohort of news items enters the platform, each encountering a user. Second, if the user shares a fake item, this news becomes visible to a subset of individuals, yielding a payoff of 1 to the untrustworthy website: from an ex ante perspective, the producer receives σ. In the next period, either the previous cohort of news is public information, or the algorithm gives it less relevance in the users' news feed in favor of the new cohort; in either case, the subsequent sharing rate of those old news items is limited, making the initial rate the most relevant for payoffs. In this context, the natural metric capturing the overall visibility of fake news is the (equilibrium) fraction of them that are shared,

$$\Delta^* \equiv \nu^* \sigma^*. \tag{1}$$

We refer to it as the *rate of diffusion* of fake news.

[11] Normalizing the inelastic return to producing truthful news to 1 is qualitatively irrelevant.

3 Equilibrium Prevalence and Rate of Diffusion

The demand for misinformation. Fake news will diffuse only when they are not verified, because the verification technology is perfect and passing on false news entails losses to users. As the prevalence of fake news changes, users' varying propensities to share and incentives to verify will lead to variations in the sharing rate of unverified news. Since this rate shapes the benefits from producing fake news, an induced *demand for misinformation* naturally emerges.

To construct this demand function, it is instructive to first examine the case in which verification is prohibitively costly or simply not available. We assume that whenever there is indifference between sharing or not, users will desire to share. Given a conjectured prevalence ν, therefore, a user of type z will share the news encountered when her expected sharing payoff is non-negative, i.e.,

$$(1-\nu)z - \nu\ell(z) \geq 0.$$

Since the propensity to share, $z/\ell(z)$, is strictly increasing, for any $\nu \in [0,1]$ there exists a unique threshold type $\bar{z}(\nu) \in [0, z_{max}]$ such that all users $z \geq \bar{z}(\nu)$ choose to share, where

$$(1-\nu)\bar{z}(\nu) - \nu\ell(\bar{z}(\nu)) = 0. \tag{2}$$

Clearly, fewer users are willing to share as the prevalence increases, so $\bar{z}(\nu)$ is strictly increasing; moreover, $\bar{z}(0) = 0$, and the fact that there are no losses at the top of the distribution (part (i) in Assumption 1) implies that $\bar{z}(1) = z_{max}$. For prevalence $\nu \in (0,1)$, the population segments in two sets, with high types determining the sharing rate of unverified news.

Suppose now that the platform users can learn the authenticity of news at a moderate verification cost t. For any given level of prevalence ν, consider the users who were willing to share when verifying was not an option, i.e., $z \geq \bar{z}(\nu)$. In this region, each type z can avoid the loss $\nu\ell(z)$ provided that the verification cost t is paid. Each user's willingness to engage in verification is then given by the change in her sharing payoff, namely,

$$\nu\ell(z) - t.$$

This expression decreases as z rises because $\ell' < 0$. Thus, high types will be willing to verify if and only if the threshold type $\bar{z}(\nu)$ is willing to do so, i.e., $\nu\ell(\bar{z}(\nu)) \geq t$.

We refer to the mapping $\nu \mapsto \nu\ell(\bar{z}(\nu))$ as the (threshold) *sharing downside* function. This function is non-negative and vanishes at $\nu = 0$; it also vanishes at 1 because the highest

type is always willing to share news. Letting

$$t^{\dagger} \equiv \max_{\nu} \nu \ell(\bar{z}(\nu)), \tag{3}$$

we conclude that no user finds it optimal to verify the news encountered when $t > t^{\dagger}$. Part (ii) in Assumption 1 then ensures that the following lemma holds.

Lemma 1. *The sharing downside function $\nu \mapsto \nu\ell(\bar{z}(\nu))$ is quasi-concave.*

By the previous lemma, for each verification cost $t < t^{\dagger}$, there exists a non-empty interval $[\underline{\nu}, \bar{\nu}]$ such that type $\bar{z}(\nu)$ is willing to verify if and only if $\nu \in [\underline{\nu}, \bar{\nu}]$. For any given level of prevalence in this region, it then follows that

$$\bar{z}_h(\nu) \equiv \ell^{-1}(t/\nu) \tag{4}$$

is the highest type willing to verify news. Observe that $\bar{z}(\nu) \leq \bar{z}_h(\nu) < z_{max}$, with equality only when $\nu \in \{\underline{\nu}, \bar{\nu}\}$. Additionally, $\bar{z}_h(\nu)$ rises in ν and falls in t since the inverse loss function $\ell^{-1}(\cdot)$ is strictly decreasing. Let us then define $\bar{z}_h(\nu)$ for all $\nu \in [0,1]$ by $\bar{z}_h(\nu) = \bar{z}(\nu)$ when $\nu \notin [\underline{\nu}, \bar{\nu}]$—for notational convenience, we omit the dependence of $\bar{z}_h(\nu)$ on t.[12]

We refer to $\bar{z}_h(\nu)$ as the *marginal type* (at prevalence ν), as it effectively corresponds to the lowest type of user who is willing to share unverified news. Indeed, if $\nu \notin (\underline{\nu}, \bar{\nu})$, it is suboptimal for type $z = \bar{z}$ to verify. However, since the payoff from verifying and sharing is $(1-\nu)z - t$, this is also suboptimal for types $z < \bar{z}$ who, in turn, were unwilling to share absent the verification option. Thus, $\bar{z}_h(\nu) = \bar{z}(\nu)$ is the last type to verify without sharing in this region. Likewise, if $\nu \in (\underline{\nu}, \bar{\nu})$ all types below $\bar{z}_h(\nu)$ verify and share up until $\bar{z}_l(\nu) \equiv t/(1-\nu)$, type below which not sharing is again optimal.

All told, the sharing rate of unverified news at prevalence ν—i.e., the mass of users who skip verification yet share at that level—is given by $1 - G(\bar{z}_h(\nu))$. We can then define the *demand for misinformation* as the locus

$$\sigma_D(\nu) = \begin{cases} 1 - G(\ell^{-1}(t/\nu)) & \text{if } \nu \in (\underline{\nu}, \bar{\nu}); \\ 1 - G(\bar{z}(\nu)) & \text{if } \nu \notin (\underline{\nu}, \bar{\nu}), \end{cases} \tag{5}$$

where $\bar{z}(\nu)$ satisfies (2). In particular, if $\nu \in [\underline{\nu}, \bar{\nu}]$, the population segments into three sets:

1. low types $z < \bar{z}_l$ neither share nor verify;

[12] The role of Assumption 1 is now apparent: relaxing part (i) sometimes leads to verification intervals of the form $[\underline{\nu}, 1]$, which does not alter our conceptual findings due to $\underline{\nu} > 0$ generically; part (ii) ensures that the verification region is convex, which simplifies the description of the findings.

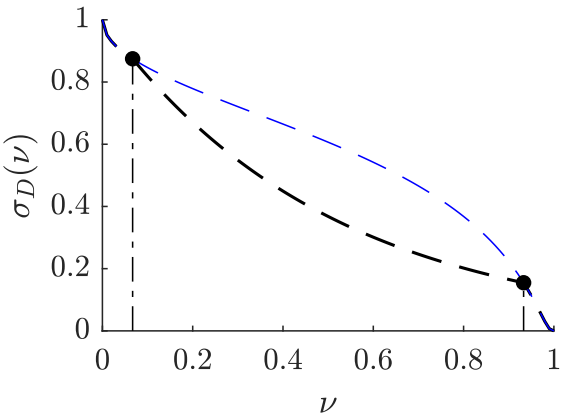

Figure 1: The demand for misinformation for $G(z) = 1 - e^{-0.5z}$ and $\ell(z) = 1/z$. The dashed blue line (top) represents the demand for news when verification costs are high, i.e., $t \geq t^\dagger = 0.5$, whereas the black line represents the demand with active verification $t = 0.25 < t^\dagger$.

2. intermediate types $z \in [\bar{z}_l, \bar{z}_h]$ verify and thus share truthful news only;
3. high types $z > \bar{z}_h$ share without verification.

Notice that if fake news prevalence $\nu \notin [\underline{\nu}, \bar{\nu}]$ the middle segment vanishes: no one verifies content, but users with high sharing gains, i.e. $z \geq \bar{z}$, still choose to share news.

Figure 1 depicts two demand functions, with the inferior (superior) one derived using a value for t that is below (above) $t^\dagger$, the value above which no one ever verifies news (see (3)). As shown in the figure, both coincide for extreme values of ν, reflecting that the inferior curve exhibits verification only for intermediate values of prevalence. Both functions are strictly decreasing due to the marginal type $\bar{z}_h(\nu)$ being increasing in ν.

The supply of misinformation. We now turn to the supply side. Recall that fake news items are costless to produce and sensitive to sharing rates: the profit of producing one unit of false content is σ, the sharing rate of unverified news. On the other hand, producing truthful news yields a deterministic revenue 1, but each unit is costly to produce. Conjecturing an equilibrium sharing rate of unverified news, σ, a producer with cost r chooses to produces fake content if and only if

$$r \geq 1 - \sigma. \tag{6}$$

The mass of producers choosing to produce fake news is then $1 - F(1-\sigma)$, which in turn determines the likelihood that each user will encounter fake content in the platform. The *supply of misinformation* is then simply defined as the locus

$$\nu_S(\sigma) \equiv 1 - F(1-\sigma). \tag{7}$$

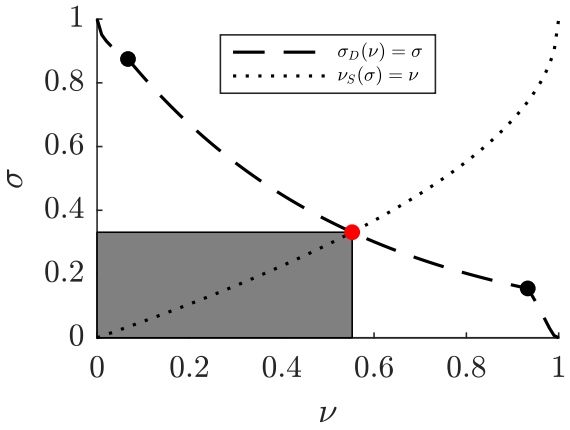

Figure 2: Equilibrium in the market for fake news: $G(z) = 1 - e^{-0.5z}$, $\ell(z) = 1/z$, $F(r) = r^2$, and $t = 0.25 < t^\dagger = 0.5$.

Notice that, as the sharing of unverified news σ rises, more producers prefer to create fake news content. As a result, the supply function $\nu_S(\cdot)$ is upward sloping with $\nu_S(0) = 1-F(1) = 0$ and $\nu_S(1) = 1 - F(0) = 0$.[13]

Equilibrium prevalence and diffusion. A competitive equilibrium arises when the supply and demand forces balance each other. As expected, the rising supply and falling demand curves, as well as their continuities, imply that *there exists a unique equilibrium.*

Proposition 1 (Existence and uniqueness).

