Debunking Rumors in Networks

By Luca P. Merlino, Paolo Pin and Nicole Tabasso*

We study the diffusion of a true and a false message (the rumor) in a social network. Upon hearing a message, individuals may believe it, disbelieve it, or debunk it through costly verification. Whenever the truth survives in steady state, so does the rumor. Communication intensity in itself is irrelevant for relative rumor prevalence, and the effect of homophily depends on the exact verification process and equilibrium verification rates. Our model highlights that successful policies in the fight against rumors increase individuals incentives to verify.
JEL: D83, D85
Keywords: Social Networks, Rumors, Verification

Information often diffuses via communication with family, friends or acquaintances. However, people transmit not only correct information, but also rumors, that is, false or imprecise information. The virality of these rumors shapes public debates, often involving significant personal and social costs.[1]

The increased reliance on online social media for news consumption and communication plays an important role in the diffusion of rumors. This has been documented in different contexts, such as fake news during the 2016 US Presidential election campaign (Allcott and Gentzkow, 2017), the dangers of childhood vaccinations (Cramer, 2018) and the origins of COVID-19 (Mian and Khan, 2020). A main concern is the degree of homophily in online social media, inducing "echo chambers" in which people are over-proportionally exposed to one particular opinion.[2]

* Merlino: ECARES, Université libre de Bruxelles, and University of Antwerp, Belgium, and , email: LucaPaolo.Merlino@uantwerpen.be. Pin: Department of Economics and Statistics, Università di Siena, Italy and BIDSA, Università Bocconi, Milan, Italy, email: paolo.pin@unisi.it. Tabasso: School of Economics, University of Surrey, UK, email: n.tabasso@surrey.ac.uk. We thank Leonie Baumann, Francis Bloch, Ugo Bolletta and Tomàs Rodgriguez-Barraquer, and (seminars) participants at CTN 2016, SAET 2016, BiNoMa 2017, LAGV 2017, the 4^{th} Conference on Network Science and Economics, the 2019 VERA workshop, Autonòma Barcelona, IMT Lucca, Navarra, Université libre de Bruxelles and Virginia Tech for valuable comments. Merlino gratefully acknowledges funding from the CNRS and the Research Foundation Flanders through grant G029621N. Pin gratefully acknowledges funding from the Italian Ministry of Education Progetti di Rilevante Interesse Nazionale (PRIN) grant 2017ELHNNJ and from Regione Toscana grant Spin.Ge.Vac.S. The bulk of the research was carried out while Tabasso was at the Università Ca' Foscari, Venice, Italy. Tabasso gratefully acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 793769. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[1] For example, believing conspiracy theories that deny the link between HIV and AIDS is associated with a less consistent use of condoms in the US (Bogart and Thorburn, 2005).

[2] While online patterns of news consumption are no more segregated than offline ones (Gentzkow and Shapiro, 2011; Halberstam and Knight, 2016), online social networks appear to be extremely homophilous (see, e.g., (Zollo et al., 2017)) and lead to more segregated communication patterns (Halberstam and

This concern has prompted demands on policy makers, news providers and online social networks to counter the diffusion of rumors (Frenkel and Isaac, 2018). For example, misinformation relating to the COVID-19 pandemic has caused responses from various platforms, from showing warnings on messages coming from debatable sources to providing links to verified and authoritative ones (Marr, 2020).

This paper proposes a tractable model to understand how the diffusion of rumors is affected by endogenous debunking and changes in homophily, as well as to derive policy implications. Crucially, our model allows for the persistence of two distinct messages about the state of the world (the truth and the rumor) when messages can be verified.

In our model, individuals receive a message from their social contacts. Society is partitioned into two groups, where individuals of each group have a prior $y > 0.5$ that the true state of the world is either 0 or 1, respectively. We say that individuals are biased towards a certain state.[3] An individual can exert verification effort, which reveals the true state with some probability. This effort is determined by their posterior belief about the state of the world upon receiving a message, calculated using Bayes rule.

Upon successful verification, individuals become aware of the true state of the world and accept it. If verification is unsuccessful, individuals' opinion matches the state towards which they are biased. Individuals communicate their opinion to their neighbors in a network, but they cannot credibly reveal their prior, nor whether they verified. The network is characterized by a degree of homophily, which is the probability with which a neighbor has the same prior as oneself.

In this model, information prevalence converges to a steady state in which the rumor always survives alongside the truth: as verification is costly, it is not perfect, so the rumor propagates. The truth to rumor ratio depends on the level of verification and the degree of homophily in the network. Indeed, an increase in communication rates increases both rumor and truth in equal proportion, such that the ratio remains constant. The degree of homophily instead creates tradeoffs. As it increases, individuals receive relatively more messages that confirm their prior. As these are verified less than opposing messages, homophily creates echo chambers and benefits the rumor. However, higher homophily also makes messages reinforcing one's prior less informative, and the opposite for messages going against one's prior. This incentivizes verification, which reduces the diffusion of the rumor. Which effect dominates depends on how verification efforts translate into verification success. We derive general conditions on the verification function that determine the impact of changes in homophily. We also discuss an example that highlights the role of the verification technology in shaping how homophily affects the quality of information.

Knight, 2016).

[3]It has been documented that people who tend to believe in fake news can clearly be identified in society and in online social communities (e.g., (Zollo et al., 2017); (Samantray and Pin, 2019)). A micro-foundation for this segregation is in Bolletta and Pin (2021).

Our model highlights that any policy which amounts to a one-time injection of truthful information, such as a time-limited information campaign, is ineffective in reducing the ratio of rumor to truth in the long run. Policies should rather incentivize individuals to verify information, thereby increasing the truth to rumor ratio. For example, strategies to combat fake news that focus on providing links to sources of verification are likely more effective than simply flagging posts as being disputed.

Our results are robust to the introduction of partisans who always hold the same opinion. Indeed, these individuals provide higher incentives to verify to the rest of the population, so as to completely offset their presence. Ignoring endogenous debunking would lead to very different conclusions.

Our paper contributes to the literature using game theory to study information diffusion in networks. This literature has two main strands. In the first, agents are Bayesian, e.g., Hagenbach and Koessler (2010) and Galeotti, Ghiglino and Squintani (2013). More related to our paper, Bloch, Demange and Kranton (2018) find that, when there are partisans who diffuse false information, other agents block messages coming from parts of the network with many partisans. Kranton and McAdams (2020) study how strategic diffusion affect media quality. In our paper, there is no strategic motive to transmit messages. However, debunking introduces a different kind of interaction: the rumor's prevalence affects the incentives of individuals to verify.

The other strand of literature considers non-Bayesian agents who learn from their neighbors. Starting from the seminal contribution of DeGroot (1974), in these models agents observe the beliefs of their neighbors and use them to update theirs using some specific rule (Banerjee and Fudenberg, 2004; Golub and Jackson, 2010).

Recently, some papers have studied the behavior of Bayesian agents when the underlying information structure is misspecified, e.g., Molavi, Tahbaz-Salehi and Jadbabaie (2018) and Banerjee and Compte (2020). In a similar spirit, in our model agents act Bayesian when they first hear a message, but they disregard additional information once they have formed an opinion. Possible interpretations are that an opinion translates into a once-for-all decision or individuals are unwilling to change their opinion. Such behavior is consistent with information avoidance, inattention, or a biased interpretation of additional information (Golman, Hagmann and Loewenstein, 2017).

In our model, individuals do not observe the diffusion of messages, but they derive them by the properties of the diffusion process assuming that their prior is correct. Since the prior might be wrong, individuals might hold wrong beliefs in the long run, as in, e.g., Compte and Postlewaite (2004).

We employ a SIS framework, a class of models introduced to study the diffusion of viruses.[4] Following the seminal work of Banerjee (1993) and Kremer (1996),

[4] SIS stands for Susceptible-Infected-Susceptible as infected individuals return to the susceptible class on recovery as the disease confers no immunity against reinfection.

some papers have studied how strategic decisions on protection affect the diffusion of a disease (Chen and Toxvaerd, 2014; Goyal and Vigier, 2015; Toxvaerd, 2019; Bizzarri, Panebianco and Pin, 2021). In particular, Galeotti and Rogers (2013) study the effect of homophily on strategic immunization. In these papers, there is a unique infectious state, whose magnitude is affected by immunization. We instead focus on the relative magnitude of truth and rumor within the overall prevalence of information, which implies different strategic considerations. In particular, while in the above papers protection is a local public good (Kinateder and Merlino, 2017), this is not true for discordant messages in our framework.

This feature also distinguishes our paper from the recent contributions on the role of costly search on social learning (Ali, 2018; Mueller-Frank and Pai, 2016). As in our paper, learning is not complete (i.e., beliefs do not converge to the truth) precisely because search (here, verification) is costly.

We study the diffusion of two messages as in Campbell, Leister and Zenou (2019) and Tabasso (2019). Contrarily to these paper, here we consider contradictory pieces of information that may be disbelieved. More broadly, our paper relates to a recent literature on opinion dynamics on random graphs (Akbarpour, Malladi and Saberi, 2018; Sadler, 2020). In those papers, agents either adopt or not without the option of external verification, and non-adoption does not create any externality. In our model, one's decision depends instead on the level of (non-)verification in the economy.

We believe we are the first to study the strategic decision of individuals to verify what they hear when there are several messages diffusing in a network.

The paper proceeds as follows. Section I introduces and discusses the model. Section II presents the main analysis. Section III discusses the policy implications. Section IV presents the model with partisans. Section V concludes. All proofs and computations for the examples are in the Appendix.

I. The Model

In this section, we formally present the model. The timeline is as follows. An individual i who hears a message at time t chooses how much effort to exert in verifying it. They then form an opinion of the true state of the world. While alive, they communicate a message in line with their opinion to their social contacts at a fixed rate.

In the following, we first describe the SIS diffusion process; then we derive the differential equations that govern the evolution of truth and rumor, given verification rates. Finally, we study how individuals chose these rates. We end the section with a discussion of the main assumptions of our model.

Diffusion Process. There is an infinite population of mass 1 of individuals, indexed by i, represented as nodes on a network. Time is continuous, indexed by t. There exist two verifiable messages $m \in \{0, 1\}$ that individuals diffuse via word of mouth. These messages pertain to the state of the world, $\Phi \in \{0, 1\}$.

Without loss of generality, we denote $\Phi = 0$ as the true state of the world, *ex-ante* unknown to the individuals. We refer accordingly to $m = 1$ as the "rumor". When individual i communicates message m to individual j, this reveals to j the set of values that Φ may take.

Society is partitioned into two groups of equal size, denoted by $b = \{0, 1\}$, where individuals of each group have a prior $y > .5$ that the true state of the world is either 0 or 1, respectively.

