Can Rumors Create Truth? Optimal Debunking of Rumors in Networks

Luca P. Merlino and Nicole Tabasso*

June 19, 2023

Abstract

We study the diffusion of a true and a false message when agents are biased towards one of the messages and able to verify messages. A recipient of a rumor who verifies it becomes informed of the truth. Consequently, a higher rumor prevalence increases the prevalence of the truth. We employ this result to discuss how a planner may optimally choose information verification rates of the population. We find that a planner who aims to maximize the prevalence of the truth may find it optimal to allow rumors to circulate.

JEL classifications: D83, D85.

Keywords: Social Networks, Rumors, Verification.

*Merlino: ECARES, Université libre de Bruxelles, Ave. F.D. Roosevelt 50, CP 114/04, Belgium; email: luca.paolo.merlino@ulb.be. Tabasso: Ca' Foscari University of Venice, Venice, Italy; email: nicole.tabasso@unive.it. Merlino gratefully acknowledges financial support from the Flemish Research Funds (FWO). Tabasso gratefully acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 793769, as well as funding from the European Union - NextGenerationEU.

1 Introduction

The diffusion of rumors, misinformation, or "fake news" has received considerable attention in recent years (e.g., Allcott and Gentzkow, 2017; Lazer et al., 2018). Yet, such information generally diffuses simultaneously with correct information, and possible interactions are often overlooked in the quest to minimize rumor diffusion. In particular, the prevalence of the truth may be the socially more important variable. This is especially the case when being aware of the truth makes a person more likely to adopt a correct behavior, while believing the rumor implies taking the same action as an uniformed agent.

The crucial question is whether there is any difference between fostering the truth and fighting the rumor. While intuition suggests that these are different sides of the same coin, in this paper we show that this is not necessarily the case.

The diffusion of information on social networks is a complex matter and various policies have been suggested to curb the spread of rumors. In the present paper, we focus on one particular aspect, namely the rate at which agents verify messages they receive. Policy makers or online social platforms can influence agents' incentives to verify through various channels. These include direct ones, such as raising *information literacy* rates or publishing guides on how to spot fake news, as done by, e.g., *The New York Times* or *Le Monde*, as well as indirect ones, by investing in education in general. Our main question of interest is the verification rate that a benevolent planner, whose goal it is to maximize the proportion of correctly informed agents in society, would set.

In our model, we describe the diffusion of information using the *SIS* (*Susceptible-Infected-Susceptible*) framework, initially developed in epidemiology, where the network is modeled as the number of meetings each agent has per period. On this network, two messages pertaining to the true state of the world diffuse via word of mouth. In particular, one is correct about the true state, and the other not (the rumor). Agents belong to one of two types, each biased towards believing one of these messages. Importantly, agents who do not verify ignore messages not in line with their bias.[1] Verification instead is able to reveal the veracity of information. Consequently, irrespective of which message agents receive, if they verify it, they become aware of the true state of the world. Finally, agents only pass on information they believe to their neighbors.

We find that, in steady state, rumor prevalence is strictly decreasing in verification rates; in fact, high enough verification rates are able to eradicate the rumor entirely. The prevalence of the truth is increasing in verification rates if the rumor dies out; but if the rumor survives, truth prevalence is actually increasing in rumor prevalence. The mechanism at work is that, as verification of a rumor reveals the truth, there are some agents who become aware of the truth after receiving a rumor. This is particularly relevant for those agents who, absent verification, would ignore the truth. Thus, an increase in verification rates may increase or decrease the prevalence of the truth.

For a planner aiming to maximize the truth's prevalence by inducing verification rates of the population, the optimal policy depends on the available budget to do so. For either a very low or very high budget, it is optimal to use all of it, if possible until the rumor is completely debunked, and beyond. However, for intermediate levels of the budget, it may be better to induce lower verification rates, which allow the rumor to circulate.

We extend our results along various dimensions. First, similar predictions obtain when the planner can target verification rates to agents who are biased against the truth. In particular, the planner may choose to diversify verification rates across groups, even if a focus on the group biased towards the rumor would allow its complete eradication. Next, similar results would

[1] This assumption captures the concept of *information avoidance* (Golman, Hagmann and Loewenstein, 2017), which we discuss in detail in Section 2.

hold if the planner's objective is to maximize the overall volume of messages, which is likely if the planner manages online social media.

We focus on the problem a planner faces when they are able to set verification rates or, alternatively, affect agents' incentives to verify messages through policy. This is a complementary problem to questions of strategic diffusion of messages (Bloch, Demange and Kranton, 2018; Kranton and McAdams, 2020; Bravard et al., 2023; Acemoglu, Ozdaglar and Siderius, 2022). Papanastasiou (2020) studies the decision of agents and a platform to verify messages in a herding model à la Banerjee (1993) to minimize the probability that there is a rumor cascade. Instead, our main question of interest is the diffusion of *truthful* messages in the presence of misinformation. A comparison of truthful and incorrect message diffusion is also the focus of Merlino, Pin and Tabasso (2023). However, there the focus is more on the diffusion of opinions, while here it is on the diffusion of messages. We discuss more in detail this difference and its implications in Section 2, after we present the model.

Verification in our paper acts very much like vaccination against a disease, as it inoculates agents against believing a rumor. This relates us to papers that focus on strategic decisions to protect one against the diffusion of a disease (Chen and Toxvaerd, 2014; Goyal and Vigier, 2015; Toxvaerd, 2019; Talamàs and Vohra, 2020; Bizzarri, Panebianco and Pin, 2021). In particular, Galeotti and Rogers (2013) employ the SIS model to investigate how a planner would allocate vaccinations among two groups in the population. In contrast to these papers, our focus is not how protection affects the harmful state, but instead its impact on the prevalence of the truth, a positive state. Furthermore, while in these papers protection is a local public good (Kinateder and Merlino, 2017, 2023), this is not true in our framework.

Related to our work, Tabasso (2019) and Campbell, Leister and Zenou (2019) study the simultaneous diffusion of two types of information. However, in these papers the two information may be held by agents simultaneously.

The paper proceeds as follows. Section 2 introduces the model. Section 3 presents the results of the diffusion process and Section 4 solves the planner's problem. Extensions are presented in Section 5 and Section 6 concludes. All proofs are in the Appendix.