(a) There exists a unique equilibrium (ν^*, σ^*) *for all verification costs* $t > 0$.

(b) There is a value $\hat{t} > 0$ *generically strictly below* $t^\dagger$ *such that, in equilibrium, a positive mass of users verify news if and only if verification costs are strictly below* $\hat{t}$.

Figure 2 displays the equilibrium in the market for fake news using the demand exhibiting non-trivial verification incentives from Figure 1. Observe that the rate of diffusion of fake news Δ^* corresponds to the rectangular shaded area, $\nu^*\sigma^*$, reminiscent of total revenue in a traditional competitive market.

By falling in the region $[\underline{\nu}, \bar{\nu}]$ where verification incentives are at play, this equilibrium does entail a non-trivial mass of users verifying the news they encounter. The wedge between $\hat{t}$ and $t^\dagger$ alluded to in part (b) of Proposition 1 states, however, that this is not always the case: non-trivial reductions in verification costs need not guarantee that verification actually occurs in *equilibrium*. We turn to this issue and other comparative statics in the next section.

[13] In reality, some fake news producers create content with purely harmful intentions: the trade-off they face is not between producing fake and truthful news, but rather to enter the market or not. If greater sharing rates of unverified news elicit more entry of such malicious producers, the supply of misinformation would remain upward sloping and it would be even more sensitive to sharing rates.

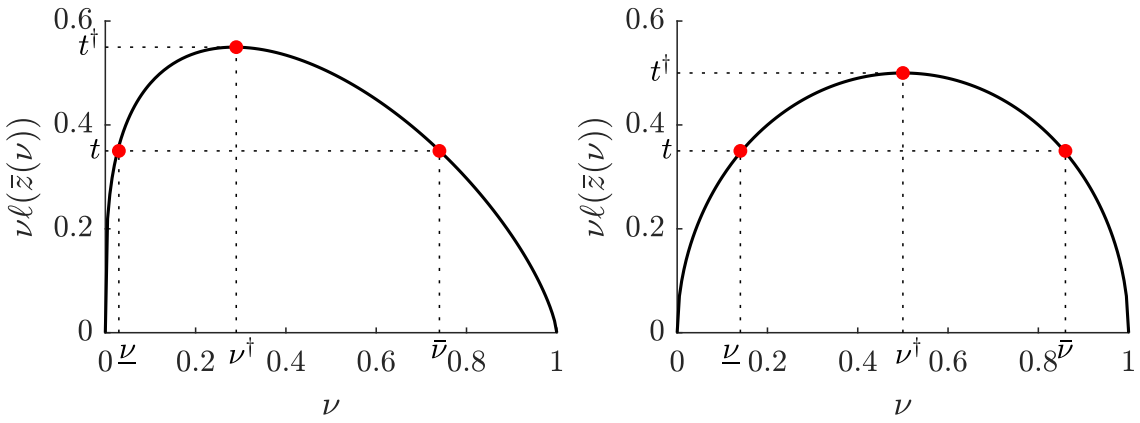

Figure 3: The downside sharing function. Left: $\ell(z) = 1/z^{2.5}$. Right: $\ell(z) = 1/z$. Both panels assume $t = 0.25$ and $z_{max} = \infty$.

4 Comparative Statics

The ability to examine equilibrium outcomes employing a framework akin to supply and demand naturally brings to light measures of sensitivity, or elasticities, as critical tools for policy analysis. In this section, we explore some comparative statics intended to highlight the importance of further applied work in this area.

Changes in verification costs. Our first observation pertains to the aforementioned wedge between the verification costs $t^\dagger$ and $\hat{t}$ of Proposition 1 being a critical element to study in assessing the efficacy of partial reductions in verification costs.

In deciding whether to verify news, recall that the threshold type at prevalence ν, $\bar{z}(\nu)$, trades off the downside of sharing false news, $\nu\ell(\bar{z}(\nu))$, against the upfront payment t. Let

$$\nu^\dagger := \arg\max_{\nu\in[0,1]} \nu\ell(\bar{z}(\nu))$$

denote the maximizer of the downside sharing function, which captures the level of prevalence at which verification incentives become *active* once lowering verification costs below $t^\dagger$: lowering t below $t^\dagger$ would induce verification only if prevalence ν was very close to $\nu^\dagger$. Figure 3 plots the downside sharing function induced by the (constant elasticity) loss family $\ell(z) = 1/z^\gamma$, $\gamma > 0$: reducing t below $t^\dagger$ increases the scope of verification by enlarging the *verification interval* $[\underline{\nu}(t), \overline{\nu}(t)]$ where $\nu^\dagger$ lies. Outside this region, either the likelihood of the news being fake is too low, or the threshold type suffers too little when passing on fake news, and thus no verification is induced.

Above $t^\dagger$, changes in verification costs t induce no verification for all levels of prevalence:

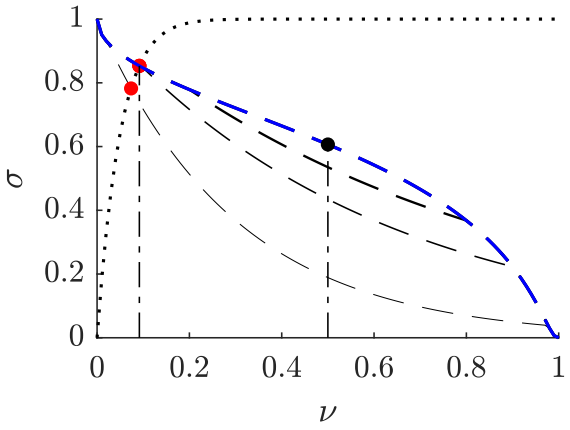

Figure 4: Different demand functions for $t = 0.5, 0.4, 0.3$ and 0.15 (moving from thick to thin dashed lines) for $G(z) = 1 - e^{-0.5z}$; $F(r) = r^{\zeta}$, with $\zeta = 0.05$; and $\ell(z) = 1/z$ (which implies that $t^{\dagger} = 0.5$). The equilibrium is affected only when t falls to 0.15

the equilibrium prevalence remains unchanged at a value that we denote ν^o. The distance between ν^o and $\nu^{\dagger}$ is therefore of key importance: reductions in t that do not generate verification intervals containing ν^o will not affect equilibrium outcomes—the value of $\hat{t}$ in Proposition 1 is then characterized as $\sup\{t \leq t_{\max} | \ \underline{\nu}(t) = \nu^o\}$ if $\nu^o < \nu^{\dagger}$, and replacing $\underline{\nu}$ by $\bar{\nu}$ in the previous set if $\nu^o > \nu^{\dagger}$. Figure 4 depicts a situation in which $t = t^{\dagger} = 0.5$, yet the equilibrium is unaffected unless verification costs fall by more than 75%.

Our model therefore warns that reductions in verification costs do not necessarily operate as traditional *demand shifters* that displace the demand at all (non-trivial) levels of prevalence. Rather, such changes are likely to reduce the demand for misinformation only in certain regions of prevalence, so lowering verification costs may be ineffective given that $\nu^o \neq \nu^{\dagger}$ generically. The mixed evidence on the effectiveness of fact-checking on sharing could be due to users' perceptions of the severity of the problem, as measured by the equilibrium prevalence, being a key variable of consideration when verification is a costly activity.[14]

Supply shifts. Reducing verification costs can lower the demand for misinformation $1 - G(\bar{z}_h(\nu))$ by weakly increasing the marginal type $\bar{z}_h(\nu)$. This, in turn, has the potential to make the demand more *sensitive* to changes in prevalence ν. Here we show that, whenever at play, verification effects can have unintended consequences for the creation and diffusion of fake news when such effects are coupled with changes in the supply side.

Specifically, as argued earlier, a central component of the platforms' response to combat

[14] In situations where users are granted access to chains of comments containing evidence that a news item is false, Friggeri et al. (2014) argue on this cost consideration that fact-checking can fail because "not all comments [are] read by users before sharing, either due to lack of interest, or because other, more recent comments are more easily viewable."

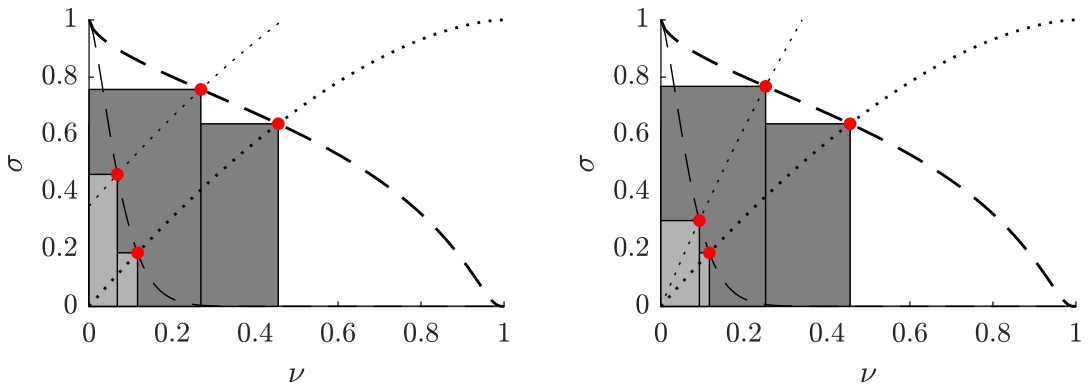

Figure 5: Left panel: introduction of an additive cost $c = 0.35$ to producing fake news; Right panel: introduction of a multiplicative penalty $\alpha = 0.5$ to fake news' sharing revenue. Both panels assume $G(z) = 1 - e^{-0.5z}$, $\ell(z) = 1/z$, $\tilde{F}(r) = r^{0.5}$, and $t = 0.05$ for the light-dashed demand.

fake news has been to directly reduce their supply. The methods utilized encompass banning repeat offenders and curtailing their ability to attract advertisers, among others. The former practice introduces costs to producing fake news that are orthogonal to content generation (e.g., change of identity and website appearance that is needed to pass screening tests); we can capture these via a scalar $c \in (0, 1)$ representing the total *expected* costs of producing fake news (i.e., a loss is incurred only if the producer is caught). On the other hand, the latter practice reduces the returns of fake news to sharing rates (e.g., there are fewer advertisers placing ads on untrustworthy websites); we can capture such an effect using a scalar $\alpha \in (0, 1)$ so that the revenue of producing fake news is $\alpha\sigma$. All told, the new supply functions are

$$\nu_S^c = 1 - F(1 + c - \sigma) \quad \text{and} \quad \nu_S^\alpha = 1 - F(1 - \alpha\sigma)$$

respectively, from where it is clear that the incentives to produce fake content fall.