Each individual has k meetings at each time t. A proportion $\beta \in [0, 1]$ of these is with individuals of the same group, while the remaining interactions are with individuals of the other group. The group one belongs to is not observable, but β is common knowledge. A meeting between two individuals is described by a link. The associated network is realized every period.

The diffusion process of information is a SIS model. Individuals may be in one of two states: either they are unaware of the debate about the state of the world, in which case they are in state S (*Susceptible*), or they may hold an opinion about its value, in which case they are in state I (*Informed*).

An individual in state S transitions into state I by hearing message m during a meeting, in which communication occurs at rate ν. We assume that ν is sufficiently small that the chance of receiving multiple messages at the k simultaneous meetings is zero, so that information transmits at rate $k\nu$. With the complementary probability, an individual in S stays in S. Individuals in state I die at rate δ and are replaced with individuals of the same type in state S. Figure 1 depicts the transmission dynamics for an individual i, who becomes informed after receiving a message from j.

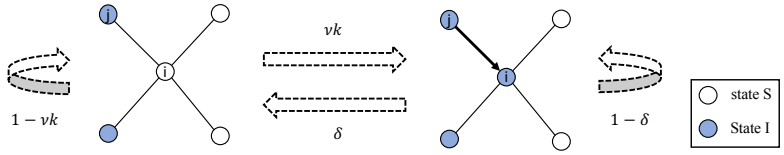

Figure 1. : The transition dynamics of player i for $k = 4$.

Information Prevalence. Individuals who hear m choose how much effort to exert in verification. In Section II.B, we show that this effort depends on the type of message an individual receives. We define l_t as the rate of verification of messages in line with one's bias (i.e., $m = b$), and h_t the verification rate of messages which go against it (i.e., $m \neq b$). Successful verification implies that an individual i knows that $\Phi = 0$ for sure; hence, i will hold this opinion. With the complement probability, the result of the verification process is inconclusive. In that case, an individual of type b holds an opinion in line with their bias (i.e., $\Phi = b$). We derive in Proposition 3 when this is the optimal behavior. Figure 2

depicts how individuals of type 0 and 1 form their opinion.

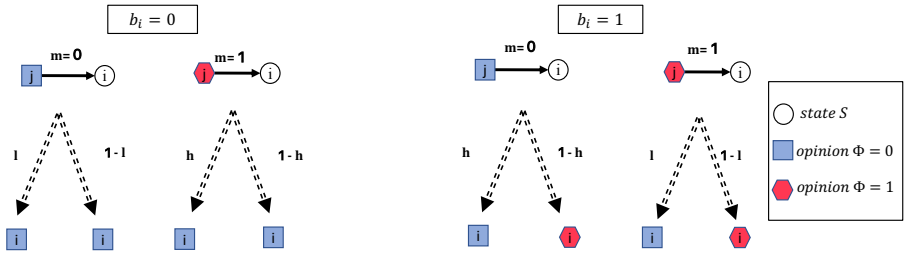

Figure 2. : A summary of the potential opinions i may hold, depending on her type, the message received, and verification success.

Denote by $\rho^b_{m,t}(\ell_t, h_t)$ the proportion of type b individuals that hold opinion in line with m at time t.[5] Note that type 0 individuals may only hold opinion 0, irrespective of verification, i.e., $\rho^0_{1,t}(\ell_t, h_t) = 0$. With some abuse of notation, we usually suppress the dependence on (ℓ_t, h_t).

The laws of motion governing the transmission of the system then are:

$$\frac{\partial \rho^0_{0,t}}{\partial t} = \frac{1}{2}(1-\rho^0_{0,t})\nu k\left[\beta\rho^0_{0,t} + (1-\beta)(\rho^1_{0,t}+\rho^1_{1,t})\right] - \frac{1}{2}\rho^0_{0,t}\delta, \tag{1}$$

$$\frac{\partial \rho^1_{0,t}}{\partial t} = \frac{1}{2}(1-\rho^1_{0,t}-\rho^1_{1,t})\nu k\left[\beta\left(h_t\rho^1_{0,t}+\ell_t\rho^1_{1,t}\right) + (1-\beta)h_t\rho^0_{0,t}\right] - \frac{1}{2}\rho^1_{0,t}\delta, \tag{2}$$

$$\frac{\partial \rho^1_{1,t}}{\partial t} = \frac{1}{2}(1-\rho^1_{0,t}-\rho^1_{1,t})\nu k\left[\beta\left((1-h_t)\rho^1_{0,t}+ +(1-\ell_t)\rho^1_{1,t}\right) + (1-\beta)(1-h_t)\rho^0_{0,t}\right] - \frac{1}{2}\rho^1_{1,t}\delta, \tag{3}$$

These expressions describe the evolution of opinions in line with $m = 0$ and $m = 1$ within the two groups $b \in \{0, 1\}$ in the *mean-field approximation* of the system, whereby information prevalence in an individual's neighborhood is the same as the prevalence in the overall population. For example, (1) describes how $m = 0$ evolves in group $b = 0$. The first term represents the mass of individuals who start holding opinion $\Phi = 0$ at time t. Indeed, the proportion of susceptible individuals of type 0, $(1 - \rho^0_{0,t})$, receive a message at rate νk. The message can come from someone of the $\rho^0_{0,t}$ individuals of their group, who holds opinion 0 and whom they meet with probability β. With probability $1 - \beta$, they meet someone of group 1, of whom $\rho^1_{0,t}$ transmit message 0 and $\rho^1_{1,t}$ message 1. The second (negative) term, indicates that a proportion δ of the informed individuals of type 0, $\rho^0_{0,t}$, die and

[5] Since $k = k_i\ \forall i \in N$, the proportion of individuals with degree k who hold opinion in line with m is identical to the overall proportion of individuals who hold that opinion.

are replaced by individuals in state S. In sum, players of group 0 always hold an opinion 0 independently of verification (see the left panel of Figure 2).

This is not the case for individuals of group 1. For example, consider how $m = 0$ evolves in group $b = 1$—equation (2). When the proportion of susceptible individuals of type 1, $(1 - \rho^1_{0,t} - \rho^1_{1,t})$, receive a message, they need to successfully verify the messages they receive to hold opinion 0. This happens with probability ℓ_t for messages in line of their bias, which they receive from $\rho^1_{1,t}$ individuals of their group, who they meet with probability β. They verify with probability h_t messages such that $m = 0$, which they can receive from $\rho^1_{0,t}$ of their group, who they meet with probability β, or from people in the other group, $\rho^0_{0,t}$, who they meet with probability $1 - \beta$. Again, informed individuals die and are replaced by individuals in state S at rate δ. The interpretation of (3) is equivalent. These transitions are depicted in the right panel of Figure 2.

Our main objects of interest are the overall prevalence of opinions 0 and 1 in the population at time t, $\rho_{m,t}$, which are

$$\rho_{0,t} = \frac{1}{2}\left(\rho^0_{0,t} + \rho^1_{0,t}\right), \tag{4}$$

$$\rho_{1,t} = \frac{1}{2}\rho^1_{1,t}. \tag{5}$$

Utility and Verification. Individuals in state I expect a (present value) lifetime utility of 1 if their opinion coincides with the true state of the world and 0 otherwise. We set the utility of being in state S to 0.

As mentioned above, when individuals first hear a message, say m, they choose how much effort to exert in verification depending on their belief the message they received is correct. We explain how beliefs are revised below. Exerting effort $\alpha \in [0, \infty)$ implies that verification is successful with probability $x(\alpha)$, in which case the individual knows that $\Phi = 0$ for sure. With the complement probability, the result of the verification process is inconclusive.

An individual of type b has a prior that $b = \Phi$ of $y > .5$. In subsection II.B, we derive the threshold of y above which it is optimal for an individual to hold the belief that $b = \Phi$ if her verification of a message is unsuccessful.

Each agent, being infinitesimal, takes as given the verification levels and information prevalence in the population. Hence, the utility of individual i who hears message m at time t is

$$U_{it} = x(\alpha_{it}) + (1 - x(\alpha_{it}))Pr_t(b = \Phi|m) - c\alpha_{it}, \tag{6}$$

where $x(\alpha_{it})$ is the probability verification is successful given a verification effort α_{it}, $Pr_t(b = \Phi|m)$ is i's expectation of being correct conditional on the message heard and c is the marginal costs of verification. As individuals observe only one message, this expectation is formed updating one's prior using the expected

prevalence of messages from the diffusion process described by equations (1)-(3).

We denote by $\Delta(x(\alpha))$ the subderivative of $x(\alpha)$, a correspondence that maps for any $\alpha > 0$ the values between the right and left derivatives of x, and by g the inverse of the subderivative of x. With some abuse of notation, we write $\Delta(x)$ instead of $\Delta(x(\alpha))$ when no confusion may arise.

ASSUMPTION 1: *We assume* $x(\cdot) : \mathcal{R}^+ \to [0, \bar{x}] \subseteq [0, 1]$ *is strictly increasing and strictly concave on* $[0, \bar{x})$*, continuous and such that* $x(0) = 0$*.*

We denote by $\bar{x}$ and $\bar{d}$ the values such that $\bar{x} \equiv \lim_{\alpha \to \infty} x(\alpha)$ and $\Delta(0) = [\bar{d}, \infty)$. Finally, we denote by α_{it}^b the equilibrium effort individual i of group b exerts in verifying a message in line with their type, i.e., $m = b$, and by α_{it}^{-b} the equilibrium effort to verify $m \neq b$. These efforts lead to the verification rates $\ell_{it} = x(\alpha_{it}^b)$ and $h_{it} = x(\alpha_{it}^{-b})$ we employed in equations (1)-(3).

Steady State and Equilibrium. The model is in steady state if equations (1), (2) and (3) are equal to zero. A steady state of the continuous dynamic system defined by these equations is *locally stable* if it satisfies *Lyapunov stability.*[6] A steady state is *positive* if the associated proportion of informed individuals, ι, is strictly positive. We remove the time subscript t to indicate the steady state value of variables.

The profile of verification efforts (α^b, α^{-b}) is an *equilibrium* if it maximizes (6) for all individuals taking as given the steady state diffusion rates of messages.

Discussion and interpretation. Before presenting the analysis of the model, we discuss its main assumptions. First, we assume that individuals' prior y about the state of the world is sufficiently high that absent verification, they believe their bias to be correct. We focus on this case because, if individuals have lower priors, they believe whichever message they first receive. In this case, rumors either die out, or verification is completely absent. We present a formal analysis of the model with lower priors y in Appendix B.