2 The Model

We start by formally introducing information, agents and the diffusion process. Next, we derive the differential equations that arise from it, and set up the planner's problem. To end the section, we discuss the main assumptions of the model.

Information. Time, indexed by t, is continuous. There exist two messages $m \in \{0, 1\}$ that diffuse simultaneously on the network. These messages convey information about the state of the world, $\Phi \in \{0, 1\}$. Without loss of generality, we assume that the true state of the world, unknown to the agents, is $\Phi = 0$. Hence, we refer to $m = 0$ as the "truth", and $m = 1$ the "rumor".

Agents. We consider an infinite population of mass 1, whose members are indexed by i. The population is partitioned into two groups, denoted by $b = \{0, 1\}$. We assume that mass $x \in [0, 1]$ of the population are of type $b = 0$ and mass $1 - x$ are of type $b = 1$. The type of an agent is determined by their *information bias*. Specifically, we assume that an agent of type b who does not verify a message is only susceptible to the message $m = b$ and ignores the other message.

A proportion $\alpha \in [0, 1]$ of the population verifies a message upon receiving it.[2] After verification, the agent becomes aware of the true state of the world, and accepts it. Thus, after receiving message m, an agent of type b believes

[2] As we discuss below, while we take verification as fixed, we discuss in Section 5 how this can be seen as a reduced form problem of one in which this is an individual choice, while the planner can affect the cost of verification.

it in the case that either, *(i)* the message is in line with their type, $m = b$, or *(ii)* the agent has verified it and therefore understands that the message is correct. Independently of their type, we assume that agents are better off if they are correctly informed about the true state of the world.

Agents are classified as being either in state S (*Susceptible*) or in state I (*Infected*). Specifically, agents are in state S if they are unaware of both messages, or if they ignore a message they have received. Figure 1 summarizes which opinion an infected agent holds depending on her type, the message received and verification.

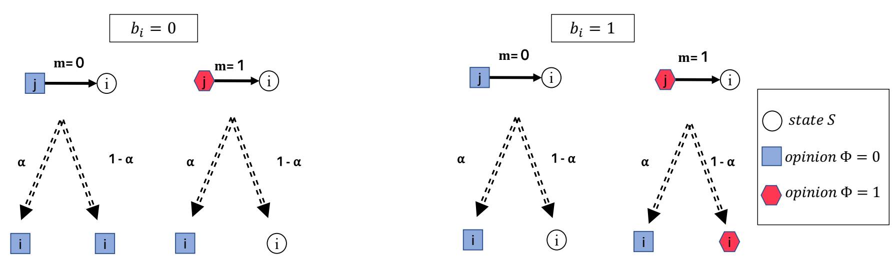

Figure 1: A summary of the potential opinions an agent i may hold, depending on her type, the message received by agent j, and verification.

Agents transition into state I when they receive a message which they do not ignore. Agents in state I die at rate δ, independently of their type and state, and are replaced by identical agents in state S.[3]

Diffusion Framework. A link between two agents i and j signifies a meeting between them. The set of meetings can be represented by a communication network. This network is realized independently in every period. Formally, we model the *mean-field approximation* of the system.

Each agent i has k meetings at t, also denoted the degree of the agent, which is constant over time. We denote by ν the per contact transmission

[3]In many scenarios, δ will conceivably be very small, given the speed at which information diffuses. Our model can accommodate arbitrarily small values of $\delta > 0$.

rate of m, which again is independent of an agent's type. It is affected, for example, by communication technology.

Information Prevalence. We define $\rho^{\alpha}_{b,m,t}$ ($\rho^{1-\alpha}_{b,m,t}$) as the proportion of type b agents at time t who believe message m after (not) having verified it, for $b \in \{0,1\}$.[4] Note that due to susceptibility to messages, it is the case that $\rho^{\alpha}_{0,1,t} = \rho^{1-\alpha}_{0,1,t} = \rho^{1-\alpha}_{1,0,t} = \rho^{\alpha}_{1,1,t} = 0$. A randomly chosen contact of an agent believes message $m \in \{0,1\}$, at time t with probability $\theta_{m,t}$ given by

$$\theta_{0,t}(\alpha) = x[\alpha\rho^{\alpha}_{0,0,t} + (1-\alpha)\rho^{1-\alpha}_{0,0,t}] + (1-x)\alpha\rho^{\alpha}_{1,0,t}, \quad (1)$$

$$\theta_{1,t}(\alpha) = (1-x)(1-\alpha)\rho^{1-\alpha}_{1,1,t}. \quad (2)$$

$\theta_{0,t}$ and $\theta_{1,t}$ are also the overall truth and rumor prevalence in the population at time t.

We assume that the per contact transmission rate, ν, is sufficiently small that an agent in state S becomes aware of message m at rate $k\nu\theta_{m,t}$ through meeting k neighbors, for $m \in \{0,1\}$. This framework allows us to model information diffusion as a set of differential equations:

$$\frac{\partial\rho^{\alpha}_{0,0,t}}{\partial t} = x\alpha(1-\rho^{\alpha}_{0,0,t})k\nu[\theta_{0,t}+\theta_{1,t}] - x\alpha\rho^{\alpha}_{0,0,t}\delta, \quad (3)$$

$$\frac{\partial\rho^{1-\alpha}_{0,0,t}}{\partial t} = x(1-\alpha)(1-\rho^{1-\alpha}_{0,0,t})k\nu\theta_{0,t} - x(1-\alpha)\rho^{1-\alpha}_{0,0,t}\delta, \quad (4)$$

$$\frac{\partial\rho^{\alpha}_{1,0,t}}{\partial t} = (1-x)\alpha(1-\rho^{\alpha}_{1,0,t})k\nu[\theta_{0,t}+\theta_{1,t}] - (1-x)\alpha\rho^{\alpha}_{1,0,t}\delta, \quad (5)$$

$$\frac{\partial\rho^{1-\alpha}_{1,1,t}}{\partial t} = (1-x)(1-\alpha)(1-\rho^{1-\alpha}_{1,1,t})k\nu\theta_{1,t} - (1-x)(1-\alpha)\rho^{1-\alpha}_{1,1,t}\delta. \quad (6)$$

These expressions keep track of how many agents enter and leave each group at each point in time. Take for example expression (3) for truth prevalence among verifying agents of type 0. The first term describes the mass of these

[4]With a slight abuse of notation we suppress the dependence of the various ρ's on the number of meetings, k, which is the same for all agents.

agents that newly believe message $m = 0$: they are the proportion of verifying type $b = 0$ agents ($x\alpha$) that did not yet believe message 0 before time t $(1 - \rho^{\alpha}_{0,0,t})$; in each period they meet k others, of whom $\theta_{0,t} + \theta_{1,t}$ are in state I, and communicate with them with probability ν. The second (negative) term, indicates that a proportion δ of the agents of this group die. The interpretation of the other expressions is similar.