Figure 5 illustrates the changes in equilibrium prevalence, ν^*, and diffusion, $\Delta^* = \sigma^*\nu^*$, when starting from the original supply $\nu_S = 1 - F(1 - \sigma)$: the left panel depicts ν_S^c while the right one ν_S^α, both in light dots; in both panels only the light-dashed (i.e., leftmost) demand function encodes verification effects. Clearly, the equilibrium prevalence with and without verification falls, and hence fewer fake news items enter the platform. However, this is not necessarily the case with the rate of diffusion Δ^* represented in the shaded rectangles: the area decreases when there is no verification (dark gray), but it *increases* when verification costs are low (light gray). Indeed, a reduction in the prevalence of fake news reduces verification incentives—captured in the marginal type falling—which leads users to proportionally increase their sharing of fake news. Thus, a policy that jointly reduces verification costs and the supply of fake news can backfire in that it may increase

the visibility of such news. This can be particularly damaging if untrustworthy websites target individuals who have a large number of followers.

Of course, the rate of diffusion at prevalence ν, $\Delta(\nu) \equiv \sigma_D(\nu)\nu$ will increase upward along the demand curve when

$$\varepsilon_D(\nu) \equiv \frac{\nu \sigma_D'(\nu)}{\sigma_D(\nu)} < -1,$$

i.e., when the demand is *elastic*. The next proposition establishes sufficient conditions on primitives enabling this to occur. To this end, let $\varepsilon_D(\nu; t)$ make the dependence on t explicit.

Proposition 2 (Elasticities). *Fix $t < \hat{t}$. Then,*

(i) if $1-G(z)$ is log-concave and $\ell(z)$ is log-convex, the demand elasticity satisfies $\varepsilon_D(\nu; t') < \varepsilon_D(\nu; t)$ for verification costs $t' < t$ and prevalence $\nu \in [\underline{\nu}(t), \bar{\nu}(t)]$;

(ii) the demand for misinformation is elastic at prevalence values $\nu \in [\underline{\nu}(t), \overline{\nu}(t)]$ if

$$\frac{zg(z)}{1-G(z)} + \frac{z\ell'(z)}{\ell(z)}\Big|_{z=\bar{z}_h(\nu)} > 0,$$

and it is elastic at prevalence values $\nu \notin [\underline{\nu}(t), \overline{\nu}(t)]$ if

$$\frac{zg(z)}{1-G(z)} + \frac{z\ell'(z)}{\ell(z)}\Big|_{z=\bar{z}(\nu)} > 1.$$

Part (i) establishes conditions under which the demand for misinformation becomes pointwise more elastic in the verification region as verification costs fall. It holds for a large class of distributions G with log-concave densities (see, e.g., Bagnoli and Bergstrom, 2005) and for geometric loss functions $\ell(z) = z^{-\gamma}$, for instance.

Part (ii) in turn states conditions under which the demand for misinformation is elastic at ν for a fixed t—we can see that the condition in the no-verification region is indeed more stringent, in line with our intuition that verification generates more sensitivity. To understand the expressions, recall that $\bar{z}_h(\nu) = \ell^{-1}(t/\nu)$, so in any non-trivial verification region,

$$\varepsilon_D(\nu) = \frac{\nu[1-G(\bar{z}_h(\nu))]'}{1-G(\bar{z}_h(\nu))} = \frac{g(\bar{z}_h(\nu))}{1-G(\bar{z}_h(\nu))} \cdot \underbrace{\frac{t}{\nu}}_{=\ell(\bar{z}_h(\nu))} \cdot \frac{1}{\ell'(\bar{z}_h(\nu))}. \tag{8}$$

Consequently, distributions G with large *hazard rates* relax our conditions, as this rate partly shapes the elasticity of demand with respect to percentage changes in the type. (The role of

the elasticity of the loss function ℓ depends on parameters.[15]) Finally, our condition in the no-verification region can be derived from expression (8) once $\bar{z}_h$ is replaced by $\bar{z}$ obeying (2).

Demand shifts. Finally, let us briefly discuss the effects of traditional demand shifters that affect demand at all points. Clearly, a strict decline in G in the sense of first-order stochastic dominance—i.e., higher types become less frequent in the population—strictly reduces the demand for misinformation at all non-trivial levels of prevalence $\nu \in (0,1)$. Thus, any equilibrium will also exhibit less prevalence and diffusion of fake news.

A similar effect takes place when the loss function ℓ increases pointwise, with the demand reduction now operating via an increase in the (extended to [0,1]) marginal type $\bar{z}_h(\nu)$. In addition, the verification region expands, as seen in $\bar{z}_h(\nu) \equiv \ell^{-1}(t/\nu)$.

If no verification takes place in equilibrium, however, it is unclear whether an increase in the loss function ℓ facilitates verification incentives. Consider a pointwise increase in the loss function ℓ when originally $\nu^o < \nu^\dagger$, i.e., when the equilibrium is below the level of prevalence at which verification incentives become active. On the one hand, the maximum value of the sharing downside function, $t^\dagger$, rises, thereby increasing the sensitivity of the demand function to reductions in t from above $t^\dagger$ at least locally around $\nu^\dagger$. On the other hand, the distance between the new values of ν^o and $\nu^\dagger$ is ambiguous. Indeed, while ν^o decreases due to the demand shift generating a downward movement along the supply curve, the direction and magnitude of the changes in $\nu^\dagger$ in general depend on supermodularity properties of the parametrized family of loss functions that underlies the change.[16]

5 Market Power

Approximately a quarter billion dollars of advertising revenue has been allocated annually to 20,000 fake news websites worldwide (Global Disinformation Index staff, 2019). While these numbers are suggestive of a fairly competitive environment, there are a few considerations that make the analysis of market power relevant. The first is ownership: parent companies such as Disinfomedia own several untrustworthy websites (Sydell, 2016). The second is popularity: outlets such as the National Enquirer are well known for exploiting clickbait-

[15] For instance, for an exponential distribution of parameter λ and a constant elasticity loss function $1/z^\gamma$,$\gamma > 0$, the left-hand side of the conditions reduces to $(\nu/t)^{1/\gamma}\lambda - \gamma$, where we used $\bar{z}_h(\nu) = (t/\nu)^{-1/\gamma}$. Whether the ratio ν/t is larger or smaller than 1 is then relevant.

[16] See Figure 3 for graphical changes in $\nu^\dagger$. Formally, it can be shown that for a smooth family $\ell(z|\varphi)$ with $\varphi \in \mathbb{R}$, if $\ell_\varphi < 0$ and $\ell(z|\varphi)$ is log-supermodular in (z,φ), then $\varphi \mapsto \nu^\dagger(\varphi)$ is increasing.

type headlines while enjoying the presence of a captive audience.[17] Finally, efforts aimed at destabilizing the financial incentives for producing fake news put pressure on the industry shrinking over time, everything else being equal.

In this section, we explore the extent to which market outcomes are affected when the supply side instead consists of a single news producer. Despite the demonstrated similarity between our approach and traditional analyses of decentralized markets, we show that there are key informational limitations that can hinder a monopolist's ability to exercise market power in the traditional way, namely, by restricting trade. Verification incentives, however, can potentially pave the way for other more sophisticated forms of market power.

Uniform policies. To make our point, we consider a monopolist that shares the cost structure of our baseline model. Specifically, (i) there is capacity for a unit mass of news only; (ii) producing fake news is costless; and (iii) producing truthful news is costly according to $r \sim F$. This stochastic cost structure can be understood as a situation in which the outlet allocates news production to reporters who vary in their skills, which naturally leads to heterogeneous production costs. Our goal is to show that the non-observability of prevalence ν will lead to the competitive outcome.

The monopolist's problem is to choose a mass $\nu \in [0,1]$ of fake news and $1-\nu$ of truthful content to be produced. Clearly, these fake news items will be optimally allocated to inefficient reporters, i.e., those with a cost $r \geq F^{-1}(1-\nu)$. Thus, a sharing rate of unverified news σ leads to total profits that amount to

$$\sigma \cdot \nu + \int_0^{F^{-1}(1-\nu)} [1-r] f(r) dr \tag{9}$$

where the first term reflects the profits from fake news (which are costless to produce), while the second reflects the profits from truthful content.

To gain intuition, let us first recast this problem into one that uses our supply and demand framework. Denote $\sigma_S(\nu)$ the inverse supply function, i.e., $\sigma_S(\nu) \equiv 1 - F^{-1}(1-\nu)$ given (7).

Lemma 2 (Monopolist's Problem). *The monopolist profits* (9) *can be written as*

$$\sigma \cdot \nu - \int_0^{\nu} \sigma_S(\nu') d\nu' + K \tag{10}$$

where K is a constant (i.e., independent of ν).

[17] According to Media Bias/Fact Check, an online media outlet "dedicated to educating the public on media bias and deceptive news practices," it could be said that the National Enquirer is "the original fake news media outlet that profits by selling fake news" (https://mediabiasfactcheck.com/national-enquirer/).

In other words, maximizing total profits (9) is equivalent to maximizing fake news' profits in our competitive benchmark, represented by a rectangular revenue net of the area below the supply curve. This finding is intuitive: while the monetary cost of producing fake news is zero, the economic cost of producing each unit of fake news is the opportunity cost of allocating a reporter to generate truthful content (so optimally producing fake news delivers the truthful counterpart as a byproduct). This opportunity cost is captured by $1 - r$, which leads to an increasing marginal cost function that coincides with our original supply curve: the first unit of fake content is very cheap due to the most inefficient reporter being allocated this task, but subsequent units are increasingly more expensive.

Traditional market power with uniform pricing is encoded in a downward-sloping demand faced by the monopolist in (10). In our model, $\sigma_D(\nu)$ has this feature, but there are two differences. First, the analog of the price from a producer's perspective, σ, is not directly controlled by the monopolist, as this value is derived from users' sharing decisions. Second, fake news prevalence, ν, is in practice *hidden* from users, as these do not really observe the contemporaneous level of prevalence in real time. The non-observability of ν then renders the determination of the equilibrium (ν, σ) effective as a simultaneous-move interaction, whereas a traditional monopoly problem is inherently sequential, with the monopolist's action taken first and being observed by buyers.

Specifically, suppose that users conjecture an equilibrium prevalence ν^M, which leads to a sharing rate $\sigma^M = \sigma_D(\nu^M)$. Critically, this latter value is constant from the monopolist perspective: taking the first-order condition in (10) using $\sigma = \sigma^M$ as given then leads to $\sigma_D(\nu^M) = \sigma_S(\nu^M)$ when the users' conjecture is correct. Thus, $\nu^M = \nu^*$ and the market equilibrium is unchanged.

Segmentation strategies. At the core of the observability issue just studied is the monopolist's inability to steer users' behavior when changing her production decision. In this last part, we explore the interplay between verification efforts and segmentation strategies that also take users' sharing decisions as fixed, and we discuss their implementation.