A key assumption of our model is that individuals are "partially Bayesian": once they first receive a message, they calculate their posterior belief that $b = \Phi$ using Bayes' rule; however, after they have formed their opinion, they do not further update the probability this is correct. One interpretation of this behavior is that after forming an opinion, individuals make a once-for-all decision. While these decisions are not always irreversible, the cost of making a wrong choice are either very substantial and/or realized only after a long delay (such as using a condom, or vaccination decisions). Another interpretation is behavioral: insofar that holding an opinion shapes one's identity, the perceived or psychological cost

[6] Formally, a steady state is locally stable if, for each neighborhood S of the steady state prevalence of messages, ρ_0^0, ρ_0^1 and ρ_1^1, there exists a neighborhood W such that each trajectory starting in W remains in S, for all $t \geq 0$ and the corresponding trajectory converges to the steady state as $t \to \infty$.

involved in changing identity may too high with respect to the benefits at stake.[7] In other words, in order to reduce the cognitive dissonance between one's belief and new information acquired in subsequent communication, individuals interpret the latter as supportive of their own opinion. This interpretation is supported by evidence of confirmation bias in online social platforms (Zollo et al., 2017).

Contrarily to the works in the social learning literature using variations of the DeGroot (1974) model, individuals do not exchange opinions with all their neighbors at every time period. Rather, ν is such that they never receive more than one message per period, and they decide then whether to verify and what to believe. Furthermore, as in Banerjee (1993), the message space is coarse, in the sense that a message contains only someone's action or opinion, i.e., $m = 0$ or $m = 1$, and not the probability they attach to their opinion being correct. This captures the idea that only actions are observable (and not beliefs) or that people transmit only imprecise information regarding their beliefs. An alternative way to model information diffusion would be for individuals to communicate their opinion about the true state of the world, i.e., $Pr(\Phi = 0)$. However, in this case verification would automatically become certifiable, as (only) successful verification leads to a posterior belief of $Pr(\Phi = 0) = 1$. Lastly, individuals do not observe their neighbors' type. If individuals observed their neighbors' opinions or types, or information were certifiable, the rumor would always die out (Prakash et al., 2012).

Finally, we make a number of simplifying assumptions to ease the exposition which are without loss of generality. In particular, our results are qualitatively unaffected by assuming a non-degenerate degree distribution $P(k)$. Our assumption that individuals in state S receive a payoff of 0 does not affect marginal considerations as transitioning into and out of this state is not a choice. Lastly, our results also hold in the limit of the death rate $\delta \to 0$. The only adjustment required is to assume that payoffs accrue at a finite time T. Thus, we are able to capture the evolution of rumors which diffuse over whole lifespans (such as the HIV-AIDS or vaccination-autism links) as well as more short-lived rumors that diffuse in a constant population.

II. Main Analysis

In this section, we solve for the equilibrium of our model. To do so, we proceed in two steps. First, we derive the steady state prevalence of truth and rumor in the population for given verification efforts. This reveals that both have positive prevalence in steady state. Second, we solve for equilibrium verification rates and

[7]Such behavior would be consistent with people filtering out negative information that contradicts their point of view or systems of beliefs in order to maintain a congruent view of the world, and hence their well-being (Taylor and Brown, 1988). This results into information avoidance, inattention, or a biased interpretation of information (Golman, Hagmann and Loewenstein, 2017). Additionally, identity is rather fixed, as stressed in the literature on inter-group conflict in social psychology (Stephan and Stephan, 2017) or identity in economics (Akerlof and Kranton, 2000).

show that: *(i)*, they depend only on whether the message received is in line or not with one's bias and, *(ii)*, if one's prior y is sufficiently high, it is optimal to hold an opinion in line with the prior if verification is not successful. Finally, we study how the truth to rumor ratio changes with homophily, and how this depends on the verification technology.

A. Steady State with Exogenous Verification

We focus now on the model with given verification efforts to understand the properties of the steady state of our model.

We introduce the *effective diffusion rate*, $\lambda = \nu/\delta$, which summarizes the effect of ν and δ. Denote by ι^b the proportion of type $b \in \{0,1\}$ individuals in state I (irrespective of their opinions) with $\iota = (\iota^0 + \iota^1)/2$ as overall information prevalence. With this notation in place, we can state the following.

PROPOSITION 1: *Assume verification rates are given. If $\lambda k > 1$, the unique locally stable steady state is positive with $\iota^0 = \iota^1 = \iota = 1 - 1/(\lambda k)$. If $\lambda k \leq 1$, only the steady state in which the prevalence of both the truth and the rumor are zero is locally stable.*

While there always exists a steady state in which the prevalence of both the truth and the rumor are zero—if no individual ever transmits an information, nobody can ever become informed,—this is stable only if $\lambda k \leq 1$. Otherwise, there is a unique positive steady state, and it is stable.

Proposition 1 establishes information prevalence within each group. However, this does not tell us the diffusion of the truth and the rumor. This is the object of the following Proposition.

PROPOSITION 2: *In the unique positive and locally stable steady state, for exogenous verification rates ℓ and h:*
i) the information prevalence of both messages, ρ_0 and ρ_1, is increasing in the effective diffusion rate, λ, and in the number of meetings, k;
ii) the truth to rumor ratio, ρ_0/ρ_1, is greater than 1, increasing in both verification rates, ℓ and h, and independent of the effective diffusion rate, λ, and the number of meetings, k;
iii) the truth to rumor ratio, ρ_0/ρ_1, is decreasing in homophily, β, if and only if individuals verify more a message against their bias, i.e., $h > \ell$.

Proposition 2 results from the steady state prevalence of truth and rumor:

$$\rho_0 = \frac{1}{2} \cdot \frac{1 + h - 2\beta(h - \ell)}{1 - \beta(h - \ell)} \iota, \tag{7}$$

$$\rho_1 = \frac{1}{2} \cdot \frac{1 - h}{1 - \beta(h - \ell)} \iota. \tag{8}$$

These equations show that the steady states of opinions inherit uniqueness and stability from ι. Since for $\lambda k \leq 1$, neither opinion is endemic, we focus in the remainder of the paper on $\lambda k > 1$. Equation (8) highlights that both opinions survive unless $h = 1$. Thus, rumors may survive in the long run even if verification is possible.

Equations (7) and (8) show that with zero verification effort, $\rho_0 = \rho_1 = \iota/2$, and the truth prevalence increases in any form of verification, while rumor prevalence decreases. Hence, for any positive amount of verification, the truth exhibits a larger prevalence than the rumor.

Proposition 2 delivers some insights about potential relationships between online social networks and the diffusion of rumors. One factor through which online social networks allegedly stimulate the diffusion of rumors is the ease with which messages can be communicated, and the number of people receiving them. Thus, one generally expects them to have increased k, ν, or both. Proposition 2 shows that our model's predictions are in line with this view. However, it stresses that the truth and the rumor equally benefit from an increase in the ease of communication due to an increase in the number of meetings k or in the effective diffusion rate λ.

This result has several implications. First, while empirical studies on the impact of online communication often focus on the diffusion of rumors alone (e.g., Zollo et al., 2017), Proposition 2 stresses that a comparison with the diffusion of truthful messages would be of a greater interest. In line with this insight, the truth to rumor ratio will be the main object of interest in the remainder of the paper.

Second, if rumors have indeed become more prevalent in relative terms, this cannot be explained by online social networks increasing communication rates *per se*. We discuss in Section III how ease of communication might have indirect effects on relative rumor prevalence.

Likewise, high degrees of homophily are commonly associated with an increased diffusion of rumors as people are likely to hear only messages in line with their bias ("echo chambers"). In fact, in our model the impact of homophily on the diffusion of truth and rumor depends entirely on the verification rates of messages. Fixing these rates, homophily indeed benefits the rumor and harms the truth if individuals are more likely to verify messages against their bias than those aligned with it. While such behavior appears intuitive, it motivates us to study endogenous verification next.

B. Endogenous Verification

Given the utility function (6), each individual i chooses a verification effort when they first hear a message such that

$$g\left(\frac{c}{1 - Pr_t(b = \Phi|m)}\right) = x(\alpha_{it}), \tag{9}$$

where g is the inverse of the subderivative of x.

Individuals hence need to calculate the probability that $b = \Phi$ conditional on having received m. We assume that they perform this calculation using Bayes' rule, given that they are aware of the transmission process and that their prior of being of type $b = \Phi$ is y. This leads to the following:

$$Pr_t(b = \Phi | m = b) = \frac{y(\beta\rho^0_{0,t} + (1-\beta)\rho^1_{0,t})}{y(\beta\rho^0_{0,t} + (1-\beta)\rho^1_{0,t}) + (1-y)\beta\rho^1_{1,t}}, \tag{10}$$

$$Pr_t(b = \Phi | m \neq b) = \frac{y(1-\beta)\rho^1_{1,t}}{y(1-\beta)\rho^1_{1,t} + (1-y)((1-\beta)\rho^0_{0,t} + \beta\rho^1_{0,t})}. \tag{11}$$

From these probabilities and (9), equilibrium verification effort depends on whether the message received is in line or not with one's bias, and not on the prior *per se*. Therefore, in equilibrium there are two verification rates. This result follows from two properties of our model. First, all individuals believe their bias is correct with the same probability, y. Second, individuals derive prevalence of messages from the diffusion process described by equations (1), (2) and (3), as they do not observe all their neighbors' opinions when they set their verification effort. Substituting (10), (11), (7) and (8) in (9), equilibrium verification efforts are thus described by

$$\ell = x(\alpha^b) = g\left(\frac{c(y(\beta + (1-\beta)h - \beta(h-\ell)) + (1-y)\beta(1-h))}{(1-y)\beta(1-h)}\right), \tag{12}$$

$$h = x(\alpha^{-b}) = g\left(\frac{c(y(1-\beta)(1-h) + (1-y)(1-\beta+\beta\ell))}{(1-y)(1-\beta+\beta\ell)}\right). \tag{13}$$

We now prove that an equilibrium with endogenous verification exists.

PROPOSITION 3: *Under Assumption 1, there exists a threshold $\overline{y}$ such that, if $y \geq \overline{y}$, an equilibrium of the model with laws of motion* (1)*,* (2)*,* (3) *and endogenous verification exists.*

Holding an opinion in line with one's bias when a message is not verified, as we assumed so far, is optimal if $Pr(b = \Phi | m \neq b) \geq .5$. By (11), in steady state this requirement translates into

$$\frac{y}{1-y} \geq \frac{1-\beta+\beta\ell}{(1-\beta)(1-h)}. \tag{14}$$

We show that there exists a threshold $\overline{y}$ such that (14) is satisfied if $y \geq \overline{y}$. Intuitively, if one's prior is sufficiently strong, lacking verification, informed individuals hold an opinion in line with their bias.

If condition (14) is not met, absent verification, players believe the first message they hear. In that case, there cannot be positive verification rates in steady state.