Steady State. We are interested in the steady state of the system, where equations (3)-(6) are equal to zero. We remove the time subscript t to indicate the steady state value of variables. We define a *positive steady state* as a steady state in which at least one type of information exhibits a positive prevalence.

Social Planner. We are interested in the problem of a social planner, whose policy tool is the verification rate α of the population. We assume that the planner has budget A available to induce verification rate α, and that, for simplicity, the unit cost of inducing verification is one.

In the benchmark model, the planner's objective is to maximize the steady state prevalence of the truth, θ_0. In Section 5, we consider alternative objectives, that is, to minimize the prevalence of the rumor, θ_1, or to maximize the overall information prevalence, $\theta_0 + \theta_1$.

Discussion of the Main Assumptions. Before continuing, let us discuss the main assumptions of our model in more detail. A key assumption is that agents' biases limit their susceptibility to messages. This is a behavioural assumption that captures succinctly the observed tendency of agents to treat information that contradicts their opinion or view of the world differently from information that confirms it, such as filtering out of negative information (Taylor and Brown, 1988). In essence, we assume that agents exhibit information avoidance by ignoring messages that contradict their types, but believing information that confirms it (Golman, Hagmann and Loewenstein, 2017). However, we assume that verification reveals the true state of the world, and agents accept it, so that information avoidance is not extreme.

Related to this, agents transmit only information they themselves believe to be true, which is either the message they received, or the truth, in case they verified the message. An interpretation of this is that agents cannot transmit their opinion but only some evidence supporting it (possibly wrong). Hence, they cannot forge such evidence if they did not receive it from one of their social contacts, or they did not acquire it via verification.

Another interpretation is that, while only one massage is true, there could be several rumors about the true state of the world. In such a world, disbelieving a message does not automatically translate in knowing the truth, as suggested by the fact that we are assuming a binary message. Indeed, we may interpret the unique incorrect message in our model as a bundle of different rumors, one of which is transmitted at a meeting. The most straightforward approach to incorporate this would be to assume that agents biased against the truth are biased in favor of all rumors, as suggested by the fact that people who tend to believe in fake news can clearly be identified in society and in online social communities (Zollo et al., 2017; Samantray and Pin, 2019).

In the model, we introduce verification as a parameter that can be costly chosen by the social planner. This captures the fact that a social planner can to a certain extent affect the cost of verification that agents face when they decide how much time or effort to dedicate to verifying messages, as in Merlino, Pin and Tabasso (2023). Here, we consider a reduced form-version of this problem, where the planner directly chooses verification rates. Another interpretation is that the verification rate represents the proportion of the population that is information literate, and that the planner may influence this rate.[5] Channels may be, e.g., through education or (digital) campaigns and guides on how to spot misinformation. There's empirical evidence that

[5]This fits well the assumption that in our model verification rates are set before messages are received. Otherwise, the verification rate would differ depending on the message received, as in Merlino, Pin and Tabasso (2023).

increased education, or sophistication, of agents makes them less susceptible to rumors (Bello and Rocco, 2021; Pennycook and Rand, 2019). Indeed, we show in Section 5 that our reduced-form model is equivalent to one in which the planner affects verification rates by subsidising education.

As this paper, Merlino, Pin and Tabasso (2023) investigate the diffusion of two contradictory pieces of information. The main difference between the two papers lies in how agents assimilate and transmit information. In the present paper, agents cannot forge new messages if they do not believe one they have received. In other words, we focus more on the diffusion of specific messages, as opposed to opinions. In Merlino, Pin and Tabasso (2023) instead, agents may hold an opinion different from the message they receive even if they do not verify it. This well describes situations in which messages are binary in nature, e.g., whether vaccines are safe or not. Hence, the two models describe somewhat different situations, which matters for policy implications.[6]

3 Diffusion of Truth and Rumor

Defining the diffusion rate λ as $\lambda = \nu k/\delta$, the conditions for an information steady state are

$$\rho^{\alpha}_{0,0} = \rho^{\alpha}_{1,0} = \frac{\lambda[\theta_0 + \theta_1]}{1 + \lambda[\theta_0 + \theta_1]}, \tag{7}$$

$$\rho^{1-\alpha}_{0,0} = \frac{\lambda\theta_0}{1 + \lambda\theta_0}, \tag{8}$$

$$\rho^{1-\alpha}_{1,1} = \frac{\lambda\theta_1}{1 + \lambda\theta_1}. \tag{9}$$

[6]For example, contrarily to what we find here, in Merlino, Pin and Tabasso (2023) increases in verification rates are always beneficial to the prevalence of the truth.

Substituting equations (7) - (9) into equations (1) and (2) respectively, the steady states for θ_0 and θ_1 are fixed points of the following expressions:

$$H(\theta_0, \theta_1) = \alpha \frac{\lambda(\theta_0 + \theta_1)}{1 + \lambda(\theta_0 + \theta_1)} + x(1 - \alpha) \frac{\lambda \theta_0}{1 + \lambda \theta_0}, \tag{10}$$

$$G(\theta_1) = (1 - x)(1 - \alpha) \frac{\lambda \theta_1}{1 + \lambda \theta_1}. \tag{11}$$

A steady state of the system as a whole is a fixed point of $\theta_0 = H(\theta_0, \theta_1)$ conditional on $\theta_1 = G(\theta_1)$, where $H(\cdot)$ and $G(\cdot)$ are strictly increasing and concave functions in their arguments with $H(0, \theta_1) \geq 0$, $G(0) = 0$, and $H(1, \theta_1), G(1) < 1$. Thus, they each cross the 45-degree line at most once, and they do so from above. As θ_1 is determined independently by (11), with a slight abuse of notation, from now on we write $H(\theta_0)$ instead of $H(\theta_0, \theta_1)$.