Our key observation—already illustrated in previous plots—is that the presence of verification incentives, by making the demand for misinformation more sensitive at intermediate levels of prevalence, can introduce *convexities*. These convexities can, in turn, be exploited by market segmentations even when these are trivial, i.e., when the resulting subpopulations preserve the original sharing propensities and differ only in their sizes.

Consider Figure 6, where both panels display a linear supply and verification incentives that tend to make the demand for misinformation more convex. Fixing a level of production ν in the interval determined by the projections of the points A and B on the x-axis, the

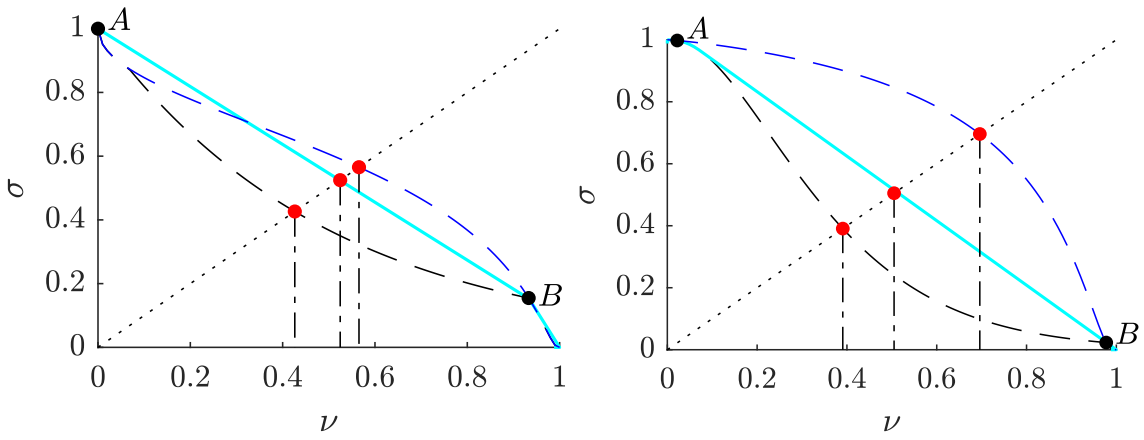

Figure 6: Trivial segmentations as the concave closure of the demand with verification. The left panel assumes an exponential distribution of types with parameter $\lambda = 0.5$ and a verification cost $t = 0.25$, while the right panel a Gamma distribution with parameters 2.5 and 1, while $t = 0.15$. In both cases, $\ell(z) = 1/z$ and $F(r) = r$, the latter assumption leading to a linear supply.

monopolist can achieve a sharing rate at the level of the intersection between the vertical line $x = \nu$ and the segment AB. Indeed, this rate can be attained by segmenting the market into two identical populations of masses α and $1-\alpha$ such that the first receives $\nu_1 \approx 0$ fake news and the second receives $\nu_2 \approx 1$ fake news, where $\nu = \alpha\nu_1 + (1-\alpha)\nu_2$.

Formally, such segmentations allow the monopolist to achieve any point on the *concave closure* of the demand for misinformation, $\sigma_D^{co}(\nu)$.[18] The optimal segmentation can then be found in two steps. First, for any given ν, what is the best way to segment the market? Given our example in the previous paragraph, the maximum payoff attainable by ν is given by $\alpha\sigma_D(\nu_1) + (1-\alpha)\sigma_D(\nu_2)$, which is exactly $\sigma_D^{co}(\nu)$. Second, what is the best aggregate production ν? Suppose that the segments have anticipated their respective levels of prevalence, resulting in total production ν^{**}. The monopolist chooses ν to maximize

$$\sigma_D^{co}(\nu^{**})\nu - \int_0^{\nu} \sigma_S(\nu')d\nu'. \tag{11}$$

Rational expectations and optimizing behavior then lead to the equilibrium characterization $\sigma_D^{co}(\nu^{**}) = \sigma_S(\nu^{**})$, i.e., to a version of the competitive equilibrium condition now involving the concave closure of the demand function. In Figure 6, this equilibrium is represented by the intermediate dot along the diagonal: the negative effects of verification efforts on profits are then partially mitigated by this segmentation technique. Moreover, when the demand without verification is convex in the first place, the monopolist strictly increases its payoff relative to a uniform policy.

[18] The concave closure of a function is the smallest concave function that is weakly larger than that function.

Let us conclude with three observations regarding this analysis. First, this type of segmentation is not equivalent to randomizing between two different levels of prevalence applied to the whole population: ν^{**} must be produced with probability 1 so that the cost of the last unit of fake news produced is exactly $\sigma_S(\nu^{**})$. Also, since the segmentations are trivial, the monopolist must rely on information orthogonal to propensity to share (e.g., geographic location after controlling for the observable characteristics defining the target population).[19]

Second, the optimization over ν in (11) considered deviations from the total production ν^{**} but not deviations on how to "split" ν^{**}. That is, implementing this segmentation requires the monopolist to overcom the temptation to send its production to the segment with the highest sharing rate, analogous to the standard commitment assumption widely used in the persuasion/information design literature (e.g., Kamenica and Gentzkow, 2011).

Finally, while the monopolist is unable to steer the behavior of the segments "in any period" exactly as it occurred in the uniform case ($\sigma_D^{co}(\nu^*)$ is fixed in (11)), a mechanism through which the monopolist signals the prevalence level corresponding to each segment is likely necessary. A reputational mechanism based on the consistency of past behavior could work, provided that each subpopulation can observe its own levels of prevalence from previous periods. Thus, a platform that preserves the opacity of the prevalence of fake news could be more successful in limiting the exercise of market power than a transparent counterpart.

6 Internal Filters

We conclude our analysis with an examination of the effects of using platform-based detection algorithms. The main concern in the public debate has pertained to the perception that removing content before reaching users is a form of censorship (Lazer et al., 2018). By contrast, we provide an economic rationale for the cautious and informed use of such approach.

We consider a situation in which an algorithm screens any news that enters the platform before it becomes visible to users. Clearly, there are direct social losses when such *filters* incorrectly eliminate truthful news. Thus, we focus on the more interesting case in which truthful news always survive, but fake news are detected with probability $\phi \in [0,1]$. In this latter case, these news items are removed from the platform so $\nu(1-\phi)$ is the effective amount of fake news that reaches the platform's users if the production of fake content is ν. Because of the public announcements that platforms have made on this topic, we assume that changes in ϕ are observable to both users and producers; moreover, we are interested on the effects of *introducing* such filters, captured by increasing ϕ starting from zero.

[19] In practice, it easy to control population sizes when targeting online in platforms such as Facebook: for instance, by setting different monetary budgets in the locations of interest.

Our analysis from Section 3 admits a straightforward adaptation to this case. To understand why, let

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

denote the *posterior chance* that a news item is fake upon encountering it given a filter quality $\phi \in [0,1]$ and a fraction $\nu \in (0,1)$ of the total content produced being false. Equipped with this posterior—which falls as ϕ rises and as ν decays—we proceed in an analogous fashion:

1. First, we construct the threshold type—i.e., the lowest type willing to share absent the verification option—as the solution $\bar{z}(\nu, \phi) \equiv \bar{z}(\psi(\nu, \phi))$ to

$$(1 - \psi(\nu, \phi))\bar{z}(\nu, \phi) - \psi(\nu, \phi)\ell(\bar{z}(\nu, \phi)) = 0.$$

2. Following analogous arguments, verification incentives will be at play if and only if this threshold type is willing to verify news, i.e., if $\psi(\nu, \phi)\ell(\bar{z}(\nu, \phi)) - t \geq 0$. In Appendix A.5, we show that the new sharing downside $\nu \mapsto \psi(\nu, \phi)\ell(\bar{z}(\nu, \phi))$ remains quasi-concave in this augmented model, which implies that there is an interval $\nu \in (\underline{\nu}, \bar{\nu})$ over which verification incentives arise. The highest type verifying news at a level of prevalence ν is then given by

$$\bar{z}_h(\psi(\nu, \phi)) \equiv \ell^{-1}\left(\frac{(1-\phi\nu)}{(1-\phi)\nu}t\right),$$

which we extend to $[0,1]\backslash[\underline{\nu}, \bar{\nu}]$ via $\bar{z}^h \equiv \bar{z}$. Obviously, this type falls as ϕ rises, and once again, it rises with ν. The mass of users who share news without engaging in verification is then given by $\sigma_D(\psi(\phi, \nu))) \equiv 1 - G(\bar{z}_h(\psi(\phi, \nu)))$.

The relevant notion of demand in this case, however, is the *effective* demand for misinformation, i.e., the downward-sloping mapping

$$\nu \mapsto \sigma_D^e(\psi(\phi, \nu))) \equiv (1-\phi)\sigma_D(\psi(\phi, \nu))), \tag{12}$$

as a piece of fake news yields a payoff $\sigma_D(\psi(\phi, \nu)))$ only when it passes the filter. Each producer then takes as given the (candidate) *effective sharing rate of unverified news*, $\sigma^e \in [0,1]$, which leads to a supply curve $\nu_S(\sigma^e) = 1 - F(1-\sigma^e)$ as in (7). A competitive equilibrium is characterized by equating the effective demand with the (inverse) supply.

The presence of a filter implies that encountering news is "good news" for the perspective of any user: we have that $\psi(\nu, \phi) < \nu$ for all $\nu \in (0,1)$; i.e., the user is now more optimistic about the veracity of the news at hand. With more optimism, there is more unverified sharing

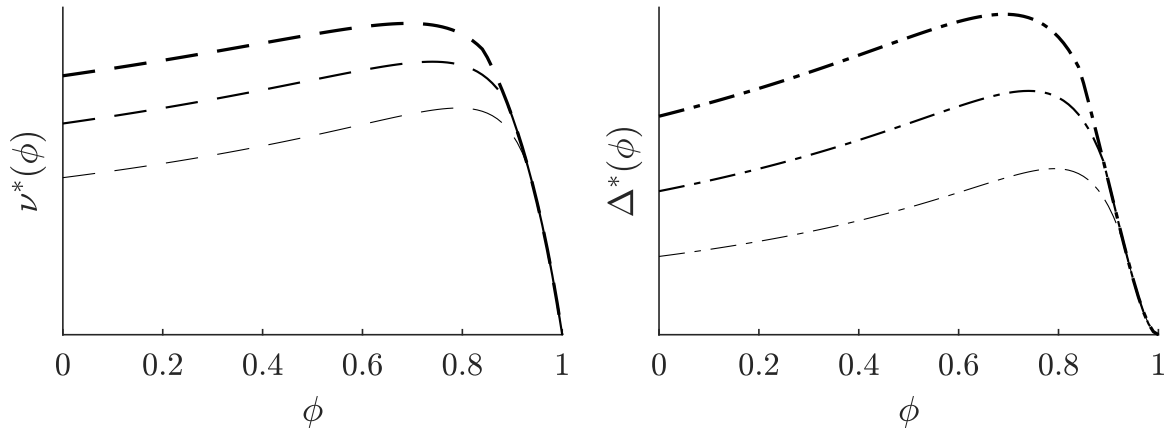

Figure 7: Both panel assume $G(z) = 1 - e^{-2.5z}$ and $F(r) = r^{2.5}$. Verification costs range from $t = 0.1, 0.15, 0.2$ moving upwards from thin to thick lines.

conditional on the encounter taking place, as reflected in the (extended) marginal type falling, $\bar{z}_h(\psi(\phi, \nu)) < \bar{z}_h(\nu)$—in other words, the users' verification incentives *relax*. Critically, this effect can outweigh the negative direct effect that the filter has on the incentives to produce news—captured by $1 - \phi$ in the effective demand (12)—resulting in both a higher prevalence and a larger diffusion of fake news. Indeed, this effect is likely stronger when the filter has low quality, but depending on the primitives of the environment it can actually prevail over an extensive region of filter qualities. Figure 7 shows that equilibrium prevalence ν^* and the resulting diffusion Δ^* fall only for high values of ϕ (trivially attaining zero when $\phi = 1$).