If the prevalence of the rumor is such that it is worthwhile to verify, people do so, thereby reducing this prevalence, until verification is no longer profitable. At that point, the truth to rumor ratio stays constant, which might entail the prevalence of the rumor to be infinitesimally small. We discuss this case and derive the above results in Appendix B.

The following propositions characterizes equilibrium verification rates.[8]

PROPOSITION 4: *If Assumption 1 holds and $y \geq \overline{y}$, in any equilibrium, individuals exert higher effort verifying messages against their bias than those in line with it ($\ell \leq h$). Equilibrium verification is independent of the number of meetings, k, and the effective diffusion rate, λ. Furthermore, there exist values on verification costs $\underline{c}$ and $\overline{c}$ such that any equilibrium takes one of the following forms:*
i) If $c \geq \overline{c}$, there is no verification and the truth to rumor ratio is 1.
ii) If $c \in [\underline{c}, \overline{c})$, individuals verify only messages against their bias.
iii) If $c < \underline{c}$, both messages are verified.
While both $\overline{c}$ and $\underline{c}$ are decreasing in y, $\overline{c}$ is independent of homophily, β, and $\underline{c}$ increasing in it. Finally, the corresponding steady state is locally stable if and only if, either it coincides with a zero steady state and $\lambda k \leq 1$, or it is positive and $\lambda k > 1$.

The intuition behind these results is the following. First, a message in line with one's bias is verified less than one against it, because receiving the latter implies a lower probability of one's bias to be correct after Bayesian updating.

When verification is very costly, no message is verified; as a result, there is an equal mass of individuals holding each of the two opinions. As verification costs decrease, individuals first verify the message against their bias, as $\ell \leq h$. For even lower verification costs, individuals verify both messages. Verification implies the truth has a higher prevalence than the rumor.

The threshold of verification cost below which both messages are verified is increasing in homophily. Indeed, when individuals are more likely to meet people with the same bias, they attach a lower informational content to messages in line with their bias, thereby triggering increased verification.

Finally, Proposition 4 provides conditions for the stability of the steady state in the endogenous equilibrium that come directly from Proposition 1, as these conditions are independent of those determining verification.

We present a simple and intuitive example of a verification function that admits explicit solutions. All computations are in Appendix C.

[8] For exogenously given verification rates, we can straightforwardly apply the arguments of Jackson and Rogers (2007) to show that the locally stable steady state is also globally stable. Here, however, we are more interested in studying what happens when verification rates become endogenous. In this case we cannot exclude that, for some particular verification function $x(\cdot)$, there are multiple equilibria, each with an associated locally stable steady state. To study the dynamics in this setting we should also define some adaptive dynamics of the verification process out-of-equilibrium, which is beyond the scope of this paper. Therefore, we stick to the notion of local stability for the steady state associated with an equilibrium.

EXAMPLE 1 (Exponential verification function with a cap): Let the verification function be:

$$x(\alpha) = \begin{cases} 1 - e^{-\alpha} & \text{if } \alpha < -log[1-\bar{x}], \\ \bar{x} & \text{if } \alpha \geq -log[1-\bar{x}]. \end{cases} \tag{15}$$

This function results from the following verification process. Consider an individual who searches for information, which consists of n realizations leading to an answer with probability p_n per realization, up to the point where all information available, $\bar{x}$, is collected. Before reaching this cap, this search process gives at least one answer with probability $1-(1-p_n/n)^n$. If $np_n = \alpha$ as the number of realizations n goes to infinity and the success of each realization goes to zero, this probability converges to $1-e^{-\alpha}$ if this is lower than $\bar{x}$, and $\bar{x}$ otherwise, leading to (15).[9]

As $\bar{x} \to 1$, (15) converges to $1-e^{-\alpha}$. In this case, Figure 3 shows which messages are verified, depending on parameters y and c, for $\beta = 0.6$. The figure shows that, if verification costs are too high, no message is verified (the light-blue region). When the costs are sufficiently low, we have two scenarios. If the prior y is relatively low, both messages are verified (the purple region), but if it is higher, only the message against one's bias is verified (the blue region). Additionally, the initial prior y has to be above the threshold $\bar{y}$ derived in Proposition 3: in the white region, this condition is not satisfied.

If instead $\bar{x} \in (0,1)$, there are regions in the (c,y) space where verification rates are at $\bar{x}$. We will discuss this case in greater detail in the next subsection, as the value of $\bar{x}$ is important to understand the effect of homophily on the truth to rumor ratio.

C. Truth to rumor ratio

The truth to rumor ratio is the fraction of the prevalence of both opinions among informed individuals, i.e., ρ_0/ρ_1. It derives from equations (7) and (8) as follows

$$\frac{\rho_0}{\rho_1} = \frac{1+h}{1-h} - 2\beta\frac{h-\ell}{1-h}. \tag{16}$$

Equation (16) highlights that homophily has two effects on the truth to rumor ratio. First, there's a direct effect: as homophily increases, individuals are more exposed to messages in line with their bias, which, as Proposition 4 shows, are verified less. Hence, the truth to rumor ratio decreases.

However, endogenous debunking implies an indirect effect through equilibrium behavior: as homophily increases, the informativeness of messages in line with

[9] Realizations might have multiple answers. In the context of this model, this is equivalent to a lower cost of effort in verification.

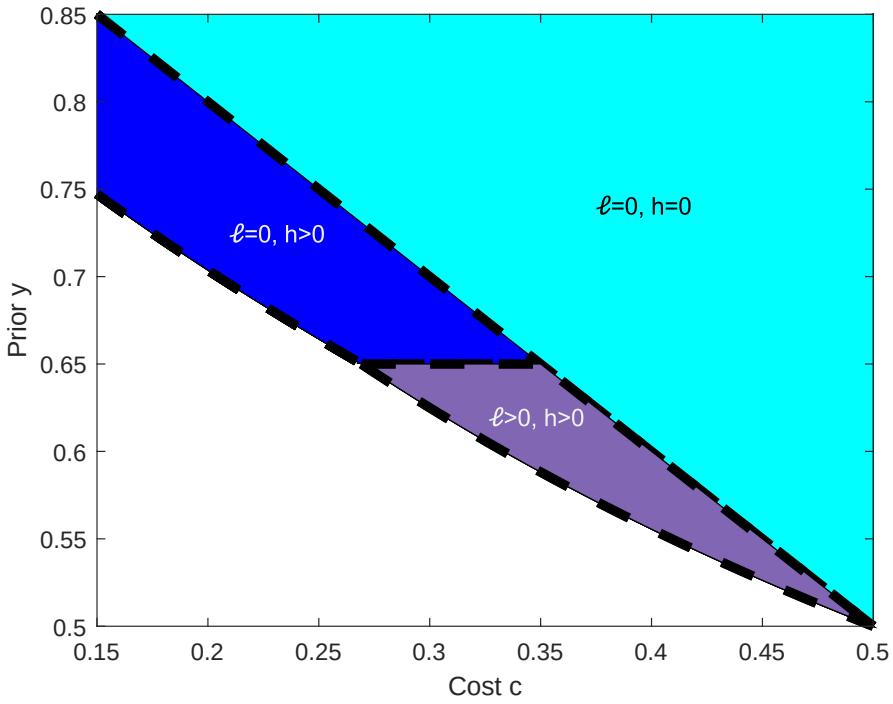

Figure 3. : Regions of the parameters c (on the x-axis) and y (on the y-axis) where different kind of equilibria exist with the exponential verification function, $\bar{x} \to 1$ and $\beta = 0.65$.

one's bias decreases, precisely because agents are more exposed to messages that are verified less. The opposite for messages against one's bias. In both cases, the probability they attach to their bias being correct decreases. As a result, individuals verify more, which increases the truth to rumor ratio.

Which of the two effects dominates is *a priori* not clear. Proposition 5 however derives two insights in this respect. Denote $L \geq 0$ as $g(L) = \ell$; hence, L represents the optimal marginal increase in the verification success of messages in favor of one's bias, $x(\alpha^b)$.

PROPOSITION 5: *In equilibrium, the truth to rumor ratio, ρ_0/ρ_1, is equal to*

$$1 + \frac{2}{y}\left(\frac{L(1-y)}{c} - 1\right)\beta.$$

Furthermore, ρ_0/ρ_1 is decreasing in homophily, β, if $c \in [\underline{c}, \overline{c}]$, i.e., when individuals verify only messages against their bias.

Proposition 5 first shows that, whenever only messages against one's bias are

verified, the indirect effect is absent. Indeed, we show in the Appendix that h does not change with homophily when $l = 0$. As a result, for verification costs such that this is an equilibrium, the truth to rumor ratio is decreasing in homophily.

Second, when both messages are verified, L is a sufficient statistic to study the truth to rumor ratio. In particular, Proposition 5 allows us to show that the total effect of homophily on the truth to rumor ratio depends on the local concavity of the verification function.

Indeed, if the verification function $x(\alpha)$ is twice differentiable, we find that the total effect of homophily on the truth to rumor ratio is

$$\frac{d}{d\beta}\frac{\rho_0}{\rho_1} = \frac{2}{y}\left(\frac{L(1-y)}{c} - 1\right) + \frac{2\beta(1-y)}{cy}\frac{dL}{d\beta}$$

Hence, the effect is positive if

$$L \geq \frac{c}{1-y} + \beta\left|\frac{dL}{d\beta}\right|. \tag{17}$$

While L is equal to $dx(\alpha)/d\alpha$ evaluated at α^b, i.e., depends on the steepness of the verification function $x(\alpha)$, $dL/d\beta$ depends on its concavity, as

$$\frac{dL}{d\beta} = \frac{\partial \ell}{\partial \beta} \cdot \left.\frac{\partial^2 x(\alpha)}{\partial \alpha^2}\right|_{x(\alpha)=\ell}.$$

Hence, when homophily increases, if verification increases faster than the decline in its marginal benefits, i.e., L is sufficiently larger than $dL/d\beta$, verification increases sufficiently that the truth to rumor ratio increases as well.

The role that steepness and concavity of the verification function play highlights why homophily may affect the truth to rumor ratio in non-obvious ways. We further elaborate on these results using the verification function we introduced in Example 1. All computations are in Appendix C.

EXAMPLE 2 (Effects of homophily in Example 1): By Proposition 5, it is interesting to study the effect of homophily on the truth to rumor ratio only whenever $y \geq \bar{y}$ and $c < \underline{c}$, such that both messages are verified.