Consequently, for each information, at most one positive steady state exists, and if it does, it is globally stable. Trivially, for any $\lambda, x, \alpha \geq 0$, there exists a steady state in which $\theta_0 = \theta_1 = 0$, which is globally stable if the positive steady state does not exist. In addition, if $\theta_1 > 0$, equation (11) allows us to explicitly derive

$$\theta_1 = (1 - \alpha)(1 - x) - \frac{1}{\lambda}. \tag{12}$$

Hence, $\theta_1 > 0$ if and only if $\alpha < 1 - 1/[\lambda(1 - x)]$. As intuition suggests, rumor prevalence is strictly decreasing in the verification rate α, and, if enough agents verify, the rumor dies out.

Regarding equation (10), the first part of it represents the influence of verifying agents—for them, receiving either message results in believing that the true state of the world is 0. The second is the additional impact on truth prevalence of those agents of type 0 who receive $m = 0$ and do not verify. If the rumor dies out, we can also explicitly derive a positive steady state of the truth,

$$\theta_0 = \alpha(1 - x) + x - \frac{1}{\lambda}. \tag{13}$$

This is positive if and only if $\alpha > (1/\lambda - x)/(1-x)$. Absent the rumor, the prevalence of the truth is strictly increasing in verification and a high enough proportion of verifiers is necessary for the truth to exhibit a positive steady state. Indeed, a higher share of verifiers implies that fewer agents biased towards the rumor ignore the truthful message when they receive it.

If both $\alpha > 0$ and $\theta_1 > 0$, then $H(0) > 0$. This means that, whenever the rumor circulates in steady state, so does the truth.

Hence, the only case in which no positive steady state for either information exists is if $1 - 1/[\lambda(1-x)] \leq \alpha \leq (1/\lambda - x)/(1-x)$, which is possible only if $\lambda \leq 2$. Otherwise, low verification rates, which benefit the rumor, lead to both rumor and truth exhibiting positive steady states, while high verification rates imply that the rumor dies out and only the truth has a positive steady state. Note that in the case of low verification rates, the truth has a positive prevalence if and only if the rumor also has a positive prevalence. In other words, the truth *only* survives because some agents heard the rumor and verified it, thus discovering the truth. In general, higher values of the diffusion rate, λ, benefit the diffusion of either type of information. Consequently, they increase the range of verification rates for which the rumor and/or the truth survive, which is why we observe a range of verification in which neither survives only for relatively low values of the diffusion rate.

The following proposition summarizes the results we have just derived.[7]

Proposition 1 *Suppose that there is some verification, i.e., $\alpha > 0$. Then,*

1. *if $\alpha \in \left(0, 1 - \frac{1}{\lambda(1-x)}\right)$, there exists a unique globally stable steady state, in which both the truth and the rumor have positive prevalence, given by* (12) *and* (10)*.*

2. *if $\lambda < 2$ and $\alpha \in \left[1 - \frac{1}{\lambda(1-x)}, \frac{\frac{1}{\lambda} - x}{1-x}\right]$, there exists a unique, and globally stable steady state and it is such that both rumor and truth die out.*

[7] As discussed above, while a steady state in which both rumor and truth die out always exists, it is unstable if a positive steady state exists. So, in this proposition we only focus on positive steady states whenever they exist.

3. *if* $\alpha \in \left(\max\{1-\frac{1}{\lambda(1-x)}, \frac{\frac{1}{\lambda}-x}{1-x}\}, 1\right]$*, there exists a unique positive steady state in which only the truth has positive prevalence, given by* (13)*. This steady state is globally stable.*

Overall, there is no steady state in which the truth dies out if the rumor survives. In addition, equation (10) highlights that, in the steady state described in point (1) of Proposition 1 the rumor in fact benefits the diffusion of the truth.

Corollary 1 *The prevalence of the truth is, ceteris paribus, increasing in the prevalence of the rumor.*

Hence, far from hurting the diffusion of the truth, the rumor creates truth. Consequently, it might be beneficial to let the rumor circulate to some degree. This observation motivates us to study the optimal level of verification next.

4 Optimal Verification

While public discussions often focus on misinformation alone, another plausible objective for a planner is to maximize the prevalence of the truth. Indeed, in many scenarios that agents face, such as how to act to minimize the chance of being infected with a disease, it is important to spread the correct guidelines on how to behave optimally. Being aware of the truth might entail taking an action that has positive externalities. For example, being aware that AIDS is a sexually transmitted disease makes it more likely to have protected sexual contacts rather than unprotected ones. Thus, in these contexts, a benevolent planner will have the objective to maximize the diffusion of the truth. This can be done by appropriately choosing the verification rate, α, for example, by implementing policies that change the costs agents face when verifying messages.

As the results of Section 3 suggest, the diffusion of the rumor plays a non-trivial role in the diffusion of the truth. The following Proposition establishes general results about the optimal use of a planner's budget A when setting the optimal verification rate α^* in order to maximize θ_0.[8]

Proposition 2 *Let the planner have budget A available to set information verification rates and let their objective be to maximize the prevalence of the truth in the population. Then, there exists a value of the diffusion rate $\bar{\lambda}$ such that*

i) For all values of the diffusion rate, λ, and the share of agents of group 0, x, there exist values $\underline{A}$ and $\bar{A}$ such that, for all $A < \underline{A}$ and for all $A > \bar{A}$, it is optimal for the planner to use all the budget available for debunking the rumor, i.e., $\alpha^* = A$.

ii) For $\lambda < \bar{\lambda}$, there exists a range of $\underline{A} \leq A \leq \bar{A}$ such that it is optimal for the planner not to use all the budget available for debunking, i.e., $\alpha^* \in (0, A)$.