Formally, we say that the demand $\sigma_D(\nu)$ is *sufficiently elastic at* $\nu < 1$ if

$$\varepsilon_D \equiv \frac{\nu \sigma_D'(\nu)}{\sigma_D(\nu)} < -\frac{1}{1-\nu}. \tag{13}$$

This condition guarantees that when the filter has a sufficiently low quality, a 1% increase in ϕ prompts more sharing of unverified news σ_D by more than 1%, leading to an increase in the effective sharing rate for a fake news prevalence ν.

Proposition 3. *Suppose that when $\phi = 0$, the demand for misinformation is sufficiently elastic at the corresponding equilibrium ν^*. Then, as ϕ rises, both the equilibrium prevalence, ν^*, and the rate of diffusion, Δ^*, initially rise and then eventually fall. The sharing rate of unverified news σ^* rises as ϕ increases.*

The conditions that ensure that the demand for misinformation is sufficiently elastic are similar to those presented in part (ii) of Proposition 2—we report them in the proof of Proposition 3 in Appendix A.5.

In general, the rate of diffusion of fake news, Δ^*, starts rising faster as ϕ increases when the distribution G falls in the monotone likelihood ratio (MLR) sense. By standard results (e.g., Athey, 2002), when G falls in this sense users have a lower propensity to share on average, so demand $\sigma_D(\cdot)$ falls. In addition, the dispersion of such propensities rises.

Proposition 4. *Consider the curve $\phi \mapsto \Delta^*(\phi)$, and suppose that the demand σ_D is sufficiently elastic for low enough qualities of the filter. Then, as G falls in the MLR sense, the fake news diffusion rate $\Delta^*(\cdot)$ shifts downward but rises faster in ϕ.*

To make the result concrete, consider the introduction of yet another type of detection algorithm: one that attempts to eliminate *social bots*, a software in the form of fake accounts whose goal is to speed up the dissemination of misinformation (Ferrara et al., 2016; Shao et al., 2018). In our model, the effect of such an algorithm can be seen as the removal of very high types z's, resulting in a drop of the distribution G in the sense described above. Proposition 4 then shows the effects of a joint intervention consisting on both removing bots and filtering news: an environment with lower sharing propensities is more susceptible to a relaxation of users' incentives once news filters are introduced, potentially leading to greater increases in the creation and diffusion of fake news.

7 Concluding Remarks

Since the 2016 U.S. presidential election it has become more clear than ever that misinformation and its diffusion on social media need for prompt and accurate verification of information on behalf of all of us. From this perspective, our goal has been to develop a model of a platform intended to examine the effectiveness of real-world policies aimed at alleviating the problem of fake news in social media, placing special emphasis on how these policies interplay with verification incentives by users. Our model highlights supply and demand forces that are central to the creation and diffusion of fake news while illustrating how key defining features of this market introduce some nuances relative to standard competitive analyses. Equally important, our work emphasizes the importance of sensitivity analysis for policy evaluation: elasticity measures analogous to those in traditional markets naturally emerge here, the knowledge of which turns out to be critical for assessing policy interventions.

Let us conclude with a discussion of some of our assumptions, as well as other potential interventions. First, the static nature is clearly a simplification: news transmission is a dynamic process, which means that users can encounter news of different vintages, verify information at different times, eliminate posts, learn from others' actions, etc. Regardless of the model assumed, however, the fraction of individuals who share without doing any

verification remains key from the perspective of producers. Furthermore, since producers can target users up to some degree of granularity, it is not unreasonable to expect that the discounted benefits of fake news diffusing through a network are a function of that "initial" fraction of individuals. To a first-order approximation, the model is likely to be capturing a critical variable shaping the supply of misinformation.

Second, our assumptions on the demand side aim to depict a simplified version of a more general world in which user heterogeneity (z, ℓ) is truly two-dimensional. Specifically, our assumptions on the loss function $\ell(\cdot)$ imply that (i) the gains and losses are negatively correlated across types and that (ii) users with a high propensity have weak incentives to verify—a combination that results in a positive demand for misinformation at all non-trivial levels of prevalence. With sufficient heterogeneity in each dimension in (z, ℓ), one would expect similar properties to arise in a more general world. In Appendix B, we examine the polar opposite one-dimensional version of our model exhibiting positive correlation. If losses $\ell(\cdot)$ are convex, the induced demand for misinformation has the same qualitative features compared to our baseline case, as users with high propensity to share also have weak incentives to verify. Yet, if losses are concave, the demand features regions of prevalence wherein there is *no* sharing of unverified news. Since it is a priori difficult to eliminate the possibility of sufficient two-dimensional heterogeneity, we view our demand side as the simplest one-dimensional version of that more realistic general setting.

Third, we have assumed that the user is fully rational: in particular, the user has perfect knowledge of the details of the environment, is Bayesian, and holds correct expectations in equilibrium. Clearly, there are behavioral/cognitive aspects that could play a role in this market, but we believe that the benchmark analyzed is of particular importance. Indeed, a sizable fraction of the efforts by platforms, fact-checking organizations, and journalist associations have been devoted to educational programs aimed at fostering user literacy in evaluating fake news.[20] This "sophisticated" benchmark, therefore, need not be too distant.[21]

Finally, our choice of policy analysis has been guided by its importance in practice and in the public debate, but other alternatives are available. It is easy to see how making the transmission of news more costly (e.g., via additional clicks) would reduce the demand for misinformation in our model, but it would also negatively affect the diffusion of truthful content. More interesting is the use of algorithms that are less invasive than those that

[20] For instance, Facebook has funded the News Integrative Initiative (https://www.journalism.cuny.edu/centers/tow-knight-center-entrepreneurial-journalism/news-integrity-initiative/) as a long-term tactic to alleviate fake news (Bernstein, 2017; Lyons, 2018).

[21] A similar trend towards educating consumers has emerged in response to *privacy* considerations, i.e., concerns related to adverse uses of the largely data about consumers collected online, a phenomenon consumers are increasingly aware of. See Bonatti and Cisternas (2020) for an application to price discrimination.

censor news: for instance, routing news to different individuals based on past behavior, as well as signals about the news at hand that are picked by the algorithm. These and other topics are the subject of ongoing research.

A Omitted Proofs

A.1 Proof of Lemma 1

First, notice that the sharing downside function $\nu\ell(\bar{z}(\nu))$ vanishes when $\nu = 0$ and $\nu = 1$, since $\ell(\bar{z}(1)) = \ell(z_{max}) = 0$. Also, $\nu\ell(\bar{z}(\nu)) > 0$ for all $\nu \in (0,1)$. Thus, $\nu\ell(\bar{z}(\nu)) > 0$ has maximizer $\nu^\dagger \in (0,1)$, and this maximizer must solve the first-order condition (FOC): $\ell(\bar{z}) + \nu\ell'(\bar{z})\bar{z}' = 0$. We will show that the FOC has a unique solution, implying that $\nu\ell(\bar{z}(\nu))$ must be single-peaked, and so it must be quasi-concave. To see this, we exploit the definition of $\bar{z}(\nu)$. Differentiate (2) with respect to ν to get $-\bar{z} + (1-\nu)\bar{z}' = \ell(\bar{z}) + \nu\ell'(\bar{z})\bar{z}'$. Hence, we must have $\bar{z}' = \bar{z}/(1-\nu)$ at an optimum of the sharing downside function. But, by (2), $\bar{z} = \nu\ell(\bar{z})/(1-\nu)$, and also $\ell(\bar{z}) = -\nu\ell'(\bar{z})\bar{z}'$ by the FOC. Altogether, for $\nu = \nu^\dagger$ we have:

$$\bar{z}' = \frac{\bar{z}}{1-\nu} = \frac{\nu\ell(\bar{z})}{(1-\nu^2)} = -\left(\frac{\nu}{1-\nu}\right)^2 \ell'(\bar{z})\bar{z}',$$

which implies that $\ell'(\bar{z}) = -[(1-\nu)/\nu]^2$, since $\bar{z}' > 0$. Dividing both sides by $\ell(\bar{z})$ and using (2) we find:

$$\frac{\bar{z}\ell'(\bar{z})}{\ell(\bar{z})} = -\left(\frac{1-\nu}{\nu}\right). \tag{A.1}$$

Because the elasticity of ℓ is decreasing in z, and $\bar{z}$ is increasing in ν, the map $\nu \mapsto \frac{\bar{z}\ell'(\bar{z})}{\ell(\bar{z})}$ is decreasing in ν. Following the same logic, the map $\nu \mapsto -\left(\frac{1-\nu}{\nu}\right)$ is strictly increasing in ν. Thus, (A.1) admits a unique solution, and hence the downside sharing function must have a unique critical point. We conclude that $\nu\ell(\bar{z}(\nu))$ is single-peaked, and thus the set $\{\nu : \nu\ell(\bar{z}(\nu)) \leq t\}$ is a non-empty closed interval. □

A.2 Proof of Proposition 1

Proof of part (a): First, since $\sigma_D(\cdot) : [0,1] \mapsto [0,1]$ is strictly decreasing in ν, its inverse function $\nu_D(\cdot) \equiv \sigma_D^{-1}(\cdot) : [0,1] \mapsto [0,1]$ is well-defined and strictly decreasing. Also, observe that this function is continuous.

Next, we examine the *excess of supply* function $\nu_S(\cdot) - \nu_D(\cdot)$, which is also continuous. Clearly, when all consumers share unverified news, i.e., $\sigma = 1$, we have $\nu_S(1) - \nu_D(1) =$

$1 - F(0) - 0 > 0$. Conversely, when no consumer share unverified news, $\nu_S(0) - \nu_D(0) = [1 - F(1)] - 1 < 0$. By the Intermediate Value Theorem (IVT), there exists $\sigma^* \in (0, 1)$ such that the excess of supply vanishes, or $\nu_S(\sigma^*) = \nu_D(\sigma^*)$. Finally, since the supply $\nu_S(\cdot)$ rises in σ, while the inverse $\nu_D(\cdot)$ falls in it, σ^* is unique because the excess of supply strictly falls in σ. □

Proof of part (b): We now show that there exists a cutoff cost $\hat{t} \leq t^\dagger$ such that our unique equilibrium entails a positive mass of consumers verifying news if and only if $t < \hat{t}$.