In the case that $0 < \ell < h = \bar{x}$, if verification costs are low, so that ℓ is large (but below $\bar{x}$), the exponential function is not as concave as it is near the origin, and $|dL/d\beta|$ is low enough to satisfy inequality (17). As verification costs increase, ℓ goes down, $|dL/d\beta|$ goes up, and inequality (17) is reversed. As verification costs increase further, also h is less than $\bar{x}$, i.e., $0 < \ell < h < \bar{x}$. We discussed this case in Example 1 (purple region in Figure 3). In this case, the truth to rumor ratio becomes $(2(1-y)-c)/c$, which is independent of homophily. This is due to a specificity of the exponential example, where, when both verification rates can freely adapt, the direct and the indirect effects of homophily balance each other

out. Note that, in all regions verification rates and the truth to rumor ratio are (weakly) decreasing in verification costs.

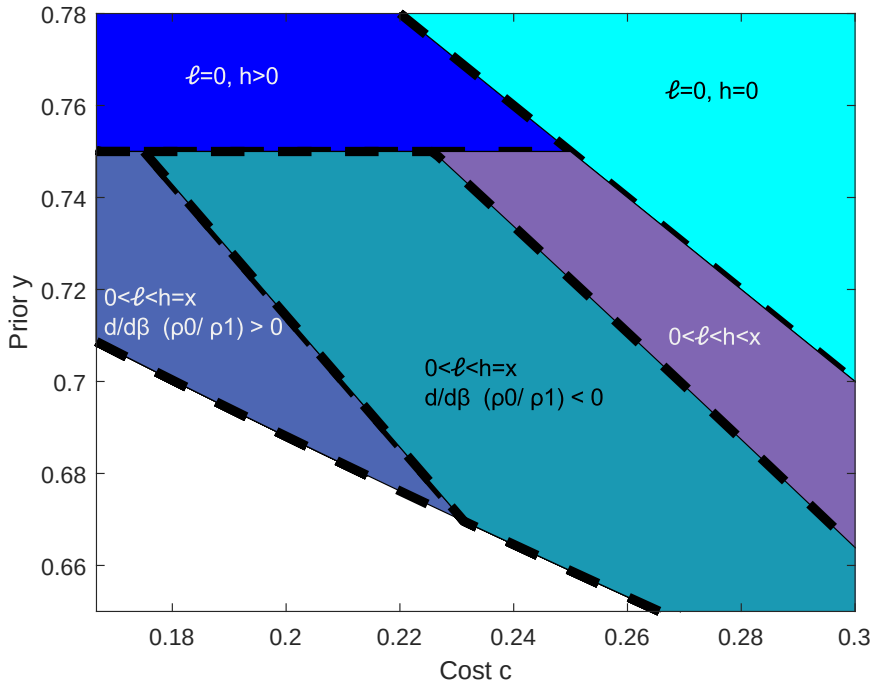

Figure 4. : Regions of the parameters c (on the x-axis) and y (on the y-axis) where different kind of equilibria exist with the exponential verification function, $\beta = 0.75$ and a cap set at $\bar{x} = 0.3$.

Figure 4 depicts these regions, illustrating how the effect of homophily on the truth to rumor ratio differs depending on the values of the prior y and verification costs, c.

In sum, even in a simple and homogeneous verification function as the one assumed here, the impact of homophily on the truth to rumor ratio depends on exact equilibrium reactions, which are difficult to predict. Despite this, our model delivers several policy implications, which we discuss in the following section.

III. Discussion and Policy Implications

Much of the ongoing discussion on how to fight the spread of rumor rests on the role of online social networks. In the following, we address the most commonly cited arguments as to how these platforms relate to the diffusion of rumors.

A. Time-limited Injections of Messages

One obvious policy in the fight against rumors is to increase the volume of truthful messages in the network for a limited time, e.g., through an information campaign. In our model, whenever the verification function delivers a unique stable equilibrium, such a policy does not affect the truth to rumor ratio in the long run. Thus, rumors cannot simply be debunked by increasing the prevalence of the truthful message in the network for a limited time.

B. Online Social Networks Display Higher Homophily

In principle, online social networks offer individuals the chance to self-select in more homogeneous groups along several dimensions than offline communication, as meeting opportunities are less constrained online. The extent to which this happens is still debated. On the one hand, there is evidence that patterns of online news consumption are no more segregated or polarized than offline ones (Gentzkow and Shapiro, 2011; Boxell, Gentzkow and Shapiro, 2018). On the other hand, online communication patterns display levels of segregation that are above the most segregated offline communication networks found by Gentzkow and Shapiro (2011), e.g., in Halberstam and Knight (2016) for Twitter and Zollo et al. (2017) for Facebook. As a result, the widespread use of online social networks might have implied an increase in homophily and resulted in echo chambers. It is usually assumed that this is one of the main culprits for an increase in the diffusion of rumors in the past decade.

Granting that online social networks have increased homophily, our preceding analysis makes the mechanism behind this argument explicit, stressing under which conditions it is valid. As meetings are more homophilous, on the one hand, individuals hear fewer discordant messages, that are verified less. On the other hand, messages confirming one's prior become less informative, so that these messages are verified more. Rumors thrive only when this first effect dominates.

C. Online Social Networks Facilitate Communication

By facilitating communication, online social networks have arguably increased the transmission rate of information. Through the lens of our model, this is captured by an increase in either the number of meetings, k, or in the diffusion rate, λ. This increases the measure of informed individuals, but it does not affect verification rates (Proposition 4).[10] The truth to rumor ratio is then unaffected by changes in these parameters. Therefore, while ease of communication leads to an absolute increase in rumor prevalence, this in itself does not imply a *relative* increase.

[10] In a more general model, the same holds for changes in the degree distribution $P(k)$.

It is possible, however, that the costs of verification, c, instead depend positively on the amount of online communication. Such *congestion* or *information overload* effects occur naturally if we consider verification costs as time costs. In that case, through increasing communication, online social networks may increase verification costs and thus affect verification rates.

To fix ideas, consider individuals to have a given time endowment $\mathcal{H}$ in each period, which they can spend on communication (νk), verification (α), and on independent leisure activities z. Assume that agents obtain utility $u(z)$ from their leisure activities, where $u(\cdot)$ is a continuous, increasing and concave function. Its concavity captures the decreasing marginal utility of private leisure. Agent i's utility at time t can then be expressed as

$$U_{it} = x(\alpha_{it}) + (1 - x(\alpha_{it}))\, Pr_i(b = \Phi|m) + u(\mathcal{H} - \nu k - \alpha_{it}). \tag{18}$$

Optimal verification effort is chosen again according to equation (9), defining $c = du/d\alpha_i$. In this case, any increase in νk increases verification costs.

We conclude that if increased communication creates congestion effects in verification, online social networks may indeed have affected verification rates and therefore the truth to rumor ratio.

D. Fact Checking to Fight Rumors

The ease with which rumors can be debunked by agents in the network has been an important aspect in discussions on policy interventions. Many established news outlets, such as *The New York Times* or *Le Monde*, publish guides on how to recognize false information, and have introduced newsrooms where fake news is identified and debunked (Roose, 2018). On various online social networks, for example *Facebook*, disputed information may be "flagged" as such (Maidenberg, 2018) by independent fact checkers.

In our model, incentives that induce higher verification rates unambiguously increases the truth to rumor ratio. Hence, fact checking should be properly designed.

We believe this is not always the case. For example, Facebook experimented with assigning "flags" to posts simply stating that they had been disputed. This practice has been abandoned in 2017, as it proved less effective to stop the spread of rumors than expected, with some evidence that it might even have promoted the spread of such posts. Instead, an alternative was suggested whereby, next to a disputed post, a user would see various links to the articles disputing it (Silverman, 2016). In the context of our model, the earlier policy would have amounted to injecting truthful messages into the network, and as such would indeed have been expected to be unsuccessful in the long run. The updated policy instead, by providing links to the sources of the dispute, can credibly claim to lower verification costs.

Overall, our model corroborates that one of the most intuitive policies to deter rumors—incentivizing individuals to verify information—is also the one most likely to succeed as long as it is carefully designed.

IV. Partisans

It is possible that the diffusion of rumors in social media is primarily driven by individuals that hold a certain opinion independently of whether they are confronted with factual evidence disproving it (Zollo et al., 2017). We now study the implication of the existence of such partisan individuals on the diffusion of messages.

We introduce partisans as people who never verify, always hold an opinion in line with their prior and transmit messages accordingly.[11]

We assume that a fraction γ of individuals of each type are partisans. This does not impact the evolution of information in group 0, as both partisans and non-partisans are always of the opinion that $\Phi = 0$. However, there are fewer individuals of type 1 who may verify any message they hear. We then derive the following Proposition.

PROPOSITION 6: *Assume a fraction γ of the population consists of partisans. The prevalence of the truth and the rumor among the remaining $1-\gamma$ individuals are unaffected by γ and remain as in the baseline model.*

This result depends on the endogeneity of verification. Indeed, the existence of partisans implies that non-partisans place a higher probability of being told the rumor. This therefore leads to a higher degree of verification among them, which offsets the negative impact that partisans have on the relative prevalence of truth and rumor. Formally, in the proof of Proposition 6 we show that the prevalence of both opinions can be re-written as in the baseline model with verification rates $\hat{x}(\alpha) = (1-\gamma)x(\alpha)$ instead of the original $x(\alpha)$'s. This leads to the equilibrium $\hat{x}(\alpha)$ being unaffected by partisans.

V. Conclusions

In this paper, we model how a correct message and a rumor diffuse in a population of individuals who seek the truth. Individuals verify the messages they hear based on the probability that what they hear is correct. They are biased in a way that, if they do not verify, they hold the opinion that adheres to their view of the world.

We find that the rumor survives for any positive cost of verification. New communication technologies increase its absolute prevalence, however, its prevalence

[11] Alternatively, for all $\beta < 1$, we can think of partisans as having a prior of 1 of being biased towards the true state of the world. If $\beta = 1$, individuals can never meet someone with the opposing bias. Hence, hearing the opposite message is conclusive proof that one's bias is wrong, contradicting the certain prior.

relative to the truth depends exclusively on verification and the degree of homophily in meetings. We show that the impact of homophily is nuanced. On the one hand, an echo chamber effect emerges: individuals are more exposed to messages in line with their bias, which are verified less. On the other, higher homophily means messages reinforcing ones prior become less informative, and the opposite for messages going against one's prior. This mechanism then leads to higher verification rates. Overall, the relative virality of rumors increases only when the first effect dominates the latter.

We employ our results to discuss policies to debunk rumors. While injections of truthful messages are ineffective in debunking the rumor in the long run, successful policy interventions revolve around incentivizing individuals to verify. In the light of our model, lowering the degree of homophily may fail to achieve this result. In sum, new communication technologies played a role in making rumors more viral if they decreased verification—due to, for example, congestion effects.