Proposition 2 establishes that it may be optimal policy for a planner to *not* fully eradicate the rumor, even if that was possible. At the same time, whenever it is optimal to eradicate the rumor, it is also optimal to spend the entire budget to induce verification. The case where a planner may purposefully allow the rumor to propagate occurs only for relatively low diffusion rates. For optimal policy, this implies that there may be scenarios where a planner wishes to allow a rumor to propagate at low diffusion rates, but the optimal decision changes to full eradication when the diffusion rates increase.[9]

Figure 2 depicts the diffusion of the truth as a function of the verification rate in an example when $\lambda < \bar{\lambda}$. In particular, due to the ambiguous role of the rumor in the diffusion of the truth, the diffusion of the truth does not have

[8] By our earlier results, each value of the verification rate induces a unique stable steady state of the truth, on which we focus.

[9] Note that, as the degree of verification necessary to fully eradicate the rumor depends positively on the diffusion rate, this might not be viable when diffusion rates are high.

a monotonic relationship with verification rates. This has counter-intuitive consequences on the optimal verification rate. When the planner's budget is very low, all of it should be spent. Similarly, when the budget is sufficiently high, the rumor should be eradicated. However, for intermediate values of the budget, the rumor should not be eradicated. As the figure shows, in this example the optimal verification rate for $A \in [\underline{A}, \bar{A})$ is $\underline{A}$, which implies that not all the available verification budget is used.

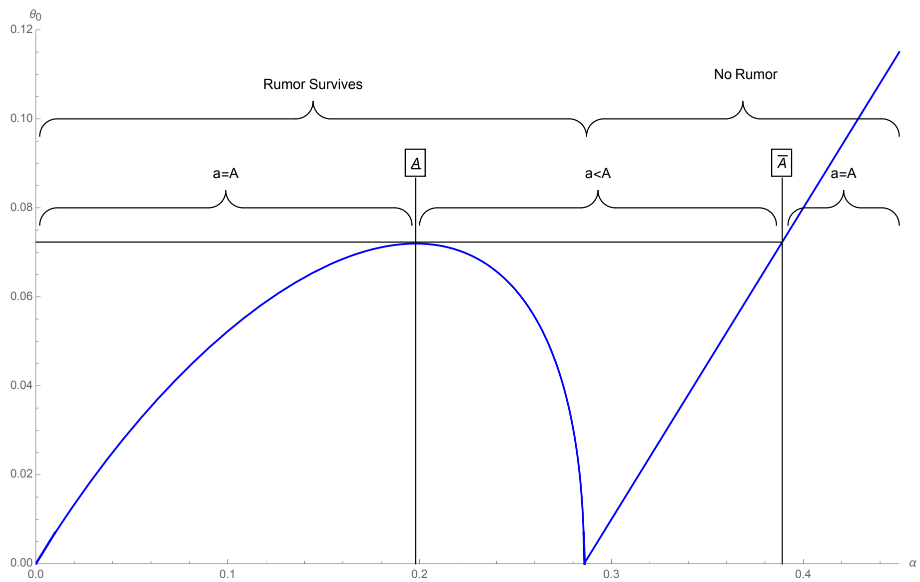

Figure 2: Steady state prevalence of the truth, θ_0, as a function of α, for $\lambda = 2$ and $x = 0.3$.

Overall, our results indicate how to optimally set information verification rates. We turn now to discuss how these results are affected by changes to a number of our assumption.

5 Discussion

5.1 Individual Verification Choices

In the preceding analysis, a central planner directly chooses the level of verification in the economy. However, we now show that this can be seen as a reduced form problem of one in which this is an individual choice, while the planner can affect the cost of verification.

In particular, we study a model in which agents decide verification before they participate in the diffusion process. The idea is that agents decide at the beginning of their lives whether to educate themselves to allow them to discern truthful from incorrect messages.[10] Education bears a cost of $\bar{c}$, and agents receive a flow utility of one if they are aware of the truth and zero otherwise. Agents are infinitely patient, so they only care about the expected proportion of their lives during which they are correctly informed. Each agent is infinitesimal, and hence unable to affect the total proportion of educated agents α in the population. Denoting by α_i the verification choice of agent i, i's expected lifetime utility is given by

$$U_i(\alpha) = \begin{cases} \frac{\lambda(\theta_0+\theta_1)}{1+\lambda(\theta_0+\theta_1)} - \bar{c} & \text{if } \alpha_i = 1, \\ x\frac{\lambda\theta_0}{1+\lambda\theta_0} & \text{if } \alpha_i = 0. \end{cases}$$

For ease of exposition, assume that $\bar{c}$ is so high that, absent an intervention by the planner, nobody would would get educated and verify. As a result, the planner's problem is to set a subsidy s such that

$$\frac{\lambda(\theta_0+\theta_1)}{1+\lambda(\theta_0+\theta_1)} - x\frac{\lambda\theta_0}{1+\lambda\theta_0} = \bar{c} - s. \tag{14}$$

The left-hand side of equation (14) is continuous in the verification rate

[10] This assumption simplifies the analysis. See Merlino, Pin and Tabasso (2023) for an alternative model in which agents decide whether to verify messages after they received them.

α, it is always positive, and smaller than $\bar{c}$ (by assumption). Therefore, the planner can induce any possible aggregate verification rate α by appropriately choosing the subsidy s. In particular, the planner sets the optimal subsidy by solving the following problem:

$$
\begin{aligned}
\max \quad & \theta_0 \\
s.t. \quad & \theta_0 = H(\theta_0) \ \text{ from equation (10)}, \\
& \frac{\lambda(\theta_0+\theta_1)}{1+\lambda(\theta_0+\theta_1)} - x\frac{\lambda\theta_0}{1+\lambda\theta_0} = \bar{c} - s, \\
& \alpha s \leq A, \\
& \alpha \in (0,1).
\end{aligned}
$$

This problem delivers the same solution as the one we have analyzed in the benchmark model. Note also that, if agents differ in their individual costs of educating themselves, i.e., we have $\bar{c}_i$ for agent i, the planner will minimize their cost of inducing a specific verification rate α by preferentially subsidizing agents with lower education costs. Hence, our analysis also extends to such a case.

5.2 Targeted Verification

In our model, we assume that message susceptibility of agents is restricted by their type, and that this restriction is overcome through verification of messages. From equations (12) and (13), we can see that if no agent was biased towards the rumor, it would die out and the truth would achieve its maximum prevalence. This raises the question whether truth prevalence could be increased by instigating verification of messages particularly among those agents whose type biases them towards believing the rumor. Online guides on how to spot misinformation, or information literacy campaigns, may be tailored and placed accordingly.