To this end, let ν^o denote the market equilibrium when the possibility of verification is not available (or, alternatively, when $t > t^\dagger$); this value exists and is unique in light of part (a). Also, define the level of prevalence

$$\nu^\dagger := \arg\max_{\nu \in [0,1]} \ \nu \ell(\bar{z}(\nu)).$$

By Lemma 1, the lowest and highest solutions to $\nu\ell(\bar{z}(\nu)) = t$, namely $\underline{\nu}(t)$ and $\bar{\nu}(t)$, respectively, are strictly increasing and decreasing in $t < t^\dagger$. Moreover, $\nu^\dagger \in [\underline{\nu}, \bar{\nu}]$, and for $t = 0$ we have $\underline{\nu}(0) = 0$ and $\bar{\nu}(0) = 1$, since $\bar{z}(1) = z_{max}$ and $\ell(z_{max}) = 0$.

Suppose now that $\nu^0 < \nu^\dagger$. We then define out verification cost of interest as

$$\hat{t} := \inf\{t \leq t^\dagger | \ \underline{\nu}(t) = \nu^o\}.$$

(By continuity, this infimum is attained, and is strictly less than $t^\dagger$.)

Indeed, let $\sigma_D(\nu; t)$ denote the demand function parametrized by t, and observe that $\sigma_D(\nu; t) \leq \sigma_D(\nu; t^\dagger)$ with strict inequality only over $(\underline{\nu}(t), \bar{\nu}(t))$. Also, let $\sigma_S(\nu) = 1 - F^{-1}(1 - \nu)$ denote the inverse supply function which, critically, is unaffected by changes in t. For $t < \hat{t}$, therefore, we have that $\nu^o \in [\underline{\nu}, \bar{\nu}]$, and so the threshold type $\bar{z}(\nu^o)$ is (strictly) willing to verify news—a positive mass of types above it will be also willing to verify news by continuity, as desired. For $t > \hat{t}$, however, $\nu^o \notin [\underline{\nu}(t), \bar{\nu}(t)]$ and, hence,

$$\sigma_D(\nu; t) = \sigma_D(\nu; t^\dagger) \ \text{ over } \ [0, \underline{\nu}(t)] \text{ and } \sigma_D(\nu; t) \leq \sigma_D(\nu; t^\dagger) < \sigma_S(\nu) \ \text{ for } \ \nu > \underline{\nu}(t),$$

i.e., the equilibrium continues to be ν^o.

If instead $\nu^o > \nu^\dagger$, it follows that $\hat{t} := \inf\{t \leq t^\dagger \mid \bar{\nu}(t) = \nu^o\}$ is our desired threshold by an analogous argument. Also, by continuity, $\hat{t} < t^\dagger$ as long as $\nu^o \neq \nu^\dagger$. This concludes the proof. □

A.3 Proof of Proposition 2

Proof of part (a). Suppose that $1-G$ is log-concave and ℓ is log convex. Fix $t < \hat{t}$, and $\nu \in [\underline{\nu}(t), \bar{\nu}(t)]$. Observe, in particular, that $\nu \in [\underline{\nu}(t'), \bar{\nu}(t')]$ for all $t' < t$. At this level of prevalence, $\sigma_D(\nu) = 1 - G(\bar{z}_h(\nu))$ with $\bar{z}_h(\nu)$ obeying $\ell(\bar{z}_h(\nu)) \equiv t/\nu$. Log-differentiate this latter expression to get:

$$\frac{\ell'(\bar{z}_h(\nu))}{\ell(\bar{z}_h(\nu))}\bar{z}_h'(\nu) = -\frac{1}{\nu} \quad \Longrightarrow \quad \nu\bar{z}_h'(\nu) = -\frac{\ell(\bar{z}_h(\nu))}{\ell'(\bar{z}_h(\nu))}.$$

Consequently,

$$\varepsilon_D(\nu;t) := \frac{\nu\sigma_D'(\nu)}{\sigma_D(\nu)} = \frac{-g(\bar{z}_h)}{1-G(\bar{z}_h)}\nu\bar{z}_h' = \frac{g(\bar{z}_h)}{1-G(\bar{z}_h)} \cdot \frac{\ell(\bar{z}_h)}{\ell'(\bar{z}_h)}.$$

Finally, differentiate this elasticity expression in t to get:

$$\frac{\partial \epsilon_D}{\partial t} = \frac{\partial}{\partial t}\left(\frac{g(\bar{z}_h)}{1-G(\bar{z}_h)}\right)\frac{\ell(\bar{z}_h)}{\ell'(\bar{z}_h)} + \frac{g(\bar{z}_h)}{1-G(\bar{z}_h)}\frac{\partial}{\partial t}\left(\frac{\ell(\bar{z}_h)}{\ell'(\bar{z}_h)}\right)$$

The first term on the right-hand side is positive due to (i) $\bar{z}_h$ falling as t rises, (ii) $z \mapsto g(z)/(1-G(z))$ being increasing by log-concavity, and (iii) $\ell/\ell' < 0$. Similarly with the second term, as $z \mapsto \ell(z)/\ell'(z)$ falls by log-convexity, while $\bar{z}_h$ decreases with t.

Proof of part (b): Consider $\nu \in [\underline{\nu}(t), \bar{\nu}(t)]$. From the proof of part (a), $\epsilon_D(\nu;t) < -1$ if and only if

$$\frac{g(\bar{z}_h)}{1-G(\bar{z}_h)} \cdot \frac{\ell(\bar{z}_h)}{\ell'(\bar{z}_h)} < -1 \Leftrightarrow z_h\frac{g(\bar{z}_h)}{1-G(\bar{z}_h)} + z_h\frac{\ell'(\bar{z}_h)}{\ell(\bar{z}_h)} > 0,$$

as desired, where we used that $\ell/\ell' < 0$.

As for $\nu \notin [\underline{\nu}(t), \bar{\nu}(t)]$, recall that the demand elasticity is determined by $\bar{z}(\nu)$ obeying $\nu\ell(\bar{z}(\nu)) \equiv (1-\nu)\bar{z}(\nu)$ in that region. Log-differentiate this latter expression to get

$$\frac{-1}{1-\nu} + \frac{\bar{z}'}{\bar{z}} = \frac{1}{\nu} + \frac{\ell'}{\ell}\bar{z}' \qquad \Longrightarrow \qquad \frac{\bar{z}'}{\bar{z}} = \frac{[\nu(1-\nu)]^{-1}}{1-\bar{z}\ell'/\ell}.$$

In particular, $1 - z\ell'/\ell > 0$ at $\bar{z}$. Hence, the demand elasticity can be written as:

$$\varepsilon_D(\nu;t) := \frac{\nu\sigma_D'(\nu)}{\sigma_D(\nu)} = \frac{-g(\bar{z})}{1-G(\bar{z})}\nu\bar{z}' = \frac{-\bar{z}g(\bar{z})}{1-G(\bar{z})} \cdot \frac{1}{1-\bar{z}\ell'(\bar{z})/\ell(\bar{z})} \cdot \frac{1}{1-\nu}. \tag{A.2}$$

Thus,

$$\frac{\bar{z}g(\bar{z})}{1-G(\bar{z})} + \frac{\bar{z}\ell'(\bar{z})}{\ell(\bar{z})} > 1 \Rightarrow \epsilon_D = \frac{-\bar{z}g(\bar{z})}{1-G(\bar{z})} \cdot \frac{1}{1-\bar{z}\ell'(\bar{z})/\ell(\bar{z})} < -1.$$

□

A.4 Proof of Lemma 2

Recall that the monopolist's profits at ν are given by $\Pi(\nu) := \sigma \cdot \nu + \int_0^{F^{-1}(1-\nu)}(1-r)f(r)dr$. Adding and subtracting $\int_{F^{-1}(1-\nu)}^1 (1-r)f(r)dr$ we can write

$$\begin{aligned}\Pi(\nu) &= \sigma \cdot \nu - \int_{F^{-1}(1-\nu)}^{\nu}(1-r)f(r)dr + \underbrace{\int_0^1 (1-r)f(r)dr}_{=:K,\text{ constant}} \\ &= \int_{F^{-1}(1-\nu)}^1 [\sigma - (1-r)]f(r)dr + K\end{aligned}$$

where in the last equality we used that $\sigma \cdot \nu$ is independent of r and that $\int_{F^{-1}(1-\nu)}^1 f(r)dr = \nu$.

Consider now the change of variables $r = 1 - \sigma_S(\nu')$ where $\sigma_S(\nu) := 1 - F^{-1}(1-\nu)$ denotes the inverse supply function. When $r = 1$ we have $\nu' = 0$, because the inverse supply $\sigma_S(0) = 0$. Likewise, when $r = F^{-1}(1-\nu)$, we have $\nu' = \nu$ given the definition of $\sigma_S(\cdot)$. Finally, $dr = -\sigma_S'(\nu')d\nu' = -[1/f(F^{-1}(1-\nu'))]d\nu'$. Altogether,

$$\begin{aligned}\int_{F^{-1}(1-\nu)}^1 [\sigma - (1-r)]f(r)dr &= -\int_{\nu}^0 [\sigma - \sigma_S(\nu')]f(1-\sigma_S(\nu'))\frac{1}{f(F^{-1}(1-\nu'))}d\nu' \\ &= \int_0^{\nu}[\sigma - \sigma_S(\nu')]d\nu' \\ &= \sigma\nu - \int_0^{\nu}\sigma_S(\nu')d\nu',\end{aligned}$$

where we have used σ is independent of ν'. This concludes the proof. □

A.5 Proof of Proposition 3

Quasi-concavity of the sharing downside function. We first prove the claim on the sharing downside function remaining quasi-concave.