In our model, there only are two opposing opinions and messages. This excludes the possibility that malevolent agents may aim to decrease the spread of the truth by creating new opposing messages over time. We think the analysis of this phenomenon is a promising avenue for future research.

Proofs

Proof of Proposition 1. We combine equations (2) and (3) to analyze the law of motion of ι^1:

$$\frac{\partial \iota_t^1}{\partial t} = \frac{1}{2}(1-\iota_t^1)\nu k\left[\beta\iota_t^1 + (1-\beta)\iota_t^0\right] - \frac{1}{2}\iota_t^1\delta.$$

Define $\vartheta^0 = \beta\iota^0 + (1-\beta)\iota^1$ and $\vartheta^1 = \beta\iota^1 + (1-\beta)\iota^0$. Then, in steady state, information prevalence in either group is

$$\iota^0 = \frac{\lambda k\vartheta^0}{1+\lambda k\vartheta^0}, \tag{A-1}$$

$$\iota^1 = \frac{\lambda k\vartheta^1}{1+\lambda k\vartheta^1}, \tag{A-2}$$

from which it follows that there exists one steady state in which $\iota^0 = \iota^1 = \iota = 1 - 1/(\lambda k)$. To show that this steady state is unique, note that, by (A-1) and (A-2), in any positive steady state it must be that

$$\begin{aligned}
\frac{\iota^0}{\iota^1} &= \frac{\lambda k\vartheta^0[1+\lambda k\vartheta^1]}{\lambda k\vartheta^1[1+\lambda k\vartheta^0]},\\
\iota^0\vartheta^1[1+\lambda k\vartheta^0] &= \iota^1\vartheta^0[1+\lambda k\vartheta^1],\\
\iota^0\vartheta^1-\iota^1\vartheta^0 &= \lambda k\vartheta^0\vartheta^1[\iota^1-\iota^0],\\
\beta\iota^0\iota^1+(1-\beta)\iota^{0^2}-\beta\iota^1\iota^0-(1-\beta)\iota^{1^2} &= \lambda k\vartheta^0\vartheta^1[\iota^1-\iota^0],\\
(1-\beta)(\iota^{0^2}-\iota^{1^2}) &= \lambda k\vartheta^0\vartheta^1[\iota^1-\iota^0].
\end{aligned}$$

If $\iota^1 > \iota^0$, then the right-hand side of this equation is positive while the left-hand side is negative, and vice versa for $\iota^1 < \iota^0$. It can only hold if $\iota^0 = \iota^1 = \iota$.
Finally, either both groups have positive information prevalence, or neither. Deriving the Jacobian of the differential system reveals that both eigenvalues are negative at the positive steady state if and only if $\lambda k > 1$. The steady state of zero information prevalence has instead two negative eigenvalues if and only if $\lambda k \leq 1$. These properties of the steady state are inherited from the associated ρ_0 and ρ_1. This concludes the proof of Proposition 1. ■

Proof of Proposition 2. To derive equations (7) and (8), note that, by (1), $\rho_0^0 = \iota^0$. By Proposition 1, $\iota^1 = \iota^0 = 1 - 1/(\lambda k)$. Plugging these values in the steady states of equations (2) and (3), equations (7) and (8) obtain.
The derivatives of (7) and (8) show that the prevalence of the truth is increasing, while the prevalence of the rumor is decreasing, in ℓ and h. Hence, the lowest value that ρ_0 can take, and the highest value of ρ_1, is at $\ell = h = 0$, when they are both equal to ι. The truth always has at least as high a prevalence as the rumor. This concludes the proof of Proposition 2. ■

Proof of Proposition 3. Suppose that the prior y is such that it is optimal for individuals who do not verify a message to hold an opinion in line with their bias. Then, the equilibrium conditions on $\ell, h \in [0, \bar{x}]$ translate into

$$\begin{aligned}
\frac{c(y(\beta+(1-\beta)h+\beta(\ell-h))+(1-y)\beta(1-h))}{\beta(1-y)(1-h)} &\in \Delta(\ell),\\
\frac{c(y(1-\beta)(1-h)+(1-y)(1-\beta+\beta\ell))}{(1-y)(1-\beta+\beta\ell)} &\in \Delta(h).
\end{aligned}$$

As a consequence of Assumption 1, $\Delta(x(\alpha))$ is an upper hemicontinuous, nonempty, closed, and convex correspondence and the values of the left and right derivative of x always exist. Define G as the correspondence from $[0, \bar{x}]^2$ that applies to the functions on the left-hand side of the above two equations a couple (ℓ, h) and then applies correspondence g to both solutions. G is a continuous convex valued correspondence from $[0, \bar{x}]^2$ to itself. By Kakutani's fixed point

theorem, there exists an equilibrium consisting of (ℓ, h) that are a fixed point of G.

For some verification functions, $\bar{x}$ may not be part of the domain, as it is only the asymptotic limit of $x(\alpha)$ as α goes to infinity. In this case, the function $x(\alpha)$ is always strictly increasing and concave. Therefore, Δ is a strictly decreasing function from $(0, \bar{d}]$ to $[0, \bar{x})$. By (12) and (13), we then have that $0 \leq \ell < h < 1$, and $\Delta(h) > 0$, which in turn implies $\Delta(\ell) > 0$. These facts imply that a fixed point of G is interior.

We now derive the values of the initial prior y such that holding an opinion in line with one's bias when a message is not verified is optimal, i.e., that $Pr(b = \Phi | m \neq b) \geq .5$. By (11), in steady-state this requirement translates into (14). The left-hand side of (14) is continuously increasing in $y \in (.5, 1)$, is 1 when $y \to .5$ and diverges to infinity as $y \to 1$. From (9), as Bayes' rule is continuous and increasing in y and $g(\cdot)$ is continuous and weakly decreasing, ℓ and h are continuous and decreasing in y. Therefore, the right-hand side of (14) is continuous and decreasing in y, it is always greater than 1 and it converges to 1 as $y \to 1$, as at that limit both h and ℓ are null. Hence, there exists a unique threshold $\bar{y} > .5$ such that for all $y \geq \bar{y}$ holding an opinion in line with one's bias when a message is not verified is optimal. This concludes the proof of Proposition 3. ■

Proof of Proposition 4. The result that in equilibrium, $\alpha^b \leq \alpha^{-b}$ follows directly from $Pr(b = \Phi | m = b) \geq Pr(b = \Phi | m \neq b)$. From equations (12) and (13), we see that neither k nor λ affect α^b and α^{-b}. We also see that there are two non–negative numbers $L \in \Delta(\ell)$ and $H \in \Delta(h)$ such that

$$L = c\left(\frac{y(\beta + (1-\beta)h - \beta(h-\ell))}{\beta(1-y)(1-h)} + 1\right), \tag{A-3}$$

$$H = c\left(\frac{y(1-\beta)(1-h)}{(1-y)(1-\beta+\beta\ell)} + 1\right). \tag{A-4}$$

To find $\underline{c}$, we need to look for the threshold at which ℓ becomes positive, setting $\ell = 0$ in the left-hand side of (A-3), and solving for this value being equal to $\bar{d}$, which is the derivative of the verification function $x(\cdot)$ at the origin. Making c explicit, this results in:

$$\underline{c} = \frac{\bar{d}\beta(1-h)(1-y)}{\beta + h(y - (1+y)\beta)},$$

which, as a first order effect, is decreasing in h and y, and increasing in β. However, from equation (13), when $\ell = 0$ we have

$$h = g\left(c\left(1 + \frac{y(1-h)}{1-y}\right)\right). \tag{A-5}$$

Therefore, h is independent of β and increasing in y, and then $\underline{c}$ is decreasing in

y and increasing in β. To find $\bar{c}$, we look for the threshold at which h becomes positive, setting $\ell = 0$ and $h = 0$ in the left-hand side of (A-4), and solving for this value being equal to $\bar{d}$. In this way, making c explicit, we find $\bar{c} = \bar{d}(1-y)$.

Finally, as λ and k do not affect verification rates, stability follows from Proposition 1. This concludes the proof of Proposition 4. ■

Proof of Proposition 5. We start from the computations in the proof of Proposition 4. Expressing equation (A-3) in terms of ℓ, we obtain

$$\ell = h - \frac{1-h}{y} + \frac{(1-h)(1-y)L}{cy} - \frac{h}{\beta}.$$

Plugging this into the truth to rumor ratio from (16), we obtain

$$\frac{\rho_0}{\rho_1} = 1 + \frac{2}{y}\left(\frac{L(1-y)}{c} - 1\right)\beta.$$

The results follow from the fact that if $L \in \Delta(\ell)$, then $g(L) = \ell$. Finally, note that from (16) the truth to rumor ratio when $\alpha^b = 0$, i.e., $\ell = 0$, is

$$\frac{\rho_0}{\rho_1} = \frac{1 + h - 2\beta h}{1-h}.$$

From equation (13), when $\ell = 0$ we have (A-5). This shows that h is independent of β and ρ_0/ρ_1 is decreasing in homophily β if α^b stays at 0. This concludes the proof of Proposition 5. ■

Proof of Proposition 6. As individuals of type 0 are always of the opinion that $\Phi = 0$, γ is irrelevant for $\rho_{0,t}$. We separately consider partisans and non-partisans of type 1. All informed partisans of this type hold opinion $\Phi = 1$; denote the corresponding prevalence as $\rho_{1,t}^{\gamma}$. We denote the prevalence of opinion b among non-partisans of type 1 as $\rho_{b,t}^{1-\gamma}$ and the proportion of informed non-partisans (partisans) as $\iota_t^{1-\gamma}$ (ι_t^{γ}). As in the benchmark model, $\iota_t^0 = \rho_{0,t}^0$, but now $\iota_t^1 = \gamma\iota_t^{\gamma} + (1-\gamma)[\iota_{0,t}^{1-\gamma} + \iota_{1,t}^{1-\gamma}]$. Hence,

$$\rho_{1,t} = \frac{1}{2}[(1-\gamma)\rho_{1,t}^{1-\gamma} + \gamma\rho_{1,t}^{\gamma}] \text{ and } \rho_{0,t} = \frac{1}{2}[\iota_t^0 + (1-\gamma)\rho_{0,t}^{1-\gamma}].$$