In this scenario, the steady state of the truth is the fixed point of

$$H(\theta_0) = x\left[\alpha_0 \frac{\lambda(\theta_0+\theta_1)}{1+\lambda(\theta_0+\theta_1)} + (1-\alpha_0)\frac{\lambda\theta_0}{1+\lambda\theta_0}\right] + (1-x)\alpha_1 \frac{\lambda(\theta_0+\theta_1)}{1+\lambda(\theta_0+\theta_1)}. \quad (15)$$

The prevalence of the rumor instead is given by $\theta_1 = (1-x)(1-\alpha_1) - 1/\lambda$ and depends exclusively on verification in the group biased towards the rumor.

Given a budget A and assuming that verification costs are the same in both groups, the planner's problem is the following:

$$\begin{aligned} \max \quad & \theta_0 & (16)\\ \text{s.t.} \quad & \theta_0 = H(\theta_0) & (17)\\ & x\alpha_0 + (1-x)\alpha_1 \leq A & (18)\\ & \alpha_0, \alpha_1 \in (0,1). & (19) \end{aligned}$$

In the following, we constrain ourselves to scenarios where $\lambda > 1/(1-x)$, as otherwise the rumor always dies out, independently of verification rates. As we want to focus on the question of budget allocation across groups, we restrict our attention to budgets $A \leq x$; this implies that for positive rumor prevalence, it is always optimal for the planner to use all their budget.[11]

Under these conditions, we can derive the optimal allocation of resources to induce targeted verification.

Proposition 3 *The planner's problem to maximise the prevalence of the truth subject to verification constraints as described in* (16)-(19) *has a unique solution. Furthermore,*

i) For all values of the diffusion rate, λ, there exist values A'' and A', with $A' < 1 - 1/(\lambda(1-x))$, such that for all $A < A'$ and for all $A > A''$, it is optimal to debunk rumors only in group 1*, i.e.,* $\alpha_0 = 0$.

[11] As this condition implies a budget sufficient to allow all agents of type 0 to verify messages, we do not perceive this as particularly stringent.

ii) For $A \in [A', A'']$, there exist combinations of the budget, A, and the diffusion rate, λ, such that it is optimal to debunk rumors also in group 0, i.e., $\alpha_0 > 0$.

As Proposition 3 highlights, the main result of our baseline model, namely that it may be optimal to allow a rumor to circulate, carries over also when the planner can target individual groups to induce verification. Here, this takes the form of diverting resources towards verification in the group biased towards the truth, despite them being insusceptible to the rumor.

5.3 Minimizing Rumor Prevalence

In general, a benevolent planner may have different objectives than maximizing the prevalence of the truth in setting verification rates. An obvious one is to minimize the diffusion of the rumor. In this case, the optimal policy is straightforward, as shown in the following corollary. Define α' as the optimal verification rate when the planner wishes to eradicate the rumor. Then,

Corollary 2 *Let a policy maker's aim be to minimize the diffusion of the rumor. They are able to fully eradicate the rumor if and only if $A \geq 1 - 1/(\lambda(1-x))$. Whenever A is lower than this value, it is optimal for the policy maker to set $\alpha' = A$.*

Notably, as communication increases (either through easier technology, or more meetings), a higher rate of message verification will be necessary to eliminate the rumor.

5.4 Maximizing Overall Information Prevalence

Online social platforms try to influence to which extent their users verify the messages they receive. There are, for example, various guides on how to spot mis- and disinformation that are published by online providers. Platforms, however, benefit from the total volume of messages that are exchanged on them, irrespective of their veracity, as higher volume translates into a higher revenue for the platform. In the context of our model, this leads to a different objective as the one we studied so far. More precisely, it is reasonable to assume that online social platform maximize the total information prevalence in the economy, i.e., $\theta = \theta_0 + \theta_1$.

Total prevalence in our model is given by the fixed point that solves

$$H(\theta) = \alpha \frac{\lambda\theta}{1+\lambda\theta} + x(1-\alpha)\frac{\lambda\theta_0}{1+\lambda\theta_0} + \theta_1 \tag{20}$$

Our next Proposition establishes how the optimal verification rate $\tilde{\alpha}^*$ of the platform relates to the optimal rate the planner would choose, α^*. To allow this comparison, we assume that both planner and platform would have the same budget A available and that they both face a cost of unity of inducing verification.

Proposition 4 *Let the platform have budget A available to set information verification rates and let their objective be to maximize overall information prevalence. Then, there exists a value of the budget $\tilde{A} > \bar{A}$ such that, for all $A \geq \tilde{A}$, the platform debunks as the planner, i.e., $\tilde{\alpha}^* = \alpha^* = A$. For all $A < \tilde{A}$ instead, the optimal verification rate of the platform is weakly lower than the one of the planner.*

The fact that a player who aims to maximize the prevalence of information overall will in general choose lower verification rates than one whose objective is only to maximize the prevalence of the truth is unsurprising. It is a direct consequence of the fact that rumor prevalence is linearly decreasing in verification, an effect which is added to the effect of verification on truth prevalence. More interesting is the fact that, for high enough budgets, both objectives are met at the same verification rate. This is due to the fact that, as verification rates increase, fewer messages are ignored. In fact, in the proof of Proposition 4 we show that overall information prevalence is maximized when all agents verify messages.

6 Conclusions

In this paper, we model how a true and a false message spread in a population of biased agents who become aware of the veracity of messages they receive if they verify them.

In this framework, we find that the presence of a false message creates truth, in the sense that a larger prevalence of the rumor leads to a larger prevalence of the truth. As a result, it is possible that increased verification rates lead to a lower prevalence of the truth. We employ this result to show that a central planner may optimally choose to allow a rumor to perpetuate in the network, even if they have sufficient resources to eradicate it. In addition, we show that the question of how to optimally allocate resources to induce verification across the differently biased groups depends on the exact value of the budget the planner has at their disposal, as well as the diffusion rate of messages. Our results challenge the intuition that making it easier to assess the veracity of information must necessarily be beneficial to society.