Claim A.1. *Fix $\phi < 1$. The mapping $\nu \mapsto \left[\frac{(1-\phi)\nu}{1-\phi\nu}\right]\ell(\bar{z}(\nu,\phi))$ is quasi-concave.*

Proof: It is clear that $\left[\frac{(1-\phi)\nu}{(1-\phi\nu)}\right]\ell(\bar{z}(\nu,\phi))$ vanishes at $\nu = 0$, and also at $\nu = 1$ since $\bar{z}(1,\phi) = z_{max}$ and $\ell(z_{max}) = 0$. Thus, its maximum must be interior, and thus obey the FOC:

$$\frac{1-\phi}{(1-\nu\phi)^2}\ell(\bar{z}) + \frac{(1-\phi)\nu}{1-\phi\nu}\ell'(\bar{z})\bar{z}_\nu = 0, \tag{A.3}$$

where $\bar{z}_\nu$ denotes the partial derivative of $\bar{z}$ with respect to ν. However, by definition of $\bar{z}$ we have $\left[\frac{(1-\phi)\nu}{(1-\phi\nu)}\right]\ell(\bar{z}(\nu,\phi)) \equiv \left[\frac{1-\nu}{(1-\phi\nu)}\right]\bar{z}(\nu,\phi)$, and so the above FOC can be expressed as:

$$-\frac{1-\phi}{(1-\nu\phi)^2}\bar{z} + \frac{1-\nu}{1-\phi\nu}\bar{z}_\nu = 0.$$

Thus, $\bar{z}_\nu = \frac{1-\phi}{(1-\nu\phi)(1-\nu)}\bar{z}$. Plugging this expression for $\bar{z}_\nu$ into (A.3) yields an analog to equation (A.1):

$$\frac{\bar{z}\ell'(\bar{z})}{\ell(\bar{z})} = -\left(\frac{1-\nu}{\nu(1-\phi)}\right).$$

By the same reasons given in the proof of Lemma 1, there is a unique value of ν that solves this equation, from where we conclude. □

Turning to the proof of the Proposition, let $\nu^*(\phi)$ denote the equilibrium prevalence given a filter of quality ϕ. This value is the unique solution to the equation

$$\underbrace{(1-\phi)\sigma_D(\psi(\nu^*(\phi),\phi))}_{\sigma_D^e(\psi(\nu^*(\phi),\phi))\equiv} = \sigma_S(\nu^*(\phi))$$

where $\sigma_S(\nu) := 1 - F^{-1}(1-\nu)$ is the inverse supply, which is upward sloping. Totally differentiating with respect to ϕ, we obtain

$$-\sigma_D + (1-\phi)\sigma_D'\left[\frac{\partial\psi}{\partial\nu}[\nu^*]'(\phi) + \frac{\partial\psi}{\partial\phi}\right] = \sigma_S'[\nu^*]'(\phi) \Rightarrow [\nu^*]'(\phi) = \frac{-\sigma_D + (1-\phi)\sigma_D'\frac{\partial\psi}{\partial\phi}}{\sigma_S' - (1-\phi)\sigma_D'\frac{\partial\psi}{\partial\nu}} \quad \text{(A.4)}$$

Since $\sigma_S' > 0 > \sigma_D'$ and $\partial\psi/\partial\nu > 0$, it follows that the sign of $[\nu^*]'$ is fully determined by the sign of

$$\chi(\phi) := -\sigma_D(\psi(\nu^*(\phi),\phi)) + (1-\phi)\sigma_D'(\psi(\nu^*(\phi),\phi))\frac{\partial\psi}{\partial\phi}(\psi(\nu^*(\phi),\phi)).$$

Moreover, evaluating the above expression at $\phi = 0$,

$$\chi(0) = -\sigma_D(\nu^*(0)) - \sigma_D'(\nu^*(0))\nu^*(0)(1-\nu^*(0)),$$

where we have used that $\psi(\nu,0) = \nu$ and $\frac{\partial\psi}{\partial\phi}(\nu,0) = -\nu(1-\nu)$. Consequently, when the demand $\sigma_D(\cdot)$ is sufficiently elastic at $\nu = \nu^*(0)$, we have that $\chi(0) > 0$, implying that a small increase in the filter's quality leads to an increase in the equilibrium prevalence $\nu^*(\phi)$. Because a change in the filter has no *direct* effect on the supply function, the new equilibrium is the consequence of an *upward* movement along the supply curve, and hence the new equilibrium displays a higher rate of diffusion of fake news, $\Delta^*(\phi) := \nu^*(\phi)\sigma^*(\phi)$: intuitively,

the small increase in ϕ shifts the effective demand $(1-\phi)\sigma_D(\cdot)$ up in the (ν, σ^e)-space around the point studied. Conversely, as the filter becomes perfect, i.e., $\phi \to 1$, the effective demand shifts left towards the origin, with $(\nu^*(\phi), \sigma^{e,*}(\phi)) \to (0,0)$ and $\Delta^*(\phi) = \sigma^{e,*}(\phi)\nu^*(\phi) \to 0$.

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

Conditions for sufficient elasticity. When will the demand be sufficiently elastic at a prevalence level ν for the case $\phi = 0$?

1. Suppose that there is verification in equilibrium, i.e., $t < \hat{t}$. Using the expression (8) for the elasticity of demand in this region, the condition for sufficient elasticity at ν is

$$\varepsilon_D(\nu) = \frac{\nu[1 - G(\bar{z}_h(\nu))]'}{1 - G(\bar{z}_h(\nu))} = \frac{g(\bar{z}_h(\nu))}{1 - G(\bar{z}_h(\nu))} \frac{t}{\nu} \frac{1}{\ell'(\bar{z}_h(\nu))} < -\frac{1}{1-\nu}. \tag{A.5}$$

   For concreteness, assume that the sharing gains z are exponentially distributed with $1 - G(z) = e^{-\lambda z}$, $\lambda > 0$, and losses are geometric, $\ell(z) = z^{-\gamma}$, $\gamma > 0$. Then, it is easy to see that the condition becomes

$$(1-\nu)\nu^{\frac{1}{\gamma}} > t^{\frac{1}{\gamma}}\gamma/\lambda.$$

   That is, sufficient demand elasticity obtains for intermediate levels of ν, given the quasi-concave nature of the function $\nu \mapsto (1-\nu)\nu^{\frac{1}{\gamma}}$.

2. If instead there no verification in equilibrium, we can use the expression (A.2) for the elasticity of demand in this region, so sufficient elasticity at ν is given by

$$\varepsilon_D(\nu; t) = \frac{-\bar{z}g(\bar{z})}{1 - G(\bar{z})} \cdot \frac{1}{1 - \bar{z}\ell'(\bar{z})/\ell(\bar{z})} \cdot \frac{1}{1-\nu}. < -\frac{1}{1-\nu}. \tag{A.6}$$

   Assuming again that the sharing gains z are exponentially distributed with $1 - G(z) = e^{-\lambda z}$, $\lambda > 0$, and losses are geometric, $\ell(z) = z^{-\gamma}$, $\gamma > 0$, the condition reduces to

$$\lambda \bar{z}(\nu) > 1 + \gamma.$$

   Since $\bar{z}(\cdot)$ is increasing, sufficient elasticity obtains for high enough $\nu > \bar{z}^{-1}\left(\frac{1+\gamma}{\lambda}\right)$.

In either case, sufficient demand elasticity is easier to meet when λ is high enough, i.e.,

when the hazard rate of G large. General conditions that apply to the any G and ℓ within the class under study can be easily obtained from (A.5)–(A.6).

A.6 Proof of Proposition 4

Let us smoothly parametrize $G(z|\theta)$, $\theta \in \mathbb{R}$, with $G(z|0) \equiv G(z)$, and assume that the density $g(z|\theta)$ is log-supermodular in (z, θ). We start with a preliminary result.

Lemma A.1. *Consider distributions G_H and G_L, with densities $g_H(z)$ and $g_L(z)$ respectively, such that $z \mapsto g_H(z)/g_L(z)$ is increasing. Then, the demand for misinformation σ_D is absolutely less elastic with G_H than G_L.*

Proof: Define the parametrized type distribution $(z, \theta) \in \mathbb{R}_+ \times \{0, 1\} \mapsto H(z, \theta) \in [0, \infty]$ by $H(z, 0) \equiv G_L(z)$ and $H(z, 1) \equiv G_H(z)$. By standard results, the density $h(z, \theta)$ is log-supermodular in (z, θ), since the likelihood ratio $h(z, 1)/h(z, 0) = g_H(z)/g_L(z)$ is monotone. Thus, the residual distribution $1 - H(z, \theta) = \int \mathbf{1}_{[z, z_{max}]}(z')h(z', \theta)dz'$ is log-supermodular in (z, θ), since the indicator $\mathbf{1}_{[z, z_{max}]}(z')$ is log-supermodular in (z, z'); see Lemmas 3–4 in Athey (2002). Altogether, $-h(z, \theta)/H(z, \theta)$ is increasing in θ, or:

$$-h(z, 1)/H(z, 1) = \frac{-g_H(z)}{1 - G_H(z)} \geq \frac{-g_L(z)}{1 - G_L(z)} = -h(z, 0)/H(z, 0).$$

Call $\zeta(\nu) \equiv \max\{\bar{z}(\nu), \bar{z}_h(\nu)\}$ (with $\bar{z}_h(\nu)$ defined as $\bar{z}(\nu)$) in the no-verification region). Next, using (5) with the distribution H, for almost every ν, the demand elasticity $\sigma_D(\cdot|\theta)$ is given by:

$$\frac{\nu\sigma_D'(\nu|\theta)}{\sigma_D(\nu|\theta)} = -\frac{h(\zeta(\nu), \theta)}{1 - H(\zeta(\nu), \theta)}\zeta'(\nu).$$

Consequently, σ_D is absolutely less elastic with $\theta = 1$ or G_H than $\theta = 0$ or G_L. □

We now prove the result. Denote $(1 - \phi)\sigma_D(\psi(\nu, \phi)|\theta)$ the effective demand given θ (i.e., with distribution $G(z|\theta)$). We will show that effective demand given θ is supermodular in (ϕ, θ). To see this, differentiate the effective demand in ϕ to get:

$$\begin{aligned}\frac{\partial[(1 - \phi)\sigma_D(\psi(\nu, \phi)|\theta)]}{\partial\phi} &= -\sigma_D(\psi(\nu, \phi)|\theta) + (1 - \phi)\sigma_D'(\psi(\nu, \phi)|\theta)\psi_\phi(\nu, \phi) \\ &= \sigma_D(\psi|\theta)\left[-1 + \frac{\sigma_D'(\psi|\theta)}{\sigma_D(\psi|\theta)}\psi_\phi(1 - \phi)\right]\end{aligned}$$

Therefore,

$$\frac{\phi}{(1 - \phi)\sigma_D(\psi)}\frac{\partial[(1 - \phi)\sigma_D(\psi|\theta)]}{\partial\phi} = \frac{-1}{1 - \phi} + \frac{\sigma_D'(\psi|\theta)}{\sigma_D(\psi|\theta)}\psi_\phi$$

Now observe that the elasticity of the effective demand in ϕ is decreasing in θ, because the demand is absolutely less elastic as θ rises (Lemma A.1 above) and also $\psi_\phi < 0$. Thus, for each ν, if the effective demand rises (respectively falls) in ϕ, then it rises (respectively falls) proportionally less (respectively more) in ϕ when $\theta > 0$. Finally, since the demand σ_D is sufficiently elastic at ν^* (i.e., when $\phi = \theta = 0$), then it will continue to be so for low enough $\phi, \theta > 0$ by continuity. □

B Other Demand Functions: Positive Correlation Between Gains and Losses

By assuming that losses are given by $z \mapsto \ell(z)$ decreasing, our model is one of negative correlation between gains and losses across types. On top of this, $\ell(z_{max}) = 0$ implies that the highest type—i.e., the one with the largest gain from sharing—never verifies.