The system describing the evolution of the prevalence of opinions is

$$\frac{\partial \rho_{0,t}^{0}}{\partial t} = \frac{1}{2}(1-\rho_{0,t}^{0})\nu k[\beta \iota_t^0 + (1-\beta)\iota_t^1] - \frac{1}{2}\rho_{0,t}^{0}\delta, \tag{A-6}$$

$$\frac{\partial \rho_{1,t}^{\gamma}}{\partial t} = \frac{1}{2}\gamma(1-\rho_{1,t}^{\gamma})\nu k[\beta \iota_t^1 + (1-\beta)\iota_t^0] - \frac{1}{2}\gamma\rho_{1,t}^{\gamma}\delta, \tag{A-7}$$

$$\begin{aligned}\frac{\partial \rho_{0,t}^{1-\gamma}}{\partial t} &= \frac{1}{2}(1-\gamma)[1-\rho_{0,t}^{1-\gamma}-\rho_{1,t}^{1-\gamma}]\nu k\Big[\beta\ell[(1-\gamma)\rho_{1,t}^{1-\gamma}+\gamma\rho_{1,t}^{\gamma}]+\\ &+\beta h(1-\gamma)\rho_{0,t}^{1-\gamma}+(1-\beta)h\rho_{0,t}^{0}\Big] - \frac{1}{2}(1-\gamma)\rho_{0,t}^{1-\gamma}\delta,\end{aligned} \tag{A-8}$$

$$\begin{aligned}\frac{\partial \rho_{1,t}^{1-\gamma}}{\partial t} &= \frac{1}{2}(1-\gamma)[1-\rho_{0,t}^{1-\gamma}-\rho_{1,t}^{1-\gamma}]\nu k\Big[\beta(1-\ell)[(1-\gamma)\rho_{1,t}^{1-\gamma}+\gamma\rho_{1,t}^{\gamma}]+\\ &+(1-h)\left(\beta(1-\gamma)\rho_{0,t}^{1-\gamma}+(1-\beta)\rho_{0,t}^{0}\right)\Big] - \frac{1}{2}(1-\gamma)\rho_{1,t}^{1-\gamma}\delta.\end{aligned} \tag{A-9}$$

By combining equations (A-8) and (A-9), we find that the evolution of $\iota_t^{1-\gamma} = \rho_{0,t}^{1-\gamma} + \rho_{1,t}^{1-\gamma}$ mirrors the one of ι_t^{γ}. Following the same analysis of the benchmark model, we find that the steady state values of ρ_0 and ρ_1 are

$$\rho_0 = \frac{1}{2}\frac{1+(1-\gamma)h+2\beta(1-\gamma)(\ell-h)}{1+\beta(1-\gamma)(\ell-h)}\iota \text{ and } \rho_1 = \frac{1}{2}\frac{1-(1-\gamma)h}{1+\beta(1-\gamma)(\ell-h)}\iota.$$

Hence, verification with partisans is as in the benchmark model with $\hat{x}(\alpha) = (1-\gamma)x(\alpha)$. The prevalence of both opinions and verification rates are therefore unchanged. This concludes the proof of Proposition 6. ■

Analysis of the Model when Priors are below the Threshold $\bar{y}$

In the benchmark model presented in this paper, we assume that priors y are high enough such that it is optimal for individuals to believe their bias is correct absent successful verification, i.e., $y \geq \bar{y}$. If this is not the case, and $0.5 \leq y < \bar{y}$, absent successful verification it is optimal for individuals to believe whichever message they receive. This leads to the following differential equations of the

system:

(A-1)
$$\frac{\partial \rho^0_{0,t}}{\partial t} = \frac{1}{2}(1-\rho^0_{0,t}-\rho^0_{1,t})\nu k\left[\beta(\rho^0_{0,t}+h_t\rho^0_{1,t})+(1-\beta)(\rho^1_{0,t}+h_t\rho^1_{1,t})\right]-\frac{1}{2}\rho^0_{0,t}\delta,$$
(A-2)
$$\frac{\partial \rho^0_{1,t}}{\partial t} = \frac{1}{2}(1-\rho^0_{0,t}-\rho^0_{1,t})\nu k\left[\beta(1-h_t)\rho^0_{1,t}+(1-\beta)(1-h_t)h_t\rho^1_{1,t}\right]-\frac{1}{2}\rho^0_{1,t}\delta,$$
(A-3)
$$\frac{\partial \rho^1_{0,t}}{\partial t} = \frac{1}{2}(1-\rho^1_{0,t}-\rho^1_{1,t})\nu k\left[\beta\left(\rho^1_{0,t}+\ell_t\rho^1_{1,t}\right)+(1-\beta)(\ell_t\rho^0_{1,t}+\rho^0_{0,t})\right]-\frac{1}{2}\rho^1_{0,t}\delta,$$
(A-4)
$$\frac{\partial \rho^1_{1,t}}{\partial t} = \frac{1}{2}(1-\rho^1_{0,t}-\rho^1_{1,t})\nu k\left[\beta(1-\ell_t)\rho^1_{1,t}+(1-\beta)(1-\ell_t)\rho^0_{1,t}\right]-\frac{1}{2}\rho^1_{1,t}\delta.$$

We can follow the same steps as in the proof of Proposition 1 to show that magnitude of the total steady state information prevalence is $\iota^0=\iota^1=1-1/(\lambda k)$, as in the benchmark model. The uniqueness and stability properties of this steady state follow.

Given the steady state properties of information prevalence in each group, the steady state condition for the system of equations (A-1)-(A-4) are

$$\rho^0_0 = \beta[\rho^0_0+h_t\rho^0_1]+(1-\beta)[\rho^1_0+h_t\rho^1_1], \quad \text{(A-5)}$$
$$\rho^0_1 = \beta(1-h_t)\rho^0_1+(1-\beta)(1-h_t)\rho^1_1, \quad \text{(A-6)}$$
$$\rho^1_1 = \beta(1-\ell_t)\rho^1_1+(1-\beta)(1-\ell_t)\rho^0_1, \quad \text{(A-7)}$$
$$\rho^1_0 = \beta[\rho^1_0+\ell_t\rho^1_1]+(1-\beta)[\rho^0_0+\ell_t\rho^0_1]. \quad \text{(A-8)}$$

From this system, it is immediate that the steady state in which all information prevalence is zero still exists. It is also easy to see that, if $\rho^0_1=\rho^1_1=0$, there exists a steady state in which $\rho^0_0=\rho^1_0$; we refer to this state as the *no rumor steady state*. In that case, $\iota^0=\rho^0_0$ and $\iota^1=\rho^1_0$. Finally, it can readily be shown that, if $h_t=\ell_t=0$—we refer to this state as the *no verification steady state*,—any steady state has the property that $\rho^0_1=\rho^1_1$ and $\rho^0_0=\rho^1_0$; however, their magnitudes are not determined.

To show that there exist no other steady states, we re-arrange equations (A-6) and (A-7) to solve for ρ^0_1:

$$\rho^0_1[1-\beta(1-h_t)] = (1-\beta)(1-h_t)\frac{(1-\beta)(1-\ell_t)}{1-\beta(1-\ell_t)}\rho^0_1$$

and further re-arranging shows that this requires

$$1-\beta(1-h_t)-\beta(1-\ell_t)=(1-h_t)(1-\ell_t)-2\beta(1-h_t)(1-\ell_t),$$

which is only satisfied if $h_t=\ell_t=0$.

When there is no verification, we can write the evolution of truth and rumor prevalence over time as

$$\begin{aligned}\frac{\partial\rho_{0,t}}{\partial t} &= \frac{1}{2}(1-\rho_{0,t}-\rho_{1,t})\nu k\rho_{0,t}-\frac{1}{2}\rho_{0,t}\delta,\\ \frac{\partial\rho_{1,t}}{\partial t} &= \frac{1}{2}(1-\rho_{0,t}-\rho_{1,t})\nu k\rho_{1,t}-\frac{1}{2}\rho_{1,t}\delta.\end{aligned}$$

This is identical to the system in Prakash et al. (2012) for equal virus strength, i.e., we have a non-hyperbolic fixed point and setting the differential equations to zero will not allow us to find a solution. Following Prakash et al. (2012), we find instead that we always have

$$\int_0^{\rho_{0,t}}\frac{1}{\nu k\rho_{0,t}}d\rho_{0,t} = \int_0^{\rho_{1,t}}\frac{1}{\nu k\rho_{1,t}}d\rho_{1,t}$$

which implies that

$$\frac{\rho_{0,t}}{\rho_{1,t}} = \frac{\rho_{0,0}}{\rho_{1,0}}.$$

Thus, whenever $0.5\leq y<\bar{y}$ and at least one verification rate is strictly positive, a positive rumor prevalence is no steady state and rumor prevalence must be decreasing over time. Given that the probability that message m is correct is given by (10) if $m=b$ and by 1 minus the expression in (11) if $m\neq b$, both of these probabilities are increasing whenever rumor prevalence decreases (and truth prevalence increases). Therefore, as long as it is optimal for individuals to exert positive effort to verify at least $m\neq b$, rumor prevalence will decrease and truth prevalence increase over time. This process will continue until it becomes unprofitable for individuals to exert any verification effort, and the truth to rumor ratio that prevails at this point will from then on remain constant.

Computations for the Examples

Computations for Example 1. Here, we compute the different possible equilibria with the exponential verification function with a cap at $\bar{x}$.

Case I. By Proposition 4 and since $\bar{d}=1$ in this example, no verification happens, i.e., $\ell_1=h_1=0$, if $c\geq 1-y$. Therefore, in all the following cases with verification, $c<1-y$ must hold.

Case II. If there is some verification, there are several possible equilibria. We first focus on cases where only one message is verified, i.e., $\ell_2 = 0$. We look for h that solves (A-4), setting $H(h_2) = 1 - h_2$. The solution is

$$h_2 = \frac{1 - y - c}{1 - y - cy},$$

which is always non negative if $c < 1 - y$. Moreover, $h_2 < \bar{x}$ if

$$c > \frac{(1 - \bar{x})(1 - y)}{1 - \bar{x}y} = c_1$$

To derive when $\ell_2 = 0$, we set the left-hand side of (A-3) lower than 1, and obtain that this requires $y \geq \beta$. In this case, it is possible to compute explicitly $\bar{y}_2 = 1/(1 + 2c)$.

Case III. If $c \leq c_1$, we have an equilibrium with $\ell_3 = 0$ and $h_3 = \bar{x}$; again, $\ell_3 = 0$ is guaranteed if $y \geq \beta$. The condition on priors requires $y \geq \bar{y}_3 = 1/(2 - \bar{x})$. Therefore, this constraint is binding only when $\beta < 1/(2 - \bar{x})$.