In our work, all agents benefit from being aware of the truth, and there are no incentives for agents to diffuse an information they themselves do not believe. The inclusion of such strategic considerations appears a promising avenue for future research.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2022. "A Model of Online Misinformation." NBER Working Paper no 28884.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social media and fake news in the 2016 election." *Journal of Economic Perspectives*, 31(2): 211–36.

Banerjee, Abhijit V. 1993. "The economics of rumours." *The Review of Economic Studies*, 60(2): 309–327.

Bello, Piera, and Lorenzo Rocco. 2021. “Education, information, and COVID-19 excess mortality.” IZA Discussion Paper.

Bizzarri, Matteo, Fabrizio Panebianco, and Paolo Pin. 2021. “Epidemic dynamics with homophily, vaccination choices, and pseudoscience attitudes.” *arXiv preprint arXiv:2007.08523*.

Bloch, Francis, Gabrielle Demange, and Rachel Kranton. 2018. “Rumors and social networks.” *International Economic Review*, 59(2): 421–48.

Bravard, Christophe, Jacques Durieu, Sudipta Sarangi, and Stéphan Sémirat. 2023. “False information from near and far.” *Games and Economic Behavior*, 137: 152–174.

Campbell, Arthur, C. Matthew Leister, and Yves Zenou. 2019. “Social Media and Polarization.” Mimeo.

Chen, Frederick, and Flavio Toxvaerd. 2014. “The economics of vaccination.” *Journal of Theoretical Biology*, 363: 105–117.

Galeotti, Andrea, and Brian W Rogers. 2013. “Strategic immunization and group structure.” *American Economic Journal: Microeconomics*, 5(2): 1–32.

Golman, Russell, David Hagmann, and George Loewenstein. 2017. “Information avoidance.” *Journal of Economic Literature*, 55(1): 96–135.

Goyal, Sanjeev, and Adrien Vigier. 2015. “Interaction, protection and epidemics.” *Journal of Public Economics*, 125: 64–69.

Kinateder, Markus, and Luca Paolo Merlino. 2023. “Free riding in networks.” *European Economic Review*, 152: 104378.

Kinateder, Markus, and Luca P. Merlino. 2017. “Public goods in endogenous networks.” *American Economic Journal: Microeconomics*, 9(3): 187–212.

Kranton, Rachel, and Davis McAdams. 2020. “Social networks and the market for news.” Mimeo.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. “The science of fake news.” *Science*, 359(6380): 1094–1096.

Merlino, Luca P., Paolo Pin, and Nicole Tabasso. 2023. “Debunking rumors in networks.” *American Economic Journal: Microeconomics*, 15(1): 467–96.

Papanastasiou, Yiangos. 2020. “Fake news propagation and detection: A sequential model.” *Management Science*, 66(5): 1826–1846.

Pennycook, Gordon, and David G Rand. 2019. “Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning.” *Cognition*, 188: 39–50.

Samantray, Abhishek, and Paolo Pin. 2019. “Credibility of climate change denial in social media.” *Palgrave Communications*, 5(1): 1–8.

Tabasso, Nicole. 2019. “Diffusion of Multiple Information: On Information Resilience and the Power of Segregation.” *Games and Economic Behavior*, 118: 219–40.

Talamàs, Eduard, and Rakesh Vohra. 2020. “Free and perfectly safe but only partially effective vaccines can harm everyone.” *Games and Economic Behavior*, 122: 277 – 289.

Taylor, Shelley E, and Jonathon D Brown. 1988. “Illusion and well-being: a social psychological perspective on mental health.” *Psychological bulletin*, 103(2): 193.

Toxvaerd, Flavio. 2019. "Rational disinhibition and externalities in prevention." *International Economic Review*, 60(4): 1737–1755.

Zollo, Fabiana, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. 2017. "Debunking in a world of tribes." *PloS one*, 12(7).

A Proofs

Proof of Proposition 2. First, note that by equation (13) the prevalence of the truth is equal to $\theta_0 = 1 - 1/\lambda$ if $\alpha = 1$. This is the highest value that θ_0 can take. By continuity of θ_0 in α, there always exists a value $\bar{A}$, such that it is optimal to set $\alpha = A$ if $A > \bar{A}$.
Next, assume that the planner's budget is not sufficiently large to fully eradicate the rumor. Hence, the steady state truth prevalence is given by equation (10). By the implicit function theorem, the effect of α on θ_0 is given by

$$\frac{d\theta_0}{d\alpha} = -\frac{-\frac{\partial H}{\partial \alpha}}{1 - \frac{\partial H}{\partial \theta_0}}.$$

As $H(\theta_0)$ is strictly concave in θ_0, we know that at the steady state, $\partial H(\theta_0)/\partial\theta_0 < 1$. Hence, $d\theta_0/d\alpha > 0$ if and only if $\partial H(\theta_0)/\partial\alpha > 0$, where

$$\frac{\partial H(\theta_0)}{\partial \alpha} = \frac{\lambda(\theta_0+\theta_1)}{1+\lambda(\theta_0+\theta_1)} - x\frac{\lambda\theta_0}{1+\lambda\theta_0} - \alpha(1-x)\frac{\lambda}{[1+\lambda(\theta_0+\theta_1)]^2}. \quad \text{(A-1)}$$

As the combination of the first two terms is always positive whenever some information survives, it is obvious from equation (A-1) that at $\alpha = 0$ it is beneficial for the truth to increase verification rates. As θ_1 is strictly decreasing in α, for given θ_0, $\partial H(\theta_0)/\partial\alpha$ is strictly decreasing in α. Thus, by continuity, setting $\alpha = A$ is optimal for low values of A, i.e., for all $A < \underline{A}$.