We now explore the polar opposite case: positive correlation and the highest type verifies. We distinguish between concave and convex losses and restrict to $z_{max} = +\infty$. We will show that with concave losses, there are regions of prevalence in which no type ever shares unverified news. However, if losses are convex, the resulting demand function is qualitatively similar to the one we study in the baseline case, as in this case the propensity to share is also negatively related to the propensity to verify.

B.1 Positive Correlation with Concave Losses: $\ell'' < 0 < \ell'$

High verification costs. In this case, no user verifies news but some still choose to share given the conjectured prevalence ν. Indeed, a user prefers to share when $(1-\nu)z - \nu\ell(z) \geq 0$, or $z/\ell(z) \geq \nu/(1-\nu)$. Since ℓ is increasing and concave, and $\ell(0) = 0$, the mapping $z \mapsto z/\ell(z)$ is increasing. Thus, for $\nu \in (0,1)$ there exists a unique consumer type $\bar{z}(\nu) \in (0, z_{max})$ such that all consumers $z \geq \bar{z}(\nu)$ choose to share, where:

$$(1-\nu)\bar{z}(\nu) - \nu\ell(\bar{z}(\nu)) = 0. \tag{B.1}$$

Clearly, $\bar{z}(\nu)$ rises as ν rises. For $\nu = 0, 1$ we let $\bar{z}(0) = 0$ and $\bar{z}(1) = z_{max}$.

Moderate verification costs. Suppose now that the users can learn the authenticity of news at a moderate verification cost t such that someones verifies. For each level of prevalence ν, consider those who are willing to share even if costs were high, i.e., $z \geq \bar{z}(\nu)$. Then, by

costly verifying news these users improve their sharing payoff by

$$\nu\ell(z) - t.$$

Unlike in the baseline model, the value of verifying—i.e., the savings from not sharing fake news—increases with z. Thus, type $z = \bar{z}$ is the most reluctant to verify, whereas $z = z_{max}$ is the most prone to verify.

Assumption 2. $\ell(z_{max}) = \infty$.

Given this assumption, $z = z_{max}$ always verifies. Now, if it is optimal to verify for $\bar{z}(\nu)$, then it will be optimal for $z \geq \bar{z}$. However, if $\nu\ell(\bar{z}(\nu)) < t$, then there exist a type $\bar{z}_h(\nu) \in (\bar{z}, z_{max})$, with $\bar{z}_h(\nu) \equiv \ell^{-1}(t/\nu)$, such that all types above $z \geq \bar{z}_h$ prefer to verify.

Because, $\ell', \bar{z}' > 0$, the function $\nu\ell(\bar{z}(\nu))$ is increasing in ν, and thus there exists a unique $\bar{\nu}$ such that $\nu\ell(\bar{z}(\nu)) < t$ if and only if $\nu < \bar{\nu}$. To summarize, if $\nu \geq \bar{\nu}$ all types $z \geq \bar{z}$ verify and share. If $\nu < \bar{\nu}$ only types $z \geq \bar{z}_h$ verify and share.

Consider now the users that were not sharing originally, i.e., $z < \bar{z}$. For them, the value of verifying is that now they may share and increase their payoff in an amount equal to $(1-\nu)z - t$, which is increasing in z. Logically, if verifying is suboptimal for type $z = \bar{z}$ then it is suboptimal for all types $z < \bar{z}$. This implies that for a level of prevalence $\nu < \bar{\nu}$, there is no "entry" of new users. However, when $\nu \geq \bar{\nu}$, entry happens: all types $z \in [\bar{z}_l(\nu), \bar{z}(\nu)]$ verify and share news, where $\bar{z}_l(\nu) \equiv t/(1-\nu)$. All told, if $\nu \geq \bar{\nu}$, the market segments in two: users $z \in [\bar{z}_l, z_{max}]$ share and verify, while users $z < \bar{z}_l$ do not share. If $\nu < \bar{\nu}$ then the market segments in three: types $z < \bar{z}$ do not share; types $z \in [\bar{z}, \bar{z}_h]$ share unverified news; and types $z > \bar{z}_h$ share and verify news.

The demand for misinformation, $\nu \mapsto \sigma_D(\nu)$, is given by:

$$\sigma_D(\nu) = \begin{cases} G(\bar{z}_h(\nu)) - G(\bar{z}(\nu)) & \text{if } \nu < \bar{\nu}; \\ 0 & \text{if } \nu > \bar{\nu}. \end{cases}$$

B.2 Positive Correlation with Convex Costs: $\ell' > 0$ and $\ell'' > 0$

High verification costs. Suppose that verification is prohibitively costly. Then, a user will share only if $(1-\nu)z - \nu\ell(z) \geq 0$, or $z/\ell(z) \geq \nu/(1-\nu)$. Since ℓ is increasing and convex, and $\ell(0) = 0$, the mapping $z \mapsto z/\ell(z)$ falls in z, and so for $\nu \in (0,1)$ all consumers $z \leq \bar{z}$ find it optimal to share, where: $(1-\nu)\bar{z}(\nu) - \nu\ell(\bar{z}(\nu)) = 0$. Clearly, $\bar{z}(\nu)$ falls as ν rises. For $\nu = 0, 1$, we let $\bar{z}(0) = z_{max}$ and $\bar{z}(1) = 0$.

Moderate verification costs. Suppose that the users can learn the authenticity of news at a moderate verification cost t. Consider again users types $z \leq \bar{z}$. Then, by costly verifying news these users may improve their sharing payoff in an amount equal to:

$$\nu\ell(z) - t,$$

implying that $\bar{z}(\nu)$ is the most prone type to verify among all $z \leq \bar{z}$.

Now, since $\ell(0) = 0$ it follows that if it is optimal to verify for type $\bar{z}(\nu)$, then it will be optimal for all types $z \in [\bar{z}_h, \bar{z}]$, where $\bar{z}_h(\nu) \equiv \ell^{-1}(t/\nu)$. However, because $\ell' > 0 > \bar{z}'$, the sharing downside function $\nu \mapsto \nu\ell(\bar{z}(\nu))$ is now non-monotone, vanishing when $\nu = 0$ and $\nu = 1$. Under mild regularity conditions, $\nu\ell(\bar{z}(\nu))$ is hump-shaped and thus, $\nu\ell(\bar{z}(\nu)) > t$ if and only if $\nu \in (\underline{\nu}, \bar{\nu})$. All told, if $\nu \in (\underline{\nu}, \bar{\nu})$ all types $z \in [\bar{z}_h, \bar{z}]$ verify and share, whereas types $z \in [0, \bar{z}_h]$ share unverified news. If $\nu \notin (\underline{\nu}, \bar{\nu})$ all types $z \leq \bar{z}$ share unverified news.

Consider now types $z > \bar{z}$, for which the value of verification is $(1 - \nu)z - t$. Logically, if verifying is optimal for type $\bar{z}(\nu)$, then it will also be for all $z > \bar{z}$. Thus, if $\nu \in (\underline{\nu}, \bar{\nu})$ all types $z \in [\bar{z}_h, z_{max}]$ share and verify news. Now, if prevalence $\nu \notin (\underline{\nu}, \bar{\nu})$ then only types $z \in [\bar{z}_l, z_{max}]$ verify news, where $z_l(\nu) \equiv t/(1 - \nu) > \bar{z}$.

Altogether, the demand for misinformation takes the form,

$$\sigma_D(\nu) = \begin{cases} G(\bar{z}_h) & \text{if } \nu \in (\underline{\nu}, \bar{\nu}); \\ G(\bar{z}) & \text{if } \nu \notin (\underline{\nu}, \bar{\nu}). \end{cases}$$

References

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

Allen, J., B. Howland, M. Mobius, D. Rothschild, and D. J. Watts (2020): "Evaluating the fake news problem at the scale of the information ecosystem," *Science Advances*, 6, eaay3539.

Athey, S. (2002): "Monotone comparative statics under uncertainty," *The Quarterly Journal of Economics*, 117, 1187–223.

Bagnoli, M. and T. Bergstrom (2005): "Log-concave probability and its applications," *Economic Theory*, 26, 445–469.

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

Bernstein, L. (2017): "Facebook, Mozilla, and philanthropists fund 14 million initiative to counter fake news," *WJLA-TV*, https://wjla.com/news/nation-world/news-integrity-initiative-seeks-to-counter-fake-news.

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

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.

Ferrara, E., O. Varol, C. Davis, F. Menczer, and A. Flammini (2016): "The rise of social bots," *Communications of the ACM*, 59, 96–104.

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

Gentzkow, M. and J. M. Shapiro (2008): "Competition and Truth in the Market for News," *Journal of Economic Perspectives*, 22, 133–154.

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

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

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

Halon, Y. (2020): "Zuckerberg knocks Twitter for fact-checking Trump, says private companies shouldn't be 'the arbiter of truth'," *Fox News*, https://www.foxnews.com/media/facebook-mark-zuckerberg-twitter-fact-checking-trump.

Henry, E., E. Zhuravskaya, and S. Guriev (2020): "Checking and sharing alt-fact," Tech. rep., Available at SSRN.

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

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

Kranton, R. and D. McAdams (2019): "Social Networks and the Market for News," Tech. rep., Duke University.

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

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

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

Mitchell, A., J. Gottfried, G. Stocking, M. Walker, and S. Fedeli (2019): "Many Americans say made-up news is a critical problem that needs to be fixed," *Pew Research Center*, 5.

Mullainathan, S. and A. Shleifer (2005): "The market for news," *American Economic Review*, 95, 1031–1053.

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

Peltzman, S. (1975): "The effects of automobile safety regulation," *Journal of Political Economy*, 83, 677–725.

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*.

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.

Shao, C., G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer (2018): "The spread of low-credibility content by social bots," *Nature Communications*, 9, 1–9.

Silverman, C. (2016): "This Analysis Shows how Fake Election News Stories Outperformed Real News on Facebook," *BuzzFeed News*, November 16.

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

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

"Fact Checking on Facebook" (2020): "Fact Checking on Facebook," *Facebook Business Help Center*, https://www.facebook.com/business/help/2593586717571940?id=673052479947730.

Global Disinformation Index staff (2019): "The Quarter Billion Dollar Question: The Quarter Billion Dollar Question: How is Disinformation Gaming Ad Tech?" https://disinformationindex.org/wp-content/uploads/2019/09/GDI_Ad-tech_Report_Screen_AW16.pdf.

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

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

Vásquez, J. (2019): "A theory of crime and vigilance," Tech. rep., Smith College.

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