Case IV. In the equilibrium where $0 < \ell, h, < \bar{x}$, the system from equations (A-3) and (A-4) can be solved substituting L with $1 - \ell$ and H with $1 - h$. If $\ell, h > 0$, this results in

$$\ell_4 = \frac{1 - c - y}{1 - y} \cdot \frac{\beta - y}{\beta} \text{ and } h_4 = \frac{1 - c - y}{1 - y} \cdot \frac{1 - (1 - c)y - c\beta}{1 - y - c\beta}.$$

Hence, $\ell_4 > 0$ needs $\beta > y$, and this also implies $h_4 > 0$ as $h_4 > \ell_4$; moreover, $h_4 < \bar{x}$ holds if the left-hand side of (A-4) is lower than $1 - \bar{x}$, which translates into

$$c > \frac{\beta(1 - \bar{x})(1 - y)}{\beta - \bar{x}(\beta - (1 - \beta)y)} = c_2$$

Note that both ℓ_4 and h_4 are decreasing in c and increasing in β for $\beta \in (.5, 1)$ and $y \in (.5, 1)$. In this equilibrium, (14) holds if

$$y \geq \bar{y}_4 = \frac{1}{2}\left(2 + c - \sqrt{c}\sqrt{4 + c - 4\beta + 4\beta^2 c - 4\beta c} - 2\beta c\right).$$

Case V. Let us consider now the equilibrium in which $0 < \ell_5 < \bar{x}$ and $h_5 = \bar{x}$. First, we set the right-hand side of (A-3) to $1 - l$, and we solve for ℓ, obtaining

$$\ell_5 = 1 - c\frac{(1 - \beta)\bar{x}y + \beta(1 - \bar{x} + y)}{\beta(cy + (1 - \bar{x})(1 - y))}.$$

which is decreasing in c and increasing in β. Note that again ℓ_5 is positive if $y < \beta$. To see that $h = \bar{x}$ is possible, we set the right-hand side of (A-4) weakly

smaller than $1-\bar{x}$, and we solve for c, obtaining

$$c \leq c_3 \quad = \frac{1-y}{2} + \frac{(1-y)\left(1-\bar{x}\beta-\sqrt{(y+1-\beta(1-\bar{x}))^2-4\bar{x}y}\right)}{2(\beta-y)}$$

This is an equilibrium only when $\ell_5 < \bar{x}$; this holds if $1-\ell_5 < 1-\bar{x}$, that is:

$$c > \frac{\beta(1-\bar{x})^2(1-y)}{\beta-\beta\bar{x}+\bar{x}y} = c_4.$$

The condition (14) on the prior is satisfied if

$$y \geq \bar{y}_5 \quad = \frac{c-\beta\bar{x}+\bar{x}-3+\beta+\sqrt{(3-\beta-c+\beta\bar{x}-\bar{x})^2+4(1-\beta c)(\beta-2\beta c+2c-\beta\bar{x}+\bar{x}-2)}}{2(\beta-2\beta c+2c-\beta\bar{x}+\bar{x}-2)}.$$

Note that this condition implies $c > c_4$.
Case VI. Finally, there is an equilibrium with $\ell_6 = h_6 = \bar{x}$ if $c \leq c_4$. In this case, condition (14) on the prior is satisfied if

$$y \geq \bar{y}_6 = \frac{1-\beta+\beta\bar{x}}{2(1-\beta)(1-\bar{x})+\bar{x}}.$$

When $\bar{x} \to 1$, $\ell, h \to 1$, and this is an equilibrium only for $c = 0$.
Summing up this example, $\ell_1 = h_1 = 0$ if $c \geq 1-y$; if instead $c < 1-y$, there are the following equilibria:

A) *i)* $\ell_1 = h_1 = 0$ if $c \geq 1-y$;

B) if $c < 1-y$ and $y \geq \beta$, $\ell = 0$, the following cases emerge:
ii) $\ell_2 = 0$ and $h = h_2 \in (0, \bar{x})$ if $c_1 < c$ and $y \geq \bar{y}_1$;
iii) $\ell_3 = 0$ and $h = h_3 = \bar{x}$ if $c \leq c_1$ and, whenever $\beta < \bar{y}_2$, $y \geq \bar{y}_2$;

C) if $c < 1-y$ and $y < \beta$, $\ell > 0$, the following cases emerge:
iv) $0 < \ell_4 < h_4 < \bar{x}$ if $c > c_2$ and $y \geq \bar{y}_4$;
v) $\ell_5 \in (0, \bar{x})$ and $h_5 = \bar{x}$ if $c_3 \leq c < c_4$ and $y \geq \bar{y}_5$;
vi) $\ell_6 = h_6 = \bar{x}$ if $c \leq c_4$ and $y \geq \bar{y}_6$.

Note that not all equilibria necessarily exist for all parameters values. For example, when $\bar{x} \to 1$, we have an exponential verification function without any cap, and we only have three possible equilibria:
i) $\ell_1 = h_1 = 0$ if $c \geq 1-y$;
ii) $\ell_2 = 0$ and $h_2 \in (0,1)$ if $c < 1-y$, $y \geq \beta$ and $y \geq \bar{y}_2$;
iii) $0 < \ell_4 < h_4 < \bar{x}$ if $c < 1-y$, $y < \beta$ and $y \geq \bar{y}_4$.
This example is depicted in Figure 3.

Computations for Example 2. From equation (16) we have the expression of

the truth to rumor ratio, as a function of ℓ and h. If we are in the corner solution where $\ell = 0$ and $h = 0$, then the truth to rumor ratio is one and remains so if β changes marginally.

For all the other cases discussed in the example, we plug the values of ℓ and h, as obtained in Example 1, into equation (16). In this way we obtain that the truth to rumor ratio is constant in homophily when $0 < \ell < h < 0$ and when $\ell = h = \bar{x}$. The truth to rumor ratio when $\ell \in (0, \bar{x})$ and $h = \bar{x}$ is

$$\frac{\rho_0}{\rho_1} = \frac{(1+\bar{x})(1-y) + cy + 2\beta((1-y)(1-\bar{x}) - c)}{(1-\bar{x})(1-y) + cy},$$

which is increasing in β if $y \leq (1 - c - \bar{x})/(1 - \bar{x})$, and decreasing otherwise. Hence, the effect of homophily on the truth to rumor ratio is positive if $y \in [y_2, (1 - c - \bar{x})/(1 - \bar{x})]$ and negative if $y \in ((1 - c - \bar{x})/(1 - \bar{x}), 1]$.

*

REFERENCES

Akbarpour, Mohammad, Suraj Malladi, and Amin Saberi. 2018. "Just a few seeds more: Value of network information for diffusion." Mimeo.

Akerlof, George A, and Rachel E Kranton. 2000. "Economics and identity." *Quarterly Journal of Economics*, 115(3): 715–753.

Ali, S Nageeb. 2018. "Herding with costly information." *Journal of Economic Theory*, 175: 713–729.

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, and Drew Fudenberg. 2004. "Word-of-mouth learning." *Games and economic behavior*, 46(1): 1–22.

Banerjee, Abhijit, and Olivier Compte. 2020. "Consensus and disagreement: Information aggregation under (not so) naïve learning." Mimeo.

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

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.

Bogart, Laura M, and Sheryl Thorburn. 2005. "Are HIV/AIDS conspiracy beliefs a barrier to HIV prevention among African Americans?" *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 38(2): 213–218.

Bolletta, Ugo, and Paolo Pin. 2021. "Polarization when people choose their peers." *Available at SSRN 3245800*.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2018. "A note on internet use and the 2016 US presidential election outcome." *Plos one*, 13(7): e0199571.

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.

Compte, Olivier, and Andrew Postlewaite. 2004. "Confidence-enhanced performance." *American Economic Review*, 94(5): 1536–1557.

Cramer, Shirley. 2018. "Moving the needle: Promoting vaccination uptake across the life course." *Retrieved from https://www.rsph.org.uk/*.

DeGroot, Morris H. 1974. "Reaching a consensus." *Journal of the American Statistical Association*, 69(345): 118–121.

Frenkel, Sheera, and Mike Isaac. 2018. "Inside Facebook's election 'war room'." *The New York Times*.

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

Galeotti, Andrea, Christian Ghiglino, and Francesco Squintani. 2013. "Strategic information transmission networks." *Journal of Economic Theory*, 148(5): 1751–1769.

Gentzkow, Matthew, and Jesse M. Shapiro. 2011. "Ideological segregation online and offline." *The Quarterly Journal of Economics*, 126(4): 1799–1839.

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

Golub, Benjamin, and Matthew O Jackson. 2010. "Naive learning in social networks and the wisdom of crowds." *American Economic Journal: Microeconomics*, 2(1): 112–149.

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

Hagenbach, Jeanne, and Frédéric Koessler. 2010. "Strategic communication networks." *The Review of Economic Studies*, 77(3): 1072–1099.

Halberstam, Yosh, and Brian Knight. 2016. "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter." *Journal of Public Economics*, 143: 73–88.

Jackson, Matthew O, and Brian W Rogers. 2007. "Relating network structure to diffusion properties through stochastic dominance." *The BE Journal of Theoretical Economics*, 7(1).

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

Kranton, Rachel, and David McAdams. 2020. "Social networks and the market for news." Mimeo.

Kremer, Michael. 1996. "Integrating behavioral choice into epidemiological models of AIDS." *The Quarterly Journal of Economics*, 111(2): 549–573.

Maidenberg, Micah. 2018. "Facebook to start fact-checking photos, videos." *Retrieved from https://www.wsj.com/articles/facebook-to-start-fact-checking-photos-videos-1536867288 on December 5, 2018.*

Marr, Bernard. 2020. "Coronavirus fake news: How Facebook, Twitter, and Instagram are tackling the problem." *Forbes.com*, March, 27.

Mian, Areeb, and Shujhat Khan. 2020. "Coronavirus: The spread of misinformation." *BMC medicine*, 18(1): 1–2.

Molavi, Pooya, Alireza Tahbaz-Salehi, and Ali Jadbabaie. 2018. "A theory of non-Bayesian social learning." *Econometrica*, 86(2): 445–490.

Mueller-Frank, Manuel, and Mallesh M Pai. 2016. "Social learning with costly search." *American Economic Journal: Microeconomics*, 8(1): 83–109.

Prakash, B Aditya, Alex Beutel, Roni Rosenfeld, and Christos Faloutsos. 2012. "Winner takes all: competing viruses or ideas on fair-play networks." 1037–1046, ACM.

Roose, Kevin. 2018. "We asked for examples of election misinformation. You delivered." *The New York Times*, November, 4.

Sadler, Evan. 2020. "Diffusion games." *American Economic Review*, 110(1): 225–70.

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

Silverman, Craig. 2016. "This analysis shows how fake election news stories outperformed real news on Facebook." *BuzzFeed.com*, Nov., 16.

Stephan, Walter G, and Cookie White Stephan. 2017. "Intergroup threat theory." In *The International Encyclopedia of Intercultural Communication.* , ed. Y.Y. Kim, 1–12. Wiley Online Library.

Tabasso, Nicole. 2019. "Diffusion of multiple information: On information resilience and the power of segregation." *Games and Economic Behavior*, 118: 219–40.

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): e0181821.