Finally, we show when $\partial H(\theta_0)/\partial\alpha$ is negative if $A \in [\underline{A}, \bar{A}]$. As $\partial H(\theta_0)/\partial\alpha$ is decreasing in α, we look at its value for the highest possible value of α such that the rumor still survives. In fact, the rumor dies out if $\alpha = 1 - 1/[\lambda(1-x)]$. The limit of $\partial H(\theta_0)/\partial\alpha$ as α approaches this value is

$$\frac{\partial H(\theta_0)}{\partial \alpha} = (1-x)\frac{\lambda\theta_0}{1+\lambda\theta_0} - \left[1 - x - \frac{1}{\lambda}\right]\frac{\lambda}{[1+\lambda\theta_0]^2},$$

which is negative if

$$\theta_0\,[1+\lambda\theta_0] < 1 - \frac{1}{\lambda(1-x)}, \tag{A-2}$$

i.e., for low values of θ_0, and positive for high ones. As $\alpha \to 1 - 1/[\lambda(1-x)]$, we find that $\theta_0 \to 1 - 2/\lambda$ and therefore condition (A-2) is satisfied whenever $\lambda < 2 + \sqrt{2 - 1/(1-x)}$. Note furthermore that $\partial H(\theta_0)/\partial\alpha$ is strictly increasing in λ, i.e., it is more likely that (A-2) is satisfied for lower values of λ. Continuity of $\partial H(\theta_0)/\partial\alpha$ in both λ and α then yields the result. ■

Proof of Proposition 3. First, given that it is optimal for the planner to use all their budget whenever $A \leq x$, their problem (16) can be rewritten as

$$\begin{aligned} \max_{\alpha_0 \in [0,1]} \quad & \theta_0 \\ \text{s.t.} \quad & \theta_0^3\lambda^2 + \theta_0^2 B + \theta_0 C + D = 0. \end{aligned} \tag{A-3}$$

where $B = \lambda(1+\lambda-2A\lambda-2\lambda x+2\alpha_0\lambda x)$, $C = \lambda(1-A-x(1-\alpha_0))(1-A\lambda-\lambda x+\alpha_0\lambda x)$ and $D = A(1-\lambda+\lambda A+\lambda x-\alpha_0\lambda x)$. Dividing by λ^2 and after some algebra, the constraint can be rewritten as $f(\alpha_0, \theta_0) = \theta_0^3\lambda^2+\theta_0^2 b+\theta_0 c+d = 0$ with $b = -(2\alpha_1(1-x) + 2x - 1 - 1/\lambda)/\lambda^2$, $c = -(1-\alpha_1)(1-x)(\alpha_1(1-x) + x - 1/\lambda)/\lambda^2$ and $d = -\theta_1/\lambda$. Note that $f(\alpha_0, \theta_0)$ is continuous in α_0. Furthermore, given that b can be either positive or negative and that $c, d < 0$, by Descartes' rule of sign, $f(\alpha_0, \theta_0) = 0$ admits at most one positive real solution. If the solution is negative, $\theta_0^* = 0$. If it is bigger than one,

$\theta_0^* = 1$. This proves existence and uniqueness.

Next, we consider the question whether it is always optimal to prioritise verification in group 1 above the one in group 0. First, note that whenever the rumor dies out, the prevalence of the truth becomes $\theta_0 = x + (1-x)\alpha_1 - 1/\lambda$, strictly increasing in α_1 and independent of α_0. As in the case of a unique verification rate of the whole population, θ_0 is maximised at $\alpha_1 = 1$, thus, there always exists a budget A'' such that for $A > A''$ it is optimal to invest it entirely in the verification of group 1 and to set $\alpha_0 = 0$. Next, the implicit function theorem allows us to study the effect of increases in α_1 by determining the sign of

$$\frac{\partial H(\theta_0)}{\partial \alpha_1} = (1-x)\left[\frac{\lambda\theta_0}{1+\lambda\theta_0} - A\frac{\lambda}{[1+\lambda(\theta_0+\theta_1)]^2}\right]. \tag{A-4}$$

It is straightforward to show that for given θ_0, equation (A-4) is strictly increasing in λ and decreasing in A and α_1. Furthermore, it is negative at $\theta_0 = 0$, positive at $\theta_0 = 1$, and strictly increasing in θ_0. At $A = 0$, the effect of increasing α_1 on the prevalence is positive, strictly so whenever $\theta_0 > 0$. By continuity of equation (A-4) in A, we can then always find a value A' such that it is optimal for the planner to set $\alpha_0 = 0$ if $A < A'$, with one caveat: If $\lambda \leq 1/(x + (1-x)\alpha_1)$, i.e., the truth only survives if the rumor does, any increase in α_1 must be such that the rumor continues to survive. In fact, as $A = 1 - x - 1/\lambda$ is necessary to eradicate the rumor, setting $\alpha_0 = 0$ is always optimal whenever $A < A' \leq 1 - x - 1/\lambda$.

Finally, consider the limit of equation (A-4) as $\alpha_1 \to 1 - 1/(\lambda(1-x))$. In this case $\theta_1 \to 0$ and $\theta_0 \to 1 - 2/\lambda$. At these values, equation (A-4) shows that truth prevalence increases as α_1 is reduced if

$$-3 + \lambda + \frac{2}{\lambda} < A,$$

which means, whenever

$$\lambda \in \left(\frac{3 + A - [(3 + A)^2 - 8]^{1/2}}{2}, \frac{3 + A + [(3 + A)^2 - 8]^{1/2}}{2} \right). \tag{A-5}$$

Due to continuity of equation (A-4) in α_1, the result follows. This concludes the proof of Proposition 3. ■

Proof of Proposition 4. From the facts that $\theta = \theta_0 + \theta_1$ and that θ_1 is decreasing in α linearly, it is immediate that maximization of θ requires a lower α than maximization of θ_0 whenever $\theta_1 > 0$. At the same time, note that the fixed point of equation (20) can be written as

$$\theta = \alpha \frac{\lambda \theta}{1 + \lambda \theta} + (1 - \alpha) \left[x \frac{\lambda \theta_0}{1 + \lambda \theta_0} + (1 - x) \frac{\lambda \theta_1}{1 + \lambda \theta_1} \right], \tag{A-6}$$

and also note that, if $\alpha = 1$, total information prevalence would be identical to the one in the standard SIS model, i.e., $\theta = 1 - 1/\lambda$. Equation (A-6) shows that, for all other values of α, θ will be lower than this. By continuity of θ, there must exist a budget $\tilde{A}$ such that above it, it is optimal for the platform to set verification rates equal to the budget. By the fact that $\theta = \theta_0$ when the rumor dies out, this choice is identical for the platform and the planner. Finally, as $\theta > \theta_0$ whenever $\theta_1 > 0$, it must be the case that $\tilde{A} > \bar{A}$. This concludes the proof of Proposition 4. ■